International Journal of Innovative Computing, Information and Control
Volume **21**, Number **1**, February **2025**

 ISSN 1349-4198
pp. 135–152

DEEP LEARNING-BASED ALGORITHM FOR COMPLEX SMALL TARGET DETECTION IN UAV AERIAL IMAGES

Shuangyuan Li[1,*], Jianglong Lin[2], Yanchang Lv[2] and Tianyu Li[2]

[1]Information Center
[2]School of Information and Control Engineering
Jilin Institute of Chemical Technology
No. 45, Chengde Street, Longtan District, Jilin 132022, P. R. China
{ linjianglong; lvyanchang; litianyu }@jlict.edu.cn
*Corresponding author: lsy@jlict.edu.cn

Received May 2024; revised September 2024

Abstract. *In recent years, with the rapid development of drone technology and object detection algorithms, drone-based object detection has found widespread applications across multiple domains. However, current algorithms often face challenges such as missed detections and false positives when processing low-resolution images and small objects. To address these challenges, this paper introduces TE-YOLOv5s, a target detection algorithm based on YOLOv5s specifically designed for UAV (Unmanned Aerial Vehicle) aerial images. Firstly, a four-scale feature fusion structure is proposed based on the original model framework of YOLOv5s in this paper. Additionally, a detection layer for diminutive targets is incorporated in order to enhance the model's capacity to identify diminutive targets effectively. Secondly, the backbone architecture of the YOLOv5s model is reconfigured, and the Transformer encoder is integrated into the C3 module to augment the model's feature extraction capability for various local information. Finally, the loss function of the model is enhanced, and the Focal-EIOU Loss function is employed to enhance both the convergence speed and the accuracy of the regression results. Through training, validating, and testing on a custom dataset, the enhanced detection model outperforms the original model in detecting diminutive targets within complex images, achieving improvements of 6.1% and 7.2% in Precision and mAP@0.5, respectively. The model exhibits superior performance in detecting diminutive targets within complex images, making it more suitable for UAV deployment and application.*
Keywords: UAV, YOLOv5s, Target detection, Loss function

1. **Introduction.** In recent times, UAVs have experienced widespread adoption in both the military sector and the civilian market due to their cost-effectiveness, notable flexibility, heightened efficiency, and rapid response capabilities. With the swift ascent of artificial intelligence, UAV aerial complex image detection technology has been found extensively applied in civilian sectors, such as circuit inspection, forest fire prevention, intelligent transportation, crop monitoring, and military fields such as personnel search and rescue, target detection, and anti-narcotics reconnaissance. In these applications, the identification of diminutive targets in aerial imagery with complex backgrounds has become an important and challenging task. Contemporary approaches utilized for identification of targets within an image are primarily divided into two classifications: conventional feature-based methods and deep learning network-based methods. Conventional target detection methods cover the two directions of feature detection and segmentation detection. However, the formulation and choice of characteristics in these methodologies heavily depend on a priori conditions, resulting in varying degrees of constraints on their

DOI: 10.24507/ijicic.21.01.135

accuracy, objectivity, robustness, and generalizability. Meanwhile, traditional target detection methods usually adopt a sliding window strategy, which leads to long computation time, low efficiency, and low precision when dealing with complex scenes [1,2]. Obviously, traditional methods cannot meet the demand for target detection.

In the swift evolution of convolutional neural networks, deep learning-based target identification algorithms are widely used because they show excellent detection performance, and it is primarily classified into single-stage and two-stage algorithms [3]. Two-stage algorithms mainly include R-CNN, Fast R-CNN, Faster R-CNN, Cascade R-CNN, etc., which use heuristics or CNN (Convolutional Neural Network) to form region suggestions, and then classify and localize the candidate regions for prediction [4-6]. The main single-stage algorithms are SSD (Single Shot MultiBox Detector), YOLO (You Only Look Once), RetinaNet, EfficientDet, etc., which directly generate anchor frames to forecast the classification and localization of the objective through the CNN, and complete the candidate region without performing the end-to-end target detection [7-10]. The two-stage algorithm has outstanding average precision, low false detection rate and leakage rate, but its algorithm model structure is complex, large computation, poor real-time, slow detection speed, and difficulty in meeting the demand for real-time UAV target identification [11]. While the single-stage target detection model has a simple structure, fast detection speed, low complexity, and its real-time performance which is more suitable for UAV target detection, among the single-stage target identification algorithms, YOLOv5s target identification algorithm is widely used because of its high real-time performance, fast detection speed, and easy optimization and deployment. In addition, we have conducted an in-depth analysis of the YOLOv5s architecture, identifying its weaknesses in handling low-resolution images, small object detection, and specific scenarios. Through our experiments, we discovered that the optimized YOLOv5s demonstrates unique advantages and better adaptability in certain scenarios (such as low-altitude drone imagery and hazy environments). Furthermore, YOLOv5s typically exhibits superior performance and efficiency on resource-constrained devices (such as embedded systems in drones and edge devices). Therefore, in this paper, YOLOv5s is used as a benchmark model for optimization and innovation to make it more suitable for mobile devices such as UAVs.

In existing two-stage and single-stage algorithms, most approaches face challenges in balancing detection speed and accuracy, particularly when detecting complex images such as those captured by drones. This often results in reduced detection performance. Moreover, most of the existing YOLOv5s models are designed for natural scene images, and there are obvious differences between natural scene images and UAV aerial images. 1) UAVs fly at a high altitude and shoot at a high angle, resulting in a substantial quantity of diminutive targets within the aerial imagery. 2) The contexts of UAV aerial images are intricate and changeable, and the dense diminutive targets cover each other and are easy to be interfered by the environment. 3) The drone's movement speed changes drastically, and the captured images are blurred. Therefore, the existing YOLOv5s detection model will still have the problem of missed detection and false detection when applied to the identification of targets in aerial images [12]. In order to enhance the detection efficacy of the YOLOv5s algorithm, and reduce the rate of leakage and false detection, in this study, we propose a target detection algorithm, TE-YOLOv5s, built upon the YOLOv5s framework. The primary contributions of this work are outlined as follows.

1) A four-scale detection model is proposed, a 160×160 diminutive target detection head is designed to increase the fusion of the algorithm's deep semantic and shallow semantic information, and the anchor frame scale, more fitting for intricate UAV aerial images, is determined through the application of K-means algorithm reclustering. This process significantly improves the algorithm's efficacy in detecting diminutive target objects.

2) To enhance the algorithm's capability in extracting feature information from the deep network and to reduce the interference caused by the complex and variable backgrounds in drone aerial images, this paper embeds the Transformer encoder module within the C3 module as a way to strengthen the correlation between the pixel blocks and improve the model's detection precision and robustness against interference.

3) The Focal-EIOU Loss is employed to tackle the issue of substantial errors associated with CIOU Loss in the regression calculation of prediction frames, to enhance the detection accuracy of the model for blurry images, optimize the model's robustness in detecting diminutive targets within complex background images, and simultaneously to improve both the convergence speed and regression precision of the model.

4) The refined algorithm undergoes training, validation, and testing procedures on a tailored proprietary dataset, and the experimental outcomes demonstrate that the enhancement of the detection model effectively optimizes the performance of the UAV aerial intricate image target detection algorithm, particularly in detecting complex diminutive targets. This improvement aligns the algorithm more closely with real-world requirements.

The subsequent sections of this paper are organized as follows: Section 2 furnishes an overview of current research in the relevant field; Section 3 details the framework of YOLOv5s, the enhanced YOLOv5s model, and the improvement methodology; Section 4 delineates the experimental methodology and scrutinizes the experimental findings; ultimately, Section 5 concludes the manuscript and delineates future prospects.

2. **Related Research.** In recent years, to enhance the accuracy and efficiency of algorithms aimed at detecting intricate diminutive targets, and to minimize both the occurrence of false positives and the false detection rate in the model pertaining to intricate diminutive targets, numerous researchers have put forth various optimization approaches for both two-stage and single-stage target identification methodologies, which have gained widespread adoption. In the enhancement of two-stage methods, Ding et al. incorporated DConv (Deformable Convolution) into Faster R-CNN, optimizing the algorithm's effectiveness to comprehend irregular geometric features, but DConv increases the computational amount of the model [13]. Wang et al. enhanced the framework of the Faster R-CNN network by adding a backbone network structure, increased the quantity of output feature images of the backbone network, and then increased the number of anchor frames and adjusted the anchor frame parameters by analyzing the histogram distribution of the target to be detected, thereby enhancing the model's performance in extracting features from intricate small targets; however, this rendered the model framework more intricate and resulted in a reduction in the model's identification speed [14]. Huang et al. designed corresponding detection heads for different target classes based on Cascade R-CNN, which increased its ability to extract edge frames and extraction precision, and improved the reliability of the model's identification results, but the model's detection head possesses a sophisticated framework and an extensive quantity of parameters [15].

For the improvement of single-stage methods, Liu et al. introduced the CBSSD method for target identification, which added the ResNet-50 network as a secondary backbone to the SSD backbone network, so that the model retains richer semantic information, but it leads to the model backbone structure is too complex [16]. Javed et al. utilized the channel pruning and the depth-wise separable convolution to improve the Tiny-YOLOv2 detector, which diminished the model parameters and increased the running speed, but the depth separable convolution reduced the precision of model identification [17,18]. Liu et al. used the DenseNet-121 network as the backbone of the model based on the RetinaNet model, which improved the precision of model identification, but the structure of the backbone network of the model was redundant [19]. Yang et al. improved the original unidirectional

information flow feature fusion network in YOLOv3 into a bidirectional fusion network to enhance the recognition precision of the YOLOv3 network for diminutive targets, and introduced the EIOU Loss, but the bidirectional fusion network increased the computation amount of the model [20]. Yu et al. incorporated the concept of Focal Loss to fine-tune the loss function within the framework of YOLOv4, and used the pruning algorithm to simplify the network structure; however, the pruning algorithm diminishes the precision of model identification [21]. Zhu et al. proposed the UavTinyDet detection network based on YOLOv5, which boosts the algorithm's performances to detection diminutive targets, but the UavTinyDet identification network necks the EFP (Expanded Feature Pyramid) module increases the complexity of the model neck structure [22].

In conclusion, while existing target identification models have enhanced the ability to identify complex diminutive targets, they are still insufficient, and the existing network framework is still difficult to coordinate the detection speed and accuracy, particularly for detecting intricate diminutive targets from the viewpoint of UAVs. Thus, there is a requirement for more extensive research and improvements to boost the speed and precision of the algorithm for identifying intricate diminutive targets in UAV aerial images. It is crucial to align it more effectively with current societal and life requirements.

3. Improved Network Model for YOLOv5s.

3.1. YOLOv5s network model overview.

YOLOv5 combines the advantages of many algorithms, which effectively balances identification precision and processing speed, with high real-time performance, and is widely used. YOLOv5 has five versions, n, s, m, l, and x, and the overall framework of its network is exactly the same [23]. Considering that UAV target detection requires high detection speed in practical applications, and the model complexity of YOLOv5s is low and the operation speed is fast, therefore, this paper carries out a research on YOLOv5s algorithm.

The framework of YOLOv5s network is partitioned into three components, Backbone, Neck and Head, and uses a regression-based approach to accomplish the target detection task quickly and efficiently [24]. Among them, Backbone downsamples the input image with different magnification and performs the features extracted from the image; Neck fuses the features of different layers through FPN and PAN; Head designs three detection layers with different scales, which are employed to predict categories and perform bounding box regression for large, medium, and small targets, respectively [25,26]. Furthermore, the YOLOv5 algorithm provides image preprocessing operations and many data enhancement methods, such as Mosaic, Random affine transform, and MixUp. Various features like Mixed Accuracy Training, Genetic Hyperparameter Evolution, and Label Smoothing Processing are also provided for model training, inference, and deployment, which enable YOLOv5s models to have higher detection performance [27,28].

3.2. Enhanced version of the YOLOv5s model.

In this research, the TE-YOLOv5s target identification algorithm is formulated, building upon the YOLOv5s algorithm from Ultralytics version 6.0. Firstly, a diminutive detection head of size 160×160 is incorporated, while preserving the original three identification scales, this modification aims to enhance the identification performance for diminutive targets. Additionally, a Transformer encoder is integrated into the C3 module of the Backbone section to enhance both the model's feature extraction capacity and computational efficiency. Finally, the adoption of the Focal-EIOU Loss, designed particularly for diminutive target identification, aims to enhance the model's prediction precision while reducing both leakage and the false detection rate. The framework of the TE-YOLOv5s model is depicted in Figure 1.

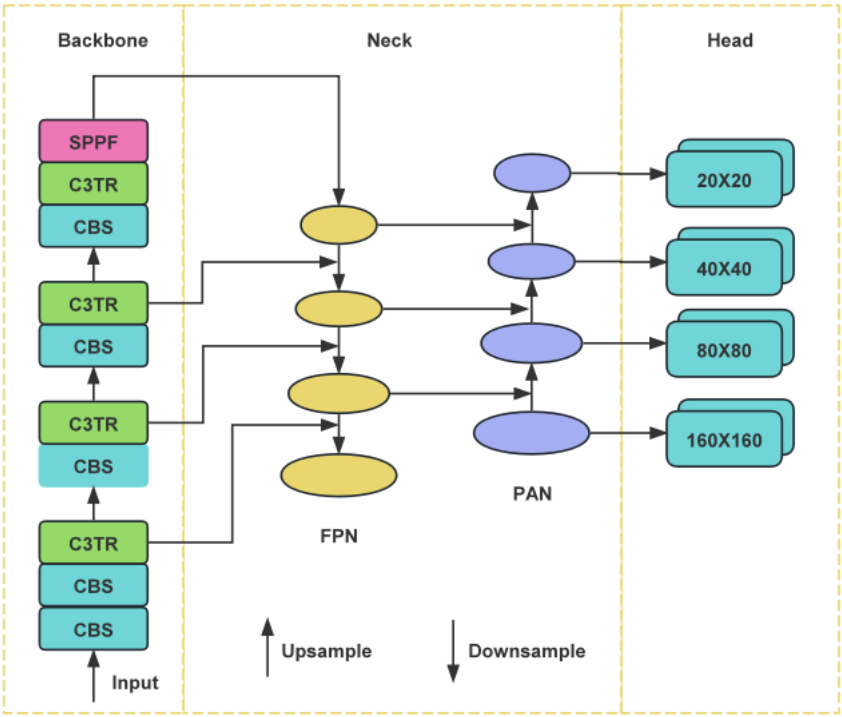


FIGURE 1. TE-YOLOv5s network framework

3.2.1. *Diminutive target detection layer.* The primordial YOLOv5s algorithm downsamples the input image 8 times, 16 times and 32 times respectively to generate predictive feature maps of 20×20, 40×40 and 80×80 to detect large, medium and small targets [29]. While there are more diminutive targets in the drone aerial images, the scale percentage is very small, after many times of downsampling, the target information within the feature images may be partially lost, resulting in sparsity, and the previous three heads can no longer effectively detect diminutive targets, which is prone to the phenomenon of missed detection. Therefore, a new very small target detection layer with a scale of 160×160 is added on top of the original 3 detection scales of the model, and the feature map of this layer can retain more positional data and elaborate characteristics of diminutive targets, enhancing the model's identification capability for diminutive targets. In addition, YOLOv5s is an anchor frame-based algorithm, and its original anchor frame scale directly affects the identification capability of the algorithm, while the original anchor frame of YOLOv5s is generated by clustering on the COCO2017 dataset, which is not applicable to the complex diminutive target identification under the viewpoint of the UAV. Therefore, to further enhance the detection performance of the algorithm, this paper regenerates the 12 anchor frame scales applicable to the small targets through the K-means++ clustering algorithm. In this algorithm, the Euclidean distance between each object and the cluster center is used as a measure of similarity, by assigning each object to the most similar cluster, and recalculating the cluster centers iteratively, this process continues until the results no longer change. The mathematical expression for the Euclidean distance is given in Equation (1).

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

where d denotes the distance; IOU represents the Intersection over Union; box denotes the current anchor box; $centroid$ indicates the cluster center; $d(box, centroid)$ represents the distance from the current anchor box to the cluster center; and $IOU(box, centroid)$ is

TABLE 1. The brief computation process of the K-means++ algorithm

The brief computation process of the K-means++ algorithm
Input: Data points $X = \{x_1, x_2, \ldots, x_n\}$, number of clusters K
Output: K cluster centers
1: Initialize cluster centers:
Randomly select the first cluster center μ_1 from the data points X.
For each data point $x_i \in X$, compute the distance $D(x_i)$ to the nearest cluster center already chosen
2: Select the remaining $K-1$ cluster centers:
Repeat until K cluster centers are chosen:
Calculate the probability for each data point x_i to be chosen as the next center:
$p(x_i) = D(x_i)^2/\Sigma D(x_j)^2$, where $D(x_j)$ is the distance of data point x_j to its nearest center.
Choose the next cluster center μ_i based on the probability distribution $p(x_i)$.
3: Run the standard K-means algorithm:
Repeat until convergence:
Assign each data point x_i to the nearest cluster center μ_j.
Update each cluster center μ_j as the mean of all data points assigned to it.
4: Return the final set of K cluster centers $\{\mu_1, \mu_2, \ldots, \mu_K\}$.

TABLE 2. Anchor frame scales

Feature drawing	20×20	40×40	80×80	160×160
Initial anchor frame scale	$(116, 90)$ $(156, 198)$ $(373, 326)$	$(30, 61)$ $(62, 45)$ $(59, 119)$	$(10, 13)$ $(16, 30)$ $(33, 23)$	$(5, 13)$ $(6, 9)$ $(9, 12)$
Improved anchor frame scale	$(28, 73)$ $(60, 35)$ $(79, 66)$	$(21, 11)$ $(17, 21)$ $(34, 81)$	$(6, 12)$ $(11, 9)$ $(10, 7)$	$(3, 4)$ $(4, 8)$ $(7, 6)$

the Intersection over Union between the current anchor box and the cluster center. Table 1 outlines the brief computation process of the K-means++ algorithm.

In this study, a re-clustering analysis using the K-means++ algorithm resulted in anchor box scales that are better suited for drone aerial imagery, as shown in Table 2.

3.2.2. *Improvement of the C3 module.* The Transformer encoder structure offers notable advantages in terms of substantial feature extraction capability and high operational efficiency, this paper takes YOLOv5s as the base model, and introduces Transformer encoder into C3 module to improve it into C3TR module. The core of Transformer encoder lies in the attention value calculation and feature weighted aggregation, through the attention value calculation to obtain the relationship between two pixels, and then use the attention value as a weighting guide to aggregate other pixel features, the final effect makes the characterization of the feature is more significant, and the amount of information is more. The Transformer encoder contains two important modules: the Multi-Head Attention Mechanism and the Fully Connected Layer MLP. Among them, the Multi-Head Attention Mechanism not only suppresses the useless information and enables the module to focus on the current feature, but also obtains the correlation between the front and

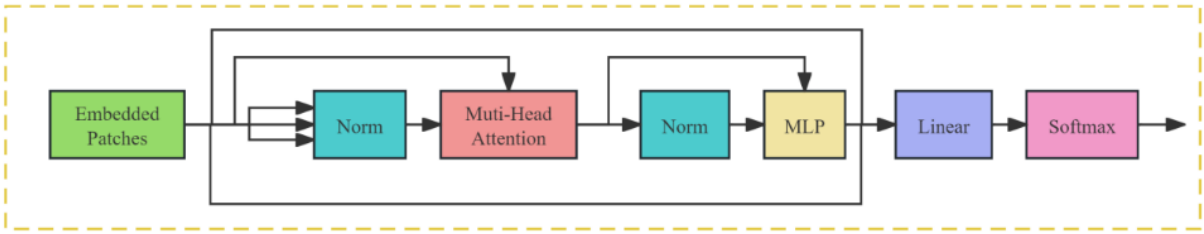


FIGURE 2. Transformer encoder structures

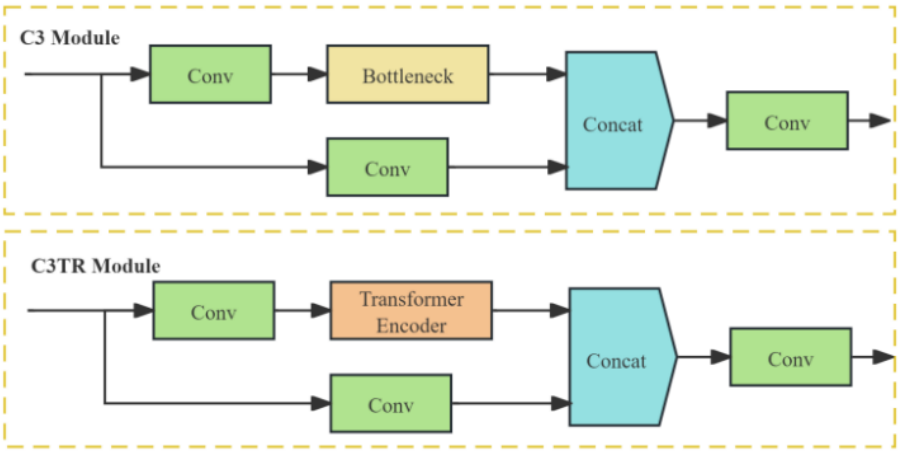


FIGURE 3. C3 module and C3TR module

back pixel blocks, while the Normalization Layer and Dropout Layers optimize the convergence speed of the network to prevent the network from overfitting [30]. The framework of Transformer encoder is illustrated in Figure 2.

The prototype C3 module is the BottleneckCSP module, which is architected with the CSP structure. The C3TR module is still architected with the CSP framework and contains three standard convolutional layers. The original bottleneck module is replaced with the Transformer encoder module, which allows it to have better detection performance while having a lower model complexity. The C3 module and the C3TR module are shown in Figure 3.

3.2.3. *Optimizing the loss function.* The YOLOv5s model utilizes the CIOU Loss to calculate bounding box loss, comprising three primary elements: Loss of bounding box position (L_{bbox}), Loss of objectness (L_{obj}), and Loss of classification (L_{cls}) [31]. CIOU Loss is calculated as shown in Equation (2).

$$L = L_{bbox} + L_{obj} + L_{cls} \tag{2}$$

CIOU Loss calculates the overlapping area, center distance and aspect ratio of the bounding box at the same time, enhancing both steadiness and convergence speed of model training. However, the true difference between the height and width of the bounding box and its confidence level is not fully characterized, which makes the model regression prediction results less accurate. For the problem of CIOU Loss, the widths and heights of the bounding box are considered separately, and Focal-EIOU Loss is employed, representing a fusion of Focal Loss and EIOU Loss. Focal Loss splits the widths and heights of the bounding box, and makes the difference with the widths and heights of the smallest external frames, respectively. EIOU Loss makes the model have a faster convergence speed and more accurate localization result by reducing the differences in the widths and

heights between the bounding box and the real frames. EIOU Loss is calculated as shown in Equations (3)-(6).

$$L_{EIOU} = L_{IOU} + L_{dis} + L_{asp} \tag{3}$$

$$L_{IOU} = 1 - IOU \tag{4}$$

$$L_{dis} = \frac{\rho^2\left(p, p^{gt}\right)}{c^2} \tag{5}$$

$$L_{asp} = \frac{\rho^2\left(w, w^{gt}\right)}{c_w^2} + \frac{\rho^2\left(h, h^{gt}\right)}{c_h^2} \tag{6}$$

where L_{IOU}, L_{dis} and L_{asp} are IOU loss, distance loss, and height and width loss, respectively. $\rho^2\left(p, p^{gt}\right)$ is the squared Euclidean distance between the center points of the predicted box p and the ground truth box p^{gt}, and c^2 is the squared length of the diagonal of the smallest enclosing box that contains both the predicted and ground truth boxes. $\rho^2\left(w, w^{gt}\right)$ represents the squared Euclidean distance between the widths of the predicted w and ground truth boxes w^{gt}. c_w^2 is the normalization parameter for the width difference between the predicted and ground truth boxes. $\rho^2\left(h, h^{gt}\right)$ is the squared Euclidean distance between the heights of the predicted h and ground truth boxes h^{gt}, and c_h^2 is the normalization parameter for the height difference between the predicted and ground truth boxes.

In a sample image, the count of anchor frames exhibiting minor regression errors is significantly lower than those with substantial errors, and anchor frames with lower quality will produce larger gradients, which will directly influence the training performance of the model. Thus, through the integration of Focal Loss based on EIOU Loss, high-quality anchor frames can be separated from low-quality anchor frames, and Focal-EIOU Loss is computed as illustrated in Equation (7).

$$L_{Focal\text{-}EIOU} = IOU^{\gamma} L_{EIOU} \tag{7}$$

where γ is the hyperparameter employed to regulate the curvature of the curve.

Focal-EIOU Loss diminishes the weights of easily-regressible samples, directing the model's attention towards samples with low overlap between the predicted and true frames. This strategy enhances the regression accuracy of the model, optimizing its capability in identifying intricate diminutive targets.

4. **Experimental Verification and Analysis.**

4.1. **Preparation of dataset and experimental environment.**

To substantiate the efficacy of the TE-YOLOv5s model, this study compares the VisDrone 2019 dataset, DOTA dataset, and UAVDT dataset, from which 2,590 aerial images in complex scenes such as low-resolution, complex background, and target-intensive are selected to produce the proprietary dataset required in this experiment, and data enhancement is carried out by means of flipping, translating, and splicing, to raise the model's robustness. Among the most commonly used public datasets in drone vision research are VisDrone 2019, DOTA, and UAVDT. These datasets are meticulously designed to cover a wide range of scenes with varying complexities, including urban areas, rural environments, highways, ports, and more. Additionally, to increase the complexity of the dataset in this study, the authors manually selected images with varying weather conditions, different lighting changes, multiple viewing angles, and multi-scale objects. This further enhanced the dataset's complexity, making it more representative of the challenging aerial images captured by drones as studied in this paper. The dataset consists of 1,942 images in the training set, 165 images in the validation set, and 483 images in the test set. It encompasses 10 label

categories, with a total of 103,995 annotations. Figure 4 shows the distribution of the training set structure. In Figure 4, it is apparent that the training set targets are rich in variety and labeled densely, and higher concentration of data points is observed in the lower-left corner of the width-height coordinate graph, or near the origin, which indicates that the aspect ratio of the targets in the dataset is less than one-tenth of that of the initial images. It aligns with the specifications for the complex and diminutive targets essential for this paper.

In this paper, we train and validate the dataset in a specific environment, and the environment configuration and hyper-parameters of the experiment are shown in Table 3.

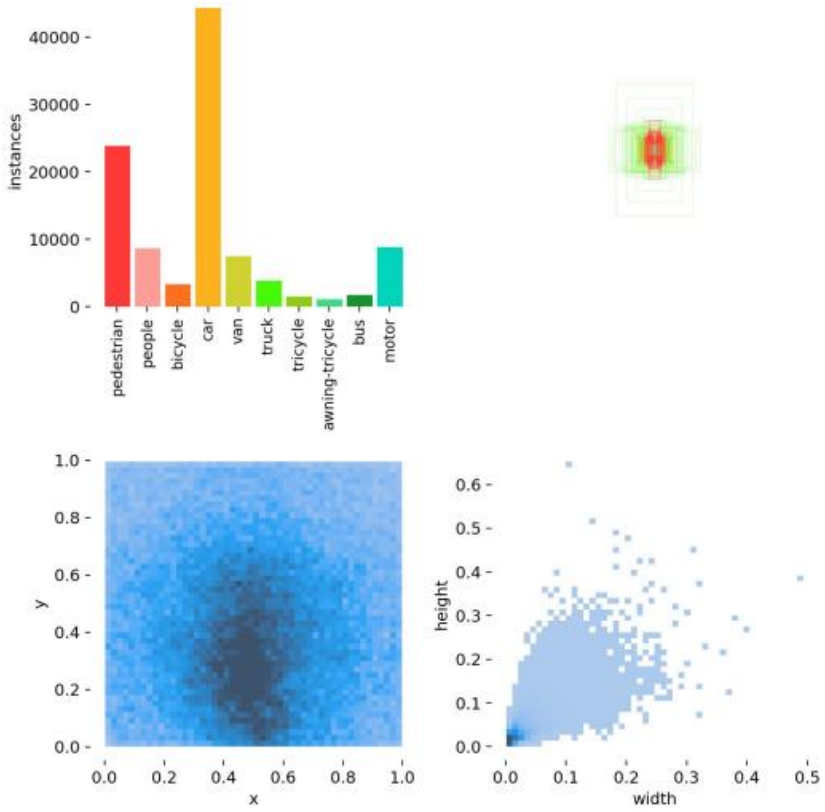


FIGURE 4. Distribution of training set structure

TABLE 3. Experimental environment and hyperparameter configuration

Name	Information
Operating system	Ubuntu 18.04
Language	Python3.8
Deep learning framework	PyTorch 1.10, CUDA 11.8
GPU	NVIDIA GeForce RTX 4090, with 24G video memory and 64G memory
Image size	640 × 640
Training batch	16
The learning rate	0.01
Epochs	200

4.2. **Evaluation indicators.** When models detect complex small targets, false identification and missed identification usually occur. Therefore, to assess the accuracy and efficiency of a model's identification capability, P (Precision), R (Recall) and mAP@0.5 are usually used as evaluation indicators [32]. The mAp (mean of Average Precision) indicator combines different categories of P (Precision) and R (Recall) and is a more comprehensive evaluation indicator.

mAP: mAP@0.5 is the average detection precision of all categories when the IOU threshold is 0.5. In target detection, a higher mAP value signifies a more effective identification capability of the model. The calculation method for mAP is depicted in Equations (8) and (9).

$$mAP = \frac{\sum_{i=1}^{c} AP_i}{c} \tag{8}$$

$$AP = \int_0^1 P(R)dR \tag{9}$$

where AP_i is the average precision of a class and c is the quantity of all categories in the dataset.

Precision: the ratio of the quantity of true targets to the total quantity of detected targets among all detected targets. P is calculated as depicted in Equation (10).

$$P = \frac{TP}{TP + FP} \tag{10}$$

Among them, TP represents true positives, indicating the quantity of targets correctly identified in the detection results; FP represents false positives, indicating the quantity of targets incorrectly identified in the detection results.

Recall: the ratio of the quantity of identified targets to the total quantity of true targets among all true targets. R is calculated as shown in Equation (11).

$$R = \frac{TP}{TP + FN} \tag{11}$$

Among them, FN represents false counterexamples, that is, the number of targets that are not identified among the correct targets.

4.3. **Analysis of ablation experiments.** To assess the influence of each enhancement on the benchmark algorithm, this paper sets up eight sets of experiments. The outcomes of the ablation experiments are presented in Table 4.

As evident in Table 4, the addition of a diminutive target detection head to YOLOv5s results in slight enhancements in precision, recall rate, and mAP@0.5 compared to the

Table 4. Ablation experiments

Method	Detect head	C3TR module	Focal-EIOU Loss	P	R	mAP@0.5
YOLOv5s				0.390	0.229	0.216
Optimized model 1	√			0.406	0.241	0.239
Optimized model 2		√		0.439	0.265	0.266
Optimized model 3			√	0.422	0.249	0.251
Optimized model 4	√	√		0.435	0.261	0.262
Optimized model 5	√		√	0.428	0.259	0.257
Optimized model 6		√	√	0.441	0.268	0.269
TE-YOLOv5s	√	√	√	0.451	0.279	0.288

baseline YOLOv5s model, indicating that the model has a smaller improvement effect. When integrating the improved C3TR module into YOLOv5s, there is a noteworthy increase in model precision (4.9%), recall rate (3.6%), and mAP@0.5 (5%) compared to YOLOv5s alone, this demonstrates the module's effective enhancement of the model's identification precision and overall performance. When Focal-EIOU Loss is used in YOLOv5s, compared with YOLOv5s, the module still has better performance improvement effect; Upon integrating all three enhanced modules into YOLOv5s simultaneously, there is a substantial improvement in model precision (6.1%), recall rate (5%), and mAP@0.5 (7.2%) compared to YOLOv5s alone. The model performance improvement effect was obvious. This experiment effectively verified that the improved module has performance impact of YOLOv5s.

4.4. **Analysis of model comparison experiments.** To validate the superiority of TE-YOLOv5s algorithm, the more popular algorithmic models in the domain of deep learning are selected for comparison experiments. The experimental results are presented in Table 5.

Table 5. Model comparison results

Models	P	R	mAP@0.5	mAP@0.5:0.95	FPS	GFLOPs	Params/M
Faster R-CNN	0.397	0.219	0.211	0.157	16.9	89.2	51.9
SSD	0.216	0.227	0.144	0.076	17.9	103.1	23.5
YOLOv3-tiny	0.321	0.150	0.112	0.049	21.5	13.5	8.70
YOLOv3	0.399	0.241	0.225	0.162	49.7	156.2	62.68
YOLOv4	0.370	0.206	0.195	0.147	56.2	112.8	47.2
YOLOv5s	0.390	0.229	0.216	0.139	156.9	16.2	7.21
YOLOv5n	0.315	0.139	0.159	0.097	127.2	4.7	1.89
YOLOv8n	0.401	0.245	0.221	0.143	105.1	8.9	3.11
YOLOv8s	0.442	0.268	0.276	0.167	136.7	29.1	12.15
CBAM-YOLOv5s	0.393	0.231	0.227	0.140	129.4	18.7	8.29
BAM-YOLOv5s	0.396	0.236	0.236	0.143	141.1	17.9	7.8
ECA-YOLOv5s	0.411	0.239	0.245	0.148	149.3	16.5	7.32
TE-YOLOv5s	0.451	0.279	0.288	0.185	147.2	28.4	7.99

As observed in Table 5, the precision of TE-YOLOv5s is 23.5% and 13% higher than SSD and YOLOv3-tiny, respectively, and the model performance is significantly improved, indicating an increase of 5.4%, 8.1%, 6.1%, 5.0%, 5.8%, 5.5% and 4.0% compared to Faster R-CNN, YOLOv4, YOLOv5s, YOLOv8n, CBAM-YOLOv5s, BAM-YOLOv5s and ECA-YOLOv5s, respectively, and the accuracy is higher, with a lower false detection rate. Meanwhile, their recall rates are all higher than the commonly used algorithmic models in the table, indicating that the model has a lower leakage rate. The mAP@0.5 of TE-YOLOv5s surpasses that of SSD and YOLOv3-tiny by 14.4% and 17.6%, respectively, the model performance has been significantly improved, it is improved by 7.7%, 9.3%, 7.2%, 6.7%, 6.1%. 5.2% and 4.3% than Faster R-CNN, YOLOv4, YOLOv5s, and YOLOv8n, CBAM-YOLOv5s, BAM-YOLOv5s and ECA-YOLOv5s, respectively, and still has a high average detection accuracy. Finally, TE-YOLOv5s exhibits a superior identification accuracy rate with less rates of leakage and false identification, and the performance of YOLOv5s has been improved, making it more appropriate for the identification of complex images from drone aerial photography. The experimental results of different algorithms on the dataset of this paper are shown in Table 6.

TABLE 6. Comparative experimental results for different target categories

Models	Pedestrain	People	Bicycle	Car	Van	Trunk	Tricycle	Awning-tricycle	Bus	Motor
Faster R-CNN	0.145	0.261	0.140	0.257	0.193	0.230	0.182	0.149	0.348	0.208
SSD	0.052	0.201	0.087	0.126	0.138	0.202	0.057	0.059	0.256	0.260
YOLOv3-tiny	0.035	0.119	0.123	0.135	0.088	0.132	0.053	0.061	0.254	0.124
YOLOv3	0.131	0.385	0.192	0.302	0.228	0.194	0.187	0.092	0.387	0.159
YOLOv4	0.114	0.375	0.176	0.332	0.175	0.094	0.171	0.037	0.288	0.151
YOLOv5s	0.199	0.357	0.199	0.345	0.316	0.177	0.149	0.105	0.371	0.212
YOLOv5n	0.114	0.175	0.049	0.129	0.237	0.188	0.203	0.117	0.207	0.166
YOLOv8n	0.242	0.291	0.183	0.237	0.319	0.231	0.144	0.199	0.222	0.145
YOLOv8s	0.119	0.396	0.198	0.437	0.281	0.401	0.124	0.241	0.365	0.199
CBAM-YOLOv5s	0.166	0.297	0.162	0.380	0.172	0.188	0.141	0.249	0.202	0.309
BAM-YOLOv5s	0.141	0.355	0.128	0.350	0.251	0.127	0.145	0.223	0.363	0.275
ECA-YOLOv5s	0.153	0.359	0.162	0.351	0.253	0.122	0.151	0.259	0.370	0.272
TE-YOLOv5s	0.189	0.378	0.231	0.415	0.176	0.304	0.159	0.280	0.401	0.351

4.5. **Analysis of experimental results.**

4.5.1. *Model performance analysis.* The performance metrics of the TE-YOLOv5s algorithm are plotted following the training, validation, and testing of the improved algorithm on a customized dataset specifically designed for complex small targets. This paper mainly analyzes the improved model performance by comparing the Precision-Confidence (P-C) curve, Recall-Confidence (R-C) curve with the benchmark model YOLOv5s, and analyzing the training and verification loss curves of TE-YOLOv5s. The P-C curve, R-C curve, training and validation loss curves are shown in Figure 5, Figure 6 and Figure 7, respectively.

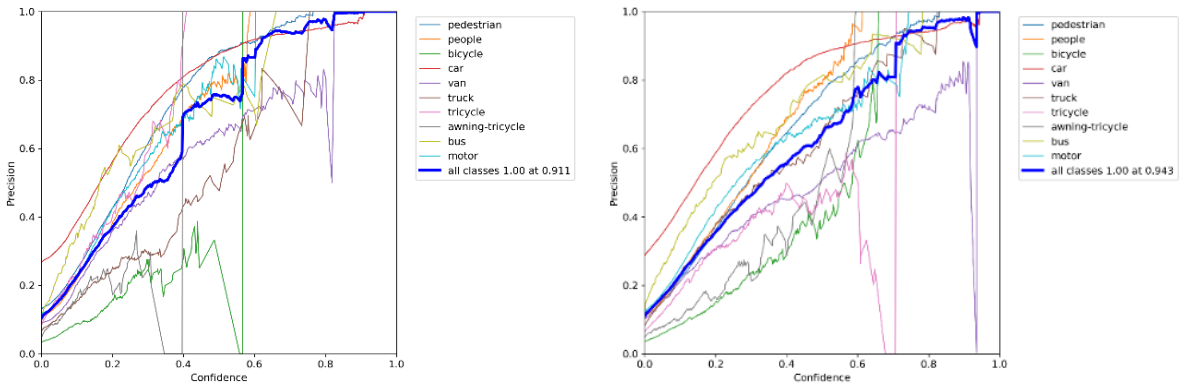


FIGURE 5. (color online) P-C curve, YOLOv5s (left) and TE-YOLOv5s (right)

As depicted in Figure 5, the curve represents the correlation between the accuracy of the model and the confidence level, the precision increases with higher confidence levels, and lower false identification rate of the model. From Figure 5, it can be seen that, under the same conditions, the precision of TE-YOLOv5s reaches 94.3%, which is 3.2% higher compared to 91.1% of YOLOv5s, indicating that the TE-YOLOv5s for the diminutive targets of complex images has lower false detection rate and better performance.

As illustrated in Figure 6, the curve represents the correlation between the recall of the model and the confidence level, with an increase in confidence level, there is a decrease in recall. The recall characterizes the leakage rate of the model, a higher recall indicates a

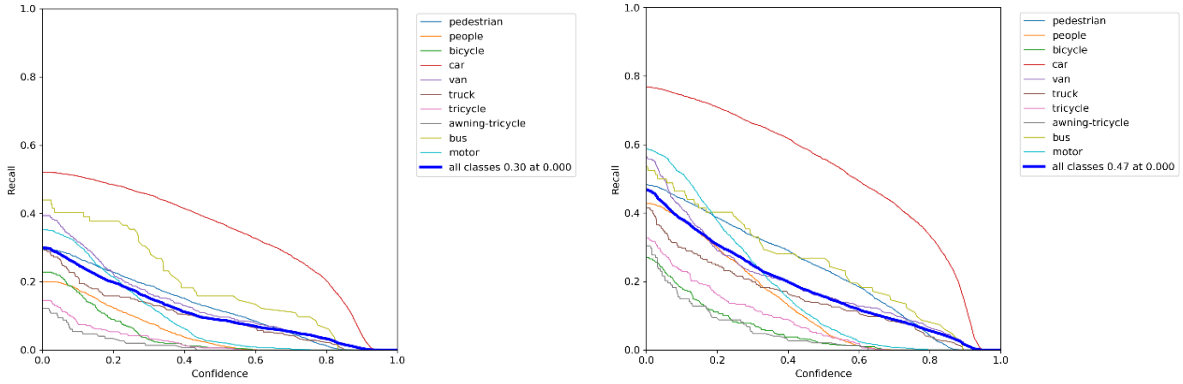


FIGURE 6. (color online) R-C curve, YOLOv5s (left) and TE-YOLOv5s (right)

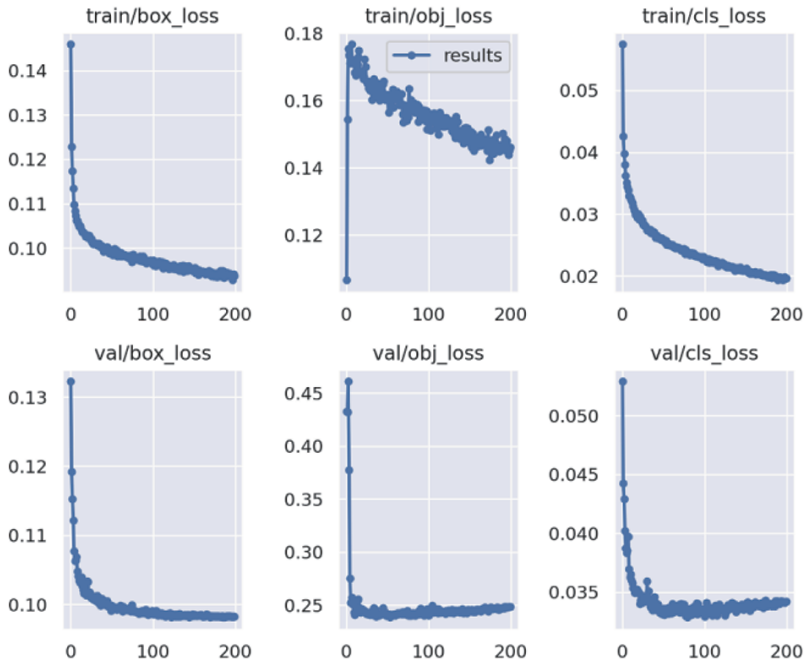


FIGURE 7. TE-YOLOv5s: Training loss curve (top), validation loss curve (bottom)

lower leakage rate, signifying better identification capability. As evident in Figure 6 under identical conditions, the recall of TE-YOLOv5s is 47%, which is 17% higher compared with 30% of YOLOv5s, indicating that TE-YOLOv5s has a much lower leakage rate and a much better performance for the small targets of the complex image.

As shown in Figure 7, the curve represents the training and validation loss curves of the model, the training loss curve represents the learning of the model in the training phase, and the validation loss curve represents the identification capability of the model on unlearned images, reflecting the robustness and generalization of the model. As observed in Figure 7, the loss value changes with epoch, and a lower loss value indicates better detection performance of the model.

Train/box loss represents the difference between the predicted bounding box and the actual bounding box when the model is being trained, and its constant decrease with epoch changes indicates that the detection accuracy of the predicted box is increasing when the model detects complex diminutive targets; train/obj loss controls the confidence

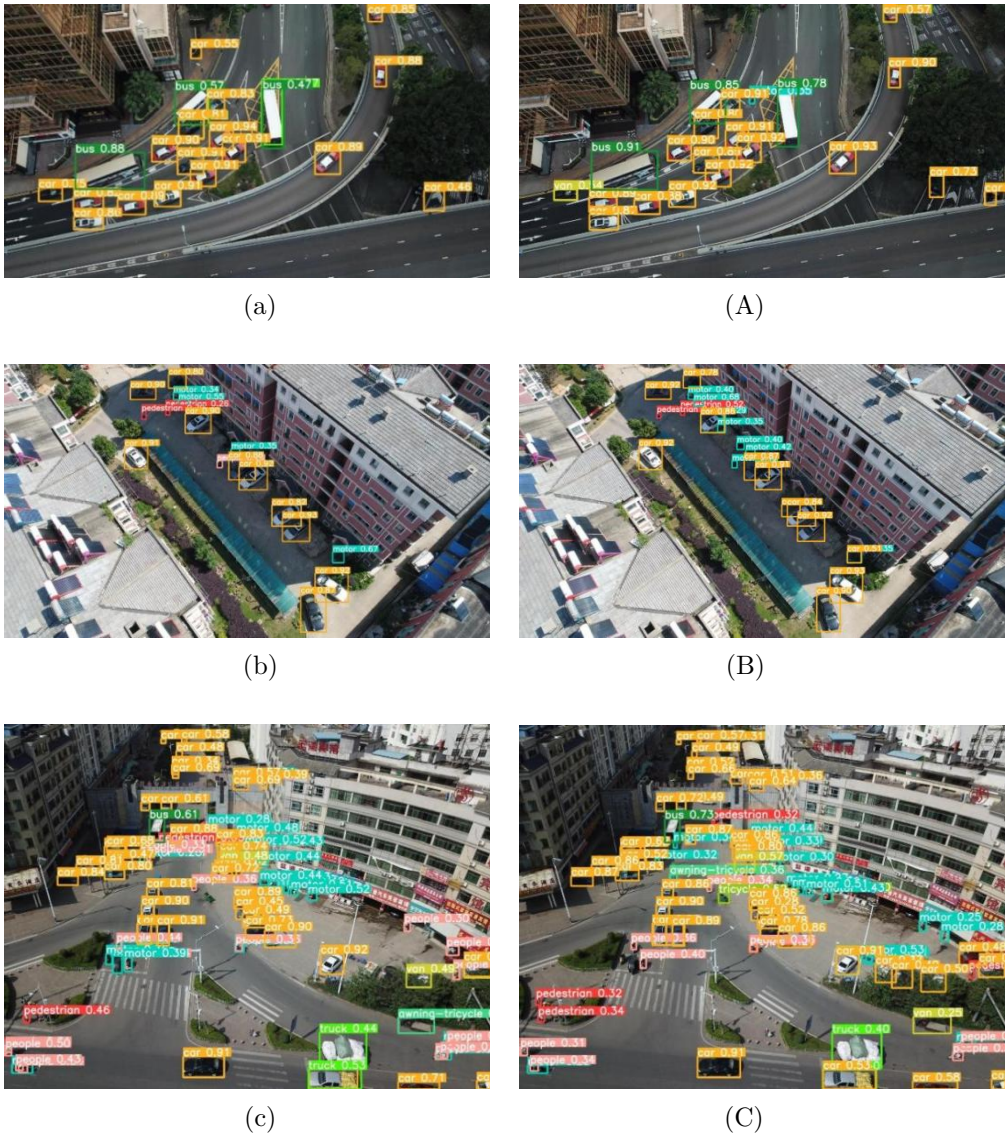
of the model in target detection, and represents the model's capability to discriminate between targets and non-targets, and its constant decrease with epoch changes indicates that the model's capability to discriminate between complex diminutive targets is increasing; train/cls loss represents the model's capability to categorize different diminutive targets, and it decreases with epoch, indicating that the model's ability to categorize different diminutive targets is rising. Moreover, the validation loss curve also decreases with epoch, and it shows that the improved algorithm has excellent identification capability for complex diminutive targets in aerial images.

4.5.2. *Comparative analysis of experimental results.* To assess the identification capability of TE-YOLOv5s in real application scenarios, aerial images under different scenes are selected in the test set to be tested on the baseline model YOLOv5s and TE-YOLOv5s, respectively. Due to the wide field of view and extensive coverage of drone aerial images, these images often contain numerous small-scale targets, complex and dynamic backgrounds, diverse lighting conditions, low resolution, and densely packed objects. Therefore, we selected a variety of scenes for our experiments to comprehensively evaluate the algorithm's performance under different conditions. The scenes are selected as ordinary scenes, complex background interference scenes, and target-intensive scenes. The test results are depicted in Figure 8.

As observed in Figure 8, by comparing Figure 8(a) with Figure 8(A), it can be analyzed that in the ordinary scene, in Figure 8(a), the target in the lower right corner is not fully detected by YOLOv5s, and the detection accuracy for the "bus" and "car" categories is noticeably lower. YOLOv5s has obvious misdetection and leakage detection phenomenon and lower accuracy; however, as shown in Figure 8(A), TE-YOLOv5s successfully detects the target in the lower right corner of the aerial image, YOLOv5s has obvious improvement in detection accuracy, and the model performance improves significantly. Comparing Figure 8(b) with Figure 8(B), it can be analyzed that in the complex background interference scene, YOLOv5s has leakage detection phenomenon on the targets, with most of the "motor" targets in Figure 8(b) not being detected, while TE YOLOv5s can accurately detect the targets not detected by YOLOv5s, and TE-YOLOv5s has lower leakage rate and higher accuracy. Comparing Figure 8(c) with Figure 8(C), in the target-intensive scenario, YOLOv5s not only missed a substantial number of "car" and "motor" targets but also exhibited false positive detections. In contrast, TE-YOLOv5s has higher accuracy rate and can accurately identify the targets missed by YOLOv5s, which reduces the leakage rate of YOLOv5s, and has better detection performance.

Based on experimental testing, YOLOv5s often encounters issues with missed and false detections when dealing with aerial images characterized by varying target scales, complex and dynamic backgrounds, densely populated targets, and low resolution. In contrast, the proposed TE-YOLOv5s model demonstrates a significant improvement in detection accuracy and a notable reduction in false detection rates when applied to drone-captured images with similar characteristics. In summary, TE-YOLOv5s has better identification capability for complex diminutive targets in drone aerial images under different environmental backgrounds.

5. **Conclusion and Outlook.** Aiming at the issues of leakage, false identification and low accuracy of the target detection model caused by the complex background, low target resolution and dense targets of UAV aerial images, this paper proposes the TE-YOLOv5s detection model with higher detection performance. Firstly, on the basis of YOLOv5s, a 160×160 detection scale for diminutive targets is added, which contains richer shallow positional information and deep semantic information, effectively reducing the leakage



(a) (A)

(b) (B)

(c) (C)

FIGURE 8. Experimental results of YOLOv5s (left) and TE-YOLOv5s (right)

detection problem of the model. At the same time, for the special characteristics of UAV aerial images, the anchor frame scale more suitable for complex small targets is obtained through the K-means algorithm reclustering, which improves the identification capability of the model. Second, the backbone network of the original model is reconstructed, and the Transformer encoder is fused into the C3 module, which suppresses the interference of useless information and enhances the feature extraction capability of the model. Third, the Focal-EIOU Loss is introduced in order that the model has faster convergence speed and regression accuracy. Finally, after training, verification and testing on a customized complex target dataset, the TE-YOLOv5s model improved detection precision by 6.1%, recall rate by 5%, and mAP@0.5 by 7.2% compared with the YOLOv5s model. It has better small target detection performance and is more appropriate for complex diminutive target detection in UAV aerial images.

In the future research, it needs to enhance the generalization capability and robustness of the model by expanding, optimizing, data enhancement and data preprocessing of the dataset, and further compress the model by adopting pruning and distillation to make it better for deployment applications. It also needs to adopt more efficient data

fusion to enhance the identification accuracy of the model to further improve the practical application value of the model so that it is more applicable to the identification of complex diminutive targets in aerial imagery from UAVs.

Acknowledgment. The authors sincerely express their gratitude to the reviewer for the thorough review and evaluation of this paper. Additionally, we appreciate your valuable contributions to the academic community. This work was supported in part by the Scientific Research Foundation of Jilin Province under Grant JJKH20230305KJ.

REFERENCES

[1] M. Y. Yang, W. Liao, X. Li and B. Rosenhahn, Vehicle detection in aerial images, *Photogrammetric Engineering and Remote Sensing*, vol.85, no.4, 2018.

[2] S. Li, Y. Lv, M. Li and Z. Wang, Detection of protective apparatus for municipal engineering construction personnel based on improved YOLOv5s, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol.14, no.7, pp.369-378, 2023.

[3] X. Feng and W. Jiang, Research on human fall detection based on Tiny-YOLOv3 algorithm, *Proc. of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 2021.

[4] L. Yu, E. Park, A. C. Berg and T. L. Berg, Visual Madlibs: Fill in the blank description generation and question answering, *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp.2461-2469, 2015.

[5] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.39, no.6, pp.1137-1149, 2017.

[6] Z. Cai and N. Vasconcelos, Cascade R-CNN: Delving into high quality object detection, *arXiv Preprint*, arXiv: 1712.00726, 2017.

[7] W. Liu, D. Anguelov, D. Erhan et al., SSD: Single shot MultiBox detector, in *Computer Vision – ECCV 2016. Lecture Notes in Computer Science*, B. Leibe, J. Matas, N. Sebe and M. Welling (eds.), Cham, Springer International Publishing, 2016.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, Focal loss for dense object detection, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.2999-3007, 2017.

[9] M. Tan, R. Pang and Q. V. Le, EfficientDet: Scalable and efficient object detection, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp.10778-10787, 2020.

[10] P. Krahenbuhl and V. Koltun, Learning to propose objectspp, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – Learning to Propose Objects*, Boston, MA, USA, pp.1574-1582, 2015.

[11] C. Candan, A low complexity two-stage target detection scheme for resource limited radar systems, *IEEE Transactions on Aerospace & Electronic Systems*, vol.49, no.1, pp.594-601, 2013.

[12] S. R. Shen, X. Zhang, W. Yan et al., An improved UAV target detection algorithm based on ASFF-YOLOv5s, *Mathematical Biosciences and Engineering (MBE)*, vol.20, no.6, pp.10773-10789, 2023.

[13] J. Ding, J. Zhang, Z. Zhan et al., A precision efficient method for collapsed building detection in post-earthquake UAV images based on the improved NMS algorithm and Faster R-CNN, *Remote Sensing*, vol.14, no.3, 663, 2022.

[14] M. Wang, X. Luo, X. Wang and X. Tian, Research on vehicle detection based on Faster R-CNN for UAV images, *2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS2020)*, Waikoloa, HI, USA, pp.1177-1180, 2020.

[15] H. Huang, L. Li and H. Ma, An improved cascade RCNN-based target detection algorithm for UAV aerial images, *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, Xi'an, China, pp.232-237, 2022.

[16] W. Liu, J. Qiang, X. Li, P. Guan and Y. Du, UAV image small object detection based on composite backbone network, *Mobile Information Systems*, vol.2022, 7319529, 2022.

[17] M. G. Javed, M. Raza, M. Ghaffar et al., QuantYOLO: A high-throughput and power-efficient object detection network for resource and power constrained UAVs, *2021 Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, Australia, pp.245-252, 2021.

[18] P. Lu, Y. Ding and C. Wang, Multi-small target detection and tracking based on improved YOLO and SIFT for drones, *International Journal of Innovative Computing, Information and Control*, vol.17, no.1, pp.205-224, 2021.

[19] J. Liu, R. Jia, W. Li, F. Ma, H. M. Abdullah, H. Ma and M. A. Mohamed, High precision detection algorithm based on improved RetinaNet for defect recognition of transmission lines, *Energy Reports*, vol.6, pp.2430-2440, 2020.

[20] Z. Yang, Z. Xu and Y. Wang, Bidirection-Fusion-YOLOv3: An improved method for insulator defect detection using UAV image, *IEEE Transactions on Instrumentation and Measurement*, vol.71, pp.1-8, Article no.3521408, 2022.

[21] Z. Yu, Y. Shen and C. Shen, A real-time detection approach for bridge cracks based on YOLOv4-FPM, *Automation in Construction*, vol.122, 103514, 2021.

[22] J. Zhu, X. Wang, Y. Liu, Q. Ji, Z. Zhao and S. Wang, UavTinyDet: Tiny object detection in UAV scenes, *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, Xi'an, China, pp.195-200, 2022.

[23] M. H. Hamzenejadi and H. Mohseni, Fine-tuned YOLOv5 for real-time vehicle detection in UAV imagery: Architectural improvements and performance boost, *Expert Systems with Application*, vol.231, 120845, 2023.

[24] Z. Ying, Z. Lin, Z. Wu, K. Liang and X. Hu, A modified-YOLOv5s model for detection of wire braided hose defects, *Measurement*, vol.190, 110683, 2022.

[25] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, Feature pyramid networks for object detection, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp.936-944, 2017.

[26] S. Li, Z. Wang, Y. Lv and X. Liu, Improved YOLOv5s-based algorithm for foreign object intrusion detection on overhead transmission lines, *Energy Reports*, vol.11, pp.6083-6093, 2024.

[27] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, Path aggregation network for instance segmentation, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp.8759-8768, 2018.

[28] L. Chen, M. Zheng, S. Duan, W. Luo and L. Yao, Underwater target recognition based on improved YOLOv4 neural network, *Electronics*, vol.10, no.14, 1634, 2021.

[29] J. Fang and P. Wang, Application of improved YOLO V3 algorithm for target detection in echo image of sonar under reverb, *Journal of Physics: Conference Series*, vol.1748, no.4, 042048, 2021.

[30] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao and F. Liu, ViT-YOLO: Transformer-based YOLO for object detection, *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, pp.2799-2808, 2021.

[31] S.-J. Ji, Q.-H. Ling and F. Han, An improved algorithm for small object detection based on YOLO v4 and multi-scale contextual information, *Computers and Electrical Engineering*, vol.105, 108490, 2023.

[32] S. Li, Y. Lv, X. Liu and M. Li, Detection of safety helmet and mask wearing using improved YOLOv5s, *Scientific Reports*, vol.13, no.1, 21417, 2023.

Author Biography

Shuangyuan Li received his Bachelor's degree in Computer Science and Technology from Jilin Institute of Chemical Technology in 2005 and his Master's degree in Computer Technology from Northeast Electric Power University in 2012. He is currently the Director of the Information Center at Jilin Institute of Chemical Technology. His research interests include AI-based target detection and cybersecurity. He has published over 50 papers in journals and conferences.

Jianglong Lin received his Bachelor of Engineering degree in Mechatronics Engineering from Lanzhou Jiaotong University, China, in 2022. He is currently a Master's student in Electronic Information at Jilin Institute of Chemical Technology, Class of 2023. His primary research interest lies in object detection based on deep learning techniques.

Yanchang Lv received his Bachelor of Engineering degree in Communication Engineering from Tianjin University Renai College, China, in 2021, and his Master's degree in Electronic Information from Jilin Institute of Chemical Technology in 2024. He is currently a Ph.D. student in Computer Science and Technology at Changchun University of Science and Technology, China. His primary research interest is object detection in artificial intelligence.

Tianyu Li received his Bachelor of Engineering degree in Automation from Linyi University, China, in 2023. He is currently a Master's student in Electronic Information at Jilin Institute of Chemical Technology, Class of 2023. His primary research interest is object detection based on deep learning techniques.