Optimized YOLOv7 for Small Target Detection in Aerial Images Captured by Drone

Yanxin Liu, Shuai Chen*, Lin Luo
School of Information and Control Engineering, Liaoning Petrochemical University, Fushun, China

***Abstract*—It is challenging to detect small targets in aerial images captured by drones due to variations in target sizes and occlusions arising from the surrounding environment. This study proposes an optimized object detection algorithm based on YOLOv7 to address the above-mentioned challenges. The proposed method comprises the design of a Genetic Kmeans (1-IoU) clustering algorithm to obtain customized anchor boxes that more significantly apply to the dataset. Moreover, the SPPFCSPC_group structure is optimized using group convolutions to reduce model parameters. The fusion of Spatial Pyramid Pooling-Fast (SPPF) and Cross Stage Partial (CSP) structures leads to increased detection accuracy and enhanced multi-scale feature fusion network. Furthermore, a Detect Head is incorporated into the classification phase for more accurate position and class predictions. According to experimental findings, the optimized YOLOv7 algorithm performs quite well on the VisDrone2019 dataset in terms of detection accuracy. Compared with the original YOLOv7 algorithm, the optimized version shows a 0.18% increase in the Average Precision (AP), a reduction of 5.7 M model parameters, and a 1.12 Frames Per Second (FPS) improvement in the frame rate. With the above-described enhancements in AP and parameter reduction, the precision of small target detection and the real-time detection speed are increased notably. In general, the optimized YOLOv7 algorithm offers superior accuracy and real-time capability, thus making it well-suited for small target detection tasks in real-time drone aerial photography.**

Keywords—Small target detection; drone aerial photography; YOLOv7; clustering algorithm; spatial pyramid pooling

I. Introduction

Modern urban areas are characterized by dense city blocks, tall buildings, high population density, and heavy traffic, and they are capable of creating complex and dynamic environments. Satellite remote sensing is subjected to limitations in capturing high-resolution and high-dynamic range information for small targets for its revisit cycles, spatial resolution, and urban canyon effects. As sensor technology has been leaping forward, Unmanned Aerial Vehicles (UAVs) equipped with various sensors have emerged as effective tools for dynamically acquiring target images. UAV aerial imaging offers several advantages (e.g., a wide field of view, strong target detection capability, high real-time performance, as well as comprehensive information acquisition). Accurate detection and recognition of small targets through UAV aerial imaging enable fine-grained monitoring and provide valuable data for data-driven decision-making. However, conventional object detection algorithms struggle to effectively localize and accurately recognize small targets due to their low resolution and high noise interference.

Deep learning-based object detection algorithms have become the mainstream method due to their optimized efficiency and accuracy. The above-mentioned algorithms typically employ two-stage or one-stage detection strategies. Two-stage detection methods generate a series of candidate object boxes, which are subsequently filtered and refined by classifiers. Examples of two-stage algorithms include Faster Region-based Convolutional Neural Network (Faster R-CNN) [1] and Region-based Fully Convolutional Network (R-FCN) [2]. One-stage detection methods utilize convolutional neural networks [3] to extract image features and perform object classification and localization based on the above-described features. Algorithms such as You Only Look Once (YOLO) [4]–[11] and Single Shot MultiBox detector (SSD) [12] offer higher accuracy and generalization capability. To be specific, YOLOv7 has been confirmed as an advanced detection algorithm in the YOLO series, surpassing previous versions for inference speed and detection accuracy. Besides, it exhibits enhanced performance in detecting targets at different scales. However, challenges remain when YOLOv7 is employed for small target detection in UAV aerial imaging. First, small targets exhibit weak feature representation, such that they turn out to be susceptible to background interference and result in issues (e.g., false positives and false negatives). Second, deep learning models require significant computational resources for training and inference, whereas UAV aerial systems are subjected to limited computing resources and storage capacity. Accordingly, improving model size and computational efficiency becomes necessary. Lastly, deep learning algorithms are dependent on large-scale, high-quality annotated datasets to enhance their generalization capability, which is challenging to obtain specifically tailored for small target detection in UAV aerial imaging.

In this study, an enhanced YOLOv7 algorithm is presented for detecting small targets in UAV aerial imaging, to tackle the above challenges and fulfill the improvement requirements. The VisDrone2019 dataset is employed as the benchmark for detection. The proposed algorithm incorporates several significant enhancements, which comprise the redesign of anchor box sizes using an optimized clustering algorithm, the reduction of unnecessary candidate boxes, the reconstruction of the Spatial Pyramid Pooling (SPP) module, the integration of group convolutions and improved pooling connections, the reduction of model parameters, and the increased detection efficiency. Furthermore, a more accurate detection head, termed Detect, is introduced for target classification and position regression. The specific contributions of this study are elucidated below:

- The design of a high-precision anchor box clustering algorithm, termed Genetic Kmeans (1-IoU), employs genetic algorithms to optimize Kmeans clustering and adopts Intersection over Union (IoU) distance as a novel distance metric. The above-described algorithm leads to higher detection accuracy while reducing the likelihood of missing small targets.
- The optimization of the SPP module, SPPFCSPC_group, by integrating group convolutions and combining the SPPF module and the CSP structures. This enhancement improves the ability exhibited by the algorithm to detect multi-scale targets and reduces model complexity while increasing object detection accuracy.
- The adoption of a more precise detection head, termed Detect, achieves higher precision and recall in target classification and localization. Accordingly, false positives are reduced significantly, and the model is endowed with the enhanced ability to distinguish between targets and the background.

The optimized YOLOv7 algorithm is assessed on 10 target categories. Comparative analysis with the baseline YOLOv7 model demonstrates a 0.18% increase in Average Precision (AP), a reduction of 5.7% in model parameters, and a 1.12 times improvement in Frames Per Second (FPS). As revealed by the experimental results, the optimized YOLOv7 algorithm achieves high precision and speed in the recognition of tiny objects during UAV aerial imagery.

II. Related Work

In object detection, small targets are commonly defined in accordance with the relative scale, with a bounding box area to image area ratio less than the square root of 0.33, or following the absolute scale, with a resolution less than 32 by 32 pixels. In UAV aerial imaging, tiny target detection requires adjustments in data format, algorithm structure, and parameter settings to tackle several challenges (e.g., small target size, weak feature representation, occlusion, deformation, high noise interference, and real-time requirements). In general, researchers address the above-mentioned challenges by implementing multi-scale detection strategies to cope with small targets of different sizes, incorporating contextual information and spatial constraints to increase the target localization accuracy, and introducing attention mechanisms to handle complex scenarios with occlusions and deformations involved.

For algorithm optimization, Zhang et al. [13] proposed YOLOv7-RAR algorithm for urban vehicle recognition. To be specific, these researchers reconstructed the backbone network using the Res3Unit structure, with the aim of enhancing the model's capability to capture more nonlinear features. Moreover, they introduced an ACmix attention mechanism to address weak target localization arising from background interference. Zhu et al. [14] developed TPH-YOLOv5 algorithm for target detection in UAV captured scenes. In the above-described method, YOLOv5 serves as the baseline model, a Transformer prediction head is employed, and a Convolutional Block Attention Module (CBAM) attention mechanism is incorporated to enhance detection performance in dense aerial target scenarios. The enhanced algorithm achieves a 7% increase in accuracy compared with the baseline YOLOv5 model. However, the above-described methods often introduced additional network layers and parameters, resulting in increased computational complexity and limiting practical applications.

For data preprocessing, augmenting the training dataset can lead to the enhanced diversity and quantity of small targets, such that the model can be endowed with the enhanced generalization capability. Optimizing anchor box strategies can reduce computational costs and improve the matching between anchor boxes and real targets, enhancing detection accuracy. For instance, Liu and Wang [15] developed a YOLO-based detection network for corn detection and used a technique for data synthesis to create simulated images of broken maize from genuine corn photographs, such that the challenge of acquiring training data for damaged corn can be addressed. In the task of insulator defect detection, Zheng et al. [16] optimized YOLOv7 algorithm using the Kmeans++ clustering algorithm to cluster insulator targets and generate anchor boxes that more significantly apply to the detection of insulator defects. In the video surveillance vehicle detection task, Pan et al. [17] designed the improved YOLOv5s algorithm using Kmeans algorithm to correct the anchor frames and coordinated the CA attention mechanism for image recognition, which provided more accurate vehicle detection results and higher efficiency in terms of processing speed. The proposed method achieved high detection accuracy and speed on NVIDIA TX2 platform. However, optimized anchor boxes may struggle to accurately differentiate targets when they are occluded or overlapped, such that the detection accuracy can be reduced.

To conform to real-time requirements, algorithm optimization techniques (e.g., network pruning and quantization) are capable of reducing model computation and memory usage, such that the inference process can be expedited. Moreover, computational complexity can be reduced using lightweight model structures. Wu et al. [18] employed pruning techniques to lightweight the YOLOv4 network for concrete crack detection, where the EvoNorm-S0 algorithm was adopted to increase the detection accuracy. The resulting model achieved a high mAP value of 92.54% and a 15.9% reduction in the inference time, such that a real-time and high-precision detection algorithm was yielded. With the aim of detecting rice diseases and pests, Jia et al. [19] improved the YOLOv7 method and used the lightweight MobileNetV3 network for feature extraction, such that the model parameters were reduced, while an accuracy of 92.3% was generated. However, lightweight structures or network pruning may reduce model capacity while adversely affecting its representation capability, particularly in complex scene tasks, such that the accuracy and generalization capability can be decreased.

Despite the advancements achieved by regulating network structures, existing network architectures still struggle to reconcile detection speed and accuracy, particularly in highly overlapping small target areas. Thus, in-depth improvements should be made to increase the speed and precision of small target recognition algorithms based on UAV aerial imagery, ultimately elevating the capabilities of small target recognition and fine-grained monitoring in UAV aerial photography.

III. Optimized YOLOv7 Algorithm

A. *Overview of YOLOv7*

YOLOv7 refers to a single-stage deep learning-based object detection framework that achieves efficient object detection by detecting all objects in a single forward pass [11]. Compared with previous versions of the YOLO series, YOLOv7 offers faster convolution operations and higher detection accuracy, enabling it to detect more fine-grained objects while maintaining high detection speed. The YOLOv7 network structure comprises three components, i.e., a backbone network, a feature pyramid pooling layer, and a Head. Fig. 1 presents the simplified diagram of the YOLOv7 network structure. Backbone utilizes multiple convolutional layers to extract rich feature information, which is employed for subsequent object detection. The neck structure introduces the Path Aggregation-FPN (PAFPN) structure, combining feature maps at different scales to endow the algorithm with the enhanced capability to recognize various-sized things. The head layer employs the RepConv structure in conjunction with the IDetect Head to predict the target class and bounding boxes from the feature maps. The YOLOv7 algorithm exhibits high speed and real-time object detection capabilities while finding wide applications in areas (e.g., real-time video surveillance, UAV aerial imaging, and autonomous driving). It is capable of expediting the realization of intelligent and automated applications in a wide variety of scenarios.

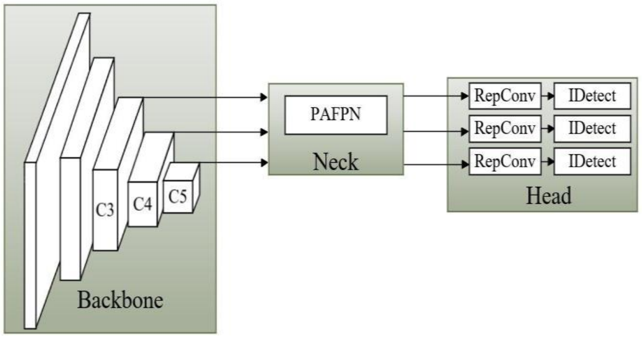


Fig. 1. Simplified diagram of YOLOv7 network structure.

However, YOLOv7 has high memory usage and may not be advantageous for mobile devices or resource-constrained systems. Additionally, the default anchor boxes of YOLOv7 are clustered based on the entire Common Objects in Context (COCO) training set, which may result in significant differences in target sizes and aspect ratios compared with the targets in specific detection scenarios. Accordingly, it is necessary to optimize and improve the YOLOv7 algorithm to better adapt to practical detection tasks and achieve superior detection performance.

B. *Overall Structure of the Optimized YOLOv7 Network*

In the optimized YOLOv7 object detection algorithm, YOLOv7 serves as the baseline model, and optimizations and improvements are introduced in three aspects (i.e., clustering anchor box sizes, SPP structure, and detection head). Fig. 2 presents the overall structure of the optimized YOLOv7. At the preprocessing stage, the Genetic Kmeans (1-IoU) clustering algorithm is proposed in this study to redefine the shape of anchor boxes. The above-described algorithm adopts genetic algorithms to optimize Kmeans clustering while employing IoU distance as a distance metric. Based on this method, the redefined anchor box shapes are more significantly consistent with the custom sample data, such that the detection accuracy can be improved, and the false positives can be reduced. The spatial pyramid structure in the feature fusion network divides the feature map into various groups via group convolution, and each group is then subjected to convolution processes independently. Moreover, the SPPF module with a serial structure is combined with the CSP structure to decrease computational costs and increase the effectiveness of the receptive field. This combination forms the SPPFCSPC_group module, which reduces the number of parameters, accelerates inference speed, and enhances the generalization ability of the model. The head layer incorporates RepConv module and Detect Head. By stacking multiple convolutional layers and sharing weights, the model can enhance its capacity to represent features and better comprehend the target's finer nuances. Moreover, RepConv module can adapt to targets of different scales and shapes. When combined with the Detect Head, it can be applied to feature maps at a wide range of levels, enhancing the model's capacity to recognize targets of all sizes and forms.

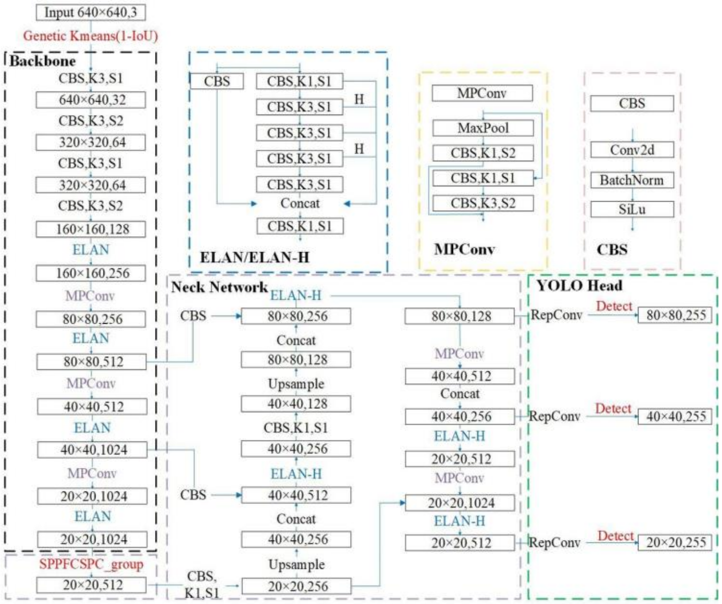


Fig. 2. Overall structure of optimized YOLOv7 network.

C. *Genetic Kmeans (1-IoU) Anchor Box Clustering Algorithms*

At the preprocessing stage of the object detection algorithm, this study proposes Genetic Kmeans (1-IoU) algorithm to recluster the anchor box shapes. Genetic Kmeans (1-IoU) algorithm utilizes genetic algorithms to optimize Kmeans clustering [20]. Following the random initialization of the population and iterative optimization through genetic operations, the problem of local optima is addressed, and clustering quality is improved [21]. Furthermore, under the presence of significant overlap between different scales and categories in the dataset employed in this study, the conventional Kmeans algorithm employs Euclidean distance as the distance measure between sample points without considering the size and overlap of the object bounding boxes. This increases the uncertainty of the model regarding the object bounding boxes. Thus, Genetic Kmeans (1-IoU) algorithm introduces IoU distance, taking into account the separation between the center points and the overlap of the two bounding boxes. To be specific, this algorithm measures the similarity between different categories by calculating the IoU distance between the cluster centers and sample points.

The specific steps of Genetic Kmeans (1-IoU) algorithm are elucidated below:

1) Randomly select k samples as the initial centers of the clusters and randomly initialize the cluster centers. Determine the IoU distance between each sample's location and the center of each cluster, then place the sample in the cluster to which it is closest. The calculation equations are written in Eq. (1) and Eq. (2).

$$d(box, centroid) = 1 - IoU(box, centroid) \tag{1}$$

$$\mathcal{L}_{IoU} = 1 - IoU = 1 - \frac{B \cap B^{gt}}{B \cup B^{gt}} \tag{2}$$

2) Transform the clustering problem into an optimization problem of assessing the objective function, which can be written as:

$$minJ_E = \sum_{j=1}^{c} \sum_{k=1}^{n_j} \| x_k^j - m_j \| \tag{3}$$

$$maxJ_B = \sum_{j=1}^{c} \left(m_j - m\right)^T (m_j - m) \tag{4}$$

where x_k^j represents the *k*-th sample that falls into the class; n_j denotes the number of samples in class *j*; m_j expresses the center of class *cj*, which is determined by Eq. (5); *m* represents the center of all samples, which is written in Eq. (6).

$$m_j = \frac{1}{n_j} \sum_{j=1}^{n_j} x_k^j \tag{5}$$

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{6}$$

3) The clustering performance of the genetic algorithm is assessed using the *fitness* value, as shown in Eq. (7). A higher fitness value indicates a greater likelihood for the individual's genes to be selected for the next generation.

$$fitness = \frac{J_B}{J_E} = \frac{\sum_{j=1}^{c} \left(m_j - m\right)^T (m_j - m)}{\sum_{j=1}^{c} \sum_{k=1}^{n_j} \| x_k^j - m_j \|} \tag{7}$$

D. SPPFCSPC_Group Structure

Based on the SPPCSPC module (as shown in Fig. 3(a)), the SPPFCSPC_group module is designed to perform feature fusion and dimensionality reduction at different scales in the Neck network structure of the improved YOLOv7 algorithm. Fig. 3(b) presents the structure of the SPPFCSPC_group module, comprising a series of group convolutions, SPPF module, and CSP structure. By partially connecting at different stages of the network and cross-linking the features of the earlier stage with the later stage, the SPPFCSPC_group module increases the performance of target recognition and the network's capacity to represent features.

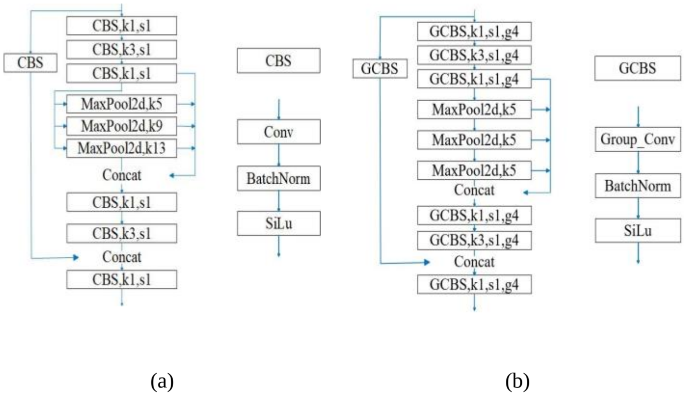


Fig. 3. Space pyramid pooling module (a) Structure of SPPCSPC module (b) Structure of SPPFCSPC_group module.

To be specific, the input image undergoes feature extraction through a series of group convolution layers. After feature extraction, the SPPF module uses group convolutions to execute multi-scale pooling operations to capture broad and specific information at various scales. The input feature map is divided into various scales by the SPPF module, and each scale undergoes a group convolution operation to obtain scale-specific feature representations [22]. The group convolutions concatenate the feature maps from multiple scales, resulting in a feature representation that contains global and local information at different scales. After feature fusion, The CSP module separates the feature map into two parts after feature fusion: one portion directly conducts the subsequent convolution operation; the other half is preprocessed before being fused with the previous component, such that the feature representation capability can be enhanced.

Fig. 4 depicts the structure of group convolution. Eq (8) and (9), respectively, indicate the number of parameters in a single convolution kernel and the total number of parameters in the convolution layer.

$$P_1 = \begin{cases} C_{in} \times K_1 \times K_2, bias = False \\ C_{in} \times K_1 \times K_2 + 1, bias = True \end{cases} \tag{8}$$

$$Parameters = C_{out} \times P_1 \tag{9}$$

where the input is expressed as C_{in}, H_{in}, W_{in}, and D_{in}; the output is denoted as C_{out}, H_{out}, W_{out}, and D_{out}.

Group convolution divides the input feature map into g groups following the channel dimension while applying a regular convolution to the respective group. The number of parameters for group convolution is represented by Eq. (10).

$$Parameters\_Group = \begin{cases} C_{out} \times (\frac{C_{in}}{g} \times K_1 \times K_2),\ bias = False \\ C_{out} \times (\frac{C_{in}}{g} \times K_1 \times K_2 + 1),\ bias = True \end{cases} \tag{10}$$

where $\frac{C_{in}}{g}$ denotes the number of channels in each group of the input feature map, i.e., the number of channels in the respective convolutional kernel. After group convolution is completed, a regular convolution is applied to the respective group. Since the respective group requires at least one convolutional kernel, the output channel number C_{out} for group convolution is at least g. If the respective group covers n convolutional kernels, the output channel number C_{out} is given by $n \times g$ $(n > 1)$, here y expresses the number of groups. In other words, the output channel number C_{out} is a multiple of the number of groups. Accordingly, group convolution requires that the input and output channel numbers be evenly divided by the number of groups g. The reduction in the number of parameters is the fundamental reason behind the decrease in the channel number of the respective convolutional kernel to $1/g$ after group convolution is completed.

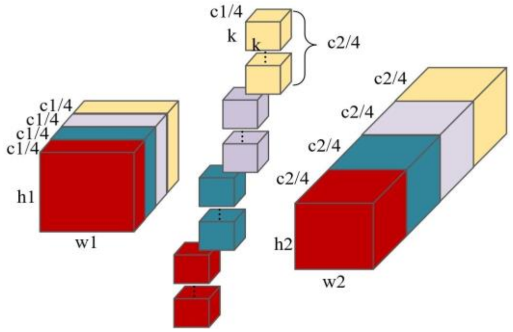


Fig. 4. Structure of group convolution.

E. Optimized YOLOv7 Detection Head

The optimized YOLOv7 algorithm utilizes the Detect Head in the Head layer to obtain prediction results. The feature map transfers the Detect Head by the optimized YOLOv7 model. Fig. 5 depicts the Detect module’s flowchart. The Detect Head utilizes a series of convolutional layers and fully connected layers to predict the position and class of the target. The above-described layers extract features through convolution operations and nonlinear activation functions and map them to the spatial coordinates and class of the target. The optimized YOLOv7 algorithm outputs the prediction results through the Detect Head, along with labels and confidence scores for the target classes. Since no extra classifiers or regressors are necessary because the Detect Head can anticipate the bounding boxes and classes of the targets directly, the model structure can be simplified, computational and memory overhead can be reduced, and the inference speed can be increased. Furthermore, the Detect Head is not dependent on feature vectors to predict the position and size of the targets, such that the bounding boxes and classes can be directly predicted. Consequently, the Detect Head enhances the precision of target localization to a certain extent.

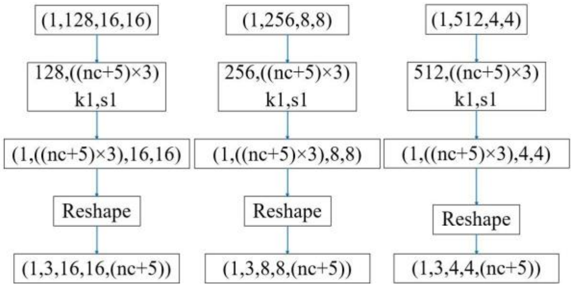


Fig. 5. Flowchart of detection head.

IV. Experimental Validation and Analysis

A. Dataset Preparation

In this study, the performance of the optimized YOLOv7 model for small object detection is assessed using the VisDrone2019 dataset. A total of 2158 samples are randomly selected to generate a custom dataset, with the aim of investigating the detection capabilities of the optimized YOLOv7 algorithm on small objects. A training set and a test set are divided into the custom dataset in 7:3 ratio. The original detection categories are further assigned to 10 classes. For simplicity, new names are assigned to the above-mentioned 10 classes in the experiment. The names and distribution of the target categories are listed in Table I. In the custom dataset, based on the definition of small objects for relative scale, small objects account for approximately 70% of the dataset. Likewise, small objects take up approximately 54% of the dataset, following the definition of small objects for absolute scale.

TABLE I. Corresponding Names and Quantity Distribution of Target Categories

Category	Models	Accuracy (%)
pedestrian	C1	79339
people	C2	27059
bicycle	C3	10480
car	C4	144867
van	C5	24956
truck	C6	12875
tricycle	C7	4812
awning-tricycle	C8	3246
bus	C9	5926
motor	C10	29647

B. Experimental Condition and Assessment Metrics

The experiments are performed on an Alienware X15 R1 laptop with the following hardware specifications: 11th Gen Intel (R) Core (TM) i7-11800H CPU (2.3GHz), 32GB RAM, NVIDIA GeForce RTX 3070 GPU with 8GB VRAM. The experiments are performed using the PyTorch deep learning framework on Windows 11 operating system. The program code is implemented in Python, utilizing libraries (e.g., CUDA, Cudnn, and OpenCV). The above-mentioned setup contributes to the training and testing of tiny item detection on the VisDrone2019 dataset. In the comparative experiments and fusion studies, the input image is configured to be 640 by 640 pixels. 50 total epochs of training are completed, with a batch size of 2. Weight decay is set to 0.0005, momentum is set to 0.937, and the initial learning rate is set to 0.01.

Common assessment metrics in object detection algorithms are employed to objectively evaluate the effectiveness of the detection models. The above-described metrics comprise AP, mean Average Precision (mAP), Number of Parameters (Params), Giga Floating-point Operations Per Second (GFLOPS), as well as FPS.

C. Comparative Analysis of Experimental Results

1) Comparison and analysis of clustering algorithm loss curves: During the training of the YOLOv7 model, the Best Possible Recall (BPR) between the default anchor boxes and each target in the custom dataset is automatically determined by the network. If the BPR falls below 0.98, the detection model uses a combination of genetic algorithm and Kmeans to recluster and generate new anchor boxes, known as Autoanchor. Autoanchor combines genetic algorithm with Kmeans clustering and utilizes Euclidean distance for mutation on the clustering results. The experiment tested three different clustering algorithms: Autoanchor using Euclidean distance, Kmeans using 1-IoU distance, and Genetic Kmeans. Table II displays the predicted anchor box forms for each clustering algorithm at various scales.

TABLE II. ANCHOR BOX SHAPES FOR THREE CLUSTERING ALGORITHMS AT PREDICTED SCALES

Branch	P3	P4	P5
Dimension	80 × 80	40 × 40	20 × 20
Autoanchor	(2,3,3,8,6,5)	(7,14,12,8,21,12)	(13,21,30,23,48,45)
Kmeans (1-IoU)	(2,3,3,7,6,5)	(6,12,11,9,11,20)	(22,12,25,25,47,39)
Genetic Kmeans (1-IoU)	(2,4,3,8,6,6)	(6,12,12,9,11,18)	(25,14,23,28,47,36)

Table III presents a comparison of assessment metrics for a variety of clustering algorithms. The models with optimized anchor box sizes achieve overall improved detection accuracy compared with the baseline network. mAP achieved by the model using Genetic Kmeans (1-IoU) as the clustering algorithm reaches 31.8%, marking improvements of 0.4% and 0.4% compared with Autoanchor and Kmeans (1-IoU), respectively. To be specific, mAP is raised by 0.9% in comparison to the original YOLOv7 algorithm. Furthermore, AP obtained by training the network with Genetic Kmeans (1-IoU) reaches 17.04%, marking improvements of 1.41% and 0.61% over the original YOLOv7 algorithm and Autoanchor, respectively. As revealed by the above-mentioned results, the detection model achieves higher detection accuracy by using 1-IoU distance and improving Kmeans clustering method with genetic algorithms to generate anchor boxes that more effectively match the sizes of detection targets in the sample dataset. In general, compared with the original YOLOv7 model, Genetic Kmeans (1-IoU) achieves the optimal performance.

TABLE III. COMPARISON OF ASSESSMENT METRICS FOR DIFFERENT CLUSTERING ALGORITHMS

Anchor	YOLOv7	Autoanchor	Kmeans (1-IoU)	Genetic Kmeans (1-IoU)
AP (%)	15.63	16.43	17.05	17.04
mAP (%)	30.9	31.4	31.4	31.8
GFLOPS	103.5	103.5	103.5	103.5
FPS	60.61	65.79	65.79	64.93
Params (M)	36.54	36.54	36.54	36.54

The trained models are further validated for loss. Fig. 6 presents the comparison of loss curves for different clustering algorithms. As depicted in Fig. 6, Autoanchor and Kmeans (1-IoU) have slightly higher losses compared with Genetic Kmeans (1-IoU), with average losses of 0.13, 0.1309, and 0.1288, respectively. In contrast, YOLOv7 has the slowest decrease in loss, with a final average loss of 0.1316. The above result suggests that the Genetic Kmeans (1-IoU) algorithm reduces the loss of the original YOLOv7 algorithm by 0.28%.

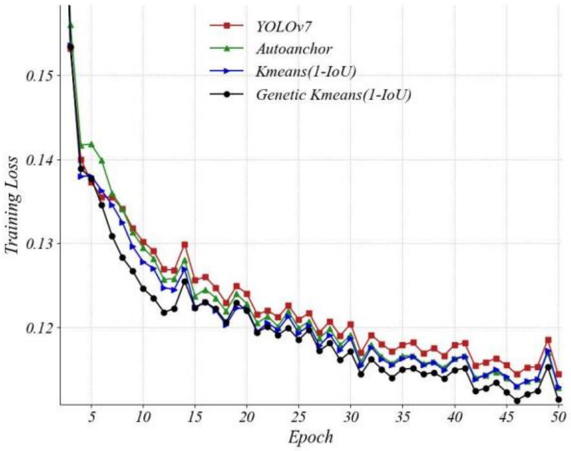


Fig. 6. Comparison of loss curves for different clustering algorithms.

2) Comparison and analysis of assessment metrics for pyramid pooling structure: A fusion experiment is performed on the pyramid pooling module to validate the effectiveness of the SPPFCSPC_group module, which utilizes grouped convolution and the SPPF structure, in small object detection from aerial images captured by drones. Starting with the SPPCSPC module, improvements are made sequentially with grouped convolution and the SPPF structure. Table IV presents the comparison of assessment metrics for the fusion study of the pyramid pooling module. As seen in Table IV, using grouped convolution decreases the module's parameter size by 5.7 M. The SPPFCSPC_group structure achieves an AP value of 15.82%, marking an improvement of 0.19% compared with the SPPCSPC structure. Moreover, FPS is increased by 3.91, validating the performance of the optimized structure for accuracy and speed.

TABLE IV. PERFORMANCE COMPARISON OF ASSESSMENT METRICS IN FUSION STUDY ON PYRAMID POOLING MODULE

Neck	SPPCSPC	SPPCSPC_group	SPPFCSPC_group
AP (%)	15.63	15.26	15.82
mAP (%)	30.9	30.8	30.8
GFLOPS	103.5	99.0	99.0
FPS	60.61	60.98	64.52
Params (M)	36.54	30.84	30.84

To further verify the effect of the SPPFCSPC_group module on the detection accuracy of small object samples in the YOLOv7 model for aerial drone images, experiments are performed to compare five different pyramid pooling structures, i.e., SPP [23], SPPF, Atrous Spatial Pyramid Pooling (ASPP) [24], Receptive Field Block (RFB) [25], and SPPFCSPC_group. Table V presents the comparison of assessment metrics for the comparative study of the pyramid pooling module.

As depicted in Table V, the YOLOv7 baseline network achieves the maximum AP value and mAP value when using the SPPFCSPC_group structure, which is 15.82% and 30.8% respectively. The adoption of group convolution reduces the model parameters to only 30.84 M, resulting in a reduction of 14.61 and 2.44 M compared with the ASPP module and the RFB module that utilize dilated convolution, respectively. As revealed by the above result, while reducing the model parameters, the model maintains a high detection accuracy and avoids weakening the information interaction between different layers by using group convolution. The above-described findings validate the effectiveness of SPPFCSPC_group module.

TABLE V. PERFORMANCE COMPARISON OF ASSESSMENT METRICS IN COMPARATIVE STUDY ON PYRAMID POOLING MODULE

Neck	SPP	SPPF	ASPP	RFB	SPPFCSPC _group
AP (%)	15.26	15.50	15.73	15.78	15.82
mAP (%)	30.6	30.7	30.6	30.5	30.8
GFLOPS	98.7	98.7	110.6	100.9	99.0
FPS	65.79	68.49	64.51	64.94	64.52
Params (M)	30.51	30.51	45.45	33.28	30.84

3) Comparison and analysis of different detection heads: The positive and negative sample allocation strategy of YOLOv7 is designed around the Lead head and the Auxiliary head, combining the positive and negative sample allocation strategies of YOLOv5 and YOLOX. To assess the impact of different detection heads on the model's detection accuracy, a comparative analysis was conducted among YOLOv5, YOLOX, the default Detect Head used in YOLOv7, Decoupled Head, and IDetect Head.

TABLE VI. PERFORMANCE COMPARISON OF DIFFERENT DETECTION HEADS IN DETECTORS

Head	IDetect	Decoupled Head	Detect Head
AP (%)	15.63	16.80	15.73
mAP (%)	30.9	24.7	30.4
GFLOPS	103.5	144.6	103.5
FPS	60.61	54.05	62.50
Params (M)	36.54	44.03	36.54

As depicted in Table VI, the Decoupled Head achieves the minimum overall precision, with a mAP value of only 24.7%, which is significantly lower than the IDetect Head (6.2%) and the Detect Head (5.7%). Besides, the Decoupled Head also has the largest parameter count, exceeding that of IDetect by 7.49 M and Detect by 7.49 M. Furthermore, when using the Detect Head, the overall AP value reaches the maximum point at 15.73%, representing a 0.1% increase compared with IDetect Head, while keeping the model size unchanged.

In order to further observe the effect of different detection heads on the model detection accuracy, the experiments are shown in Fig. 7 as scatter plots of the P-R curves on the VisDrone dataset for detectors using different detection heads. The precision P is represented by the vertical axis, while the recall rate R is represented by the horizontal axis. The area enclosed by the curve and the coordinate axes represents the AP value, where a curve closer to the top right corner indicates a better detection model. As seen in Fig. 7, the recall rate steadily raises but the accuracy declines as the number of epochs rises. Decoupled Head shows a rapid decline in precision when the recall rate reaches 20%, suggesting a lower performance of the detector. The P-R curve of Detect largely envelops the curve of IDetect, demonstrating higher precision and recall. This indicates that Detect has stronger adaptability to different scenes, lighting conditions, and variations in target morphology. Thus, through experimental analysis and comparison, Detect achieves optimal classification performance, making it the preferred detection head for the detection model.

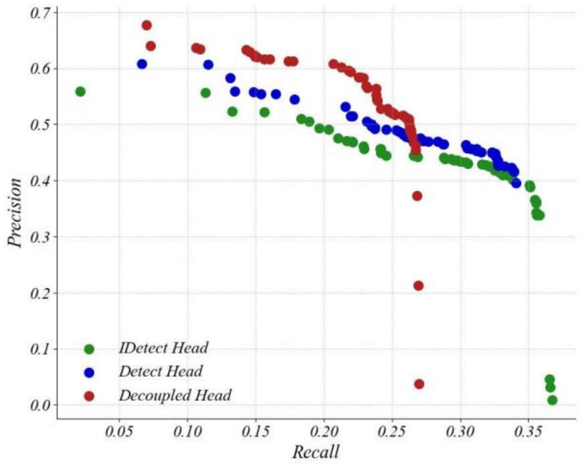


Fig. 7. P-R scatter plot of detectors with different detection heads on the VisDrone dataset.

4) Comparative analysis of fusion studies: To validate the effect of the suggested improvements on the detection model, fusion studies were conducted by testing the components of the improvement method. Table VII compares the results of the fusion studies. The fusion experiments were performed based on the YOLOv7 baseline model, and the improvements were incrementally added to observe their effects on the research objectives and assess their importance. First, IDetect Head was replaced with the Detect Head. Then, Genetic Kmeans (1-IoU) clustering algorithm was added. Finally, the

SPPFCSPC_group module was added on top of the previous modifications.

TABLE VII. COMPARATIVE RESULTS OF FUSION STUDIES

Method	Baseline	+Detect Head	+Genetic Kmeans (1-IoU)	+SPPFCSPC _group
AP (%)	15.63	15.73	15.13	15.81
mAP (%)	30.9	30.4	30.0	30.3
GFLOPS	103.5	103.5	103.5	99.0
FPS	60.61	62.50	59.52	61.73
Params (M)	36.54	36.54	36.54	30.84

Table VII shows that the proposed improvements have achieved performance gains in small object detection from aerial images. First, replacing the Detect Head resulted in higher detection accuracy with a 0.1% increase in AP and a 2.4 FPS improvement in detection speed, suggesting the good performance of the Detect Head in the context of this study. Second, when the Genetic Kmeans (1-IoU) algorithm and SPPFCSPC_group module were added on top of the Detect-based model, AP reached its maximum value at 15.81%, which is an improvement of 0.8% compared with YOLOv7. Additionally, the model's parameter count decreased to 30.84 M, which is a reduction of 5.7 M compared with YOLOv7, while achieving an FPS of 61.73, which is a 1.12 improvement over YOLOv7. The above-mentioned results demonstrate that the model may concentrate on positive anchor boxes of high quality by upgrading the original anchor boxes leading to increased detection accuracy. Moreover, the improvements in the form of group convolution and the SPPF module effectively reduce the model's parameter count while increasing the inference speed, thus conforming to the requirements for real-time detection.

5) Comparison and analysis of different models: Table VIII lists the comparative experimental results of a variety of algorithms. Two lightweight models (i.e., YOLOv5s and YOLOv7-Tiny) are tested and compared through the experiments. As depicted in Table VIII, YOLOv5s achieves a higher AP value than YOLOv7-Tiny by 1.45%, whereas its mAP value is 70.7% lower than that of YOLOv7-Tiny. As indicated by the above results, YOLOv5s exhibits high performance in certain categories, while YOLOv7-Tiny exhibits overall higher performance. Among other larger models, the YOLOR-P6 algorithm achieves the minimum detection accuracy, with an AP value of only 10.21% and an mAP value of 20.9%. Optimized YOLOv7 achieves the maximum AP value, with improvements of 0.89%, 3.63%, 5.6%, and 0.18% compared with YOLOv3-SPP, YOLOv5l, YOLOR-P6, and YOLOv7, respectively. mAP value of Optimized YOLOv7 is 30.3%, with improvements of 1.0%, 4.8%, and 9.4% compared with YOLOv3-SPP, YOLOv5l, and YOLOR-P6, respectively.

TABLE VIII. COMPARATIVE EXPERIMENTAL RESULTS OF DIFFERENT ALGORITHMS

Models	AP (%)	mAP (%)	GFLOPS	FPS	Params (M)
YOLOv3-SPP	14.92	29.3	155.7	54.05	62.61
YOLOv5l	12.18	25.5	114.3	59.52	44.66
YOLOv5s	10.20	16.6	16.4	100.00	7.08
YOLOR-P6	10.21	20.9	80.4	63.29	36.87
YOLOv7-Tiny	8.75	17.3	13.2	70.42	6.04
YOLOv7	15.63	30.9	103.5	60.61	36.54
Optimized YOLOv7	15.81	30.3	99.0	61.73	30.84

To provide a more intuitive comparison of different models in detecting the same samples, the comparison of AP values for the respective category is presented in Fig. 8. For some categories with fewer samples, difficult discrimination, and occlusion, such as 'people', 'truck', and 'bicycle', the improvement in accuracy and speed of the models is limited. However, in general, there is a balance between the reduction in model parameters and the improvement in accuracy.

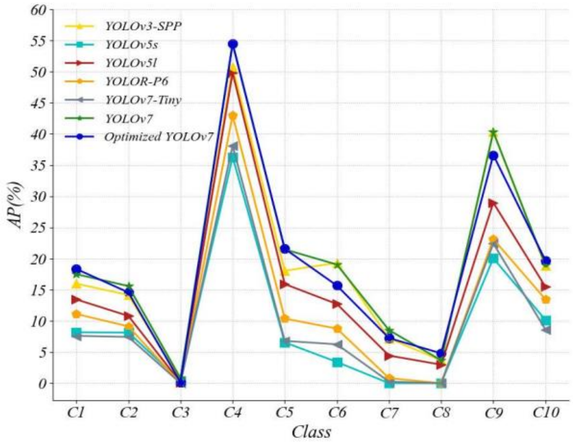
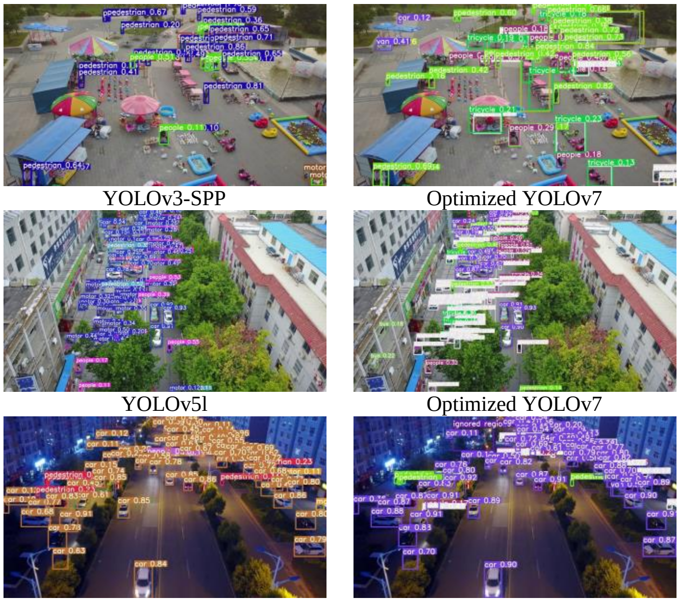


Fig. 8. Comparison of AP values for the respective category.

The detection results of YOLOv3-SPP, YOLOv5l, YOLOR-P6, and the optimized YOLOv7 algorithm are compared for three different scenarios: slight category differences, dense object distribution, and low-light conditions at night. As depicted in Fig. 9, detection algorithms are more prone to false positives and false negatives when there exists slight category differences and dense object distribution. Under low-light conditions at night, the visibility of small objects declines significantly, such that the blurred details and edges are generated, adversely affecting the effective feature extraction of the detection network. In contrast to other algorithms, the optimized YOLOv7 algorithm is effective in mitigating the above described interference factors and demonstrates outstanding detection performance in various scenarios.



YOLOv3-SPP Optimized YOLOv7

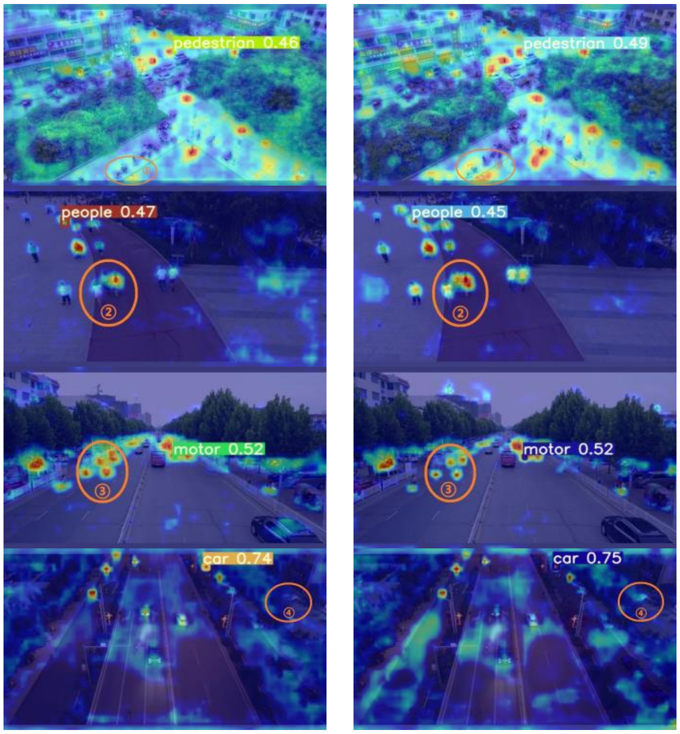
YOLOv5l Optimized YOLOv7

YOLOR-P6 Optimized YOLOv7

Fig. 9. Comparison of test outcomes among various algorithms.

6) GradCAM heatmap visualization analysis: In this study, the heatmaps of the 102nd layer of the detection network are visualized using GradCAM. The highlighted regions in the heatmaps represent the areas that the network considers relevant to the target categories. The 102nd layer represents the P3 branch of the model, i.e., the feature layer specifically developed for small object detection. This visualization presents a more intuitive insight into the network's attention and decision-making process in detecting small objects. In the VisDrone dataset, categories (e.g., 'pedestrian', 'people', and 'motor') are considered small objects for their relatively small sizes. Moreover, 'car' can still be considered a small object category in scenarios with long distances and significant occlusions even it exhibits a larger relative size. During the experiment, feature visualization is performed for the above-mentioned four small object categories, and Fig. 10 presents the visualization of the detection image heatmaps.

As depicted in Fig. 10, the highlighted regions in the heatmaps represent the detected object positions, suggesting that the network can clearly concentrate on small targets. Furthermore, the intensity of colors in the heatmaps represents the degree of network attention. Compared with the YOLOv7 algorithm, the optimized YOLOv7 algorithm exhibits stronger intensity in the highlighted regions (① and ②) when detecting 'pedestrian' and 'people', suggesting that the optimized YOLOv7 algorithm accurately focuses on the target objects while exhibiting a higher level of attention towards small targets. In the visualization image for the 'motor' category, as indicated by the label (③), when a significant overlap exists between 'people' and 'motor', the YOLOv7 algorithm tends to produce false positives. However, the optimized YOLOv7 algorithm displays a more distinctive and accurate highlighting in the area representing the 'motor' target, suggesting improved attention towards the detection targets. For the 'car' category, under a long distance or tree occlusion, the attention intensity of YOLOv7 turns out to be weaker, such that potentially missed detections are conducted. In contrast, the optimized YOLOv7 algorithm achieves the notably enhanced color intensity in the heatmap, marked as (④), suggesting an improvement in detecting small objects that are previously missed.



YOLOv7 Optimized YOLOv7

Fig. 10. Visualization of detection image heatmaps.

As revealed by the experimental analysis, the optimized YOLOv7 algorithm demonstrates significant advantages in accuracy and speed for locating tiny objects in aerial photographs taken by UAVs. In the comparative experiments, the proposed Genetic Kmeans (1-IoU) clustering algorithm allows the model to more effectively cluster the anchor box sizes for small targets. Moreover, the optimized SPPFCSPC_group module, utilizing group convolution, effectively reduces the model parameters. The integration of the SPPF module with the CSP structure enhances both the speed of inference and the precision of detection. Lastly, the use of the Detect Head improves the model's confidence in target detection. The optimized YOLOv7 algorithm is capable of recognizing small-sized objects in UAV aerial images more significantly, even in sophisticated backgrounds. Furthermore, fusion experiments are used to confirm the effectiveness of the proposed methods.

V. Conclusion

In this study, an optimized YOLOv7 algorithm is proposed to address the challenges of detecting small-sized and heavily occluded objects in aerial images captured by UAVs. The proposed method comprises three key steps. At the preprocessing stage, an anchor box clustering algorithm is designed to achieve anchor boxes that better suit the dataset, increasing the accuracy of object detection and reducing the rate of missed detections for small targets. In the feature fusion network, SPP structure based on group convolution is introduced to reduce model parameters and computational complexity. The inference speed of the model is enhanced by adopting a serial pyramid pooling method. Lastly, a detection head that is more tailored to the custom dataset is employed to refine the detection layers. With this method, more accurate detection of small-sized and low-count categories of objects can be achieved. Experimental findings show that compared with the standard YOLOv7, the suggested approach achieves an AP improvement of 0.18%, reduces the model size by 4.5 GFLOPS, decreases the network parameter size by 5.7 million, and increases FPS by 1.12. Accordingly, the proposed method enhances the applicability of the YOLO algorithm for locating tiny targets in aerial photographs that UAVs have recorded.

ACKNOWLEDGMENT

This work was supported by the Scientific Research Project of the Liaoning Provincial Department of Education (No.LJKZ0398, No.LJKZ0423) and the Scientific Research Foundation of Liaoning Petrochemical University (No.2017XJJ-044, No.2017XJJ-012). The authors also would thank reviewers for improving this article.

REFERENCES

[1] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[2] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems, 29:1–9, 2016.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 779–788, 2016.

[5] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7263–7271, 2017.

[6] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. CoRR, abs/1804.02767, 2018.

[7] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. CoRR, abs/2004.10934, 2020.

[8] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, Yiduo Li, Bo Zhang, Yufei Liang, Linyuan Zhou, Xiaoming Xu, Xiangxiang Chu, Xiaoming Wei, and Xiaolin Wei. Yolov6: A single-stage object detection framework for industrial applications, 2022.

[9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. CoRR, abs/2107.08430, 2021.

[10] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks. CoRR, abs/2105.04206, 2021.

[11] Chien Yao Wang, Alexey Bochkovskiy, and Hong Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7464–7475, 2023.

[12] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision– ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37, 2016.

[13] Yuan Zhang, Youpeng Sun, Zheng Wang, and Ying Jiang. Yolov7-rar for urban vehicle detection. Sensors, 23(4), 2023.

[14] Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. Tphyolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 2778–2788, October 2021.

[15] Zechuan Liu and Song Wang. Broken corn detection based on an adjusted yolo with focal loss. IEEE Access, 7:68281–68289, 2019.

[16] Jianfeng Zheng, Hang Wu, Han Zhang, Zhaoqi Wang, and Weiyue Xu. Insulator-defect detection algorithm based on improved yolov7. Sensors, 22(22), 2022.

[17] Yi Pan, Zhao Zhu, Yan Hu, and Qing Wang. Video surveillance vehicle detection method incorporating attention mechanism and yolov5. International Journal of Advanced Computer Science and Applications, 14(6), 2023.

[18] Peirong Wu, Airong Liu, Jiyang Fu, Xijun Ye, and Yinghao Zhao. Autonomous surface crack identification of concrete structures based on an improved one-stage object detection algorithm. Engineering Structures, 272:114962, 2022.

[19] Liangquan Jia, Tao Wang, Yi Chen, Ying Zang, Xiangge Li, Haojie Shi, and Lu Gao. Mobilenet-ca-yolo: An improved yolov7 based on the mobilenetv3 and attention mechanism for rice pests and diseases detection. Agriculture, 13(7), 2023.

[20] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[21] Lambora A, Gupta K, Chopra K. Genetic algorithm-A literature review[C]//2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon). IEEE, 2019: 380-384. Yuxin Wu and Kaiming He. Group normalization. CoRR, abs/1803.08494, 2018.

[22] Wu Y, He K. Group normalization[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9):1904–1916, 2015.

[24] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. CoRR, abs/1706.05587, 2017.

[25] Songtao Liu, Di Huang, and Yunhong Wang. Receptive field block net for accurate and fast object detection. CoRR, abs/1711.07767, 2017.