Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.060873

ARTICLE

YOLO-S3DT: A Small Target Detection Model for UAV Images Based on YOLOv8

Pengcheng Gao* **and Zhenjiang Li**

School of Cyber Security, Gansu University of Political Science and Law, Lanzhou, 730070, China

* Corresponding Author: Pengcheng Gao. Email: gpc7039@gsupl.edu.cn

Received: 12 November 2024; Accepted: 13 December 2024; Published: 06 March 2025

ABSTRACT: The application of deep learning for target detection in aerial images captured by Unmanned Aerial Vehicles (UAV) has emerged as a prominent research focus. Due to the considerable distance between UAVs and the photographed objects, coupled with complex shooting environments, existing models often struggle to achieve accurate real-time target detection. In this paper, a You Only Look Once v8 (YOLOv8) model is modified from four aspects: the detection head, the up-sampling module, the feature extraction module, and the parameter optimization of positive sample screening, and the YOLO-S3DT model is proposed to improve the performance of the model for detecting small targets in aerial images. Experimental results show that all detection indexes of the proposed model are significantly improved without increasing the number of model parameters and with the limited growth of computation. Moreover, this model also has the best performance compared to other detecting models, demonstrating its advancement within this category of tasks.

KEYWORDS: Target detection; UAV images detection; small target detection; YOLO

1 Introduction

The primary objective of target detection is to identify both the class and location of a specific target with a video or image. With the development of deep convolutional neural network technology, the method of target detection is gradually getting rid of the traditional way of artificially designing feature classifiers, and evolving into the way of learning target features from image samples to realize the detection task more efficiently and in real-time [1]. Deep learning-based methods for target detection have garnered significant interest among researchers. As an intersection of computer vision, image processing, and machine vision, target detection has been widely used in various scenarios, including industrial automation, autonomous driving, remote sensing image detection, Unmanned Aerial Vehicle (UAV) image detection, and other various domains [2].

In the MS COCO [3] dataset, targets with dimensions smaller than 32 × 32 pixels are defined as small targets. Small target detection is a subcategory of target detection and is used in various detection tasks, especially in the field of UAV image detection. Due to the rapid development of the UAV field in recent years, the cost and performance of the equipment have been greatly improved, which has led to the application of UAVs in various types of work, including safety inspection, disaster detection, rescue search, etc. [4]. Since UAV photography flights are far away from the object to be photographed and mostly use wide-angle lenses, there are a large number of small-sized targets in UAV images, which are characterized by weak features and little information, making it usually difficult for the model to accurately detect them. When the UAV is in

complex environments such as low light, more occlusion, and dense targets, it will be more challenging to detect the targets in the images in real-time and accurately [5].

To address the challenges associated with UAV image detection, numerous researchers have made many contributions to improve the capability of small target detection. Zhang et al. [6] combined the advantages of one-stage and two-stage target detection algorithms and fused the bottom-layer features with the top-layer features while generating the target candidate region, which improves the accuracy of small target detection, but this method increases the processing time. Ying et al. [7] integrated multi-scale features by fusing global attention and pixel feature attention, solved the problem of feature information loss of small-size targets at low resolution and improved the model performance and robustness, but the slow processing speed cannot achieve real-time detection. Bai et al. [8] proposed an end-to-end multi-task generative adversarial network, which generates super-resolution images from fuzzy images into the generator through up-sampling operation so that more details can be captured for detection. However, this method suffers from training difficulties. Based on the Single Shot multi-box Detector (SSD) as the basic network, Maktab et al. [9] used modules such as super resolution, deconvolution, and feature fusion to improve the model's ability to detect small targets, but it has limitations in detecting when facing complex environments. Liu et al. [10] proposed a feature enhancement module combined with Spatial Pyramid Pooling (SPP) to assign larger weights to low-latitude feature maps to improve feature extraction for small targets, but this may also lead to loss of information in other feature maps. Kim et al. [11] reduced the information loss by modifying the backbone of You Only Look Once (YOLO) with an efficient channel attention module and using transposed convolution instead of up-sampling, but the method is difficult to extract features under complex backgrounds, and there is a significant decrease in recall in the experiments. Wang et al. [12] proposed an improved model YOLOX_w based on YOLOX-X, utilizing a Slicing Aided Hyper Inference (SAHI) algorithm for data augmentation, and adding additional detection heads as well as an attention module to make the model focus on the key features, the model improves the detection accuracy but the parameters of the model and the number of operations are greatly increased. Wang et al. [13] optimized YOLOv8s by integrating Weighted Intersection over Union (WIoU), BiFormer, and proposing a Focal FasterNet Block module to extract multi-level features for enhanced detection of small-scale objects. However, it resulted in a significant reduction in detection speed compared to the baseline model. Lee et al. [14] proposed an improved RetinaNet that incorporated deformable convolutions into the backbone and optimized the detection head part and pyramid layers to enhance performance. However, the experiments only included comparisons with Faster R-CNN and YOLOv5, which makes it difficult to determine with certainty whether the improvements are effective.

To summarize, there are still two primary challenges with algorithms for the detection of small targets that require resolution in this field. First, crucial information is easily lost or missed during the process of feature extraction. Second, previously proposed algorithms exhibit excessive complexity, which makes it difficult to be deployed in terminals. These two challenges have severely limited the application of target detection models on embedded devices such as UAVs to achieve real-time and high-precision detection tasks. To address them, this paper develops a novel model with high accuracy and lightweight based on the original YOLOv8n model. The main work of this paper has the following four aspects: Firstly, the Neck and Head parts of the model are redesigned. A 160 × 160 detection head is added for better performance of small-sized targets, and comparative experiments are conducted on combinations of different-scale heads to obtain the best result. Secondly, a module called 3DSPPF is proposed to extend the receptive field and achieve better detection of small targets by adding an extra feature extraction branch to the Spatial Pyramid Pooling Fast (SPPF) module. Thirdly, replace the original dynamic sampler with DySample module. A new dynamic sampling scheme based on point sampling is proposed, which effectively improves the properties

of the model. Finally, due to the dense number of targets and the small size of targets in images, some factors of the Task-Aligned Assigner (TAA) are optimized to adjust the predicted score and the number of positive samples in the sample selection process to obtain a dynamic assignment strategy of positive and negative samples suitable for the detection of such targets.

Through the improvement of the above four aspects, we propose a target detection model called YOLO-S3DT. This model has excellent performance in various detection indicators. In this work, 30% of the training set from the public dataset of VisDrone2019 [15] is used for training. The results indicate that compared with YOLOv8n in the same experimental environment, the value of mAP50 and mAP50-95 have respectively increased by 5.7% and 2.8%, and the model parameters decreased by about 0.08 M, while the computation increases by less than 3GFlops, thus achieving the lightweight and high performance of the model.

This paper is organized into four sections. Section 2 mainly introduces the structure of the original baseline model and introduces four improvement strategies based on the model. Section 3 presents the dataset selected for the experiment, details regarding the experimental environment configuration, and specifies the evaluation indices employed. Section 4 encompasses four sets of experiments, including ablation experiments and comparison experiments, etc., and analyzes the experimental results in detail. Additionally, we offer an intuitive comparison and analysis of sample detection before and after implementing improvements. The final section summarizes the work completed in this paper.

2 Models and Methods

2.1 Baseline Model

Ultralytics improved on the previous YOLOv5 in many aspects and finally released a new version, known as YOLOv8, in early 2023 [16]. This version of the YOLO model is extensively utilized across various computer vision tasks, including but not limited to object detection, instance segmentation, pose detection, rotating object detection.

The general structure of YOLOv8 is illustrated in Fig. 1. It replaces the C3 module with the C2f module. In the Neck part, the convolution operation after up-sampling in the Pyramid Attention Network (PAN) structure is removed. In the Head part, two independent branches are used to achieve target classification and the regression of the prediction bounding box respectively, and different loss functions are used for these two tasks to get better results. Meanwhile, YOLOv8 uses TAA to dynamically allocate samples instead of the previous IoU matching strategy or single-sided proportional allocation, which gives the model better detection accuracy and robustness.

2.2 Improved Model

In the task of detecting aerial images, the original YOLOv8 network structure struggles to achieve high detection accuracy due to the significant number and density of small-sized targets. Additionally, due to the deployment in the terminal equipment, the size of the model should be appropriate, and a model that is overly complex cannot satisfy the requirement of real-time detection. To address these challenges, this paper selects YOLOv8n, the lightest model in the YOLOv8 series, as the basic model, improves the model from four aspects: small target detection layers, a redesigned SPPF module called 3DSPPF, DySample up-sampling module and TAA parameters optimization. A novel model called YOLO-S3DT, which is derived from the initials of these four improvements, is proposed.

The framework of YOLO-S3DT is shown in Fig. 2. Firstly, the network structure has been enhanced to better accommodate small-sized targets by incorporating a 160 × 160 detection head. According to the experimental results, the 80 × 80 detection head in the original model is eliminated, retaining only 160

× 160, 40 × 40 and 20 × 20 detection heads. Under the premise of improving the detection ability, the number of parameters is reduced. Secondly, a novel 3DSPPF module is proposed to obtain more rich target feature information by adding a feature extraction branch to the original SPPF module. Thirdly, a lightweight dynamic upsampler, called DySample, is utilized in the Neck part to improve the performance without adding additional parameters and operations of the model. Finally, according to the characteristics of UAV images, relevant factors of TAA are optimized, which effectively improves the detection accuracy and the rate of recall for small targets. The following four subsections detail the principles and specifies of each improvement.

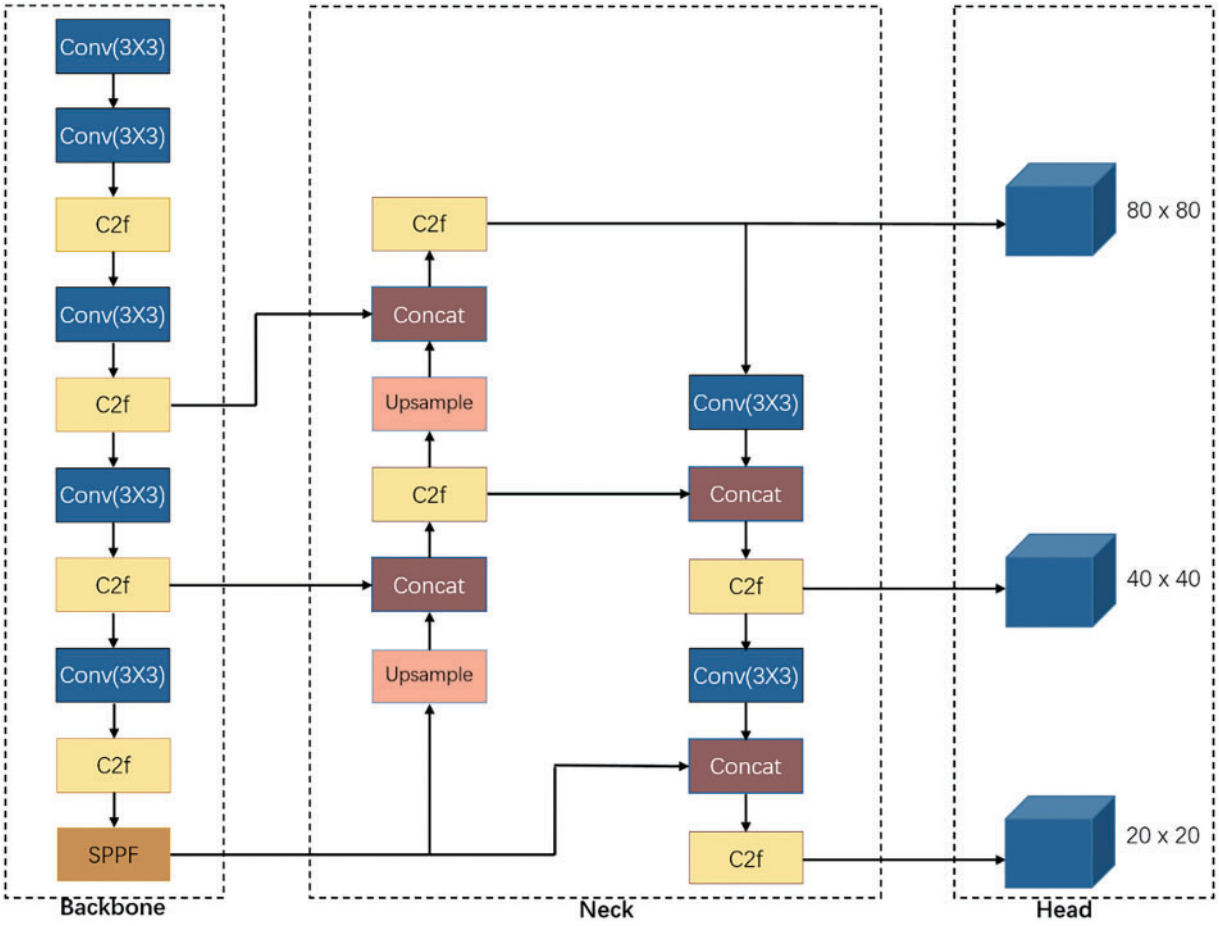

Figure 1: The general framework of baseline mode

2.2.1 Improvement of Small Target Detection Layer

For large-size targets to be detected, their semantic information often appears in the deeper feature map, while for small-size targets to be detected, their feature information is displayed in the shallower feature map [17]. With the deepening of the network, the network cannot learn the corresponding features, leading to missed detection. The detection head sizes of the original YOLOv8 model are 80 × 80, 40 × 40, and 20 × 20, respectively, corresponding to the detection of small, medium, and large targets. In aerial images, many detectable targets are situated at considerable distances from the camera, consequently, they occupy a minimal number of pixels within the image. This results in smaller targets being inadequately detected by the existing 80 × 80 detection head.

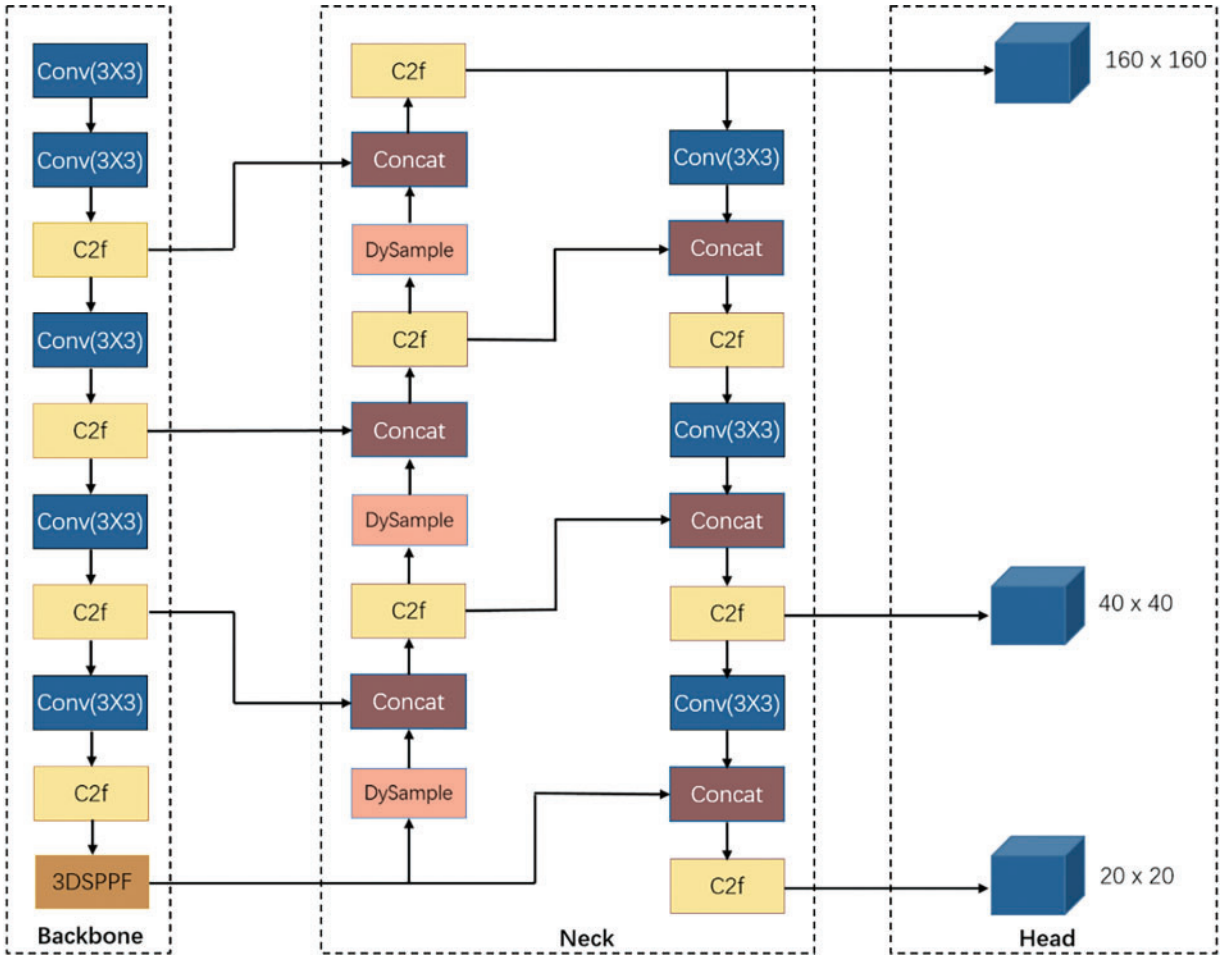

Figure 2: The framework of YOLO-S3DT

To solve the above problems, the original structure of the network is improved by adding a 160 × 160 scale head, as shown in Fig. 2. The specific improvement method is to add an up-sampling layer after the 15th layer, a C2f module, and carry the result and the output feature map of the 2nd layer into the concat layer, and through the 18th layer, another C2f layer, to obtain a feature map with the larger size for detecting smaller targets, and then through the convolution and concat operation, the feature map is reduced to the size of 80 × 80, which is used to combine with the original structure of YOLOv8n. Subsequently, this paper compares the combinations of different detection heads through a series of experiments under the same experimental environment, and finally removes the 80 × 80 detection head and retains only three other heads in the model after comprehensively considering the factors of detection accuracy, parameters, Flops, etc. From the subsequent experimental results, the improved detection head is employed to detect smaller targets. The experimental results indicate that the improved network can more effectively leverage shallow information, resulting in a significant enhancement of its ability to learn small target features.

2.2.2 Improvement of SPPF Module

The original SPPF module primarily comprises three max pooling layers, each featuring a kernel size of 5, a stride of 1, and a padding of 2. The resulting feature maps from each layer are concatenated and then fed

into the Convolution Batch normalization SiLU (CBS) module, which includes a composite unit consisting of a 1 × 1 convolutional layer, batch normalization, and the SiLU activation function.

The enhanced SPPF achieves comparable performance to the original SPP [18] by employing three smaller pooling windows. While max pooling effectively captures salient features within local regions while preserving feature map dimensions, it also suffers from information loss by retaining only one maximum value per region. Furthermore, in cases where strong features appear multiple times within a region, their strength information may be compromised due to the single maximum value retention characteristic of max pooling.

To better extract features and avoid losing important information, we extend an additional branch to reconstruct new feature maps. The main method is to pass the output of each max pooling layer through a 3 × 3 convolutional layer for secondary feature extraction, and then the three additional feature maps are fused with the feature maps extracted by the original SPPF module to obtain richer feature information. In order to increase the improvement effect and reduce the slow inference speed problem caused by the improvement, we also make two fine optimizations in this module:

1. Depthwise Separable Convolution (DSC) [19] is chosen to replace traditional convolution blocks. The two processes, Depthwise Convolution and Pointwise Convolution, separate the depth and spatial dimensions of feature graphs, thus reducing the computational load and the need for computing resources.

2. By convolutional dilation operation [20], the receptive field is extended to capture multi-scale contextual information without increasing parameters.

With the above improvements, the new SPPF module can capture important features more comprehensively and also has the characteristics of lightweight. Since three DSC modules are used in the improvement process, so it is designed as 3DSPPF. The structure of improved module is presented in Fig. 3.

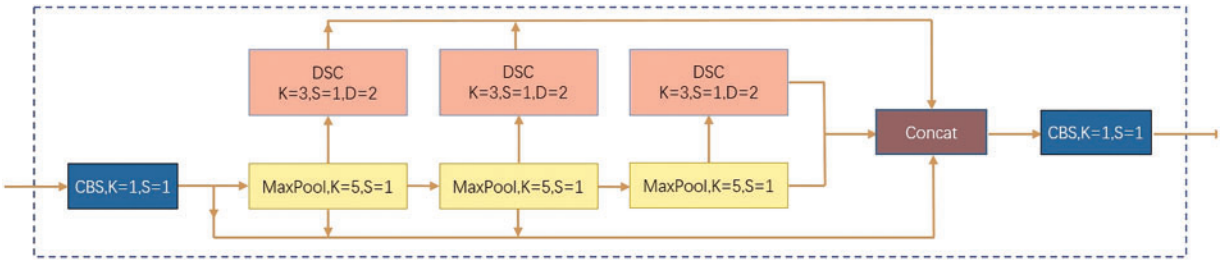

Figure 3: The structure of 3DSPPF module

2.2.3 Improvement of Up-Sampling Module

DySample [21] is used to replace the original up-sampling module, and the improved up-sampling module can effectively improve the detection accuracy compared with the former. DySample uses the feature map X with the original size $C \times H \times W$ to generate the offset O, which is added with the original sampling grid g to obtain the sample set S. Then, the Grid Sample function in PyTorch is used to up-sample the generated sample set S and the input feature map X to generate the up-sampled feature graph .

In the improvement of this module, the author solves the problem of initial disordered sampling positions caused by ignoring the location relation in the nearest initialization by separating and evenly distributing initial sampling positions through bilinear initialization. Two strategies, static scope factors, and dynamic scope factors, are proposed to limit offset range to reduce overlap. The static offset coefficient reduces the offset range by multiplying offset O by a factor of 0.25, which is simple to implement but has

limited effect. Dynamic offset coefficients are obtained by formula (1), and the generated coefficients are distributed in the range of 0 to 0.5 with 0.25 as the center point, which is a relatively complicated calculation process but can achieve better enhancement results. At the same time, the authors use the group-wise up-sampling strategy for reference and divide the feature map into 4 groups along the channel dimension to generate 4 groups of offsets. The sampling process of the fully improved module is shown in Fig. 4.

$$O = 0.5 sigmoid\left(linear_1(X) \cdot linear_2(X)\right) \tag{1}$$

In Fig. 4a, it is demonstrated that the sampling set generated by the sampling generator and the original feature map X are used for grid sample operation to generate the feature map X'. In Fig. 4b, the static offset coefficients were multiplied by 0.25 after subjecting X to a bilinear initialization operation. The result is then reshaped by a Pixel Shuffling [22] to obtain offset O. The dynamic offset coefficients are transformed by two bilinear initialization branches, one of which is multiplied by 0.5 and a sigmoid function and multiplied by the output of the other branch to obtain an intermediate variable with the dimensions $2gs^2 \times H \times W$. The result is then reshaped by a Pixel Shuffling to obtain offset O.

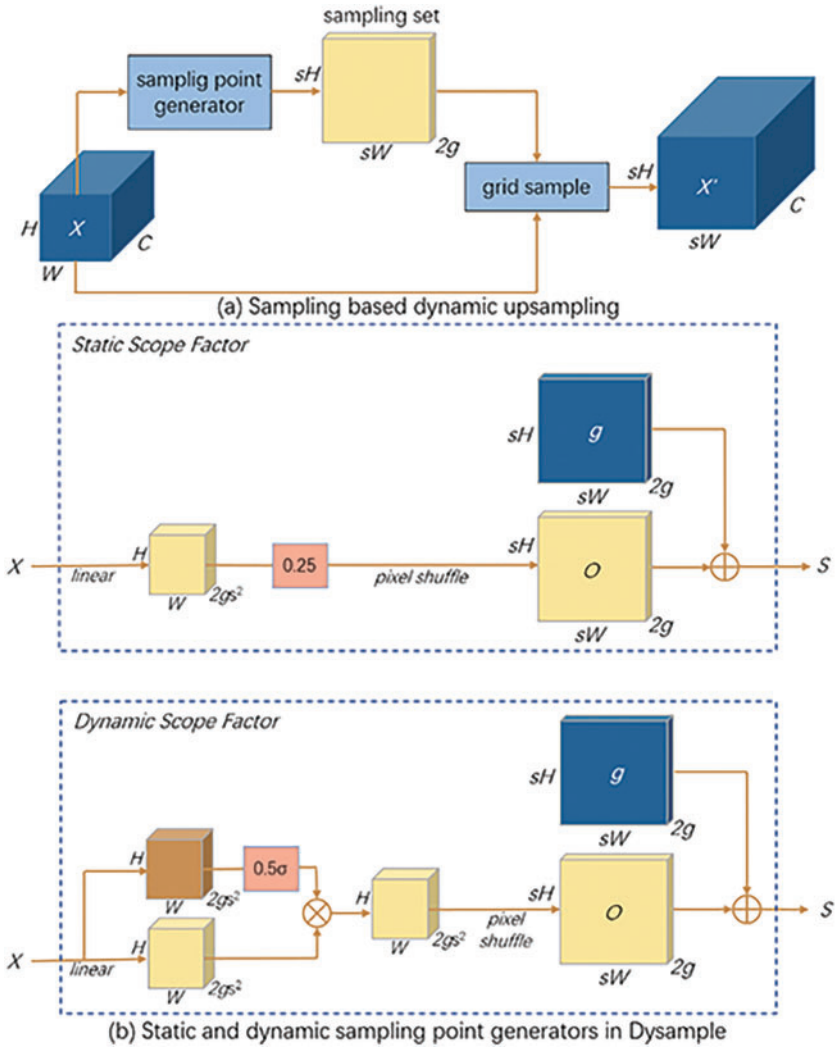

Figure 4: The sampling process of Dysample

2.2.4 Optimization of Task-Aligned Assigner Parameters

YOLOv8 matches positive samples through TAA allocation strategies and calculates class loss and position loss based on positive samples. Consequently, the quality of positive samples directly affects the model's capacity to learn important features. The process of matching positive samples with TAA involves the following four steps:

1. Calculate the alignment score. Get the score s corresponding to the actual target category label among all predicted category scores, and calculate the CIoU value of the target location label and all location prediction information to compute the alignment score:

$$t = s^{\alpha} + \mu^{\beta} \tag{2}$$

where α and β are weight hyperparameters in Eq. (2).

2. Preview positive samples. Select the prediction point whose center is inside the ground truth box as initial matching samples.

3. Further select positive samples. According to the alignment fraction t, the topK prediction points are further selected as the matching positions of positive samples.

4. Filter positive samples. If a prediction point matches multiple ground truth boxes, the ground truth box with the largest CIoU is selected as the positive sample to match this prediction point.

Since the targets to be detected in UAV images are usually characterized by small size, high density, and high occlusion, the following two hypotheses are proposed in order to make the TAA assignment strategy better adapted to this type of task:

1. Due to the fact that targets in images are often blocked by each other, the topK value should be reduced to avoid introducing more low-quality bounding boxes, which would have a negative impact on training.

2. From the above, the value of t is a key coefficient in deciding whether the sample can be selected as a positive sample or not, and it can be seen from the Eq. (2) that the value of t is determined by the sum of the scores of classification and localization after weighting. Because the small targets have fewer pixel points, the accuracy of localization in this task cannot be well reflected, while the correctness of classification affects the target detection effect to a greater extent, and it can be observed from the experimental results that small target detection often occurs in the case of misdetection. Therefore, α and β in the formula should be adjusted to increase the weight of the classification score and decrease the weight of the localization score.

In the subsequent experiments, we optimize the three parameters involved: topK, α and β. We conduct several sets of comparative experiments, and the results demonstrate that this methodology obviously improves the detection effect without increasing parameters and computational complexity, which confirms the correctness of hypotheses above. A detailed analysis of this process is presented in the following paper.

3 Environment and Evaluation Indicator

3.1 Experimental Environment

To exclude the influence of different experimental environments on the results, all experiments in this paper are established in the same environment, and the main configurations are as follows: the processor is 12-core Intel(R) Xeon(R) Platinum 8352 V. The graphics card is RTX 4090 with 24 GB of video memory. The memory size is 90 GB. The Python version is 3.10.8, the PyTorch version is 2.1.2, and the CUDA version is 12.1. The pre-training weights of the YOLOv8n model provided by ultralytics are utilized in the experimental configuration. The values of the main parameters used in the experiment are shown in Table 1.

Table 1: Values of main parameters

Param.	Val.
Epochs	200
Batch	16
Image size	640 × 640
Workers	12
Pretrained	True
Optimizer	SGD
lr0	0.01
lrf	0.01
Momentum	0.937

3.2 Evaluation Index

To show the improvement effect of the YOLOv8 model in this paper more effectively, Precision, Recall, mAP50, mAP50-95, Parameters, and GFLops are adopted as the evaluation indexes of the model performance. The first four indexes are mainly utilized to reflect the detection ability of the model, and the last two indexes are mainly utilized to reflect the complexity of the model.

The related formulas for Precision, Recall, and mAP (mean Average Precision) are shown below:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$AP = \int_0^1 P(R)\, dR \tag{5}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} (AP)_i \tag{6}$$

Among them, Eqs. (3) and (4) are used to calculate Precision and Recall, the two most fundamental indexes that can reflect the performance of the model, respectively. Eq. (5) is utilized to assess the average detection ability of the model for a single category of targets, while Eq. (6) reflects the average detection ability of the model for all categories of targets. In this paper, mAP50 and mAP50-95 are selected as two evaluation indicators, where the former represents the average detection ability of prediction results with IoU > 0.5. and the latter represents the mAP value of the IoU in the range of 0.5 to 0.95.

Parameters and GFLops indicate the number of parameters used in the model as well as the complexity of its operations. Consequently, they serve as indexes for assessing both the portability and real-time performance of the model.

3.3 Dataset

The quality of the dataset significantly influences the training process of the model, and a high-quality dataset can enhance the model to effectively extract important features from images. In this paper, approximately 30% of the data volume of the training set in the VisDrone2019 dataset is randomly selected for training purposes during the experimental phase.

As shown in Table 2, there are 2192 images in the processed training set after processing the original dataset. The validation set is unprocessed, with a total of 548 images. The training set covers 10 different categories. The images in this dataset, taken by a variety of drones, cover a wide range of areas, including:

1. Different cities, including 14 Chinese cities thousands of kilometers apart.

2. Different environments, including different weathers and light conditions.

3. Different densities, including scenes with sparse targets and crowded targets.

Overall, the dataset covers a wealth of targets to be detected and diverse scenarios, which enhances the generalizability of the trained model and underscores its practical application value.

Table 2: Details of the processed VisDrone2019 dataset

Scenario	Images of the training set	Images of the validation set	Categories
Drone	2192	548	Motor, Bus, Awning-tricycle, Tricycle, Truck, Van, Car, Bicycle, Person, Pedestrian

4 Experimental Results and Analysis

The experimental content covered in this section mainly includes four parts. First, the influence of the detection head under different combinations on the detection results is tested, and finally, the structure of the improved model is determined according to the experimental results. Then, based on the hypothesis in Section 2.2.4, the influence of different values of topK, α and β in the TAA on the detection results is tested, and a series of suggestions for small target detection are given according to the results. Moreover, based on the baseline model, the small target detection head, 3DSPPF module, DySample module and optimized TAA parameters are added respectively, and ablation experiments are conducted and analyzed. Finally, this paper selects 6 common models for comparative experiments with the YOLO-S3DT model and analyses the results.

4.1 Comparative Experiments of Detection Heads

As discussed in Section 2.2.1, the detection heads of the YOLOv8 model are configured at sizes of 80 × 80, 40 × 40, and 20 × 20 pixels, respectively. When addressing aerial image detection tasks, these existing detection heads struggle to effectively detect small targets. Therefore, this paper adds a 160 × 160 detection head to make better use of the information in shallow features to improve the learning ability of these targets. Meanwhile, to facilitate a more intuitive comparison of the impacts arising from various combinations of 160 × 160 heads and other heads on the detection outcomes, this paper conducts several sets of comparative experiments. The results are presented in Table 3 below.

From the experimental results, the original model has the worst detection effect compared with other combinations of 160 × 160 detection heads, with only 32.2% for mAP50 and 18.3% for mAP50-95 under the same experimental environment. In the four comparative experiments with a 160 × 160 head, the Precision, Recall, mAP50, and mAP50-95 all present different degrees of improvement, which directly proves that the addition of the small-size target detection head has a positive effect on the experimental results. Among them, by removing the 20 × 20 detection scale to reduce the depth of the network, the number of parameters in the model can be effectively reduced, and the results show that the number of parameters in the improved model is less than one million, which is only one third parameter in the original network, and all the detection indexes are higher than those in the original network. However, compared to the other three models with

a 160 × 160 layer, this combination has a limited improvement in detection. The other three combinations are closer in terms of experimental results, probably because they have the same network depth and an effective detection head for small targets. The model with four detection scales has the highest Precision, Recall, and mAP50 values, but this model also has the highest number of parameters and Flops of all the models. The last two combinations in the table are also very close in results and overall better than the other three. Considering the parameters and Flops, although the number of parameters of the former is reduced by about 60,000 compared to the latter, the floating-point arithmetic of the former is one GFlops higher than that of the latter. Given that aerial image detection is primarily used in terminal devices with limited processor performance, it is preferable to select detection scales with fewer floating-point computations and minimal increase in the number of parameters. Therefore, we decided to eliminate the 80 × 80 detection head in the final improvement while retaining the other three as the Head part of the improved model.

Table 3: Experimental results of detection head improvement

The scale of detection layer	Params/M	Flops/G	P/%	R/%	mAP50/%	mAP50-95/%
80 × 80, 40 × 40, 20 × 20	3.01	8.1	42.8	32.6	32.2	18.3
160 × 160, 80 × 80, 40 × 40, 20 × 20	2.93	12.4	46.1	34.8	35.1	19.7
160 × 160, 80 × 80, 40 × 40	0.99	10.6	45.6	32.8	34.0	19.0
160 × 160, 80 × 80, 20 × 20	2.76	11.7	45.9	34.7	35.1	19.9
160 × 160, 40 × 40, 20 × 20	2.82	10.8	45.3	34.8	35.0	19.8

4.2 Comparative Experiments of TAA Parameter Optimization

The screening of positive samples in YOLOv8 is implemented using TAA. The quality of positive samples directly affects the model's ability to learn features, as both category loss and position loss calculations are based on these samples. Section 2.2.4 proposes two hypotheses that inform the setting of relevant parameters within TAA. A series of comparative experiments are performed under the consistent condition. The results are presented in Table 4.

Table 4: Comparison results of TAA parameter optimization

Parameter setting	P/%	R/%	mAP50/%	mAP50-95/%
topK = 10, α = 0.5, β = 6.0	42.8	32.6	32.2	18.3
topK = 13, α = 0.5, β = 6.0	42.5	32.7	31.9	18.1
topK = 15, α = 0.5, β = 6.0	43.1	32.4	31.9	18.0
topK = 7, α = 0.5, β = 6.0	44.4	33.0	33.2	18.9
topK = 5, α = 0.5, β = 6.0	44.8	34.3	33.8	19.0
topK = 3, α = 0.5, β = 6.0	45.7	34.4	34.7	19.7
topK = 1, α = 0.5, β = 6.0	46.5	34.6	35.0	19.3
topK = 1, α = 1, β = 6.0	45.2	35.4	35.2	19.2
topK = 3, α = 1, β = 6.0	47.2	35.3	35.6	19.8
topK = 5, α = 1, β = 6.0	45.3	35.2	35.2	19.5
topK = 3, α = 1.5, β = 6.0	46.2	35.6	35.3	19.2
topK = 3, α = 1, β = 5.0	44.4	34.8	35.0	19.2

Note: The parameters of TAA in the baseline model are set to topK = 10, α = 0.5, β = 6.0.

The value of topK under the original parameter is set to 10. The values of mAP50 and mAP50-95 of the model decrease to a certain extent when the value of topK is increased. However, when it is reduced, i.e., topK is set as 7, 5, 3, and 1, all indicators of the model increase clearly. This result confirms the first hypothesis proposed before, i.e., that in detection scenarios with dense targets and strong occlusion, too large of topK would introduce more low-quality samples, which would negatively affect the learning of features by the model. When topK is set to 1, Precision, Recall, and mAP50 of the model are the highest among the first seven comparative experiments in Table 4, and mAP50-95 is slightly lower than the result when the topK is set as 3. To further increase α, which is set as 1, for the models with topK values of 1, 3 and 5, it can be seen that the detection performance of each model is still significantly improved. This result also confirms the second hypothesis, which states that the size of the object to be detected in aerial images is tiny, and only accounts for a small number of pixels. As a result, the localization will not have a large deviation. In this type of application scenario, the model's misdetection is more serious, so the proportion of the classification scores should be increased to select samples that are more accurately classified. In the comparison of these three experiments, the model with a topK value of 3 and an α value of 1 (bolded row in the table) achieves the best result, with mAP50 and mAP50-95 values of 35.6% and 19.8%, respectively. It is worth noting that the detection ability does not change obviously when topK = 1, α = 1. A possible explanation is that reducing the value of topK, while effective in reducing the number of poor-quality samples, also significantly reduces the number of positive samples selected. By raising the value of α, the weight of classification score in sample selection can be increased to filter out more high-quality samples. However, if the value of topK is too small, some samples that are beneficial to model training may be filtered out. When the value of α is further increased, i.e., α = 1.5, and all else being equal, the performance of the model appears worse. A similar situation occurs when α is no longer increased and the value of β is decreased, i.e., β = 5.

It can be seen that these parameters need to be fine-tuned according to the detection task. The model will be adversely affected by them being set to inappropriate values. Through the above series of comparative experiments, the best experimental results appear when the values of topK, α and β are set to 3, 1.0 and 6.0, respectively. By reducing the topK value, the detection effect can be significantly improved in the detection of small targets in aerial images, but setting it too small will excessively limit the number of positive samples. The values of α and β correspond to the scores of classification and positioning, respectively. In this type of detection, both values can be adjusted to increase the proportion of accurate classification, thereby further enhancing the detection effectiveness.

4.3 Ablation Experiments

To validate that each improvement strategy proposed in this paper has an enhancement effect on the performance of the target detection model for aerial images, this paper takes the YOLOv8n model as the benchmark, adds corresponding improvement measures to it in turn, and conducts a series of ablation experiments under the same experimental environment. The results are presented in Table 5.

Methods A to D correspond to adding the improved detection heads, the improved SPPF module, the improved up-sampling module, and the optimized parameters of TAA based on the baseline model, respectively. From the results, all these four improvement strategies play a positive role in improving the performance of the model, and the indexes, such as Precision and Recall have been improved to varying degrees. Among them, method D has the greatest improvement in mAP50 and mAP50-95, with increases of 3.4% and 1.6% respectively over the original model. The second is method A, which has a 2.8% improvement in mAP50 and a 1.6% improvement in mAP50-95. Methods B and C similarly improve the model with only a small increase in parameters and Flops. Methods E to G add the other three strategies in turn, based on the addition of small-sized combination of detection head. Although Flops increases to a certain

extent compared with YOLOv8n (about 2.8 G), other evaluation indexes show improvement compared with method A, which indicates that combining different improvement strategies is more effective than using method A alone. A similar situation occurs in Methods H and I, where different three methods are combined and produce a positive effect in both experiments. The Ours in the table represents the final YOLO-S3DT model obtained by incorporating all the improvements outlined in this paper. In a series of ablation experiments, this method achieves the best result. Precision and Recall are increased by 4.6% and 5.3%, mAP50 and mAP50-95 are increased by 5.7% and 2.8%, respectively. Parameters are decreased by approximately 0.08 M, and the computation amount is increased by less than 3 GFlops.

Table 5: Results of ablation experiments

Method	Detection heads	3DSPPF	DySample	TAA	Params/M	Flops/G	P/%	R/%	mAP50/%	mAP 50-95/%
YOLOv8n	–	–	–	–	3.01	8.1	42.8	32.6	32.2	18.3
A	√	–	–	–	2.82	10.8	45.3	34.8	35.0	19.8
B	–	√	–	–	3.11	8.2	44.3	32.8	32.8	18.5
C	–	–	√	–	3.02	8.1	44.2	32.6	32.7	18.6
D	–	–	–	√	3.01	8.1	47.2	35.3	35.6	19.8
E	√	√	–	–	2.92	10.9	46.1	35.1	35.3	19.9
F	√	–	√	–	2.83	10.9	46.2	35.3	35.4	20.3
G	√	–	–	√	2.82	10.8	47.1	37.0	36.9	20.3
H	√	√	√	√	2.93	11.0	46.1	35.9	35.6	20.3
I	–	√	√	√	3.12	8.2	45.3	36.1	35.8	20.1
Ours	√	√	√	√	2.93	11.0	47.4	37.9	37.9	21.1

Note: The combination of detection head is 160 × 160, 40 × 40 and 20 × 20.

4.4 Comparative Experiments

To further validate the advancement and effectiveness of the YOLO-S3DT model proposed in this paper, 6 target detection models from different YOLO series [23] and 5 additional target detectors are separately selected for comparative experiments. The results are presented in Tables 6 and 7.

Table 6: Results of comparative experiments with YOLO series

Model	Params/M	Flops/G	P/%	R/%	mAP50/%	mAP50-95/%
YOLOv3t [15]	12.13	18.9	35.3	22.6	21.4	11.4
YOLOv4-csp [24]	9.14	20.9	41.3	37.9	33.1	16.7
YOLOv5n	2.50	7.1	44.4	32.1	31.9	18.2
YOLOv6n [25]	4.23	11.8	38.9	28.8	28.1	15.9
YOLOv7t [26]	6.04	13.3	40.3	34.7	30.5	14.9
YOLOv8s	11.13	28.5	46.3	35.3	35.2	20.4
YOLO-S3DT	2.93	11.0	47.4	37.9	37.9	21.1

Note: The depth and width of the YOLOv4-csp model are set to 0.33 and 0.5.

Table 7: Results of comparative experiments with other detectors

Model	Params/M	Flops/G	mAP50/%	mAP50-95/%
SSD [27]	26.29	62.7	9.6	5.0
RetinaNet [28]	37.70	170.1	13.3	7.9
EfficientDet [29]	3.87	5.23	14.9	7.7
Faster R-CNN [30]	137.10	370.2	16.9	7.4
CenterNet [31]	32.67	70.22	31.2	14.9
YOLO-S3DT	2.93	11.0	37.9	21.1

In relation to the parameters, YOLO-S3DT has only 2.93 million, which is just 430,000 more than the smallest YOLOv5n, while the parameters of YOLOv3t and YOLOv8s both exceed 10 million, and YOLOv4-csp exceeds 9 million, which are 4.1, 3.8 and 3.1 times of the improved model respectively. In regards to computation, the improved model has a low complexity compared to other models, with a value of 11.0 Flops, the second lowest among all models. This is about 4 GFlops higher than YOLOv5n, and only 38.6%, 52.6%, and 58.2% of YOLOv8s, YOLOv4-csp, and YOLOv3t, respectively. In terms of the comprehensive performance of the model, the improved model observably outperforms other models in all the results. Among them, YOLOv5n outperforms YOLO-S3DT in Parameters and Flops. However, the values of Precision, Recall and mAP are significantly lower than those of YOLO-S3DT, especially the mAP50, which is 6% lower than that of YOLO-S3DT.

As shown in Table 7, the number of parameters of YOLO-S3DT is significantly lower than other models, with only EfficientDet being comparable. EfficientDet exhibits the lowest value of Flops which is approximately half of YOLO-S3DT. While it demonstrates commendable lightweight performance, its detection efficacy is subpar, ranking only slightly above RetinaNet. Among the other models, the highest mAP value is CenterNet, whose mAP50 and mAP50-95 reach 31.2% and 14.9%, respectively, but still has a significant gap with YOLO-S3DT.

Overall, the model proposed in this paper has fewer parameters and lower computational complexity, making it easy to deploy on various terminal devices for detection tasks. Additionally, the model outperforms other models in terms of detection accuracy, further confirming the superiority of the developed model.

4.5 Comparisons of Detection Effect

To visually reflect the detection capabilities of various models, we select samples from 4 different scenarios and use trained models to detect them. The comparisons are presented in Fig. 5. Each row represents the detection results in the same scenario. The left, middle, and right images are the detection outputs of YOLOv8n, YOLOv8s, and YOLO-S3DT models, respectively.

The first detection scenario is an intersection on a city road, with sufficient light and basically no obstruction, the UAV is directly above the target for long-distance shooting, and the scene contains a large number of small-sized targets, especially motorcycles and pedestrians with only a few pixel points. Based on the results, YOLOv8n performs well in detecting cars in the image. However, it tends to miss smaller motorcycles and sometimes misidentifies vans as cars. In contrast, YOLOv8s performs better in detecting cars, and also shows improved detection of motorcycles and vans compared to the former. However, there are still noticeable omissions. The model proposed in this paper has better detection ability for different types of targets in this scene. Fig. 5c shows that YOLO-S3DT detects a higher number of motorcycles and pedestrians compared to the other two models.

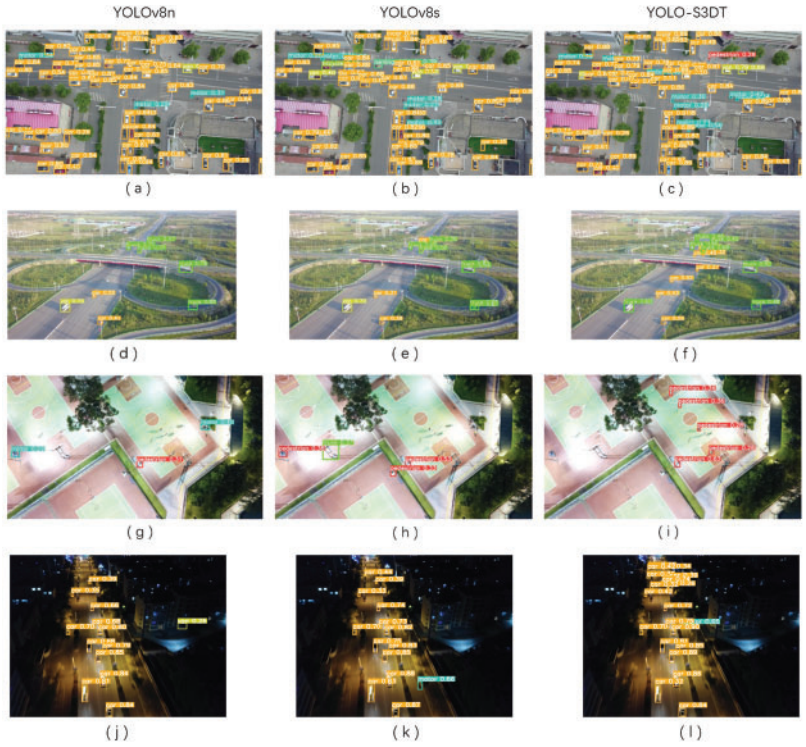

Figure 5: Comparisons of detection effect in four scenarios

The second scenario for detection is an expressway. Although there are relatively few targets in this scene, the UAV is far away from the small targets to be detected. Additionally, the high brightness of the image causes overexposure, which increases the difficulty of detection. As shown in Fig. 5d and e, both models are unable to detect cars located in the middle of the picture and further away. The improved model recognizes these targets more accurately and reduces the rate of missed detections.

The third detection scenario is under outdoor stadiums, where the pictures are taken under complex lighting conditions. Different areas within the image may have varying degrees of over or underexposure due to differences in light intensity. Due to the difficulty of detecting this type of scene, none of the three models completely detect all the targets in the image, and all of them have a high leakage rate. However, Fig. 5g and h demonstrates that the unimproved YOLOv8 models exhibit more obvious misdetections. For instance, YOLOv8n detects the facility next to the basketball court and the people around it as motorcycles, while YOLOv8s mistakenly detects a basketball hoop as a truck. In contrast, the YOLO-S3DT shows no misdetections and correctly identifies more targets.

The fourth detection scenario is a city road at night. Due to the small aperture of the UAV's lens and the weak light at night, the targets to be detected in the image have different degrees of blurriness and darkness, and the trees on both sides of the road also obscure the targets parked under them, which makes this type of scene more challenging to detect as well. In Fig. 5j and k, YOLOv8n and YOLOv8s are able to detect the cars on the road better, but struggle with detecting parked cars on the far side of the road due to occlusion. The results in Fig. 5l demonstrate that the improved model is obviously more sensitive to this class of targets and has better detection capability.

5 Conclusions

Due to the considerable distance between Unmanned Aerial Vehicles (UAVs) and their targets during operation, coupled with complex application scenarios, challenges such as false detections and missed detections frequently arise when identifying small targets within aerial imagery. In this paper, the YOLOv8n model is improved from four aspects: the detection head, the up-sampling module, the SPPF module, and the parameter optimization of positive sample screening. The YOLO-S3DT model is proposed.

Firstly, a 160 × 160 detection head is integrated into the Head component of the original network to enhance the sensitivity of the model to small-size targets. The 80 × 80 head is removed through the experiments of different combinations of detection heads for comprehensive comparison, which reduces the parameters of the model under the premise of guaranteeing the improvement of performance. Secondly, the structure of the SPPF module is redesigned by incorporating a branch for feature map reconstruction. Three DSC modules and the convolution dilation operation are used to obtain richer feature information without a large increase in computation. Additionally, the DySample is employed in the Neck component to replace the previous up-sampling block, so that the initial sampling positions are uniformly distributed to solve the problem of initial disordered sampling points, and it effectively improves the detection accuracy. Finally, the optimization of relevant parameters in TAA reduces the negative impact on the model caused by involving too many low-quality samples in the training process. Furthermore, considering the characteristics inherent in small-target detection tasks, the proportion of classification scores in the positive sample selection process is increased. Experimental results validate the effectiveness of these proposed optimizations.

The ablation and comparative experiments demonstrate not only that these methods positively contribute to improving detection capability but also that our approach exhibits significant advantages over other models regarding lightweight and accuracy. Consequently, it proves more suitable for application tasks within UAV photography scenarios characterized by limited computing resources.

Acknowledgement: We are grateful to the editors and reviewers of this journal for many helpful comments.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Pengcheng Gao and Zhenjiang Li; methodology, Pengcheng Gao; software, Pengcheng Gao; validation, Pengcheng Gao; formal analysis, Pengcheng Gao; investigation, Pengcheng Gao; resources, Pengcheng Gao; data curation, Pengcheng Gao and Zhenjiang Li; writing—original draft preparation, Pengcheng Gao; writing—review and editing, Pengcheng Gao and Zhenjiang Li; visualization, Pengcheng Gao; supervision, Zhenjiang Li; project administration, Zhenjiang Li. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

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

References

1. Fu J, Zhang Z, Sun W, Zou K. Improved YOLOv8 small target detection algorithm in aerial images. J Comput Eng Appl. 2024;60(6):100–9.
2. Peng C, Zhu M, Ren H, Emam M. Small object detection method based on weighted feature fusion and CSMA attention module. Electronics. 2022;11(16):2546. doi:10.3390/electronics11162546.

3. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference; 2014; Zürich, Switzerland. p. 740–55.
4. Shen S, Zhang X, Yan W, Xie S, Yu B, Wang S. An improved UAV target detection algorithm based on ASFF-YOLOv5s. Math Biosci Eng. 2023;20(6):10773–89. doi:10.3934/mbe.2023478.
5. Zhao GQ, Pan ZS, Li YB. Small target detection in UAV aerial image based on high-performance feature extraction and task decoupling. Microelect Comput. 2024;41(4):55–63.
6. Zhang S, Wen L, Bian X, Lei Z, Li SZ. Single-shot refinement neural network for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018; Salt Lake City, UT, USA. p. 4203–12.
7. Ying X, Wang Q, Li X, Yu M, Jiang H, Gao J, Liu Z, Yu R. Multi-attention object detection model in remote sensing images based on multi-scale. IEEE Access. 2019;7:94508–19. doi:10.1109/ACCESS.2019.2928522.
8. Bai Y, Zhang Y, Ding M, Ghanem B. SOD-MTGAN: small object detection via multi-task generative adversarial network. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; Munich, Germany. p. 206–21.
9. Maktab Dar Oghaz M, Razaak M, Remagnino P. Enhanced single shot small object detector for aerial imagery using super-resolution, feature fusion and deconvolution. Sensors. 2022;22(12):4339. doi:10.3390/s22124339.
10. Liu Z, Gao X, Wan Y, Wang J, Lyu H. An improved YOLOv5 method for small object detection in UAV capture scenes. IEEE Access. 2023;11(4):14365–74. doi:10.1109/ACCESS.2023.3241005.
11. Kim M, Jeong J, Kim S. ECAP-YOLO: efficient channel attention pyramid YOLO for small object detection in aerial image. Remote Sens. 2021;13(23):4851. doi:10.3390/rs13234851.
12. Wang X, He N, Hong C, Wang Q, Chen M. Improved YOLOX-X based UAV aerial pho-tography object detection algorithm. Image Vis Comput. 2023;135:104697. doi:10.1016/j.imavis.2023.104697.
13. Wang G, Chen Y, An P, Hong H, Hu J, Huang T. UAV-YOLOv8: a small-object-detection model based on improved YOLOv8 for UAV aerial photography scenarios. Sensors. 2023;23(16):7190. doi:10.3390/s23167190.
14. Lee SS, Lim LG, Palaiahnakote S, Cheong JX, Lock SSM, Ayub MNB. Oil palm tree detection in UAV imagery using an enhanced RetinaNet. Comput Electron Agric. 2024;227:109530. doi:10.1016/j.compag.2024.109530.
15. Zhu P, Wen L, Du D, Bian X, Fan H, Hu Q, Ling H. Detection and tracking meet drones challenge. IEEE Trans Pattern Anal Mach Intell. 2021;44(11):7380–99. doi:10.1109/TPAMI.2021.3119563.
16. Hussain M. YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection. Machines. 2023;11(7):677. doi:10.3390/machines11070677.
17. Redmon J. YOLOv3: an incremental improvement. arXiv:1804.02767. 2018.
18. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16. doi:10.1109/TPAMI.2015.2389824.
19. Chollet F. Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017; Honolulu, HI, USA. p. 1251–8.
20. Yu F. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122. 2015.
21. Liu W, Lu H, Fu H, Cao Z. Learning to upsample by learning to sample. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023; Paris, France. p. 6027–37.
22. Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA. p. 1874–83.
23. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA. p. 779–88.
24. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection. arXi:2004.10934. 2020.
25. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: a single-stage object detection framework for industrial applications. arXiv:2209.02976. 2022.

26. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of- the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023; Boston, MA, USA, CVPR. p. 7464–75.
27. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. Ssd: single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference; 2016; Amsterdam, The Netherlands. p. 21–37.
28. Lin T, Goyal P, Girshick RB, He K, Dollar P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017; Venice, Italy. p. 2999–3007.
29. Tan M, Pang R, Le QV. EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; Seattle, WA, USA. p. 10781–90.
30. Jiang H, Learned-Miller E. Face detection with the faster R-CNN. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017); 2017; Washington, WA, USA. p. 650–7.
31. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Seoul, Republic of Korea. p. 6569–78.