Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/MBE

MBE, 20(11): 19040–19064.
DOI: 10.3934/mbe.2023842
Received: 16 July 2023
Revised: 11 September 2023
Accepted: 28 September 2023
Published: 11 October 2023

Research article

An infrared small target detection model via Gather-Excite attention and normalized Wasserstein distance

Kangjian Sun[1], **Ju Huo**[1,*], **Qi Liu**[1] **and Shunyuan Yang**[2]

[1] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 150001, China
[2] School of Astronautics, Harbin Institute of Technology, Harbin 150001, China

* **Correspondence:** Email: torch@hit.edu.cn.

Abstract: Infrared small target detection (ISTD) is the main research content for defense confrontation, long-range precision strikes and battlefield intelligence reconnaissance. Targets from the aerial view have the characteristics of small size and dim signal. These characteristics affect the performance of traditional detection models. At present, the target detection model based on deep learning has made huge advances. The You Only Look Once (YOLO) series is a classic branch. In this paper, a model with better adaptation capabilities, namely ISTD-YOLOv7, is proposed for infrared small target detection. First, the anchors of YOLOv7 are updated to provide prior. Second, Gather-Excite (GE) attention is embedded in YOLOv7 to exploit feature context and spatial location information. Finally, Normalized Wasserstein Distance (NWD) replaces IoU in the loss function to alleviate the sensitivity of YOLOv7 for location deviations of small targets. Experiments on a standard dataset show that the proposed model has stronger detection performance than YOLOv3, YOLOv5s, SSD, CenterNet, FCOS, YOLOXs, DETR and the baseline model, with a mean Average Precision (mAP) of 98.43%. Moreover, ablation studies indicate the effectiveness of the improved components.

Keywords: infrared small target detection; YOLOv7; anchor update; Gather-Excite attention; normalized Wasserstein distance

1. Introduction

Infrared detection technology is one of the main means to obtain modern information. Compared with visible detection systems, the infrared detection system has the advantages of strong penetration, long detection distance and all-weather visibility. Therefore, infrared detection technology attracts more and more researchers and is widely used in military [1], medical [2], meteorological [3] and other fields. With the gradual opening of low-altitude airspace, unmanned aerial vehicles (UAVs) can be used to collect and track ground targets by carrying infrared equipment. How to effectively detect small targets from the aerial view has significant theoretical significance and engineering demand, as well as social value and economic significance.

In recent years, with the rapid development of deep learning technology, the target detection method has also changed from the traditional method based on manually designed features to the deep neural network (DNN) method based on automatically learned features [4,5]. The deep learning-based target detection methods are generally divided into two-stage methods and one-stage methods [6]. The two-stage methods generate region proposals and then classify them. The classic models are the region-convolutional neural network (R-CNN) series [7], including Fast R-CNN [8], Faster R-CNN [9], Mask R-CNN [10] and so on. They have high detection accuracy, but their detection speed is slow. It is difficult to apply in real-time detection scenarios. The one-stage methods do not have the stage of generating region proposals. They directly generate the final detection results through one stage, so they have a faster detection speed. The classic models are the YOLO series [11], including YOLOv3 [12], YOLOv5 [13], YOLOX [14] and so on.

YOLOv7 [15] is a novel model of the YOLO series, which surpasses most known target detectors in terms of accuracy and speed. Since 2022, YOLOv7 has been implemented in some real-world detection tasks. Soeb et al. [16] created a leaf image dataset from Bangladesh and used YOLOv7 for disease diagnosis. This study provided a solution for precision agriculture applications. Li et al. [17] improved YOLOv7 by embedding gamma correction, improved convolutional block attention module and Alpha GIOU. The improved model was used for the damages detection of aeroengine blades. Driver abnormal behavior is a serious threat to public safety. Liu et al. proposed the CEAMYOLOv7 model for distraction behavior recognition. The global attention mechanism (GAM) was introduced into YOLOv7 to enhance the network's capability to extract key features. The channel expansion (CE) method was also proposed for data augmentation. Moreover, the lightweight processing made the model easier to be deployed. More projects based on YOLOv7 are still being explored [18].

Although the above models show impressive performance in related works, the task of infrared small target detection is still a challenge. On the one hand, due to the long observation distance there is little shape and texture information of infrared small targets. On the other hand, due to the complex background infrared small targets may be obscured and overlapped [19,20]. To detect infrared small targets, researchers have developed some pioneering works. Zhang et al. [21] incorporated target shape reconstruction into the detection of infrared small targets and proposed the ISNet model. Based on Taylor finite difference (TFD)-inspired edge block and two-orientation attention aggregation (TOAA) block, the model can effectively extract edge features and aggregate cross-level features. Additionally, the authors established a new large-scale benchmark, IRSTD-1k, to validate the effectiveness of the proposed idea. To handle the problem of the loss of targets in deep layers, Li et al. [22] proposed a dense nested attention network (DNA-Net). Specifically, the dense nested interactive module (DNIM) and the cascaded channel and spatial attention module (CSAM) were designed to achieve repetitive fusion and enhancement between feature layers. Additionally, an infrared small target dataset, namely NUDT-SIRST, was developed. Results on a set of proposed evaluation metrics showed that the proposed method achieved better performance. A multi-level TransUNet (MTU-Net) in [23] was proposed to detect space-based infrared tiny ships. The Vision Transformer (ViT) Convolutional

Neural Network (CNN) hybrid can extract multi-level features. Wu et al. also proposed a copy-rotateresize-paste (CRRP) data augmentation method that alleviates the problem of sample imbalance. Additionally, the authors designed a FocalIoU loss to achieve target localization and shape description. Establishing the largest space-based infrared tiny ship detection dataset NUDT-SIRSTSea was a significant work. In 2022, Lin et al. [24] comprehensively considered the detection performance and practical deployment, and proposed a light-weight infrared small target detection network LIRDNet. This model combined cross-scale feature fusion module (CFM) and bottleneck attention module (BAM). The experimental results demonstrated that the CFM and BAM modules further improved the detection performance with a low amount of parameters and computations. Liu et al. [25] proposed a lightweight model for ship detection in SAR images. Authors added the coordinate attention into the backbone of YOLOv7-tiny, and improved the SPP block and the loss function. Compared with the original model, the precision of the proposed model was increased by 4.6%. This work had not yet been deployed on edge devices. Similarly, Guo et al. [26] also proposed a lightweight SAR ship target detection based on YOLO, namely LMSD-YOLO. This model has better multi-scale adaptation capabilities and has been successfully deployed on mobile platforms. However, there are still difficulties in implementing target detection directly from large-scale SAR images. Zhou et al. [27] improved YOLOv5 to make the model to perform the small target detection task. It is worth noting that authors used the Super-Resolution Generative Adversarial Network (SRGAN) to generate super-resolution images and input images into the improved detection model. Experiments verified that the super-resolution reconstruction for images can improve the detection accuracy of small targets. The disadvantage is that the process of super-resolution reconstruction is very time-consuming.

In this paper, the recent YOLOv7 model as the baseline is used for infrared small target detection. To make the model better adapt to this task domain, we make targeted improvements to YOLOv7 and propose a new detection model namely ISTD-YOLOv7. Our main contributions are summarized as follows:

1) An improved YOLOv7 model (namely, ISTD-YOLOv7) is proposed for infrared small target detection.
2) The update of anchors can make the model to converge better and faster. Feature context and spatial location information can be efficiently exploited by GE attention. NWD can alleviate the sensitivity location deviations of small targets.
3) The performance of ISTD-YOLOv7 is compared with existing models. Ablation studies are performed to investigate the impact of each component. Experiments on a public dataset demonstrate the superiority of the proposed model in infrared small target detection.

The remainder of this paper is organized as follows: Section 2 briefly introduces the YOLOv7 model. Section 3 describes the mechanism of the improved components and presents the improved model. Experimental results and analysis are given in Section 4. Section 5 summarizes the work of this paper.

2. YOLOv7 model

YOLOv7, as one of the latest representative models of the YOLO series, was proposed by Wang et al. [15] in 2022. Compared to previous YOLO series, the main contributions of YOLOv7 are that authors proposed the model re-parameterization, model scaling, extended efficient layer aggregation networks (E-ELAN), etc. This series of architectural alterations makes YOLOv7 not only more accurate, but also faster. The concise network structure of YOLOv7 is shown in Figure 1 [15]. More details of the component blocks can be found in [15].

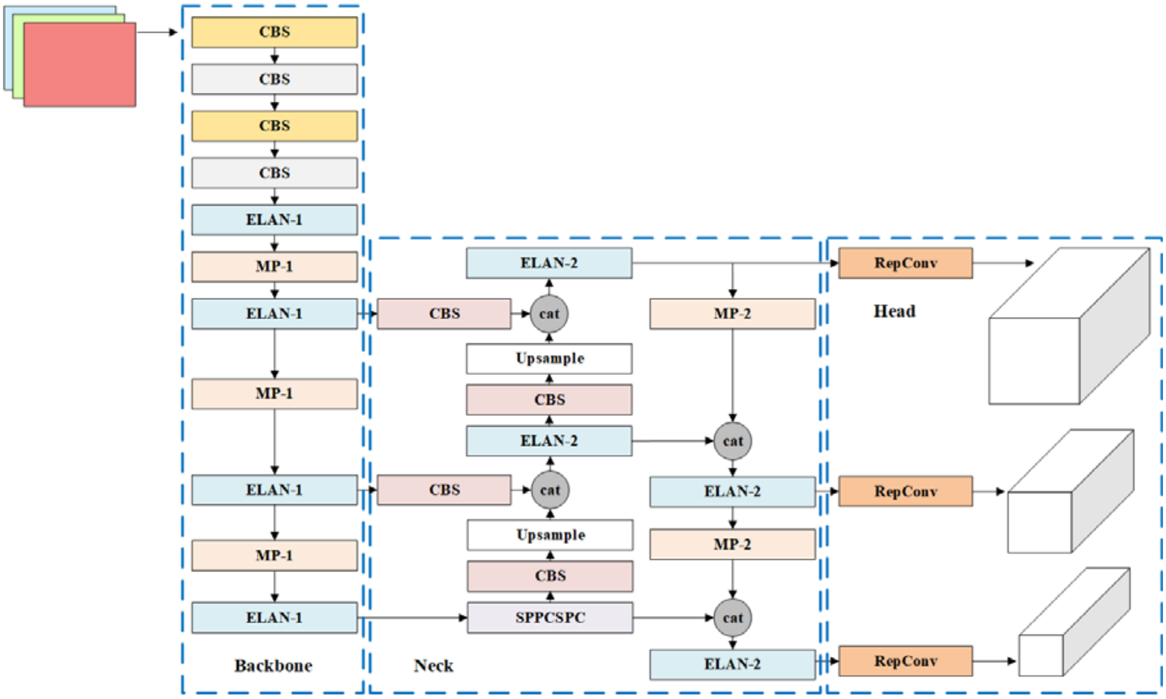


Figure 1. YOLOv7 model [15].

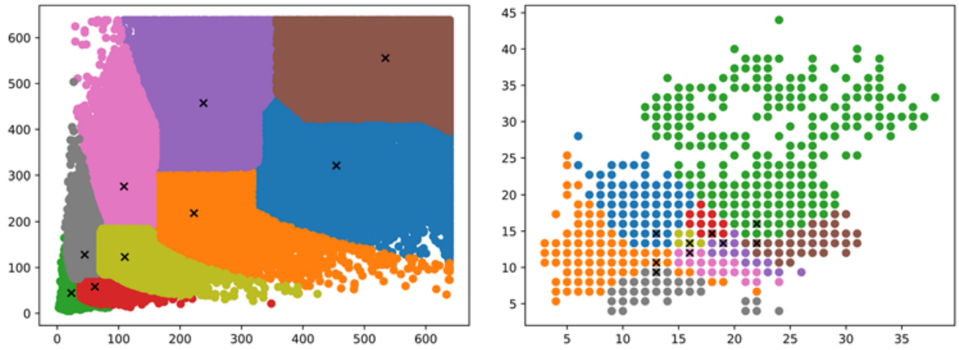
First, the model resizes the input images to (640 × 640) pixels. Then, the images are input to the backbone network for feature extraction. The backbone network of YOLOv7 consists of several CBS blocks, ELAN blocks and MP blocks. The obtained features of different scales are fused by the neck network. The neck network adopts the structure of the path aggregation feature pyramid network (PAFPN). Then, the head (prediction) network adjusts the number of channels of feature maps based on the RepConv blocks. Finally, the bounding box information confidence and category probability are output.

3. Proposed ISTD-YOLOv7 model

3.1. Anchor update

The sizes of the anchors are obtained by clustering the width and height of the ground-truth boxes of the training samples. Whether the anchors are reasonable or not greatly affects the detection performance of the model. Generally speaking, the anchors of YOLOv7 are obtained by clustering based on the VOC dataset or the COCO dataset in the training process. VOC dataset provides 20 classes of targets, including person, horses, bicycles, motorbike and more [28]. The COCO dataset focuses on scene understanding and provides 80 classes of targets. These targets are mainly obtained from everyday scenes [29]. The VOC dataset and the COCO dataset are common large-scale datasets in target detection. However, the sizes of targets in these datasets are significantly different from those in infrared small target datasets.

In this paper, in order to make YOLOv7 converge better and faster, the K-means method is used to re-cluster the sizes of the targets based on the selected dataset. The number K of clustering centers is set to 9. The selected dataset in this paper is described in Section 4. Figure 2 shows the clustering results of the VOC dataset and the selected dataset. It can be seen that the distribution of cluster centers varies greatly. The target size of the VOC dataset can be several hundred pixels, while the target size of the selected dataset is obviously much smaller. Table 1 gives the results of anchors. The anchor update can provide a reasonable prior for the detection model.



(a) VOC dataset

(b) Selected dataset

Figure 2. Results of clustering.

Table 1. Results of anchors.

Dataset	Anchor (pixels)		
VOC dataset	(23, 44),	(61, 58),	(44, 128),
	(110, 122),	(108, 276),	(222, 218),
	(238, 457),	(454, 320),	(534, 555).
Selected dataset	(12, 9),	(12, 10),	(13, 14),
	(16, 11),	(16, 13),	(18, 13),
	(18, 14),	(22, 13),	(21, 16).

3.2. Gather-Excite attention

For images or feature maps, the context information of the space can improve the representational capability of the network. In 2018, the Gather-Excite (GE) attention mechanism was proposed by Hu et al. [30]. This mechanism defines two operators: gather operator and excite operator. Figure 3 shows the operation process of the two operators [30]. The gather operator ξ_G extracts features from local spatial locations, defined as shown in Eq (1). The excite operator ξ_E maps features to the original scale, defined as shown in Eq (2).

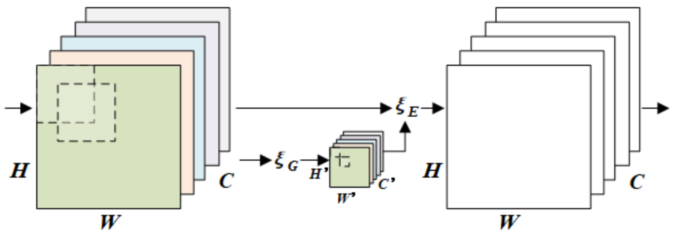


Figure 3. GE block [30].

$$\xi_G : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{H'\times W'\times C} \tag{1}$$

where *H*, *W* and *C* represent the height, width and channel of any input *x*, *e* represents the extent ratio, $H'=H/e$, $W'=W/e$. A global extent ratio using global average pooling is used in this paper.

$$\xi_E(x,\hat{x}) = x \odot f(\hat{x}) \tag{2}$$

$$f : \mathbb{R}^{H'\times W'\times C} \rightarrow [0,1]^{H\times W\times C} \tag{3}$$

where $\hat{x}$ represents the output after processing by ξ_G, $\odot$ represents the Hadamard product, *f* represents a map relationship.

In this paper, three GE attention blocks are added at three output branches of the backbone network of YOLOv7 respectively. The diagram is shown in Figure 4. Infrared small targets have the characteristics of small size and dim signal. Therefore, location information is essential for the detection of small targets. By adding GE attention blocks to the backbone feature extraction network of YOLOv7, the model can more efficiently exploit feature context and spatial location information for infrared small targets.

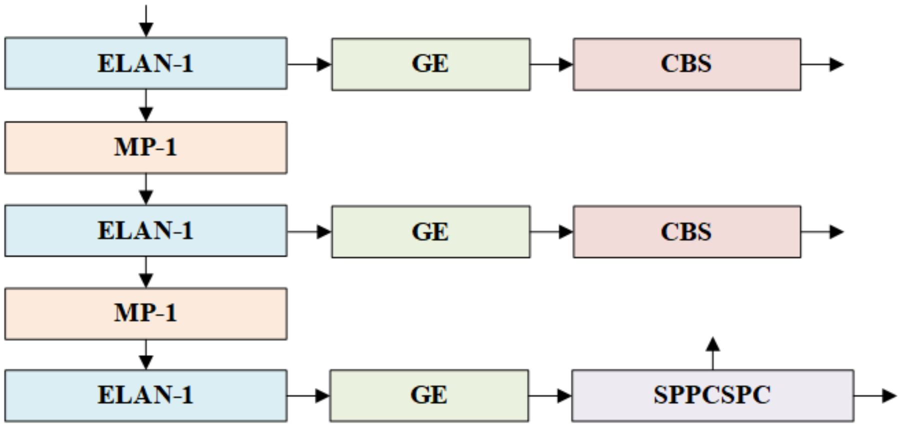


Figure 4. Diagram of adding location.

3.3. Normalized Wasserstein distance

The sensitivity of IoU metric to targets with different scales is quite variant. For small targets, a slight location change may lead to a significant change in IoU. However, for targets with normal size, the change of IoU is slight for the same location deviation [31]. Figure 5 gives a specific analysis. For a small target, a location deviation leads to an IoU drop from 0.47 to 0.02. However, for a normal target, the same location deviation only leads to an IoU drop from 0.83 to 0.49.

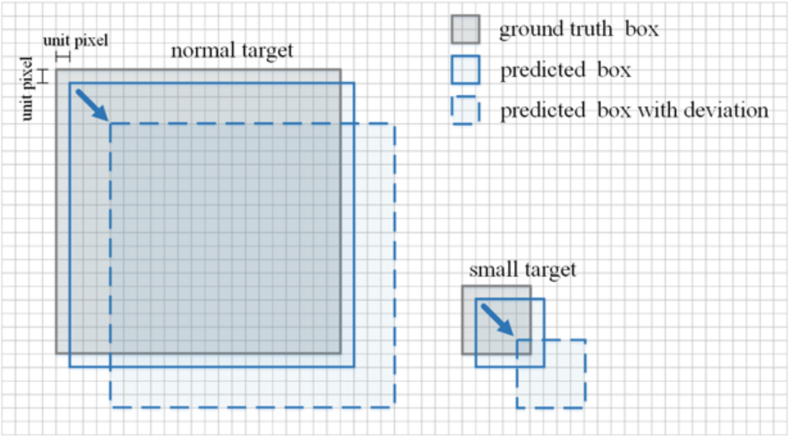
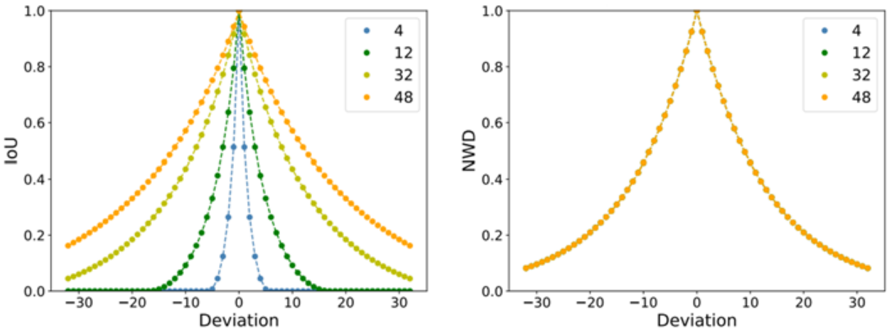


Figure 5. Sensitivity analysis of IoU.

Wang et al. [31] proposed a novel metric method based on the Wasserstein distance. Specifically, the bounding box is modeled as the 2D Gaussian distribution, and then the similarity between the corresponding Gaussian distributions is calculated by using the proposed metric, namely the Normalized Wasserstein Distance (NWD). Figure 6 [31] shows the deviation curves of IOU and NWD under different target sizes. As the target size becomes smaller, the IoU-deviation curves decrease faster, while the NWD-deviation curves remain overlapped and smooth. Compared with IOU, NWD is insensitive to location deviations of small targets. Some research has been presented in the literature regarding the theoretical and empirical benefits of using NWD [32–34].



(a) IoU-deviation curve (b) NWD-deviation curve

Figure 6. Deviation curves of IoU and NWD [31].

Specifically, for a bounding box (*cx*, *cy*, *w*, *h*), the intrinsic elliptic of the bounding box can be expressed as:

$$\frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} = 1 \tag{4}$$

where (*cx*, *cy*), *w* and *h* represent the center coordinate, width and height of the bounding box respectively. (μ_x, μ_y), σ_x and σ_y represent the center coordinates of the ellipse, the length of the *X*-axis and the length of the *Y*-axis respectively. Therefore, $\mu_x = cx$, $\mu_y = cy$, $\sigma_x = w/2$ and $\sigma_y = h/2$. The probability density function of the 2D Gaussian distribution is as follows:

$$f(\boldsymbol{x} | \mu, \Sigma) = \frac{\exp\left(-\frac{1}{2} (\boldsymbol{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x} - \mu) \right)}{2\pi |\Sigma|^{\frac{1}{2}}} \tag{5}$$

where $\boldsymbol{x}$, μ and Σ represent the coordinate (x, y), mean and co-variance of the distribution respectively. When $(\boldsymbol{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x} - \mu) = 1$, the bounding box can be modelled as a 2D Gaussian distribution $N(\mu, \Sigma)$ with:

$$\mu = \begin{bmatrix} cx \\ cy \end{bmatrix}, \ \Sigma = \begin{bmatrix} \frac{w^2}{4} & 0 \\ 0 & \frac{h^2}{4} \end{bmatrix} \tag{6}$$

For Gaussian distributions N_a and N_b which are modeled from bounding boxes (cx_a, cy_a, w_a, h_a) and (cx_b, cy_b, w_b, h_b), the Wasserstein distance is shown in Eq (7). After normalization, the final form of NWD metric is obtained, namely Eq (8).

$$W_2^2(N_a, N_b) = \left\| \left(\left[cx_a, cy_a, \frac{w_a}{2}, \frac{h_a}{2} \right]^{\mathrm{T}}, \left[cx_b, cy_b, \frac{w_b}{2} \cdot \frac{h_b}{2} \right]^{\mathrm{T}} \right) \right\|_2^2 \tag{7}$$

$$NWD(N_a, N_b) = \exp\left(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C} \right) \tag{8}$$

In this paper, NWD is integrated into YOLOv7 to replace IoU. The specific improvement part is the loss function of YOLOv7. NWD-based regression loss can not only solve the issue that YOLOv7 is sensitive to the location deviation of small targets, but also still provide gradient to optimize the network in some cases. The improved loss function of YOLOv7 is as follows:

$$L_{ISTD\text{-}YOLOv7} = 1 - NWD(N_p, N_g) \tag{9}$$

where N_p and N_g represent the Gaussian distribution model of prediction box *p* and ground-truth box *g* respectively.

3.4. ISTD-YOLOv7 model

In order to more effectively detect small targets in infrared image data, we propose the ISTD-YOLOv7 model, which can maintain good performance. The diagram of the model is shown in Figure 7. First, the infrared images enter the backbone network consisting of convolution groups to extract features. After that, these features enter designed GE blocks. GE blocks are added at three output branches of the backbone network to exploit feature context and spatial location information. Then, the neck network with PAFPN structure is used for feature fusion, producing better semantic information. Finally, the feature maps of various scales enter the head network to produce the prediction results.

The purpose of the training process is to continuously reduce the difference between the prediction results and ground truth boxes. In this paper, the prediction results are iteratively optimized by the NWD-based loss function. The NWD metric is insensitive to location deviations of small targets. For the testing process, we use the trained model for inference and obtain the prediction results. The size of the small target is re-clustered to obtain anchors. The predicted bounding boxes are adjusted based on updated anchors. Then, the final detection result is obtained after non-maximum suppression (NMS) [35].

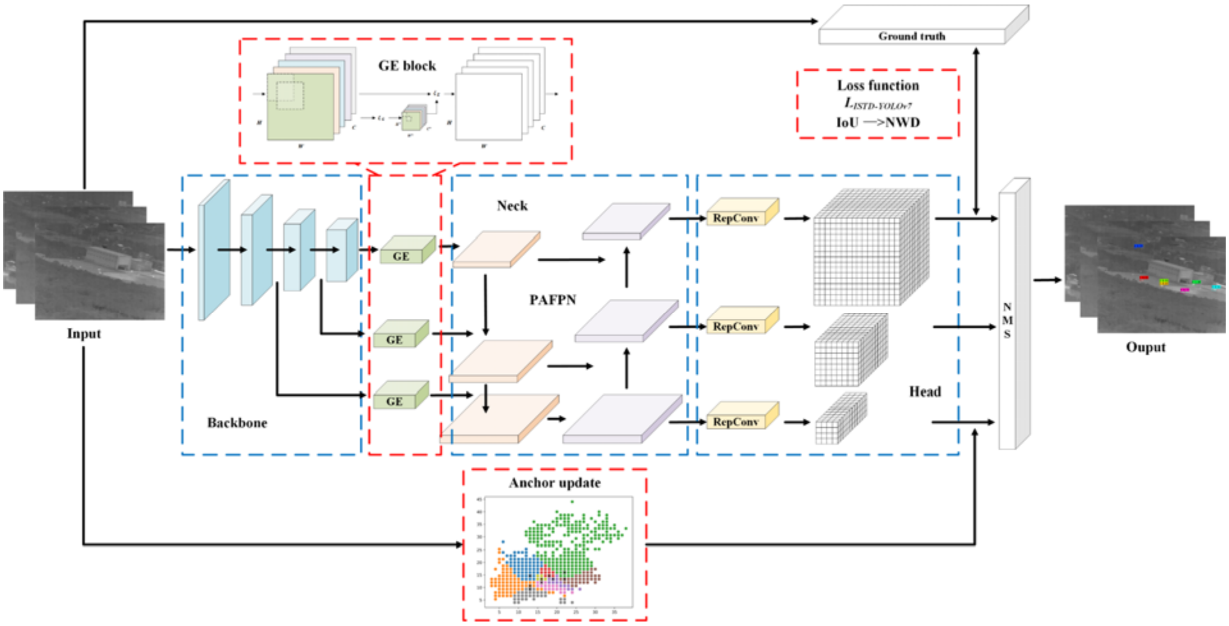


Figure 7. ISTD-YOLOv7 model.

4. Experimental results and analysis

4.1. Experiments platform

All experiments are run on a computer with an Intel(R) Core(TM) i9-12900KF (64 GB DDR5) CPU, one NVIDIA GeForce RTX 3090Ti (24 GB) GPU and the Microsoft Windows 10 system. The deep learning framework is PyTorch 1.7.1. The stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, a weight decay of 0.0005 and a momentum of 0.937 is chosen to reduce the loss function. The batch size is 32 and the number of epochs is 300.

4.2. Dataset

The dataset in this paper was published by Fu et al. [36] and has been used in some official competitions. All images in this dataset were taken by a UAV equipped with an infrared camera. The dataset includes 21,750 images, 8 classes and 89,174 targets, where targets are some vehicles under ground background. More details of the dataset are given in Table 2. We randomly divided the training set, validation set and testing set in the ratio of 8:1:1. The main challenges of this dataset focus on the complex environment interference and complex imaging conditions. It can provide material bases for the research of infrared image characteristics, infrared small target detection and tracking.

Table 2. Details of dataset.

Resolution	Depth	Format	Memory
(640 × 480) pixels	8 bit	.bmp	≈300 k

4.3. Evaluation indices

In order to evaluate the detection performance of the model, some evaluation indices are selected in this paper, including: Precision, Recall, F1 score, Average Precision and mean Average Precision. These indices are all in the range of [0,1], and the larger the values are, the better the results will be. Their equations are as follows [37,38]:

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

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

$$F1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

$$AP = \int_0^1 P(R) \cdot \mathrm{d}R \tag{13}$$

$$mAP = \frac{1}{C}\sum_i^C AP_i \tag{14}$$

where *TP* represents true positive, *FP* represents false positive and *FN* represents false negative. The confusion matrix is given in Table 3. *C* represents the number of classes. *P* represents Precision, *R* represents Recall, *F*1 represents F1 score, *AP* represents Average Precision and *mAP* represents mean Average Precision. *mAP* is the mean of *AP*s of all classes and enables the evaluation of the overall detection accuracy of the model.

Table 3. Confusion matrix.

	Predicted result = Positive	Predicted result = Negative
Actual result = Ture	*TP* (True Positive)	*FN* (False Negative)
Actual result = False	*FP* (False Positive)	*TN* (True Negative)

The above indices can evaluate the pixel-level performance. Some research [21–24] has demonstrated that target-level performance is also important for infrared small target with limited shape and texture information. The probability of detection and the false-alarm rate are defined as follows:

$$P_d = \frac{T_{correct}}{T_{All}} \tag{15}$$

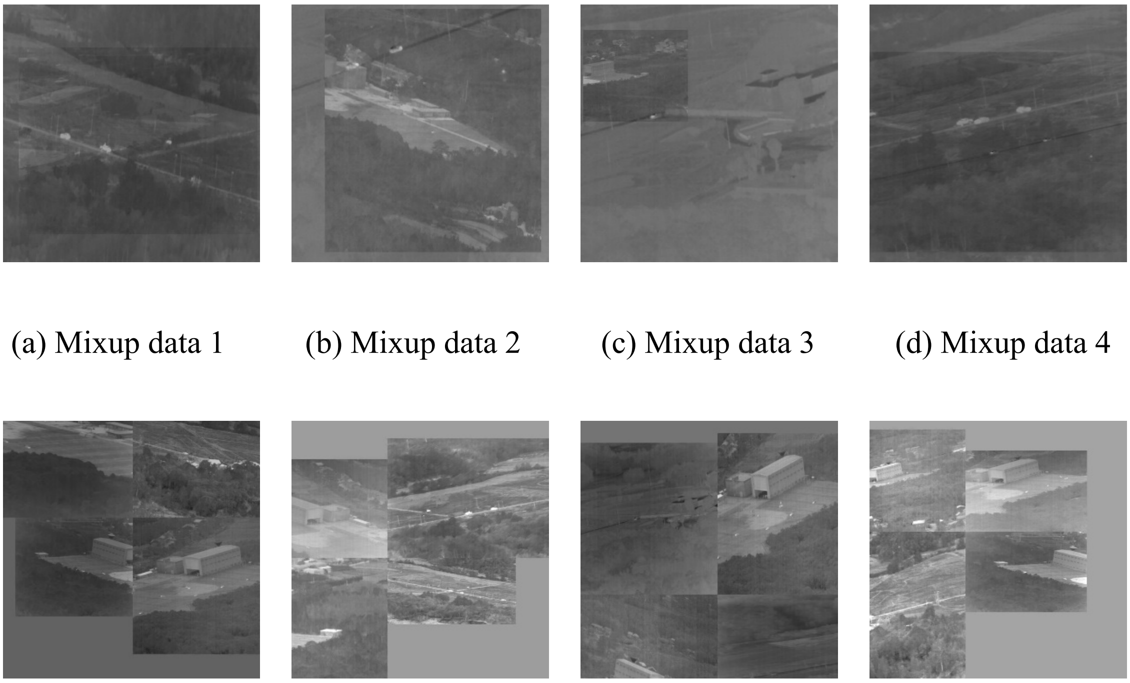
$$F_a = \frac{P_{false}}{P_{All}} \tag{16}$$

where P_d represents the probability of detection, F_a represents the false-alarm rate. $T_{corerect}$ represents the correctly predicted target number, T_{All} represents all target number. Targets are correctly predicted if the centroid deviation of the targets is smaller than the threshold $T_{distance}$. In this paper, $T_{distance}$ is set to 3 [21–24]. P_{false} represents the falsely predicted pixels, P_{All} represents all image pixels. Pixels are incorrectly predicted if the centroid deviation of the targets is larger than the threshold $T_{distance}$.

4.4. Comparison with the baseline model

In this section, the performance of ISTD-YOLOv7 and YOLOv7 is compared from three aspects: training process, verification process and testing process. Before training the two models, data augmentation technologies are used to enhance the data randomly. Taking two data augmentation methods, Mixup and Mosaic, as examples, Figure 8 shows the infrared image results obtained after processing by the two methods. Mixup uses simple linear interpolation on two random infrared images to construct new training samples, as shown in Figure 8(a)–(d). Mosaic randomly intercepts four infrared images and merges them into one infrared image as new training data, as shown in Figure 8(e)–(h). Data augmentation technology can greatly enrich the training data, improve the generalization capability of the model and make the network more robust.

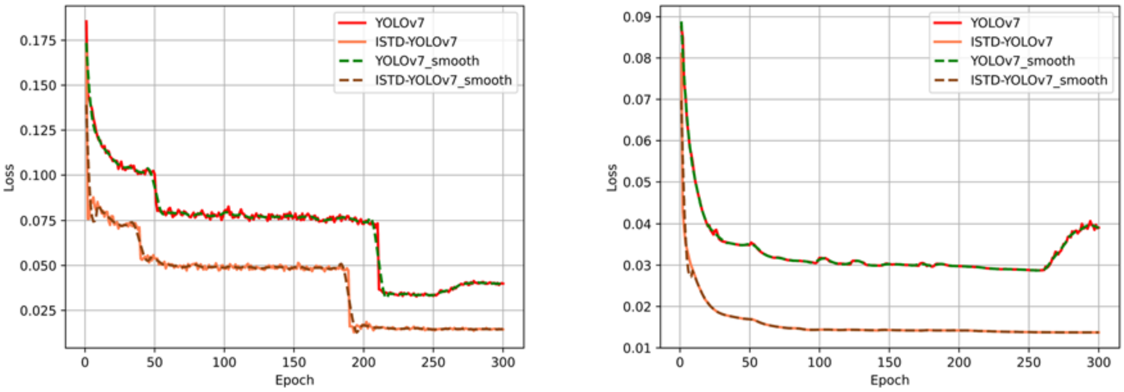
Figure 9 shows the convergence curves of ISTD-YOLOv7 and YOLOv7 on the training set and the verification set respectively. The red line is the original data of ISTD-YOLOv7, the coral line is the original data of YOLOv7, the green line is the smoothed data of ISTD-YOLOv7, and the brown line is the smoothed data of YOLOv7. It can be seen from Figure 9(a) that, in the training process, the convergence curve of ISTD-YOLOv7 is located below the convergence curve of YOLOv7. It shows that the convergence accuracy of ISTD-YOLOv7 is better than that of YOLOv7. In addition, it can be seen that the convergence speed of ISTD-YOLOv7 is better than that of YOLOv7. Specifically, ISTD-YOLOv7 escapes from local optima more quickly, achieving global optima at about 190 iterations, while YOLOv7 needs more than 210 iterations to achieve convergence. Similarly, it can be seen from Figure 9(b) that in the verification process, the convergence curve of ISTD-YOLOv7 is more stable and flatter, and the whole is located below the convergence curve of YOLOv7. It is worth noting that, after 250 iterations, the convergence curve of YOLOv7 shows a significant rise. It means that the YOLOv7 model is overfitting, while the ISTD-YOLOv7 model can better characterize this hard dataset of infrared small targets.



(a) Mixup data 1 (b) Mixup data 2 (c) Mixup data 3 (d) Mixup data 4

(e) Mosaic data 1 (f) Mosaic data 2 (g) Mosaic data 3 (h) Mosaic data 4

Figure 8. Results of data augmentation.



(a) Training set loss curve (b) Validation set loss curve

Figure 9. Loss curve of YOLOv7 and ISTD-YOLOv7.

On the basis of comparing the training process and the verification process, the performance of the two models is evaluated on the testing set. The testing set contains 2175 infrared small target images. The number of targets in each class is shown in Figure 10. Table 4 compares the evaluation results of the two models on the testing set. Note that the best result in this paper is marked in bold. From Table 4, it can be found that ISTD-YOLOv7 has improvements compared with YOLOv7 in precision (from 97.52% to 98.80%), recall (from 96.23% to 96.87%), F1 (from 96.87% to 97.83%) and mAP (from 97.44% to 98.43%). These are made possible by the application of improvements enhancing the feature extraction capability of the network for limited information, improving the recall of the model and making ISTD-YOLOv7 detect more precisely.

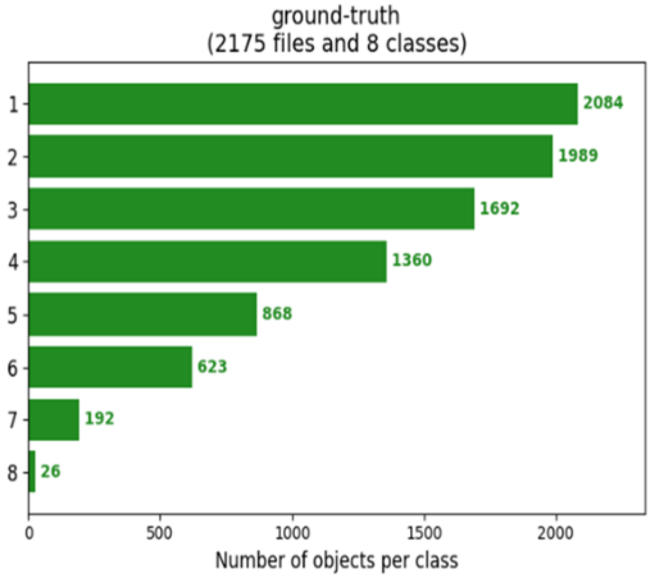


Figure 10. Information about testing set.

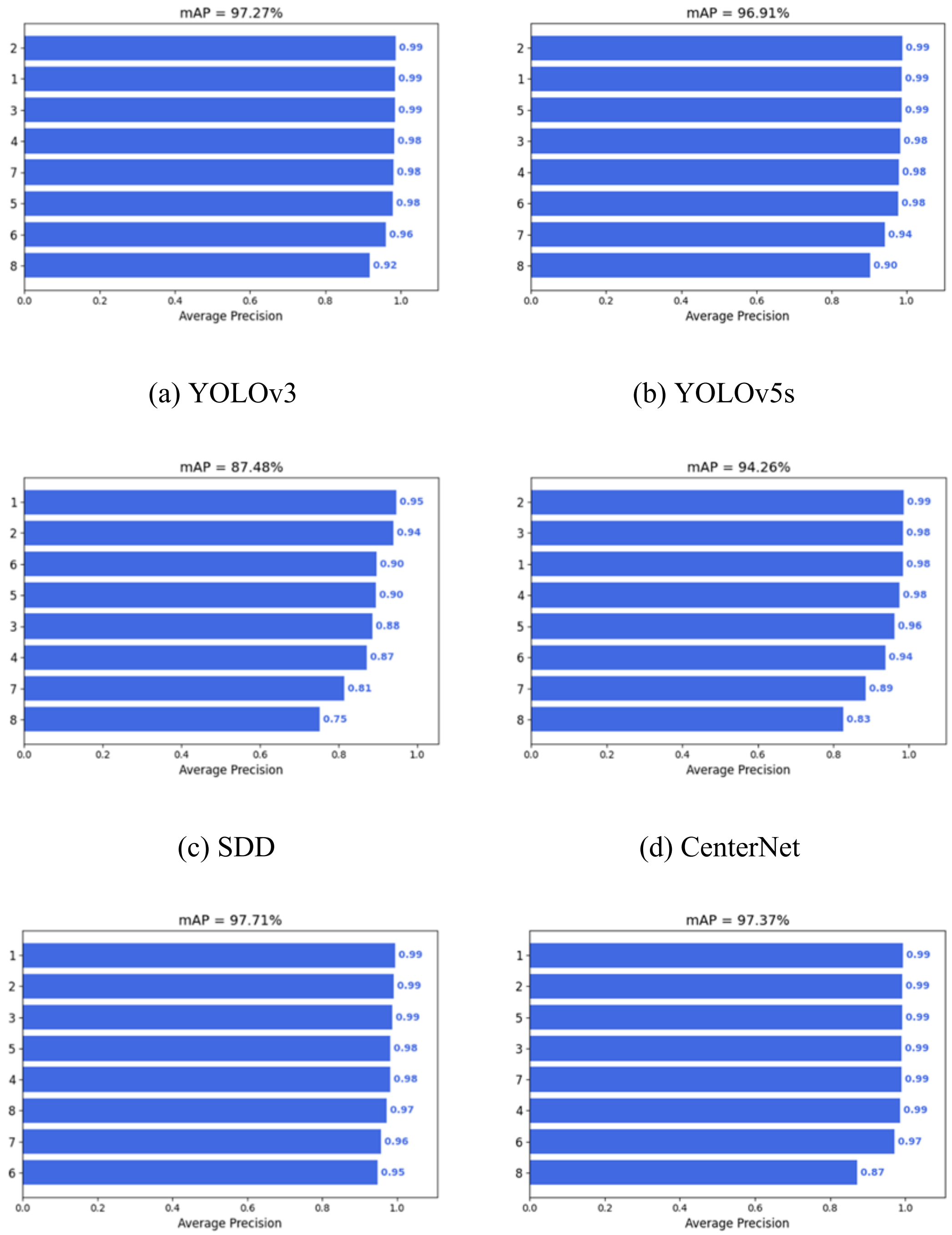
Table 4. Evaluation results of YOLOv7 and ISTD-YOLOv7.

Model	P (%)	R (%)	F1 (%)	mAP (%)
YOLOv7	97.52	96.23	96.87	97.44
ISTD-YOLOv7	**98.80**	**96.87**	**97.83**	**98.43**

4.5. Comparison with classical models

In this section, ISTD-YOLOv7 are compared with other state-of-the-art detection models. YOLOv3 [12], YOLOv5s [13] and YOLOXs [14] are also from the YOLO family, but they have not been tested on the dataset of this paper. SSD [39] is the anchor-based model. CenterNet [40] and FCOS [41] are the anchor-free models. DETR [42] is the first detection model based on a transformer.

Figure 11 shows the AP value of each class of different models. The ordinate indicates the class and the abscissa indicates the AP value. The AP values of each model are sorted from large to small and then displayed from top to bottom. The index AP comprehensively considers the balance between precision and recall under different confidence levels. ISTD-YOLOv7 is the only model with AP values over 96% in all classes. It proves that our model has a better overall detection effect on the given dataset. In addition, it is not difficult to find that the AP values of the eighth class of all models except FCOS are all the lowest. This is because the number of the eighth-class targets in the training set is fewer, and the models cannot learn the feature information of this class more fully. Nevertheless, the AP value of our model in the eighth class is more than 96%, while the AP value of SSD model in the eighth class is only more than 75%. mAP is the mean of all classes of AP and cannot reflect the above potential results.



(a) YOLOv3

(b) YOLOv5s

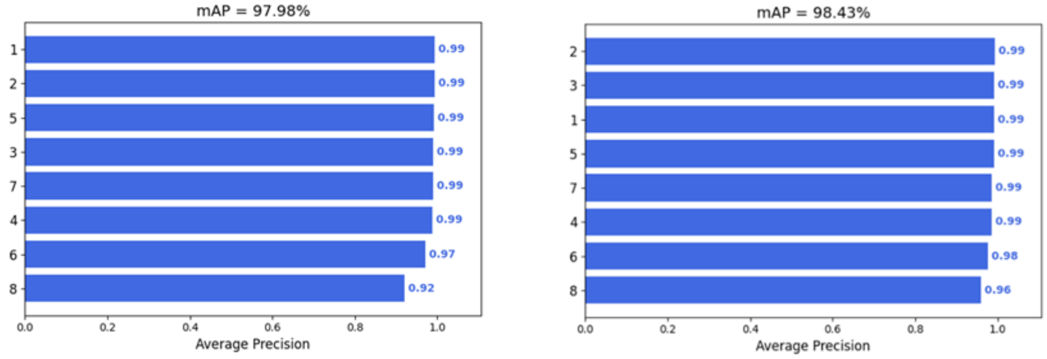
(c) SDD

(d) CenterNet

(e) FCOS

(f) YOLOXs

Continued on next page



(g) DETR

(h) ISTD-YOLOv7

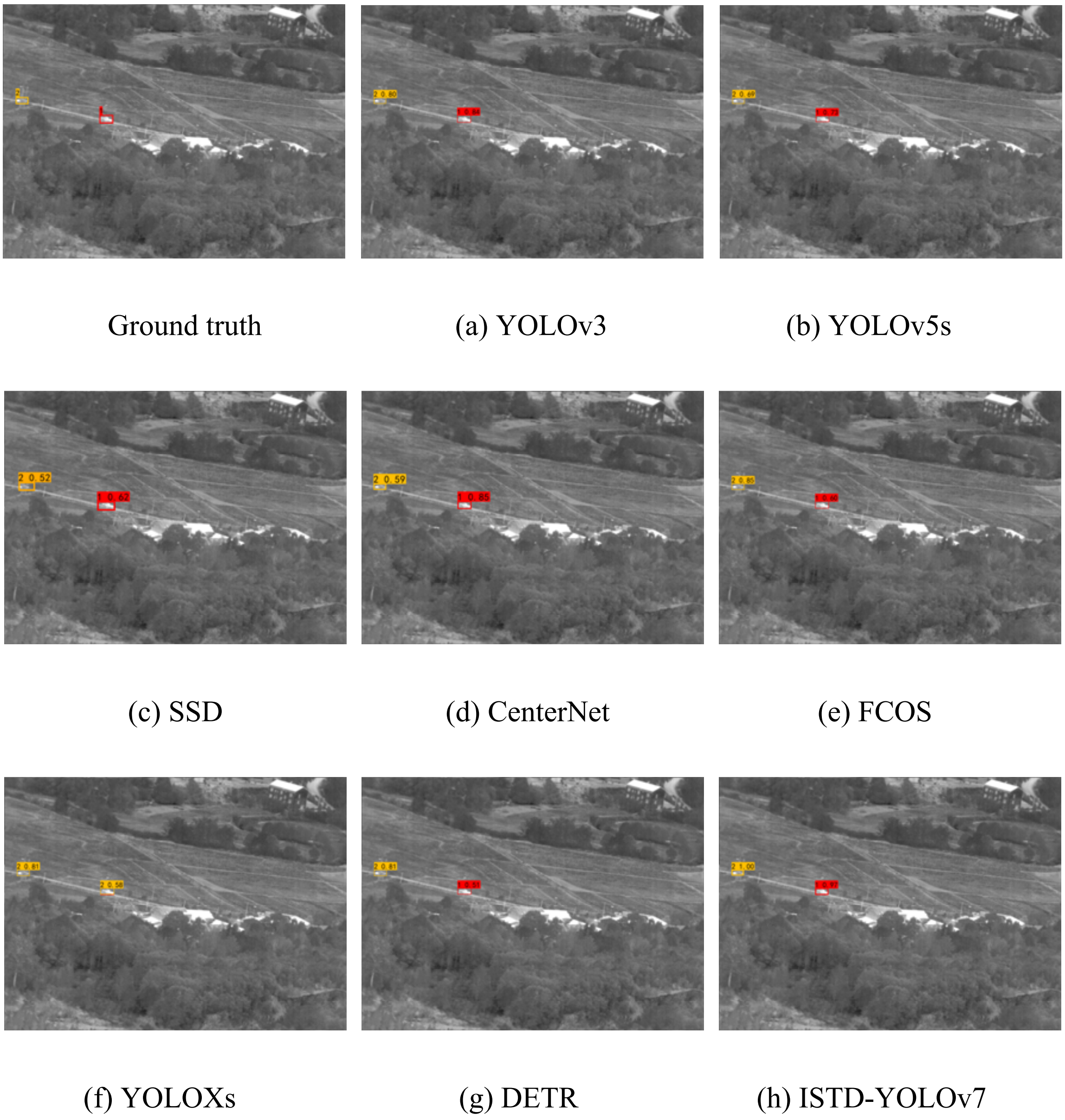
Figure 11. AP of each class of different models.

More quantitative results are given in Table 5. In terms of precision, ISTD-YOLOv7 obtains the best result of 98.80%. YOLOv3 obtains the best recall of 97.45%, and ISTD-YOLOv7 ranked second. F1 and mAP are two comprehensive indices, and our model significantly outperforms the comparison models. Moreover, in term of the target-level performance, Pd is the ratio of correctly predicted targets and all targets, and Fa is the ratio of false predicted target pixels and all the pixels in the image. Our model achieves 94.66% on Pd and 94.08 × 10^{-6} on Fa. The performance of SSD is not satisfactory on the given dataset. These findings show that ISTD-YOLOv7 performs better overall than comparison models regarding its capacity to detect infrared small targets. This is attributed to YOLOv7's own network structure and our focused improvements to it. Facing the infrared small targets in complex scenes, the updated anchors, GE attention and NWD-based loss in ISTD-YOLOv7 substantially improve the convergence performance and feature extraction capability of the network and alleviate the sensitivity to the location deviation of small targets.

Table 5. Evaluation results of different models.

Model	P (%)	R (%)	F1 (%)	mAP (%)	Pd (%)	Fa (10^{-6})
YOLOv3	97.15	**97.45**	97.30	97.27	93.93	115.88
YOLOv5s	97.72	95.00	96.35	96.91	92.77	127.63
SSD	92.78	41.81	57.65	87.48	77.02	1245.72
CenterNet	96.31	92.15	94.19	94.26	93.45	112.54
FCOS	98.30	80.26	88.37	97.71	92.20	130.51
YOLOXs	96.90	96.25	96.60	97.37	93.19	118.65
DETR	97.35	96.83	97.09	97.98	93.15	119.75
ISTD-YOLOv7	**98.80**	96.87	**97.83**	**98.43**	**94.66**	**94.08**

The ground truths and the qualitative results of all models are provided in Figures 12–15. The qualitative results show the class and the confidence of the detected target in different colors. Here, "Target 1" to "Target 8" respectively represent eight different infrared small vehicles. Limited to space, we only show some typical results of different methods. Image 1 is selected from the day outfield scene, Image 2 is selected from the day infield scene, Image 3 is selected from the night outfield scene and Image 4 is selected from the night infield scene. In Image 1, YOLOXs has obvious false detection cases. In Image 2, only CenterNet and ISTD-YOLOv7 detect all targets, while other models have different degrees of missed detection phenomena. Further analysis of missed detection phenomena shows that, because the “Target 7” is very weak and almost submerged in the background, it is more difficult to detect. In this case, ISTD-YOLOv7 can still detect it with a confidence of 0.78. SSD is the model with the most severe missed detection phenomena, only detecting “Target 1”. It can be seen that eight models detect all infrared small vehicles in Image 3. ISTD-YOLOv7 detects targets with significantly high confidence levels. In Image 4, SSD and FCOS have missed detection phenomena. ISTD-YOLOv7 is not affected by white noise in complex scenes during the detection process, and the confidence level of the detection results on “Target 1”, “Target 2” and “Target 3” is 1.00. The qualitative results more intuitively prove the superiority of our model.



Ground truth

(a) YOLOv3

(b) YOLOv5s

(c) SSD

(d) CenterNet

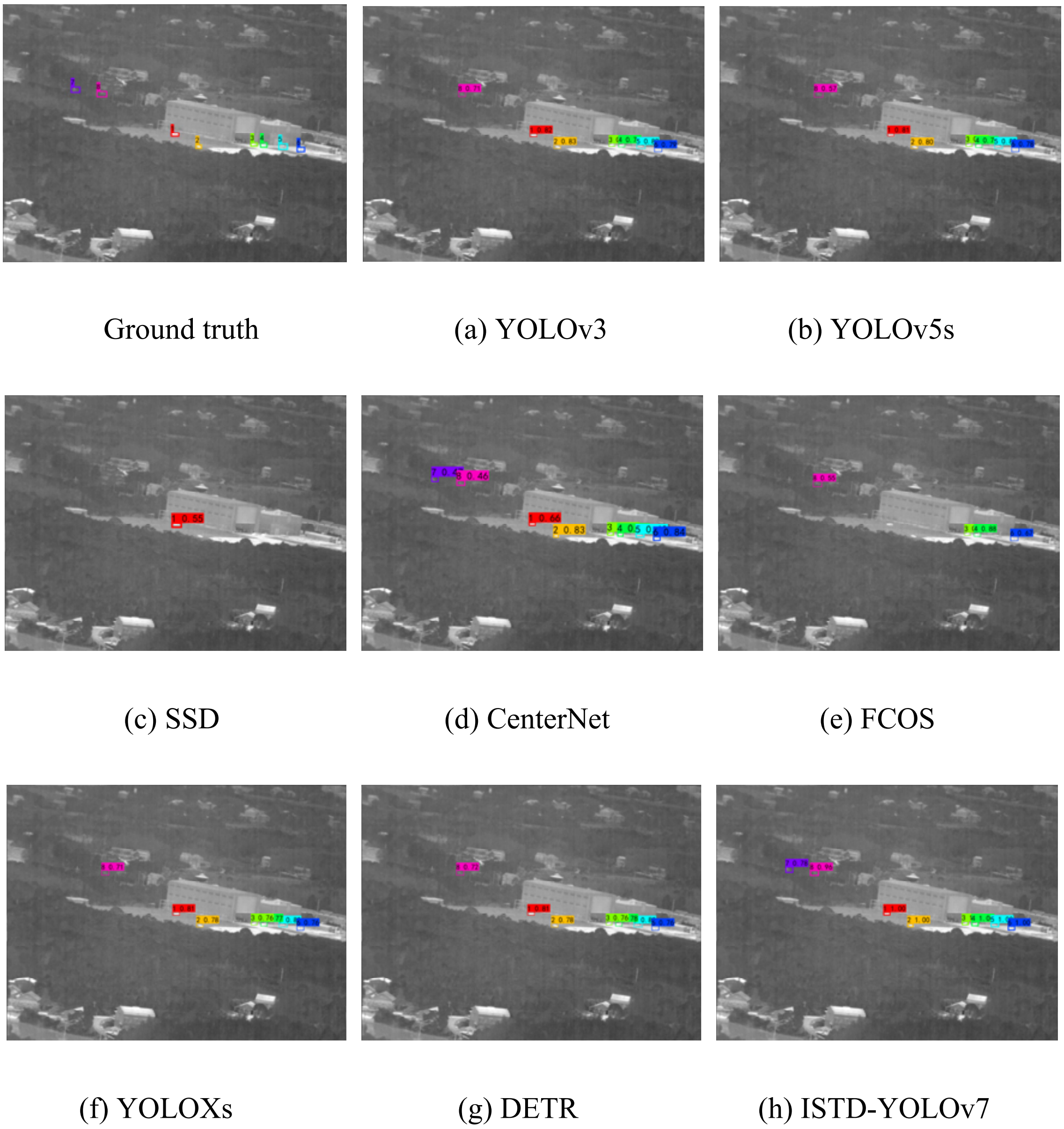
(e) FCOS

(f) YOLOXs

(g) DETR

(h) ISTD-YOLOv7

Figure 12. Visual results of Image 1.



Ground truth

(a) YOLOv3

(b) YOLOv5s

(c) SSD

(d) CenterNet

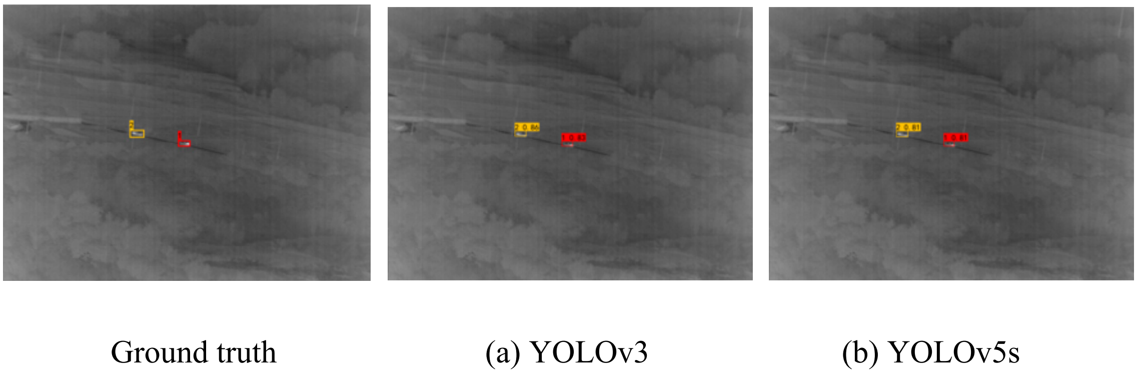
(e) FCOS

(f) YOLOXs

(g) DETR

(h) ISTD-YOLOv7

Figure 13. Visual results of Image 2.

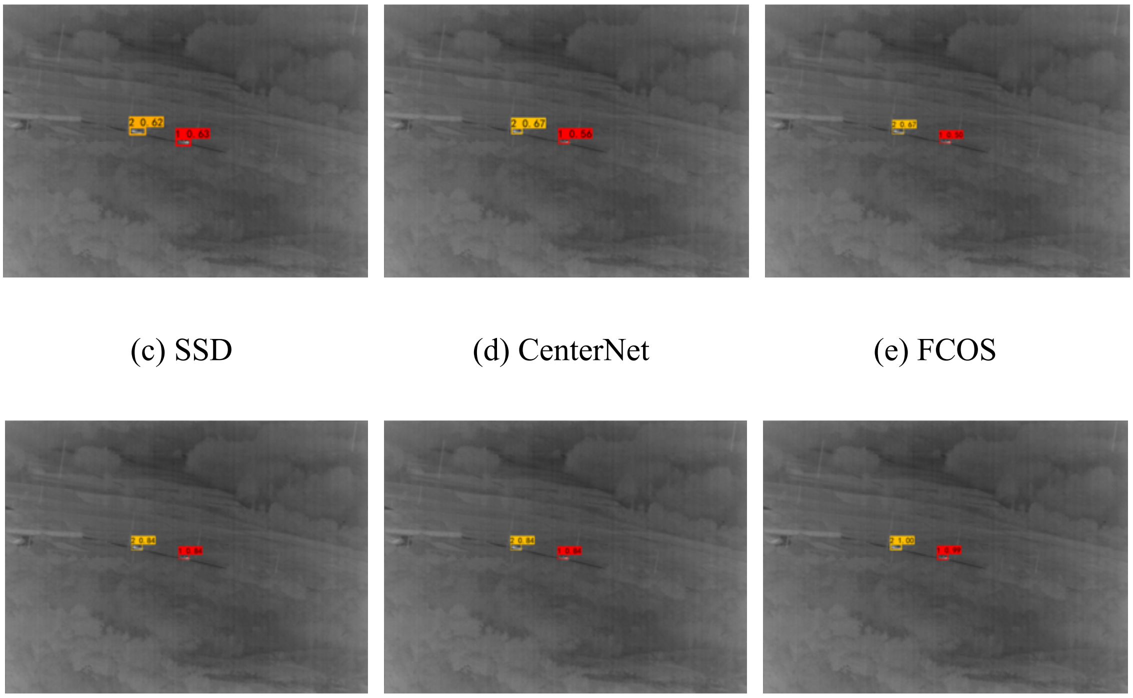


Ground truth

(a) YOLOv3

(b) YOLOv5s

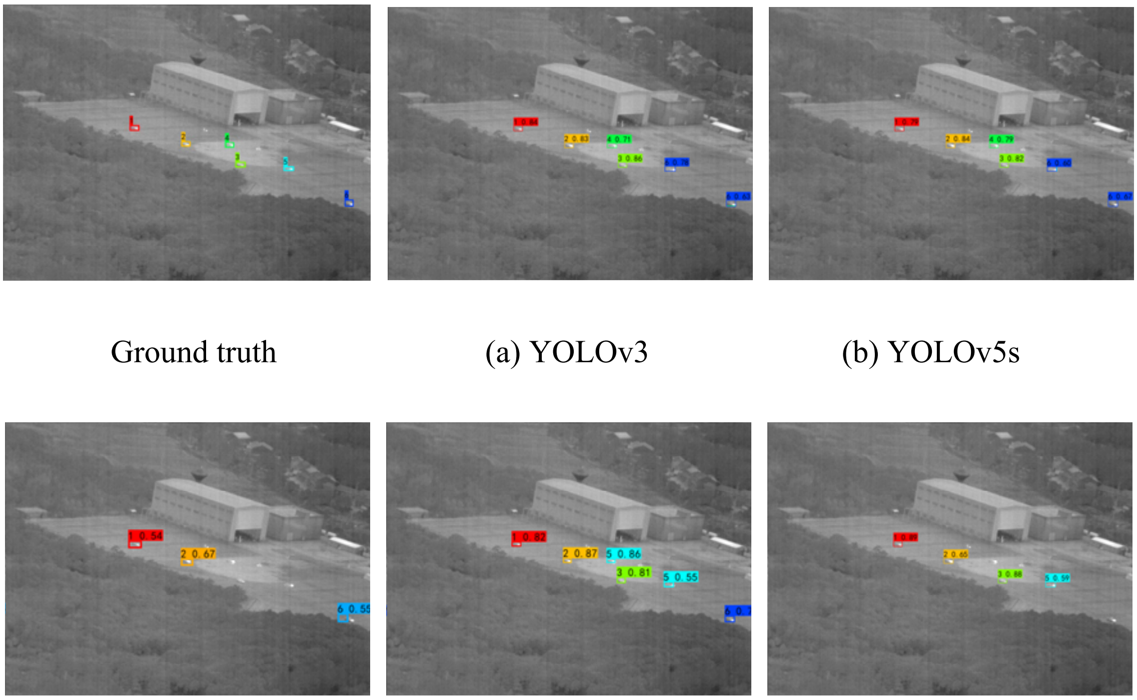
Continued on next page



(c) SSD (d) CenterNet (e) FCOS

(f) YOLOXs (g) DETR (h) ISTD-YOLOv7

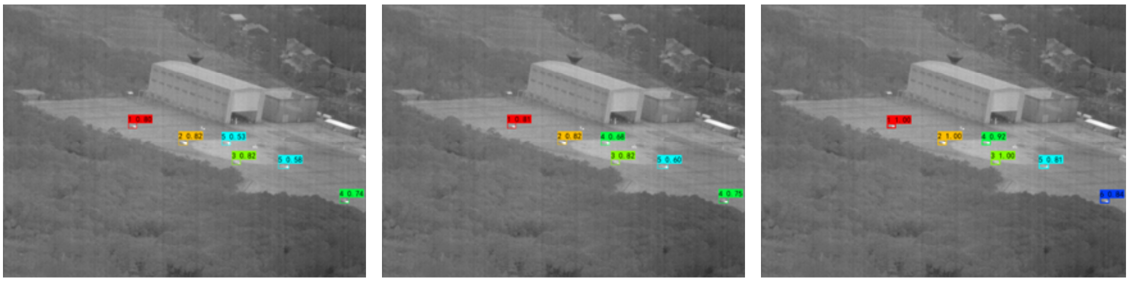
Figure 14. Visual results of Image 3.



Ground truth (a) YOLOv3 (b) YOLOv5s

(c) SSD (d) CenterNet (e) FCOS

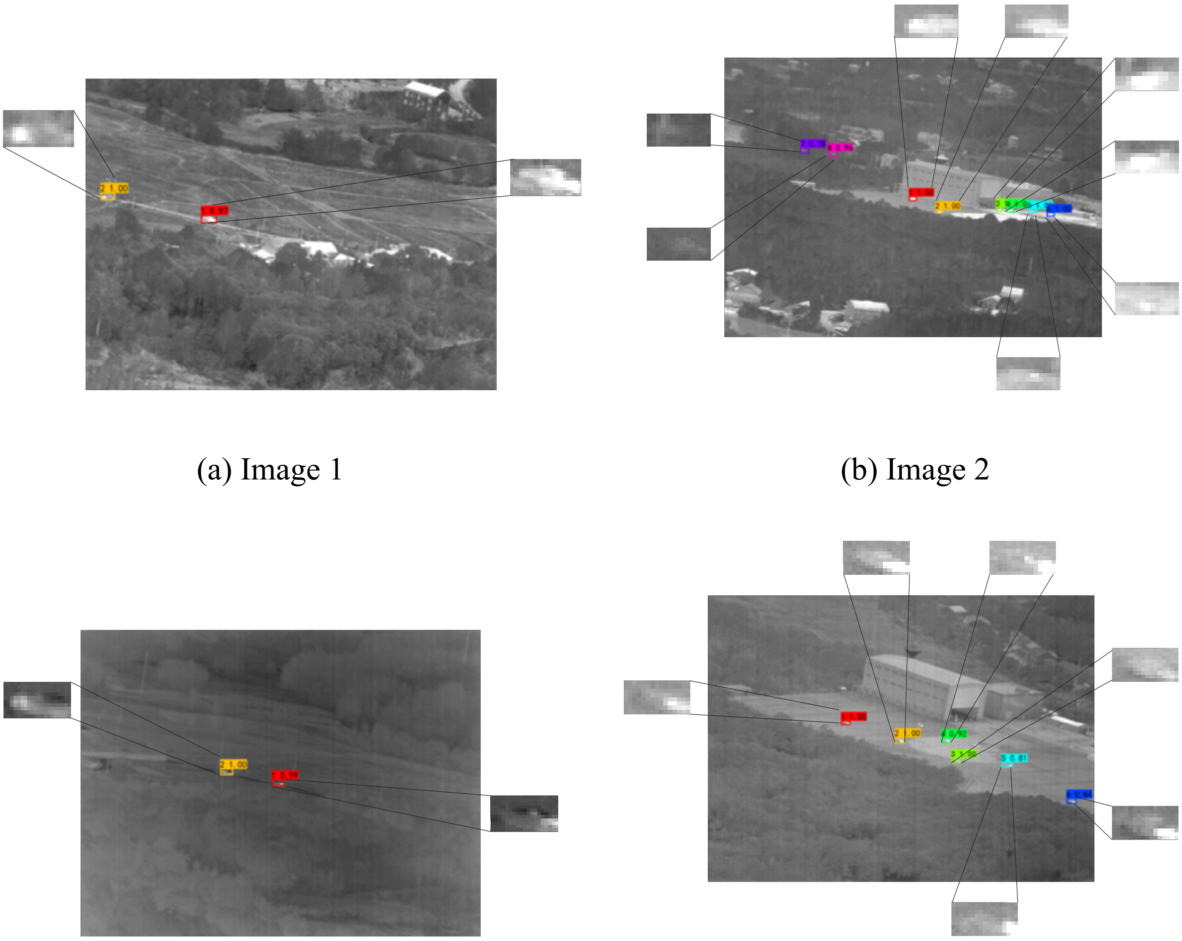
Continued on next page



(f) YOLOXs (g) DETR (h) ISTD-YOLOv7

Figure 15. Visual results of Image 4.

To further discuss the detection results of our model, we crop and enlarge the obtained targets on Images 1–4, as shown in Figure 16. It is not difficult to find that our model detects all the targets in the four images. The displayed cropped targets are potentially helpful for situation analysis and target attack on the battlefield.



(a) Image 1 (b) Image 2

(c) Image 3 (d) Image 4

Figure 16. Results of obtained targets by ISTD-YOLOv7.

4.6. Complexity analysis

In this section, the model parameters, floating-point operations per second (FLOPs) and frames per second (FPS) are also calculated. Spatial complexity determines the number of parameters in the model, and time complexity can be measured using FLOPs. FPS is used to evaluate the detection speed, which is tested on one 3090Ti GPU. According to Table 6, it can be seen that YOLOv5s has lower parameters, smaller computations, and faster inference speed. YOLOXs ranks second overall. ISTD-YOLOv7 has 37.232 M parameters, 105.234 G FLOPs and 36 FPS. In terms of FPS, YOLOv5s, SSD and YOLOXs have significant advantages. ISTD-YOLOv7 ranks in the middle on various evaluation indices. In summary, our model achieves better detection performance within an acceptable time. However, our model is not lightweight enough and does not have an advantage in complexity, which is the limitation of current work.

Table 6. Params, FLOPs, and FPS of different models.

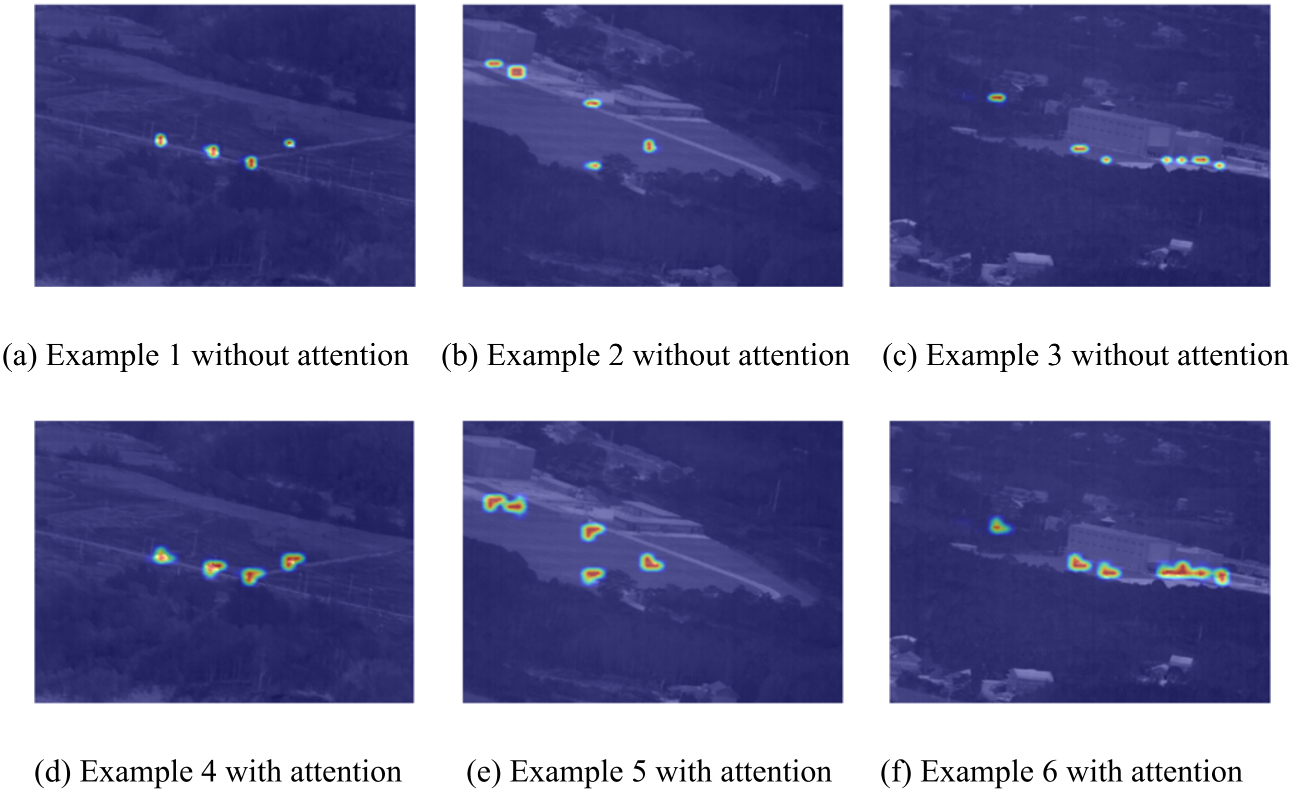
Model	Params	FLOPs	FPS
YOLOv3	61.561 M	155.380 G	48
YOLOv5s	**7.082 M**	**16.537 G**	**79**
SSD	24.547 M	276.251 G	72
CenterNet	32.665 M	109.714 G	49
FCOS	32.127 M	161.410 G	25
YOLOXs	8.968 M	26.927 G	77
DETR	36.762 M	73.642 G	24
ISTD-YOLOv7	37.232 M	105.234 G	36

4.7. Ablation studies

In this section, ablation studies are carried out to verify the effectiveness of the improved components. Table 7 shows the results of ablation studies. Compared with the baseline model, the detection performance of all four improved other models is improved. Moreover, ISTD-YOLOv7 obtains the best results on all indices. It indicates that the three components improve the performance of the model in small target detection from different aspects, and the gain effect of the hybrid model increases the most. Specifically, resetting anchors of the small target dataset can make the model better adapt to the given task. In this way, the bounding box can fine-tune the high-quality anchor to obtain the detection results. Figure 17 shows heat maps before and after adding the GE attention blocks. Figure 17(a)–(c) are heat maps without attention, and Figure 17(d)–(f) are heat maps with attention. The darker the color, the more significant the target area is. It is not difficult to find that adding the attention mechanism can make the model focus more on the local characteristics of infrared small targets and ignore irrelevant background information. NWD-based loss can better eliminate the performance gap between training and testing, and is suitable for small target detectors. The NWD metric can handle the problem that small targets are easy to be falsely predicted because the IoU metric is sensitive to the location deviation of the small targets.

Table 7. Results of ablation studies.

Model	P (%)	R (%)	F1 (%)	mAP (%)
YOLOv7	97.52	96.23	96.87	97.44
YOLOv7+Anchor update	98.42	96.34	97.37	97.85
YOLOv7+GE attention	98.05	96.72	97.38	98.14
YOLOv7+NWD	97.65	96.80	97.22	98.26
ISTD-YOLOv7	**98.80**	**96.87**	**97.83**	**98.43**



(a) Example 1 without attention (b) Example 2 without attention (c) Example 3 without attention

(d) Example 4 with attention (e) Example 5 with attention (f) Example 6 with attention

Figure 17. Comparison of heat maps without and with attention.

5. Conclusions

Infrared small targets are dim and have low signal-to-noise ratio. In complex weather and terrain scenes, infrared vehicles are easily overlooked, and most current models cannot effectively detect them. In this paper, ISTD-YOLOv7 based on YOLOv7 is proposed for infrared small target detection. In order to improve YOLOv7 to adapt this task, we have adopted a series of targeted improvements.

ISTD-YOLOv7 includes anchor update and GE attention as well as the NWD loss function. On a public infrared small target dataset, a series of experimental results reveal that ISTD-YOLOv7 is superior to comparison models (YOLOv3, YOLOv5s, SSD, CenterNet, FCOS, YOLOXs, DETR and YOLOv7), and the improvements are effective. Compared with the baseline model, the mAP of ISTD-YOLOv7 improved from 97.44% to 98.43%. The major causes of the high detection performance are as follows: the update of anchor provides a more reasonable prior. Spatial location is more important for the detection of small targets, so GE attention is chosen to make the model more efficiently exploit feature context information. The NWD loss function contributes to solving the sensitivity of the IoU metric to small target location deviation.

It should be mentioned that there are still limitations to this work. First, there is a problem of the class imbalance in the dataset used. Second, our model is still not lightweight enough. For future research, we will use a Generative Adversarial Network (GAN) [43] to increase samples for training. In addition, we will reduce the parameters and computations of the model as much as possible for deployment applications.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61473100).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. B. Jiang, X. Ma, Y. Lu, Y. Li, L. Feng, Z. Shi, Ship detection in spaceborne infrared images based on Convolutional Neural Networks and synthetic targets, *Infrared Phys. Technol.*, **97** (2019), 229–234. https://doi.org/10.1016/j.infrared.2018.12.040
2. A. Özdil, B. Yılmaz, Automatic body part and pose detection in medical infrared thermal images, *Quant. InfraRed Thermogr. J.*, **19** (2021), 223–238. https://doi.org/10.1080/17686733.2021.1947595
3. F. Prata, Detection and avoidance of atmospheric aviation hazards using infrared spectroscopic imaging, *Remote Sens.*, **12** (2020), 2309. https://doi.org/10.3390/rs12142309
4. C. Gao, L. Wang, Y. Xiao, Q. Zhao, D. Meng, Infrared small-dim target detection based on Markov random field guided noise modelling, *Pattern Recognit.*, **76** (2018), 463–475. https://doi.org/10.1016/j.patcog.2017.11.016
5. M. Qi, L. Liu, S. Zhuang, Y. Liu, K. Li, Y. Yang, et al., FTC-Net: Fusion of transformer and CNN features for infrared small target detection, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **15** (2022), 8613–8623. https://doi.org/10.1109/JSTARS.2022.3210707
6. N. Nguyen, T. Do, T. Ngo, D. Le, An evaluation of deep learning methods for small object detection, *J. Electr. Comput. Eng.*, **2020** (2020), 3189691. https://doi.org/10.1155/2020/3189691
7. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81
8. J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, *IEEE Trans. Multimedia*, **20** (2017), 985–996. https://doi.org/10.1109/TMM.2017.2759508

9. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
10. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 386–397. https://doi.org/10.1109/TPAMI.2018.2844175
11. P. Jiang, D. Ergu, F. Liu, Y. Cai, B. Ma, A review of YOLO algorithm developments, *Procedia Comput. Sci.*, **199** (2022), 1066–1073. https://doi.org/10.1016/j.procs.2022.01.135
12. J. Redmon, A. Farhadi, YOLOv3: An incremental improvement, preprint, arXiv:1804.02767.
13. S. Shen, X. Zhang, W. Yan, S. Xie, B. Yu, S. Wang, An improved UAV target detection algorithm based on ASFF-YOLOv5s, *Math. Biosci. Eng.*, **20** (2023), 10773–10789. https://doi.org/10.3934/mbe.2023478
14. Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, YOLOX: Exceeding YOLO series in 2021, preprint, arXiv:2107.08430.
15. C. Wang, A. Boschkovskiy, H. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, preprint, arXiv:2207.0269.
16. M. Soeb, M. Jubayer, T. Tarin, M. Mamun, F. Ruhad, A. Parven, et al., Tea leaf disease detection and identification based on YOLOv7 (YOLO-T), *Sci. Rep.*, **13** (2023), 6078. https://doi.org/10.1038/s41598-023-33270-4
17. S. Li, J. Yu, H. Wang, Damages detection of aeroengine blades via deep learning algorithms, *IEEE Trans. Instrum. Meas.*, **72** (2023), 1–11. https://doi.org/10.1109/TIM.2023.3249247
18. S. Liu, Y. Wang, Q. Yu, H. Liu, Z. Peng, CEAM-YOLOv7: Improved YOLOv7 based on channel expansion and attention mechanism for driver distraction behavior detection, *IEEE Access*, **10** (2022), 129116–129124. https://doi.org/10.1109/ACCESS.2022.3228331
19. F. Chen, C. Gao, F. Liu, Y. Zhao, Y. Zhou, D. Meng, et al., Local patch network with global attention for infrared small target detection, *IEEE Trans. Aerosp. Electron. Syst.*, **58** (2022), 3979–3991. https://doi.org/10.1109/TAES.2022.3159308
20. Y. Dai, Y. Wu, F. Zhou, K. Barnard, Attentional local contrast networks for infrared small target detection, *IEEE Trans. Geosci. Remote Sens.*, **59** (2021), 9813–9824. https://doi.org/10.1109/TGRS.2020.3044958
21. M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, J. Guo, ISNet: Shape matters for infrared small target detection, in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 867–876. https://doi.org/10.1109/CVPR52688.2022.00095
22. B. Li, C. Xiao, L. Wang, Y. Wang, Z. Lin, M. Li, W. An, et al., Dense nested attention network for infrared small target detection, *IEEE Trans. Image Process.*, **32** (2023), 1745–1758. https://doi.org/10.1109/TIP.2022.3199107
23. T. Wu, B. Li, Y. Luo, Y. Wang, C. Xiao, T. Liu, et al., MTU-Net: Multilevel TransUNet for space-based infrared tiny ship detection, *IEEE Trans. Geosci. Remote Sens.*, **61** (2023), 1–15, Art no. 5601015. https://doi.org/10.1109/TGRS.2023.3235002
24. Z. Lin, B. Li, M. Li, L. Wang, T. Wu, Y. Luo, et al., Light-weight infrared small target detection combining cross-scale feature fusion with bottleneck attention module, *J. Infrared Millimeter Waves*, **41** (2022), 1102–1112. https://doi.org/10.11972/j.issn.1001-9014.2022.06.020
25. Y. Liu, X. Wang, SAR ship detection based on improved YOLOv7-Tiny, in *Proceedings of the 2022 IEEE 8th International Conference on Computer and Communications*, (2022), 2166–2170. https://doi.org/10.1109/ICCC56324.2022.10065775

26. Y. Guo, S. Chen, R. Zhan, W. Wang, J. Zhang, LMSD-YOLO: A lightweight YOLO algorithm for multi-scale SAR ship detection, *Remote Sens.*, **14** (2022), 4801. https://doi.org/10.3390/rs14194801
27. X. Zhou, L. Jiang, C. Hu, S. Lei, T. Zhang, X. Mou, YOLO-SASE: An improved YOLO algorithm for the small targets detection in complex backgrounds, *Sensors*, **22** (2022), 4600. https://doi.org/10.3390/s22124600
28. VOC dataset, Available from: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/.
29. COCO dataset, Available from: http://cocodataset.org/#download.
30. J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-Excite: Exploiting feature context in convolutional neural networks, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*, (2018), 9423–9433.
31. J. Wang, C. Xu, W. Yang, L. Yu, A normalized Gaussian Wasserstein distance for tiny object detection, preprint, arXiv:2110.13389.
32. C. Xu, J. Wang, W. Yang, H. Yu, L. Yu, G. Xia, Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark, *ISPRS J. Photogramm. Remote Sens.*, **190** (2022), 79–93. https://doi.org/10.1016/j.isprsjprs.2022.06.002
33. H. Lai, L. Chen, W. Liu, Z. Yan, S. Ye, STC-YOLO: Small object detection network for traffic signs in complex environments, *Sensors*, **23** (2023), 5307. https://doi.org/10.3390/s23115307
34. Z. Zheng, N. Chen, J. Wu, Z. Xv, S. Liu, Z. Luo, EW-YOLOv7: A lightweight and effective detection model for small defects in electrowetting display, *Processes*, **11** (2023), 2037. https://doi.org/10.3390/pr11072037
35. J. Hosang, R. Benenson, B. Schiele, Learning non-maximum suppression, in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 6469–6477. https://doi.org/10.1109/CVPR.2017.685
36. R. Fu, H. Fan, Y. Zhu, B. Hui, Z. Zhang, P. Zhong, et al., A dataset for infrared time-sensitive target detection and tracking for air-ground application, *China Sci. Data*, **7** (2022), 206–221. https://doi.org/10.11922/sciencedb.j00001.00331
37. C. Chen, G. Yuan, H. Zhou, Y. Ma, Improved YOLOv5s model for key components detection of power transmission lines, *Math. Biosci. Eng.*, **20** (2023), 7738–7760. https://doi.org/10.3934/mbe.2023334
38. M. Huang, Y. Wu, GCS-YOLOV4-Tiny: A lightweight group convolution network for multi-stage fruit detection, *Math. Biosci. Eng.*, **20** (2023), 241–268. https://doi.org/10.3934/mbe.2023011
39. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, et al., SSD: Single Shot MultiBox Detector, in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2
40. K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian, CenterNet: Keypoint triplets for object detection, in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 6568–6577. https://doi.org/10.1109/ICCV.2019.00667
41. Z. Tian, C. Shen, H. Chen, T. He, FCOS: Fully convolutional one-stage object detection, in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9626–9635. https://doi.org/10.1109/ICCV.2019.00972

42. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *Proceedings of the Computer Vision—ECCV 2020*, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

43. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, preprint, arXiv:1406.2661.