Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.047541

ARTICLE

MSC-YOLO: Improved YOLOv7 Based on Multi-Scale Spatial Context for Small Object Detection in UAV-View

Xiangyan Tang[1,2], Chengchun Ruan[1,2,*], Xiulai Li[2,3], Binbin Li[1,2] and Cebin Fu[1,2]

[1]School of Computer Science and Technology, Hainan University, Haikou, 570228, China

[2]Hainan Blockchain Technology Engineering Research Center, Hainan University, Haikou, 570228, China

[3]School of Cyberspace Security (School of Cryptology), Hainan University, Haikou, 570228, China

*Corresponding Author: Chengchun Ruan. Email: ccruan@hainanu.edu.cn

Received: 08 November 2023 Accepted: 27 February 2024 Published: 25 April 2024

ABSTRACT

Accurately identifying small objects in high-resolution aerial images presents a complex and crucial task in the field of small object detection on unmanned aerial vehicles (UAVs). This task is challenging due to variations in UAV flight altitude, differences in object scales, as well as factors like flight speed and motion blur. To enhance the detection efficacy of small targets in drone aerial imagery, we propose an enhanced You Only Look Once version 7 (YOLOv7) algorithm based on multi-scale spatial context. We build the MSC-YOLO model, which incorporates an additional prediction head, denoted as P2, to improve adaptability for small objects. We replace conventional downsampling with a Spatial-to-Depth Convolutional Combination (CSPDC) module to mitigate the loss of intricate feature details related to small objects. Furthermore, we propose a Spatial Context Pyramid with Multi-Scale Attention (SCPMA) module, which captures spatial and channel-dependent features of small targets across multiple scales. This module enhances the perception of spatial contextual features and the utilization of multiscale feature information. On the Visdrone2023 and UAVDT datasets, MSC-YOLO achieves remarkable results, outperforming the baseline method YOLOv7 by 3.0% in terms of mean average precision (mAP). The MSC-YOLO algorithm proposed in this paper has demonstrated satisfactory performance in detecting small targets in UAV aerial photography, providing strong support for practical applications.

KEYWORDS

Small object detection; YOLOv7; multi-scale attention; spatial context

1 Introduction

In recent years, the rapid advancements in unmanned aerial vehicles (UAVs) and computer vision technologies have opened up new possibilities for various applications, including object detection and recognition. The ability to detect and identify objects accurately and efficiently from aerial imagery has become increasingly important in fields such as surveillance [1], disaster management [2], agriculture [3], obstacle avoidance [4], and urban planning [5]. Unmanned aerial systems offer a unique perspective and the capability to cover large areas quickly, making them an ideal platform for remote sensing and data collection.

Object detection in aerial images presents several challenges due to factors such as varying object scales, occlusions, cluttered backgrounds, and perspective distortions. Traditional methods relying solely on manual or semi-automatic image interpretation are time-consuming and prone to human errors. In addition, since UAVs usually take images at different heights and angles, the scale of the target to be detected varies drastically, which will cause a large number of small objects to be enriched in the image and bring a major challenge to target detection. To overcome these limitations, researchers have turned their attention towards leveraging deep learning techniques for automatic object detection. Among various deep learning architectures, the You Only Look Once (YOLO) series [6–10] has emerged as a popular choice for real-time object detection in images and videos. The YOLO series, with its unique approach of dividing the input image into grid structures, forecast bounding boxes and class probabilities for every individual cell, have demonstrated impressive accuracy and efficiency. However, applying the YOLO framework to aerial imagery poses specific challenges that necessitate modifications and improvements to ensure optimal performance.

This research paper aims to address these challenges by proposing novel enhancements to the YOLOv7 framework [11] for object detection in aerial images using drone-captured imagery. Our proposed method combines the strength of the YOLO architecture with adaptations tailored to handle the unique characteristics of aerial imagery, such as varying object scales and complex backgrounds. We hypothesize that by incorporating specific modifications, we can improve the accuracy, robustness, and efficiency of object detection using drone-acquired aerial images. The workflow utilizing MSC-YOLO is depicted in Fig. 1.

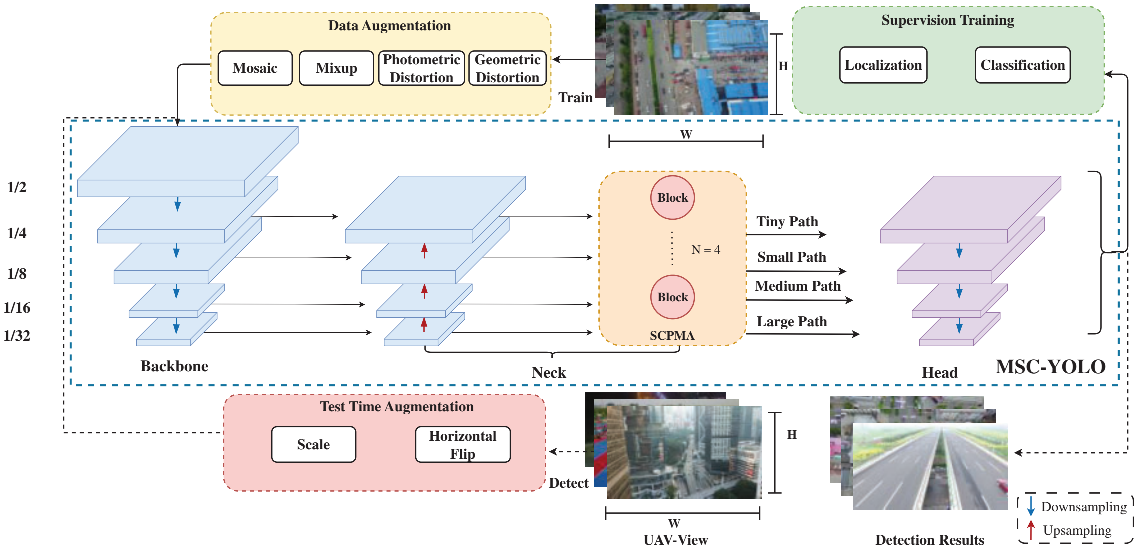


Figure 1: Overview of the working pipeline using MSC-YOLO

To validate our proposed approach, we conduct extensive experiments using publicly available aerial datasets, which are Visdrone2023 [12] and UAVDT [13], respectively, comparing the performance of our modified MSC-YOLO method against existing object detection algorithms. Our results demonstrate the effectiveness of our approach in accurately detecting and localizing the objects of interest, particularly in the challenging aerial environment, even for small target detection, highlighting its practicality for real-world applications.

The primary contributions of this paper are outlined below:

- Shallow feature maps contain more small object features. In the neck part, we introduce shallow feature maps for feature fusion and combine the upsampling process of the “Small” prediction head branch to form the “Tiny” prediction head to handle tiny objects.

- We propose the Spatial-to-Depth Convolutional Combination (CSPDC), which contains only non-stride convolutions, as a replacement for the original downsampling layer, to reduce the loss of detailed features.
- We propose the Spatial Context Pyramid with Multi-Scale Attention (SCPMA) and integrate it into YOLOv7, which enhances the model's ability to effectively fuse features of various scales and utilize spatial context information to improve the performance of small object detection.
- We evaluate MSC-YOLO on two datasets, Visdrone2023 and UAVDT. Experimental results show that MSC-YOLO achieves the level of state-of-the-art (SOTA) object detectors and significantly improves the detection performance of small objects in the dataset. It is shown to be an effective and competitive small object detector for UAV-view.

The structure of this paper is organized as follows: In Section 2, we discuss the existing related work. In Section 3, we introduce the architecture of the YOLOv7 model as well as the design philosophy behind the MSC-YOLO model and explain the implementation of each component of our model. Section 4 provides a detailed description of the experiments for MSC-YOLO, evaluating our proposed small object detection model from both quantitative and qualitative perspectives, and verifies each proposed component with ablation studies. In Section 5, we discuss why the detection performance of MSC-YOLO is superior to YOLOv7 and the current limitations faced by the algorithm. In Section 6, we summarize the design methods and results of this paper, as well as future work.

2 Related Works

To demonstrate the rationale behind the design of our model for detecting small objects, in this section, we will introduce the general approach for detecting small objects, which includes data augmentation, general object detectors, and detectors for small objects on UAV-captured images.

2.1 Data Augmentation

The imbalance in sample quantities within a dataset can have a significant impact on object detection results. When a dataset contains a considerably higher number of large-sized objects compared to small-sized objects, the model's ability to learn features from small object samples becomes limited. In such cases, the utilization of data augmentation strategies can enhance the detection performance of small objects within the dataset. Some researchers have proposed methods for data augmentation using multiple images clustered together [14], namely MixUp [15], Copy paste [16], CutMix [17] and Mosaic [9]. RRNet [18] proposed an adaptive enhancement method called AdaResampling. Although it shares some similarities in principle with [16], its uniqueness stems from the use of prior segmentation maps. These segmentation maps guide the sampling process for pasting locations. Furthermore, this method applies scale adjustments to the objects to be pasted, further minimizing scale differences. To obtain a larger sample of small objects without acquiring additional data, both Zhang et al. [19] and Wang et al. [20] adopted operations based on partitioning and resizing to acquire more small object training samples. DS-GAN [21] has developed an innovative pipeline for data augmentation, integrating techniques such as object segmentation, image reconstruction, and image blending, with the aim of producing premium synthetic data for pre-existing diminutive objects.

2.2 Object Detection

With the advancement of deep learning, significant breakthroughs have been made in object detection methods based on deep neural networks. These deep learning-based object detection methods can be divided into two types, one-stage detectors and two-stage detectors, according to whether they generate candidate boxes first for subsequent detection.

YOLO is a distinguished one-stage object detection approach that reformulates the task of object detection as a regression problem. It predicts the positions and categories of objects by making predictions on a grid applied to the image. YOLO9000, YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7, and YOLOv8 [7–11] have been launched one after another, continuously improving the detection accuracy and speed. Additionally, SSD (Single Shot MultiBox Detector) [22] is another commonly used one-stage object detection algorithm. It predicts the positions and categories of objects at multiple scales by utilizing multiple feature maps to facilitate detection at different scales. YOLOX [23] switches the YOLO series detectors to anchor-free mode for the first time and performs other advanced detection technologies, namely decoupling heads and the leading label assignment strategy SimOTA.

Two-stage object detectors are a popular class of models used in computer vision for accurate and robust object detection. These detectors consist of two main stages: Region proposal generation and object classification. In the initial phase, a collection of prospective bounding boxes for objects is produced, followed by the subsequent phase, which involves categorizing these candidates into various classes of objects. One of the most well-known two-stage object detectors is the Faster R-CNN (Region Convolutional Neural Network) architecture introduced by Ren et al. in 2015 [24]. Faster R-CNN revolutionized object detection by combining a region proposal network (RPN) with a region-based convolutional neural network (R-CNN) [25]. Another notable advancement in two-stage object detection is the Cascade R-CNN architecture proposed by Cai et al. in 2018 [26]. Cascade R-CNN introduces a cascade of classifiers, where each stage progressively refines the detection results by filtering out false positives and focusing on harder examples. In recent years, there has also been an increased interest in exploring novel backbone architectures for two-stage object detectors. For example, the Feature Pyramid Network (FPN) introduced by Lin et al. in 2017 [27] has become a popular choice for enhancing the feature representation across different scales.

2.3 Small Object Detection on UAV-Captured Images

Developing object recognition techniques for images captured by drones is intricate because of the variability in flight dynamics, extensive scene ranges, and the variety of objects present. Existing approaches address these challenges through various strategies. ClusDet [28] integrates object clustering and detection in an end-to-end framework. Zhang et al. [29] proposed a difficult region estimation network to identify high-density areas for detection. GLSAN [30] determines an adaptive region to enhance precision in areas of high density. DMNet [31] employs a crop strategy guided by density maps to balance foreground and background information. DSHNet [32] addresses performance degradation in long-tail scenes. HawkNet [33] aggregates multi-scale features effectively. CDMNet [34] employs a streamlined dual-task network for density estimation. TPH-YOLOv5 [35] merges transformer-centric prediction heads with YOLOv5, enhancing detection performance in scenarios with dense objects and varying object sizes. SODNet [36] refines the detection of smaller objects without compromising real-time efficiency. QueryDet [37] introduces an innovative query strategy to accelerate the processing speed of feature pyramid-driven object detection systems. DCRFF as introduced by Mittal et al. in 2022 [38] leverages receptive field principles and merges feature maps to enhance the performance of deep object detection methods in low-altitude aerial imagery. ARSD [39] outlines a two-stage adaptive region detection schema. It first utilizes a comprehensive detection network for preliminary object localization. After this, it implements an object clustering method with constant point density and a tailored selection algorithm to pinpoint object-rich sub-zones. These approaches collectively contribute to the advancement of object detection on drone-captured images.

3 Methodology

To address the challenge of detecting small objects in mid-low altitude aerial photography by drones, we propose an improved YOLOv7 algorithm (MSC-YOLO) based on multi-scale spatial context. In this section, we first introduce the architecture of YOLOv7 and the overall framework of MSC-YOLO. Subsequently, the key components of MSC-YOLO are detailed, including the prediction head for tiny objects, the CSPDC module, and the SCPMA module.

3.1 Overview of YOLOv7

YOLOv7 is a brand new version of the YOLO series of detectors. It is based on a collection of some existing tricks and module reparameterization and dynamic label allocation strategies. The Efficient Layer Aggregation Networks (ELAN) proposed by it can make a deeper model learn and converge more effectively by controlling the shortest and longest gradient path. The author of YOLOv7 referred to the above structure and proposed Extended-ELAN (E-ELAN). E-ELAN adopts the feature aggregation and feature transfer process similar to ELAN, only uses the group convolution, expansion module, and shuffling module similar to ShuffleNet [40] in the calculation module, and finally fuses the features through the aggregation module. By adopting this method, more diverse features can be obtained, and at the same time, the calculation and utilization efficiency of parameters can be improved. The overall architecture of YOLOv7 is illustrated in Fig. 2.

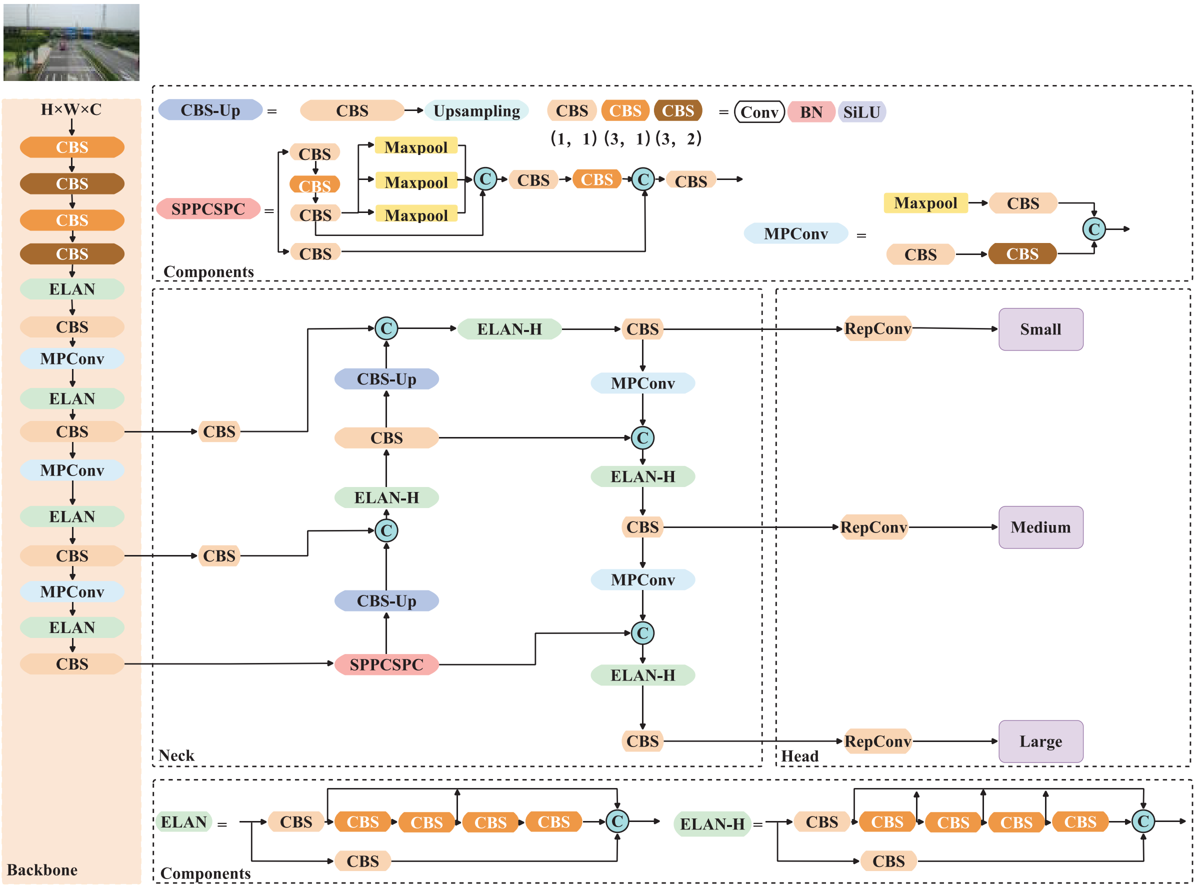


Figure 2: Overall architecture of the YOLOv7

3.2 Overview of MSC-YOLO

To solve the problem of sharp target scale changes and small object detection difficulties in UAV aerial scenes, in this section, we use the tiny object prediction head, modified down-sampling CSPDC module, and SCPMA module, propose improved YOLOv7 (MSC-YOLO) algorithm. The overall architecture of MSC-YOLO is illustrated in Fig. 3.

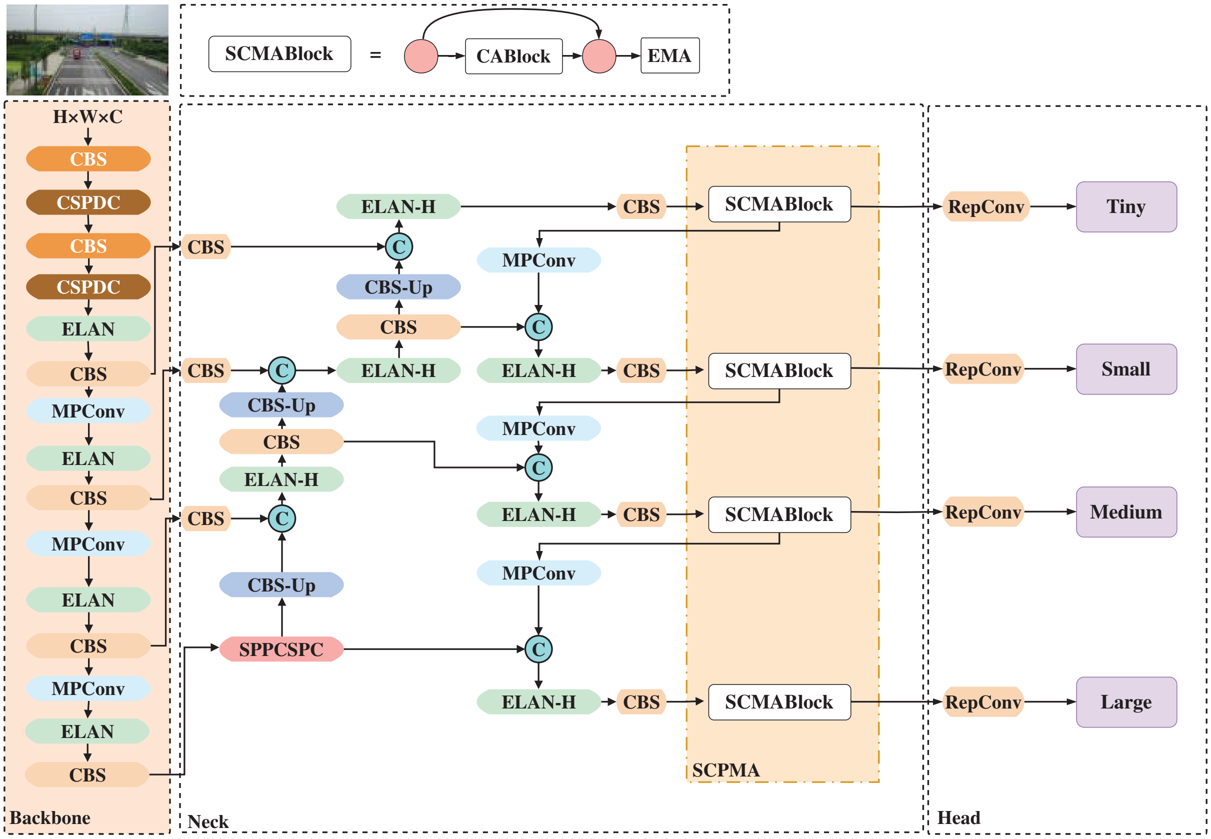


Figure 3: Overall architecture of the MSC-YOLO

3.3 Prediction Head for Tiny Objects

In the experimental section, we introduce the categories of the datasets used and the proportion of small objects. Among them, there are a large number of tiny objects that are difficult to detect. Therefore, we add a fourth prediction head based on the original three prediction heads. The added prediction head is upsampled for the third time in the Neck, and at the same time, the shallower features of the Backbone are fused across connections to make it have a smaller receptive field, which enhances model's capability to identify both small and multi-scale objects. Although the extra prediction head increases the parameters and Giga Floating-point Operations Per Second (GFLOPs) of the model, it notably enhances the efficiency of detecting small objects.

3.4 Spatial-to-Depth Convolutional Combination Module

In object detection, deep learning models such as convolutional neural networks (CNN) are usually used to extract image features. Downsampling can reduce the size of the image, thereby reducing the amount of computation, and can preserve important feature information. Target objects may appear in different scales in the image, some objects may be small, while others may be large. Through downsampling, images of different scales can be generated, so that objects can be detected at different scales. However, downsampling reduces the resolution of an image, resulting in a certain degree of information loss. Smaller objects may be blurred or lost during downsampling, affecting detection accuracy.

Therefore, we propose to use the CSPDC downsampling module to optimize downsampling and reduce the loss of feature information caused by downsampling with a step size of 2. This is a modified version of SPD-Conv [41]. In the process, we utilize the intermediate feature map, denoted as F, which has dimensions $S \times S \times C_1$. Here, S indicates the width and height of this map, while C_1 stands for the channel count. Firstly, the number of channels of the feature map is adjusted by convolution, sliced into four sub-feature maps of size $S/2 \times S/2 \times C_2$ and then spliced into a feature map of $S/2 \times S/2 \times 4 * C_2$ by the dimension of the channels, and finally adjusted back to the original number of channels by the convolutional layer. The whole process is shown in Fig. 4. We describe its algorithm as shown in Algorithm 1.

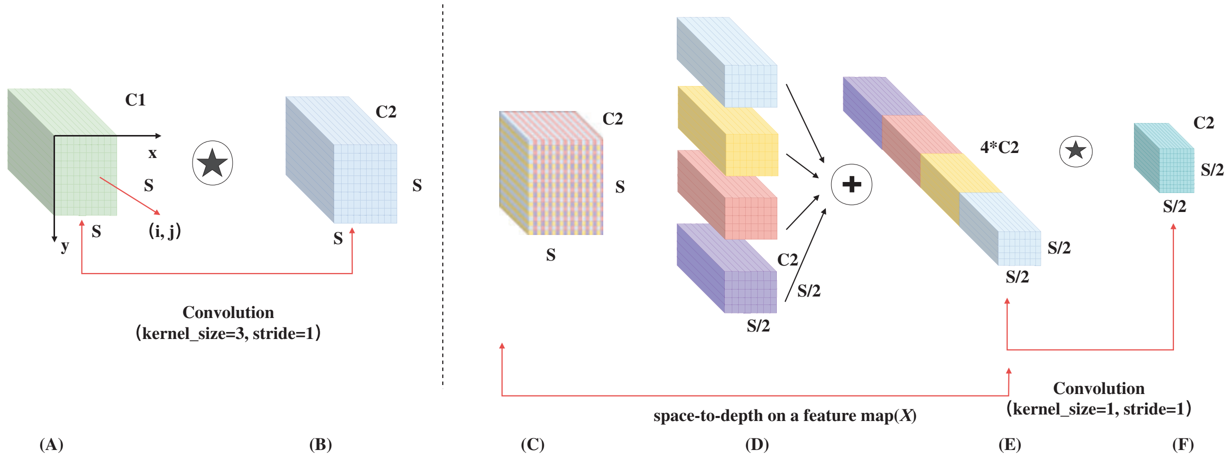


Figure 4: Overall architecture of the CSPDC

```
**Algorithm 1:** CSPDC algorithm
**Input:** the feature map $F$ of size $S \times S \times C_1$
**Output:** the feature map $F'$of size $S/2 \times S/2 \times C_2$
$F \leftarrow S \times S \times C_1, scale \leftarrow 2$
Function CSPDC($F, scale$)
        $f \leftarrow Conv\,(F, kernel\_size = 3, stride = 1, C_2)$
        $f'_{0,0} \leftarrow f\,[0\colon S\colon scale, 0\colon S\colon scale]$
        $f'_{1,0} \leftarrow f\,[1\colon S\colon scale, 0\colon S\colon scale]$
        $f'_{0,1} \leftarrow f\,[0\colon S\colon scale, 1\colon S\colon scale]$
        $f'_{1,1} \leftarrow f\,[1\colon S\colon scale, 1\colon S\colon scale]$
        $F' \leftarrow cat\left(f'_{0,0}, f'_{1,0}, f'_{0,1}, f'_{1,1}\right)$
        $F' \leftarrow Conv\,(F', kernel\_size = 1, stride = 1, C_2)$
        Return $F'$
End Function
```

3.5 Spatial Context Pyramid with Multi-Scale Attention

Upon aggregating feature maps from different levels, the feature pyramid still retains spatial local information. To further enhance the features, we introduce the Spatial Context Pyramid (SCP) [42], which augments the features by learning global spatial context within each level. The core component of SCP named the Context Aggregation Block (CABlock), resembles a self-attention mechanism, as depicted in the left half of Fig. 5. Three 1 ×1 convolutional branches serve as linear transformations, representing contextual relations, contextual content, and contextual importance. The aggregated spatial context information is subsequently fed into the Efficient Multi-Scale Attention (EMA) to form the Spatial Context with Multi-scale Attention Block (SCMABlock). EMA [43] is an efficient multi-scale attention module that groups original features into sub-features along the channel dimension, reducing computational overhead while retaining channel-wise information. Moreover, the setup of parallel sub-networks captures spatial and channel dependencies in a multi-scale manner. After incorporating comprehensive global information to adjust the channel weights within each parallel pathway, the results from these two parallel branches are further combined using interactions across dimensions, aiming to recognize pixel-to-pixel associations. The structure of EMA is illustrated in the right half of Fig. 5. Aerial images captured by drones often cover vast areas, where features of small objects may not be distinct, especially in complex background scenarios. SCMABlock enables the model to learn and utilize more crucial features in such contexts. We describe its algorithm as shown in Algorithm 2.

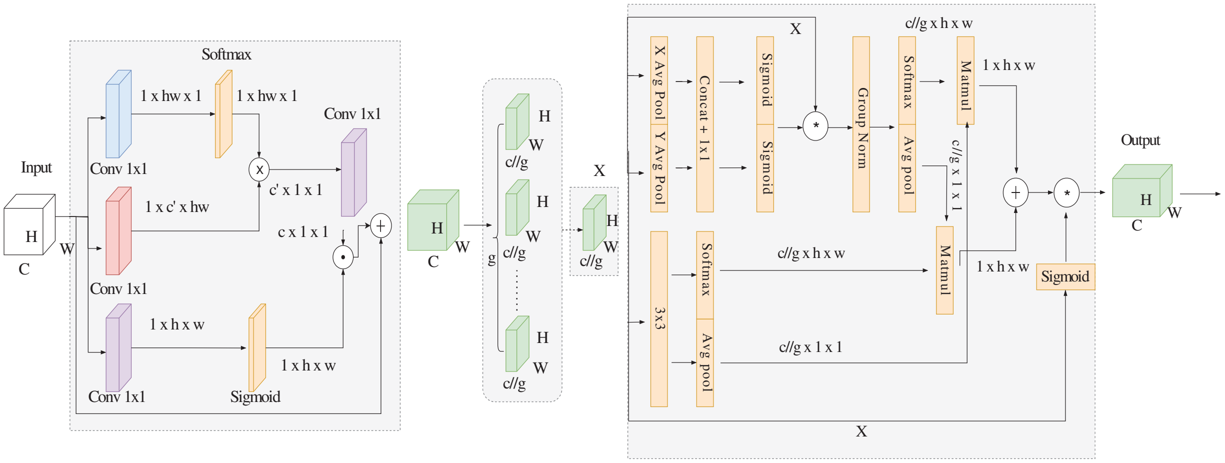


Figure 5: The SCMABlock's architecture of the SCPMA

Algorithm 2: SCPMA algorithm

Input: the feature maps of level i in the feature pyramid P_i, each contains N_i pixels. The linear transformations w_a, w_k and w_v are implemented using three convolutional blocks with 1×1 convolutional kernels.

Output: the feature maps of level i in the feature pyramid Q_i

```
For $i = 0, 1, \ldots, level_i$ do
    $P_i^{\prime} \leftarrow P_i^j + a_i^j \cdot \sum_{j=1}^{N_i} \left[ \frac{\exp\left(w_k P_i^j\right)}{\sum_{m=1}^{N_i} \exp\left(w_k P_i^m\right)} \cdot w_v P_i^j \right], a_i^j \leftarrow \frac{\exp\left(w_a P_i\right)}{\sum_{n=1}^{N_i} \exp\left(w_a P_i^n\right)}$
```

(Continued)

Algorithm 2 (continued)

```
    $P_i'$ feature groups By G groups, set $X$ as a sub_feature_group;
    $X \leftarrow \left\{X_0, X_g, \ldots, X_{G-1}\right\}, X_g \in R^{C//G\times H\times W}$
    Compute $X_h$, $X_w$ by applying Y Avg Pool and X Avg Pool on $X$
    Concatenate $X_h$ and $X_w$, Conv1 $\times 1$, then split it back into $X_h$ and $X_w$
  $X_1 \leftarrow X * Sigmoid\,(X_h) * Sigmoid\,(X_w)$ by Group Norm;
  $X_2 \leftarrow$ Apply Conv3 $\times$ 3 to $X$;
  Compute Attention Weights:
    $X_2' \leftarrow$ Softmax($X_1$ global average pooling) Matmul reshape $X_2$
    $X_1' \leftarrow$ Softmax($X_2$ global average pooling) Matmul reshape $X_1$
  Calculate Final Attention-Weighted Feature Map:
    $Q_i \leftarrow Sigmoid\left(X_1' + X_2'\right) * X$
  Reshape the final $Q_i$ to the original shape
End For
Return $Q_0, \ldots, Q_i$
```

4 Experiments and Evaluations

In this section, we introduce the small object datasets used, namely Visdrone2023 and UAVDT. We will also provide details of the implementation and the main evaluation metrics. To better evaluate our model, we assess the effectiveness of the small object detection achieved by our model from both quantitative and qualitative perspectives.

4.1 Datasets

The Visdrone2023 and UAVDT datasets are selected for use based on their extensive application and representativeness in the field of drone object detection. The two popular benchmark datasets encompass a diverse range of drone aerial photography scenes, effectively demonstrating the performance of the proposed method in handling small object detection.

The Visdrone2023 dataset encompasses aerially-acquired images from diverse environments, subjected to varying weather and illumination circumstances. The training set includes 6,471 images, with 548 images for validation and 1,610 images for the final testing of models. The image scale of the dataset is about 2,000 ×1,500 pixels. It includes diverse scenarios such as dense and sparse scenes, strong daylight, cloudy weather, and nighttime conditions. There are ten object categories in the dataset. The objects exhibit significant variations in scale, with a considerable proportion of small and densely packed targets. Considering that the official UAVDT dataset does not divide into training and test sets, we maintain the same division ratio as the methods we compared with, specifically 23,258 images for training and 15,069 images for testing. The images used for testing do not have scenes similar to the training pictures, and all scenes are shot by drones. The resolution of each image is approximately 1,080 ×540 pixels. The images have been annotated with bounding boxes, encompassing three distinct categories, namely, car, truck, and bus.

According to official calculations, 97.1% of the targets occupy 1% or less of the total image area. Following the definition of small objects in the COCO dataset [44], small objects are those whose area occupies 1% or less of the total area or whose area is smaller than 32 × 32 pixels. Therefore, these 97.1% of targets can be classified as small objects. Similarly, 74.87% of the UAVDT datasets belong to small-sized objects, and it is a challenging task to accurately identify small objects on both datasets.

4.2 Implementation Details and Evaluation Metrics

We use PyTorch 1.8.1 as the base environment for implementing MSC-YOLO and train the final model using the high-computational graphics card NVIDIA RTX3090 GPU. Pretrained weights of YOLOv7 on the COCO dataset are employed to accelerate the convergence during the training phase. Since MSC-YOLO and YOLOv7 share a significant portion of the backbone and neck architecture, some of the weights are transferred to our model. Furthermore, due to hardware computational limitations, the input images are set to 1,280 ×1,280. During the training phase, we utilize the Adam optimizer with an initial learning rate of $3e^{-4}$, coupled with a cosine learning rate adjustment strategy. To address the memory consumption of large-sized input images, we set batch size to 2, momentum to 0.937, and weight decay to $5e^{-4}$. The first two epochs are used for warmup, and a total of 100 epochs are trained.

To better utilize the object features in the existing datasets, we employ data augmentation techniques such as horizontal flipping, mosaic, mixup, paste_in (image copy-paste), scale, and cutmix, among others. These techniques expand the distribution of object features in the dataset, enhancing the model's generalizability in detecting complex scenes.

We adopt mean average precision (mAP) and mean average precision at IOU threshold 0.5 ($\text{mAP}_{0.5}$) as the performance evaluation metrics for object detection. By utilizing precision (P) and recall (R), we can calculate the AP, mAP, and $\text{mAP}_{0.5}$ as outlined below:

$$AP = \int_0^1 P(R)\, dR \tag{1}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i, IOU = 0.5{:}\ 0.05{:}\ 0.95 \tag{2}$$

$$mAP_{0.5} = \frac{1}{N}\sum_{i=1}^{N} AP_i, IOU = 0.5 \tag{3}$$

where N represents the number of classifications of the targets.

4.3 Quantitative Evaluation

Based on the model implementation and hyperparameter settings in Section 4.2, we evaluate MSC-YOLO using various metrics including mAP, $\text{mAP}_{0.5}$, Parameters and Frames Per Second (FPS). Additionally, we employ the COCO evaluation metric mAP_S to assess the algorithm's ability to detect small objects. To demonstrate the effectiveness of our proposed algorithm, we compare it with SOTA methods such as one-stage and two-stage approaches. To save time in configuring the baseline models, some of them, including Faster-RCNN, Cascade-RCNN, YOLOX, YOLOv5X, and YOLOv6, are implemented using the open-source project MMdetection [45]. All models were trained on the Visdrone2023 dataset, and the experimental results are presented in Tables 1 and 2. Furthermore, to demonstrate the robustness of our proposed algorithm, we conduct experiments on the UAVDT dataset, and the results are shown in Table 3.

Table 1: Comparison of detection results on the Visdrone2023 validation set between MSC-YOLO and current existing methods

Method	$mAP_{0.5}$ (%)	mAP (%)	mAP_S (%)	Params (M)	FPS
Faster-RCNN [24]	42.0	25.4	15.4	41.4	27.9
Cascade-RCNN [26]	41.8	26.2	15.3	69.2	22.6
ClusDet [28]	53.2	28.4	19.1	–	3.7
Zhang et al. [29]	58.0	30.3	–	–	–
GLSAN [30]	55.8	32.5	–	–	–
DMNet [31]	49.3	29.4	21.6	–	3.4
DSHNet [32]	51.8	30.3	–	–	–
HawkNet [33]	44.3	25.6	19.9	–	10.6
CDMNet [34]	52.9	31.9	23.8	–	–
DCRFF [38]	57.0	35.0	17.6	–	–
ARSD [39]	57.9	35.0	–	–	–
QueryDet [37]	48.1	28.3	–	–	2.8
YOLOX-L [23]	52.5	31.4	23.7	54.2	18.3
YOLOv6L [10]	57.4	36.4	26.7	58.5	21.9
YOLOv5X	57.3	36.5	**27.2**	86.3	24.3
TPH-YOLOv5++ [46]	**61.9**	**41.4**	–	–	20.1
YOLOv7 [11]	59.2	35.5	27.1	**37.2**	**43.1**
MSC-YOLO (**Ours**)	**62.4**	**38.5**	**29.6**	**40.5**	**37.0**

Note: – indicates that the paper does not provide this data. Bold font indicates the top two rankings for this metric.

Table 2: Comparison of detection results on the Visdrone2023 test-dev set between MSC-YOLO and current existing methods

Method	$mAP_{0.5}$ (%)	mAP (%)	mAP_S (%)	Params (M)	FPS
Faster-RCNN [24]	23.0	13.3	7.0	41.4	28.3
Cascade-RCNN [26]	22.9	13.6	7.4	69.2	23.5
YOLOX-L [23]	42.3	24.3	15.1	54.2	19.4
YOLOv6L [10]	47.3	**28.7**	17.3	58.5	21.9
YOLOv5X	47.6	28.5	**17.6**	86.3	23.8
YOLOv7 [11]	**48.2**	27.4	16.8	**37.2**	**44.8**
YOLOv8 [47]	45.3	27.9	17.0	46.6	**55.6**
MSC-YOLO (**Ours**)	**52.1**	**30.4**	**19.1**	**40.5**	38.0

Table 3: Comparison of detection results on the UAVDT test set between MSC-YOLO and current existing methods

Method	$mAP_{0.5}$ (%)	mAP (%)	mAP_S (%)	FPS
ClusDet [28]	26.5	13.7	9.1	–
Zhang et al. [29]	–	17.7	–	–
GDFNet [48]	26.1	15.4	8.9	–
GLSAN [30]	30.5	19.0	–	–
DMNet [31]	24.6	14.7	9.3	–
DSHNet [32]	30.4	17.8	–	16.4
CDMNet [34]	35.5	20.7	13.9	–
SODNet [36]	29.9	17.1	11.9	45.0
UFPMP-Net [49]	38.7	24.6	–	–
TPH-YOLOv5 [35]	**41.3**	**26.9**	–	25.1
YOLOv7 [11]	40.0	24.2	**27.9**	**53.8**
MSC-YOLO (**Ours**)	**42.9**	**27.2**	**28.5**	**45.7**

4.3.1 Experimental Results on the Visdrone2023 Dataset

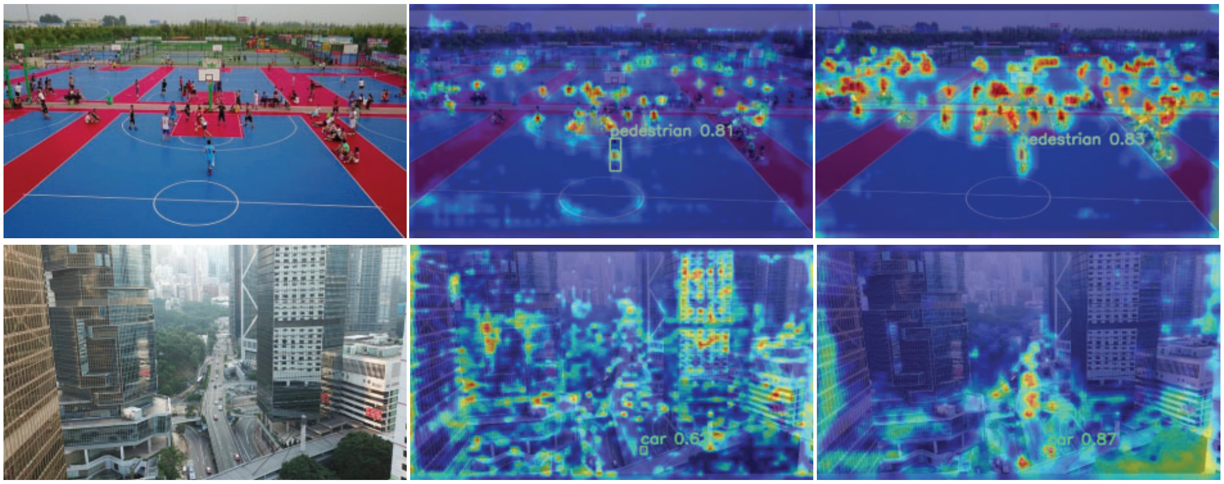
The experimental results on the validation set of Visdrone2023 are shown in Table 1. Our MSC-YOLO achieves mAP, $mAP_{0.5}$, and mAP_S scores of 38.5%, 62.4%, and 29.6%, respectively. This represents an improvement of 3%, 3.2%, and 2.5% over YOLOv7, respectively. Furthermore, our MSC-YOLO outperforms TPH-YOLOv5++ [46] by 0.5% in terms of $mAP_{0.5}$. It is worth noting that our algorithm demonstrates significant advantages over TPH-YOLOv5++ in terms of FPS, with an increase of 84.1% in FPS. While achieving the highest detection capability, MSC-YOLO only increases the parameters by 8.9% compared to YOLOv7. Compared to other methods, our method achieves the best performance with the lowest number of parameters. Considering the difference in object scale distribution between the validation set and the test set, our experiments on the Visdrone2023 test set are shown in Table 2, MSC-YOLO achieves the best detection performance among all the methods, and compared to YOLOv7 our improvement improves the mAP by 3% and the mAP_S by 2.3%, which outperforms the newest detector YOLOv8 in the YOLO series. Overall, MSC-YOLO achieves impressive detection performance while maintaining real-time efficiency compared to other baseline models.

4.3.2 Experimental Results on the UAVDT Dataset

We also conduct experiments on UAVDT, as shown in Table 3. Relative to existing methods, our MSC-YOLO establishes a novel benchmark across all three evaluation metrics, with mAP, $mAP_{0.5}$, and mAP_S of 27.2%, 42.9% and 28.5%, respectively. MSC-YOLO achieves a performance gain of nearly 3%, and $mAP_{0.5}$ is 1.6% higher than TPH-YOLOv5. However, MSC-YOLO does not have a significant improvement in the mAP_S metric. It may be due to the large difference in the scale variation of small objects between the UAVDT and Visdrone2023, where the size variation of objects in the UAVDT dataset is well-balanced, which leads to the small performance improvement. Overall, MSC-YOLO has significantly improved compared to the baseline method.

4.4 Qualitative Evaluation

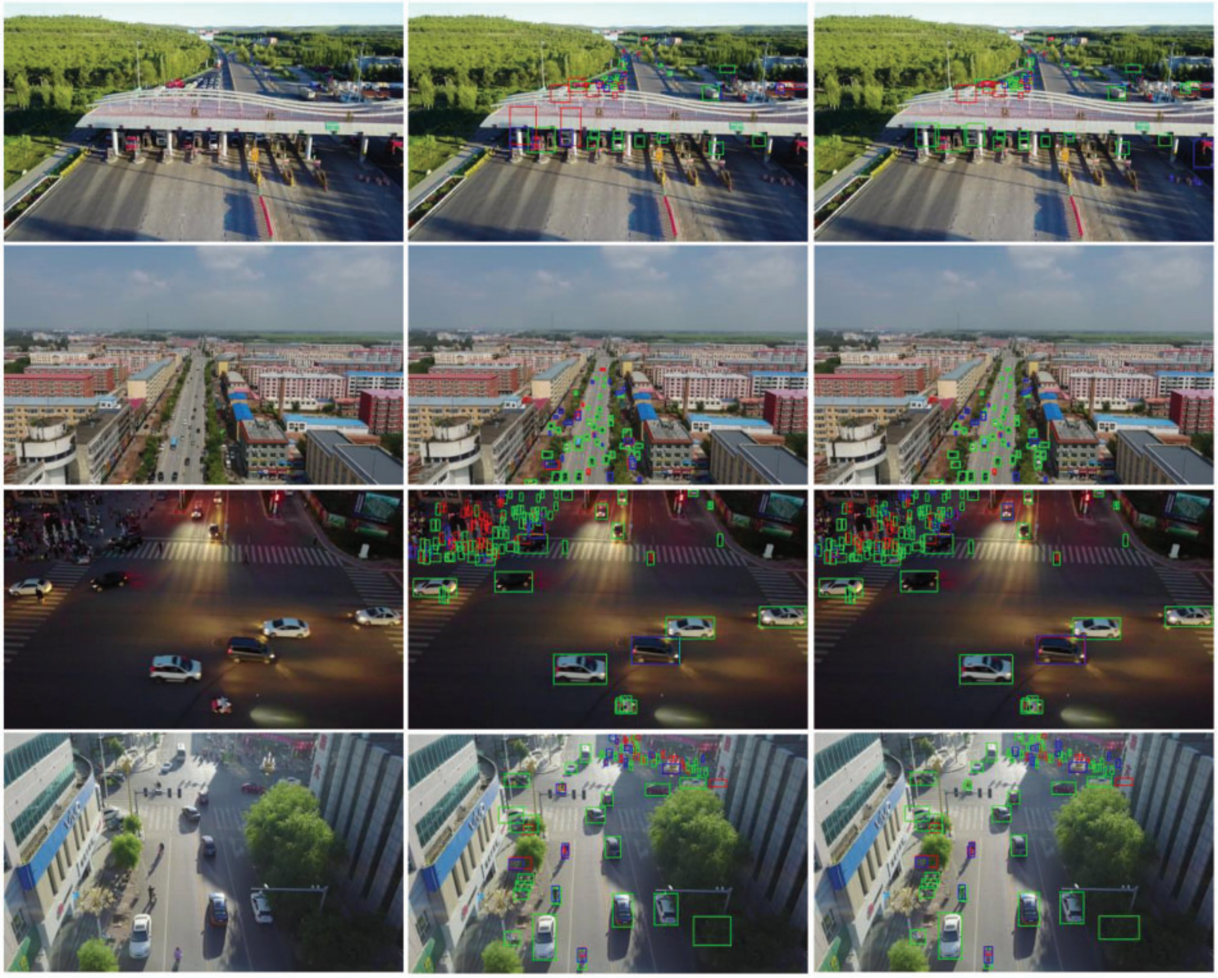
The qualitative results for the validation and test sets of the Visdrone2023 dataset are shown in Figs. 6 and 7. We use the Gradient-weighted Class Activation Mapping (Grad-CAM) [50] method to visualize the features of the detection results of the two models, so that we can observe which areas of the image the network pays attention to after passing through the backbone and head before identifying and locating the target, as shown in Fig. 6. We compare the attention results of the baseline YOLOv7 and our method. It can be observed that there are a large number of pedestrians scattered in the distance of the basketball court in Fig. 6a, YOLOv7 can only focus on some pedestrian targets, and the features of pedestrians in the distance are almost ignored, while MSC-YOLO has a smaller receptive field and multi-scale spatial context attention capability, our method enhances the attention capability of small targets. In addition, the second row of Fig. 6 shows that MSC-YOLO can suppress numerous irrelevant features in complex background target detection, and YOLOv7 incorrectly focuses on the buildings around the road, which may be due to the fact that the black-and-white color of the buildings is similar to the features of the road and vehicles.



(a) original image (b) YOLOv7 (c) MSC-YOLO

Figure 6: Visualization of model-generated feature maps. (a) Images of basketball courts and city roads. (b) Attention heatmap of YOLOv7. (c) Attention heatmap of MSC-YOLO

Fig. 7 shows the comparison of our detection results with the baseline method. The first row is a picture taken by a drone at a low altitude. YOLOv7 is poorly robust to trucks with a lot of occlusion, our model detects the front end of the trucks well and detects trucks farther down the road. The second row is a typical medium and high altitude UAV captured image, MSC-YOLO is significantly better than YOLOv7 for detecting vehicles on the far side of the road. The third and fourth rows are images under different lighting conditions containing dense crowds. MSC-YOLO still shows effectiveness under such complex lighting conditions, improving the detection accuracy of crowds in front of shopping malls in the images.



(a) original image (b) YOLOv7 (c) MSC-YOLO

Figure 7: Visualization of the detection results of YOLOv7 and MSC-YOLO. The crimson rectangle highlights instances where the object is overlooked, the azure rectangle signifies cases of misdetection, and the green rectangle confirms that the object is accurately detected. (a) Original image to be detected. (b) Detection results of YOLOv7. (c) Detection results of MSC-YOLO. It can be seen that MSC-YOLO has obvious improvement effects on small targets in different lighting conditions and dense areas

4.5 Ablation Study

The effectiveness of each proposed component is validated through ablation experiments on the Visdrone2023 test set, and the results are listed in Table 4. The state-of-the-art detector YOLOv7 is chosen as the baseline. We gradually add the tiny object prediction head P2, CSPDC, and SCPMA to the baseline. The ablation experiments are conducted with the same training configuration to ensure a fair comparison. After performing Test-Time Augmentation (TTA), we ultimately increased the $\text{mAP}_{0.5}$ score of MSC-YOLO to 52.1%.

Table 4: Results of ablation studies on the VisDrone2023 test-dev set

Baseline YOLOv7	With P2	With CSPDC	With SCPMA	$mAP_{0.5}$ (%)	mAP (%)	mAP_S (%)
✓				48.2	27.4	16.8
✓	✓			49.8 (+1.6)	28.9 (+1.5)	17.8 (+1.0)
✓	✓	✓		50.6 (+0.8)	29.5 (+0.6)	18.4 (+0.6)
✓	✓	✓	✓	**51.5 (+0.9)**	**30.4 (+0.9)**	**19.2 (+0.8)**

4.5.1 Ablation Study on Each Category

To delve deeper into the effectiveness of various improvement strategies across all categories on the Visdrone2023 dataset, we determine the AP for each category, illustrated in Fig. 8. "YOLOv7+ P2" refers to the prediction head with a small receptive field added based on YOLOv7. In the Visdrone2023 dataset, humans walking or standing are defined as "Pedestrians", while humans relying on transportation are defined as "People". The additional prediction head increases "pedestrian" AP by 1.8%, but "People" AP only increases by 0.6%. This result shows that the additional prediction head has better performance in dense environments. By adding the SCPMA module, notable enhancement in AP is observed for all categories except the Van category, specifically increasing Tricycle's AP by 2%. Therefore, the introduction of the spatial context pyramid with multi-scale attention can tap into more spatial context features to enhance the detection potential for small objects.

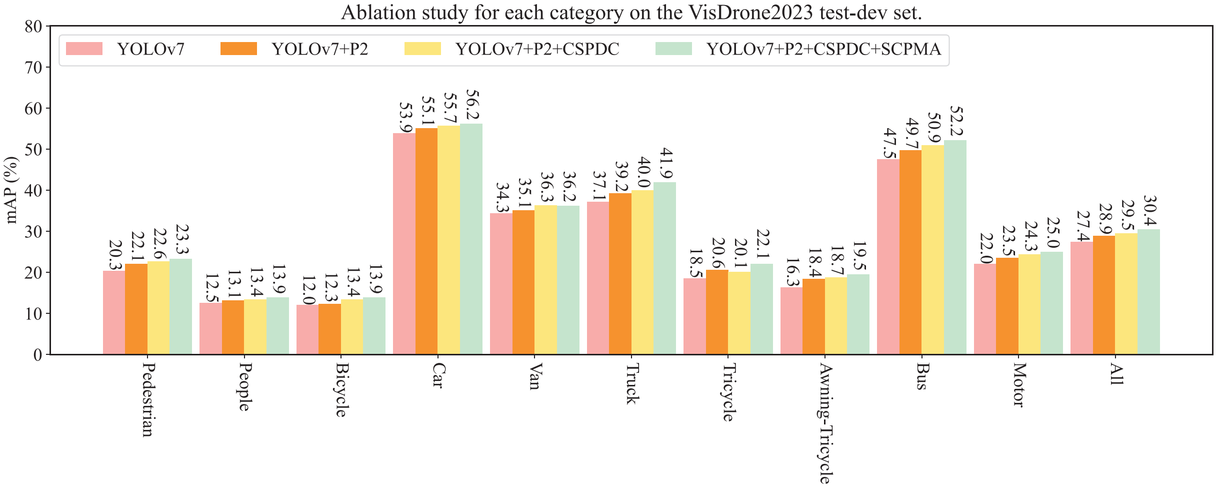


Figure 8: Results of ablation studies for the detection of each category on the Visdrone2023 test-dev set using the proposed improvement strategies

4.5.2 Effect of Different Numbers of Prediction Heads

The YOLO series object detectors are designed with three prediction heads, representing three different object size scales (Small, Medium, and Large). This design choice is based on the fact that the COCO dataset, which the YOLO series is trained on, does not contain a large number of objects with highly variable sizes. Therefore, the three prediction heads effectively match the object size distribution in the dataset. However, the Visdrone2023 and UAVDT datasets, with their high image resolutions and a significant number of small object instances, prompt us to investigate the impact of the number of prediction heads on detection results. We conduct experiments by varying the number of prediction heads, Table 4 demonstrates that after adding the extra small object prediction head P2, the $mAP_{0.5}$ and mAP_S metrics have improved by 1.6% and 1%, respectively.

4.5.3 Effect of CSPDC

We replace the original downsampling layers with CSPDC, the original downsampling layers have a convolutional kernel of size 3 and a stride of 2. Table 4 shows the changes in model detection performance after replacing the CSPDC module. As described in Section 3.4, CSPDC performs complex operations on the input feature map, adding a certain amount of computational overhead. However, the $mAP_{0.5}$ has increased from 49.8% to 50.6%, and mAP_S has increased by 0.6%. This improvement justifies the use of the CSPDC module.

4.5.4 Effect of SCPMA

Table 4 validates the effectiveness of our proposed SCPMA in enhancing spatial context feature perception and multi-scale feature information utilization. The outputs from the Neck part, i.e., P_2, P_3, P_4, and P_5, are fed into SCPMA. The GFLOPS increased from 174.9 to 182.9, and the number of parameters increased from 40.3M to 40.52M. As a result, MSC-YOLO achieves improvements of 0.9% $mAP_{0.5}$, 0.9% mAP, and 0.8% mAP_S. It demonstrates that by adding only a small computational overhead and a slight increase in the number of parameters, the detection performance of the model improved significantly.

5 Discussion

In this section, we will discuss why MSC-YOLO outperforms YOLOv7 in terms of performance and the limitations of the algorithm.

5.1 Analysis of Main Component Performance Gains

Based on the results from Tables 1–3 and Fig. 8, as well as subsequent ablation experiments, we summarize the reasons as follows:

(1) Prediction Head P2. The receptive fields of the three prediction heads in YOLOv7 have limited effects on detecting high-resolution images taken by drones. The MSC-YOLO introduces shallower feature maps with a size of 160 × 160, which, after feature fusion in the Neck part, can contain more features of small targets. In Fig. 8, the addition of the "Tiny" prediction head results in the greatest increase in AP for "pedestrian", but only a 0.6% increase for "person". Therefore, we visualize the attention effects after adding the "Tiny" prediction head as shown in Fig. 9. It can be observed that the receptive field of the feature map in the "Small" prediction head is limited, and the smaller receptive field of the "Tiny" prediction head matches well with the distant crowd in the basketball court. The further fusion of the "Tiny" and "Small" prediction heads enables the model to learn a sufficient number of features.

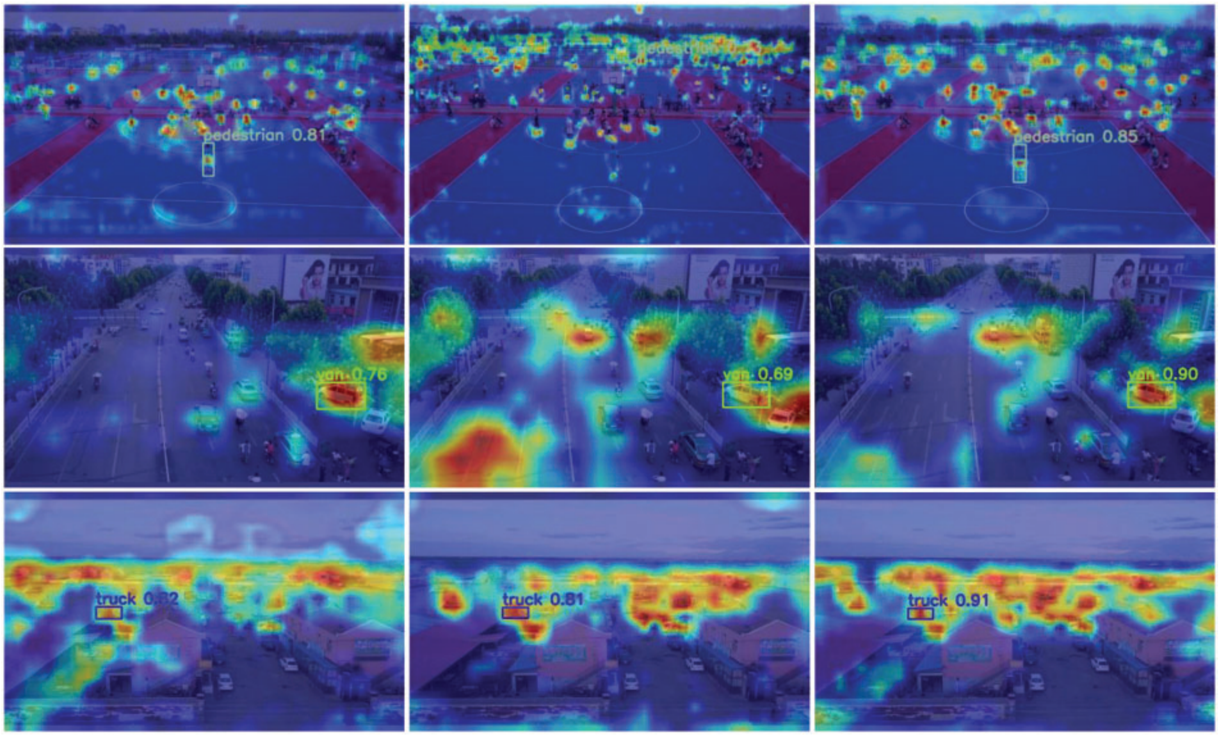


Figure 9: Visualization of the proposed component attention feature map, where deeper red indicates a greater contribution to the prediction

(2) CSPDC. The downsampling operation with a 3 × 3 convolutional kernel and a stride of 2 reduces the feature map size by half but loses too many detailed features. Our proposed CSPDC minimizes the reduction of spatial feature information. It first extracts features, then adjusts all spatial information across spatial and channel dimensions through spatial-depth operations, and finally returns to the original number of channels. In Fig. 8, the effectiveness of CSPDC has primarily reflected in the categories “Bicycle”, “Van”, and “Bus”, with the AP increasing by 1.1%, 1.2%, and 1.2%, respectively. The first image in the second row of Fig. 9 shows the attention result of YOLOv7 when detecting “Van”, the middle image shows the attention effect with the addition of the “Tiny” prediction head, and the third image shows the attention result after adding CSPDC. Although the detection effect in the middle image focuses on more features related to “Van”, it also pays attention to more irrelevant features. The effectiveness of CSPDC is due to the model learning the true features of “Van” by reducing the feature loss caused by downsampling.

(3) SCPMA. Considering the characteristic of the attention mechanism to enhance the model’s focus on important features while suppressing irrelevant ones, we propose SCPMA to enhance the model’s attention effect. SCPMA first performs a coarse self-attention, then calibrates multi-scale spatial context and channel features, ultimately achieving attention effects based on multi-scale spatial contexts. In Fig. 8, the effect of SCPMA is mainly reflected in the three categories: “Truck”, “Tricycle”, and “Bus”, with the AP increasing by 1.9%, 2.0%, and 1.3%, respectively. Therefore, we continue to visualize the results. The first image in the third row of Fig. 9 is the attention result of YOLOv7 when detecting “Truck”, the middle picture shows the attention effect with the addition of the “Tiny” prediction head and CSPDC, and the third image shows the attention effect after adding SCPMA, where the contribution of the red parts is significantly better than that of YOLOv7.

5.2 Limitations

In this study, we delve into the application of the enhanced YOLOv7 algorithm for detecting small objects in drone-captured aerial imagery. While this approach has demonstrated notable advancements in certain areas, it confronts several pivotal challenges that warrant further investigation. A primary concern is sustaining high-resolution object detection in computationally constrained environments, particularly where low-power devices are involved. Consequently, future research will prioritize enhancing the algorithm's detection speed without compromising accuracy, a critical aspect for fulfilling the real-time surveillance demands of drones. In addition, when objects are obscured by other objects in the environment, especially dense small objects, the detection capability of the algorithm is greatly reduced. This limitation underscores the necessity for more precise and robust object recognition techniques. Another prominent issue is category misidentification, exemplified by the model's occasional confusion between similar categories, such as "cars" and "vans". This may be attributed to the model's overfitting or underfitting of specific category characteristics. In future work, it will be necessary to more carefully adjust and optimize the learning process of category features to improve classification accuracy. At the same time, as drones are increasingly used in modern society, due to time and resource constraints, this study is not able to verify a larger number of datasets, such as small object detection in drone firefighting and rescue scenarios. Therefore, extending the diversity of training datasets and exploring more versatile model architectures represent crucial avenues for future research.

6 Conclusion

To address the challenges associated with suboptimal small object detection due to low resolution and inconspicuous features, this paper presents an enhanced approach to YOLOv7 based on multi-scale spatial context. By incorporating minute object prediction heads, a finer-grained receptive field is obtained to cater to smaller objects. The proposed methodology introduces the Spatial-to-Depth Convolutional Combination module and the Spatial Context Pyramid with Multi-Scale Attention, which collectively mitigates the loss of intricate features of small objects. This facilitates an increased emphasis on pivotal features while diminishing the interference of background information. Furthermore, comprehensive evaluations are conducted on the Visdrone2023 and UAVDT datasets. The detection performance of MSC-YOLO on the Visdrone2023 and UAVDT datasets has improved by 3.0% in terms of the mAP compared to YOLOv7, ultimately raising the mAP to 30.4% on the Visdrone2023 test-dev set and 27.2% on the UAVDT test set. These results underscore the efficacy of MSC-YOLO in the realm of small object detection from the perspective of unmanned aerial vehicles. In future work, we will focus on designing more lightweight models and introducing model pruning techniques to increase the model's FPS to make it more suitable for real-time drone scenarios.

Acknowledgement: We thank the School of Computer Science and Technology of Hainan University and the Hainan Blockchain Technology Engineering Research Center for their resources and support.

Funding Statement: This work was supported by the Key Research and Development Program of Hainan Province (Grant Nos. ZDYF2023GXJS163, ZDYF2024GXJS014), National Natural Science Foundation of China (NSFC) (Grant Nos. 62162022, 62162024), the Major Science and Technology Project of Hainan Province (Grant No. ZDKJ2020012), Hainan Provincial Natural Science Foundation of China (Grant No. 620MS021), Youth Foundation Project of Hainan Natural Science Foundation (621QN211).

Author Contributions: The author's contributions to the paper are mainly manifested in the following aspects: literature review: X. Li, C. Fu; algorithm design: X. Tang and C. Ruan; experimental design and implementation: C. Ruan, X. Tang; results analysis and interpretation: X. Tang, C. Ruan, B. Li; manuscript drafting: X. Tang, C. Ruan, X. Li. All authors have reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The Visdrone2023 dataset are available on https://github.com/VisDrone/VisDrone-Dataset. The UAVDT dataset are available on https://drive.google.com/file/d/1m8KA6oPIRK_Iwt9TYFquC87vBc_8wRVc/view.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Gu, T. Su, Q. Wang, X. Du, and M. Guizani, "Multiple moving targets surveillance based on a cooperative network for multi-UAV," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 82–89, 2018. doi: 10.1109/MCOM.2018.1700422.

[2] T. Xiong *et al.*, "Multi-drone optimal mission assignment and 3D path planning for disaster rescue," *Drones*, vol. 7, no. 6, pp. 394, 2023. doi: 10.3390/drones7060394.

[3] W. H. Maes and K. Steppe, "Perspectives for remote sensing with unmanned aerial vehicles in precision agriculture," *Trends Plant Sci.*, vol. 24, no. 2, pp. 152–164, 2019. doi: 10.1016/j.tplants.2018.11.007.

[4] L. Zheng and C. Song, "Fast near-duplicate image detection in riemannian space by a novel hashing scheme," *Comput. Mater. Contin.*, vol. 56, no. 3, pp. 529–539, 2018.

[5] I. Bisio, C. Garibotto, H. Haleem, F. Lavagetto, and A. Sciarrone, "Traffic analysis through deep-learning-based image segmentation from UAV streaming," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6059–6073, 2022. doi: 10.1109/JIOT.2022.3223283.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 779–788.

[7] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 6517–6525.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.

[9] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv preprint arXiv:2004.10934, 2020.

[10] C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," arXiv preprint arXiv: 2209. 02976, 2022.

[11] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Vancouver, BC, Canada, 2023, pp. 7464–7475.

[12] P. Zhu *et al.*, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7380–7399, 2021. doi: 10.1109/TPAMI.2021.3119563.

[13] H. Yu *et al.*, "The unmanned aerial vehicle benchmark: Object detection, tracking and baseline," *Int. J. Comput Vision*, vol. 128, no. 5, pp. 1141–1159, 2020. doi: 10.1007/s11263-019-01266-1.

[14] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, and K. Cho, "Augmentation for small object detection," arXiv preprint arXiv:1902.07296, 2019.

[15] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.

[16] G. Ghiasi *et al.*, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2918–2928.

[17] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 6022–6031.

[18] C. Chen *et al.*, "RRNet: A hybrid detector for object detection in drone-captured images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Seoul, Korea (South), 2019, pp. 100–108.

[19] X. Zhang, E. Izquierdo and K. Chandramouli, "Dense and small object detection in UAV vision based on cascade network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Seoul, Korea (South), 2019, pp. 118–126.

[20] X. Wang, D. Zhu, and Y. Yan, "Towards efficient detection for small objects via attention-guided detection network and data augmentation," *Sens.*, vol. 22, no. 19, pp. 7663, 2022. doi: 10.3390/s22197663.

[21] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes and A. Del Bimbo, "A full data augmentation pipeline for small object detection based on generative adversarial networks," *Pattern Recognit.*, vol. 133, no. 3, pp. 108998, 2023. doi: 10.1016/j.patcog.2022.108998.

[22] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. ECCV*, Amsterdam, the Netherlands, 2016, vol. 9905, pp. 21–37. doi: 10.1007/978-3-319-46448-0.

[23] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017. doi: 10.1109/TPAMI.2016.2577031.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 580–587.

[26] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, 2018, pp. 6154–6162.

[27] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 936–944.

[28] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proc. IEEE/CVF International Conf. Comput. Vis.*, Seoul, Korea (South), 2019, pp. 8310–8319.

[29] J. Zhang, J. Huang, X. Chen, and D. Zhang, "How to fully exploit the abilities of aerial image detectors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Seoul, Korea (South), 2019, pp. 1–8.

[30] S. Deng *et al.*, "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2020. doi: 10.1109/TIP.2020.3045636.

[31] C. Li, T. Yang, S. Zhu, C. Chen, and S. Guan, "Density map guided object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Seattle, WA, USA, 2020, pp. 737–746.

[32] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2021, pp. 3257–3266.

[33] H. Lin, J. Zhou, Y. Gan, C. M. Vong, and Q. Liu, "Novel up-scale feature aggregation for object detection in aerial images," *Neurocomputing*, vol. 411, no. 11, pp. 364–374, 2020. doi: 10.1016/j.neucom.2020.06.011.

[34] C. Duan, Z. Wei, C. Zhang, S. Qu, and H. Wang, "Coarse-grained density map guided object detection in aerial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 2789–2798.

[35] 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 *Proc. IEEE/CVF International Conf. Comput. Vis. Workshops*, Montreal, BC, Canada, 2021, pp. 2778–2788.

[36] G. Qi, Y. Zhang, K. Wang, N. Mazur, Y. Liu and D. Malaviya, "Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion," *Remote Sens.*, vol. 14, no. 2, pp. 420, 2022. doi: 10.3390/rs14020420.

[37] C. Yang, Z. Huang, and N. Wang, "QueryDet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 13658–13667.

[38] P. Mittal, A. Sharma, R. Singh, and V. Dhull, "Dilated convolution based RCNN using feature fusion for Low-Altitude aerial objects," *Expert. Syst. Appl.*, vol. 199, no. 5, pp. 117106, 2022. doi: 10.1016/j.eswa.2022.117106.

[39] Y. Wan *et al.*, "ARSD: An adaptive region selection object detection framework for UAV images," *Drones*, vol. 6, no. 9, pp. 228, 2022. doi: 10.3390/drones6090228.

[40] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "ShuffleNet V2: Practical guidelines for efficient CNN architecture design," in *Proc. ECCV*, Munich, Germany, vol. 11218, 2018, pp. 122–138.

[41] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," in *Proc. Euro. Conf. Mach. Learning Knowl. Disc. Databases*, Grenoble, France, vol. 13715, 2022, pp. 443–459.

[42] Y. Liu, H. Li, C. Hu, S. Luo, Y. Luo and C. W. Chen, "Learning to aggregate multi-scale context for instance segmentation in remote sensing images," arXiv preprint arXiv:2111.11057, 2021.

[43] D. Ouyang *et al.*, "Efficient multi-scale attention module with cross-spatial learning," in *IEEE Int. Conf. Acou., Speech Signal Process.*, Rhodes Island, Greece, 2023, pp. 1–5.

[44] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Zurich, Switzerland, vol. 8693, 2014, pp. 740–755.

[45] K. Chen *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," arXiv preprint arXiv:1906.07155, 2019.

[46] Q. Zhao, B. Liu, S. Lyu, C. Wang, and H. Zhang, "TPH-YOLOv5++: Boosting object detection on drone-captured scenarios with cross-layer asymmetric transformer," *Remote Sens.*, vol. 15, no. 6, pp. 1687, 2023. doi: 10.3390/rs15061687.

[47] Y. Zeng, T. Zhang, W. He, and Z. Zhang, "YOLOv7-UAV: An unmanned aerial vehicle image object detection algorithm based on improved yolov7," *Electronics*, vol. 12, no. 14, pp. 3141, 2023. doi: 10.3390/electronics12143141.

[48] R. Zhang, Z. Shao, X. Huang, J. Wang, and D. Li, "Object detection in UAV images via global density fused convolutional network," *Remote Sens.*, vol. 12, no. 19, pp. 3140, 2020. doi: 10.3390/rs12193140.

[49] Y. Huang, J. Chen, and D. Huang, "UFPMP-Det: Toward accurate and efficient object detection on drone imagery," in *Proc. AAAI Conf. Artif. Intell.* , 2022, pp. 1026–1033.

[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020. doi: 10.1007/s11263-019-01228-7.