文章编号：1001-9014(2024)04-0582-06

DOI:10.11972/j.issn.1001-9014.2024.04.019

Research on fast detection method of infrared small targets under resource-constrained conditions

ZHANG Rui[1], LIU Min[1], LI Zheng[2*]

(1. School of Opto-Electronic and Comunication Engineering, Xiamen University of Technology, Xiamen 361024, China;
2. Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China)

Abstract: Infrared small target detection is a common task in infrared image processing. Under limited computational resources. Traditional methods for infrared small target detection face a trade-off between the detection rate and the accuracy. A fast infrared small target detection method tailored for resource-constrained conditions is proposed for the YOLOv5s model. This method introduces an additional small target detection head and replaces the original Intersection over Union (IoU) metric with Normalized Wasserstein Distance (NWD), while considering both the detection accuracy and the detection speed of infrared small targets. Experimental results demonstrate that the proposed algorithm achieves a maximum effective detection speed of 95 FPS on a 15 W TPU, while reaching a maximum effective detection accuracy of 91.9 AP@0.5, effectively improving the efficiency of infrared small target detection under resource-constrained conditions.

Key words: infrared UAV image, fast small object detection, low impedance, loss function

资源受限条件下的红外小目标快速检测方法研究

张 瑞[1], 刘 敏[1], 李 争[2*]

(1. 厦门理工学院 光电与通信工程学院,福建 厦门 361024;
2. 中国科学院上海技术物理研究所,上海 200083)

摘要:红外小目标检测是红外图像处理中的一项常见任务。在计算资源受限的条件下,传统的红外小目标检测方法面临着检测率和检测精度的平衡问题。本文针对 YOLOv5s 模型提出了一种在资源受限条件下快速红外小目标检测方法,该方法增加了一个小目标检测头,并用 Normalized Wasserstein Distance(NWD)度量取代了原来的 Intersection over Union(IoU)度量,同时考虑了红外小目标的检测精度和检测速率。实验结果表明,改进后的算法在 15 W TPU 上实现了最大 95 FPS 的红外小目标有效检测速度,同时达到了最大 91.9 AP@0.5 的检测精度,有效提高了资源受限条件下的红外小目标检测效率。

关 键 词:红外无人机;快速小目标检测;低功耗;损失函数

中图分类号:TP18　　**文献标识码**:A

Introduction

The application range of unmanned aerial vehicles (UAVs) is constantly expanding, encompassing areas such as military reconnaissance, outdoor photography, power line inspection and other fields. However, at the same time, it has also given rise to a series of social issues. These include concerns about privacy infringement due to UAVs being used for illicit filming and the potential threat to national security posed by the military application of UAVs. Therefore, the research of anti-UAV technology has important practical significance. Infrared UAV target detection technology is a technique that uses infrared imaging to continuously monitor UAVs. It enables UAV target detection based on infrared radiation, and also has obvious advantages in low-light conditions[1]. In recent years, this technology has become an important research direction and provides an effective

Received date: 2023-11-06, **revised date**: 2024-04-25　　**收稿日期**:2023-11-06,**修回日期**:2024-04-25

Biography:ZHANG Rui (1991-), male, Lianyungang, associate professor, Doctor, research area focuses on infrared imaging and intelligent perception. E-mail: 1546853078@qq.com

[*] **Corresponding author**:E-mail: lizheng_sitp@163.com

complement to radar target detection technology.

However, infrared UAV target detection faces challenges. The distance between the UAV and the sensor makes small infrared targets lack distinctive texture and shape features, hampering the detection. Additionally, background clutter and noise, like clouds and buildings, can cause confusion with obstacles[2]. Infrared images have high noise, poor spatial resolution, and are sensitive to environmental temperature changes. These challenges add complexity to infrared UAV target detection. Infrared small UAVs are shown in Fig. 1.

In recent years, there has been a continuous emergence of object detection methods based on deep learning, which has achieved impressive detection performance. These methods can be categorized into two types based on how they handle input images. The first type is the two-stage detection method, such as the region-based R-CNN and its variants[3]. The second type is the one-stage detection methods, including RetinaNet[4], SSD[5], and YOLO series[6]. During the flight of UAVs, real-time transmission of infrared images is required for the infrared cameras. In scenarios that demand high real-time performance, methods of the YOLO series, known for their fast speed and high accuracy, have been widely adopted. Among them, YOLOv5 stands out as an advanced detector with strong real-time processing performance and low hardware computing requirements, allowing for easy deployment on mobile devices. Therefore, using YOLOv5 can significantly enhance the detection accuracy and real-time performance of infrared small UAV[7].

This paper proposes a UAV detection method based on an improved YOLOv5s model to address the challenges of detecting small targets. The original YOLOv5 structure only includes three feature detection heads, which are not effective in extracting the feature information of small target UAVs captured by infrared cameras at long distances. To address this issue, this paper adds a feature detection head suitable for detecting small targets by YOLOv5. Additionally, the Intersection over Union (IoU) in the original YOLOv5 model is not a good metric for small target detection tasks. Therefore, this paper replaces it with a more suitable metric for small targets called the Normalized Gaussian Wasserstein Distance (NWD)[8]. This metric calculates the similarity between bounding boxes using the Gaussian distribution of their corresponding boxes.

1 Method to improve YOLOv5

YOLOv5 can be categorized into five architectures based on the depth and width of the model: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Among these, to balance detection speed and accuracy, we choose to improve YOLOv5s. The YOLOv5 network structure consists of three main components. CSP-Darknet53 serves as the backbone feature extraction network, extracting features from the input image. CSP-Darknet53 is an improved version of Darknet53 from YOLOv3, utilizing the Cross-Stage Partial (CSP) network strategy to reduce parameters and computation, thus enhancing the inference speed. In the middle part, YOLOv5 combines two modules: SPPF and PANet[9]. The SPPF module increases receptive fields and diversifies the feature pyramid, while the PANet module achieves bottom-up and top-down feature fusion, thereby improving the object detection capability. In the head part, YOLOv5 adopts the head structure of YOLOv4. This structure outputs predictions such as class probabilities, confidence scores, and bounding boxes on the feature maps.

1.1 Small object detection head

The YOLOv5 model uses a backbone network that undergoes five downsampling stages, producing five feature maps (P1-P5) with resolutions of 1/2, 1/4, 1/8, 1/16, and 1/32 of the input image size, respectively. The neck network combines multi-scale features in a top-down and bottom-up manner without changing the feature map sizes. The detection head operates on the P3-P5 feature maps for object detection. This design is based on the relationship between the feature layer size and the receptive field size in YOLOv5. The receptive field refers to the size of the input image region corresponding to each output unit in a convolutional neural network. A larger receptive field captures more object features, making it suitable for detecting larger objects. On the contrary, a smaller receptive field can only capture a limited

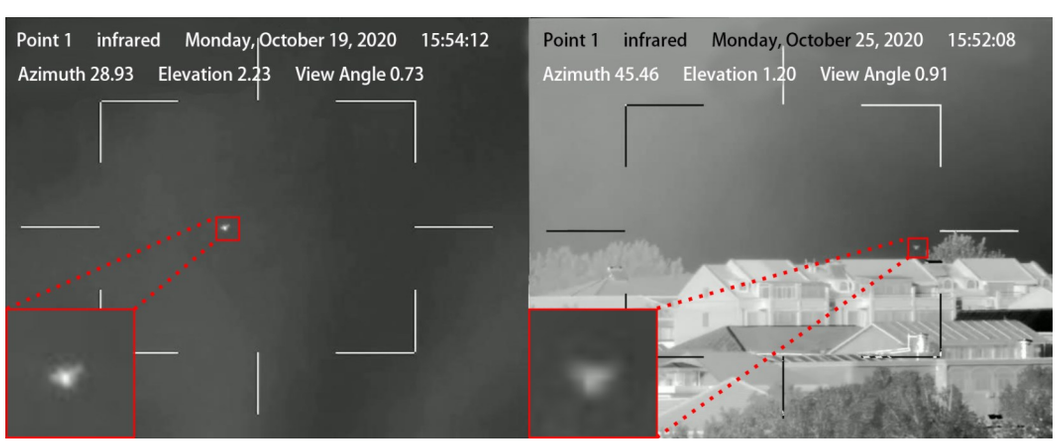

Fig. 1 Infrared small UAVs
图 1 红外小型无人机目标

number of object features, making it suitable for detecting smaller objects. A smaller receptive field implies that each pixel in the feature map is influenced by a smaller region in the original image. This enables more precise localization of object positions and boundaries, unaffected by irrelevant regions. Additionally, a smaller receptive field corresponds to a larger feature map size, preserving more spatial information and avoiding the loss of fine details on small objects. Therefore, a smaller receptive field is better suited for detecting smaller objects.

We have added a new detection head for small objects after the P2 feature layer in the YOLOv5 model. The detection head operates at a resolution of 160×160 pixels in the P2 layer, which corresponds to two downsampling operations in the backbone network. Each pixel in the P2 layer has a receptive field of 10×10 pixels, which is the smallest receptive field among the P2-P5 feature extraction layers. Additionally, we have assigned different loss function weights to the P2-P5 feature layers based on the target sizes. Specifically, we have assigned a weight of 4 to the feature layers for P2 and P3, a weight of 1 to the feature layer for P4, and a weight of 0.4 to the feature layer for P5. The purpose of this weighting scheme is to enhance the focus on small and tiny objects while reducing overfitting to large objects. The weighted formula for the object loss function is shown in Eq. (1). Although adding the new detection head increases the computational and memory overhead of the model, it significantly improves the detection performance for small objects. The improved YOLOv5s network architecture is shown in Fig. 2. Small target detection head is shown in Fig. 3.

$$L_{obj} = 4 \cdot L_{P2} + 4 \cdot L_{P3} + 1 \cdot L_{P4} + 0.4 \cdot 4 \cdot L_{P5} \quad . \quad (1)$$

1.2 Normalized Gaussian Wasserstein Distance

In YOLOv5, IoU is used as an indicator to measure the degree of matching between predicted bounding boxes and real bounding boxes. It is obtained by calculating the ratio of the intersection area and the union area of the two. However, in the UAV images obtained by infrared devices, the overlapping part of the bounding boxes of small targets is often very small, which will result in lower IoU values. As shown in Fig. 4, IoU is very sensitive to the scale of small targets. For the infrared small UAV target with a size of 6×6 pixels, only a small position deviation can cause IoU to decrease from 0.53 to 0.06, thereby affecting the accuracy of label allocation and reducing the performance of detection[8]. So, for small target objects, IoU does not measure their matching degree effectively.

We used a metric known as Normalized Gaussian Wasserstein Distance named NWD, which is suitable for small object detection. It is insensitive to the scale of targets, allowing for better assessment of similarities between small objects. Specifically, this method models the object bounding boxes as two-dimensional Gaussian distributions and calculates the NWD between the predicted and the ground truth distributions, as shown in Eq. (2). Modeling the bounding boxes with Gaussian distributions enables the representation of the object's position as the mean point and its shape as the standard de-

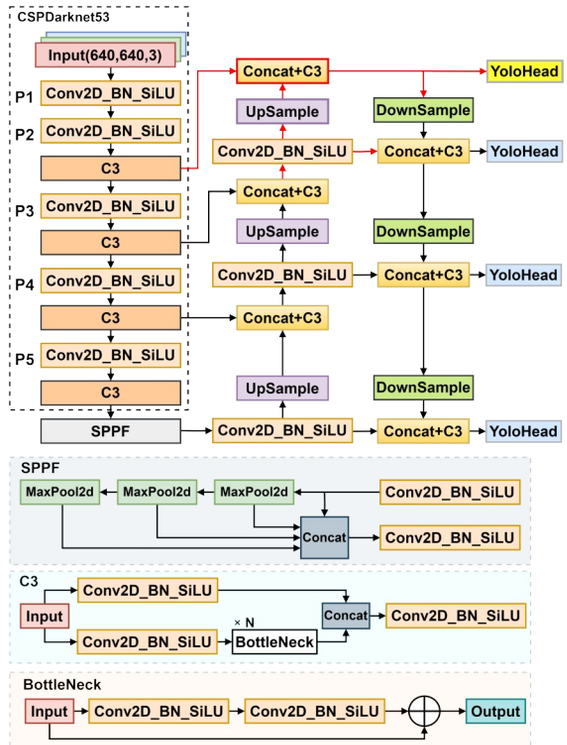

Fig. 2 Improved YOLOv5s network architecture
图2 改进的 YOLOv5s 网络架构

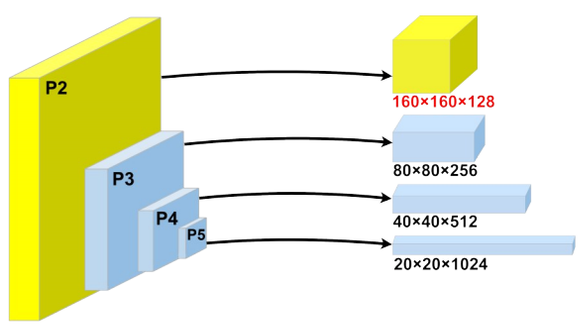

Fig. 3 Small target detection head
图3 小目标检测头

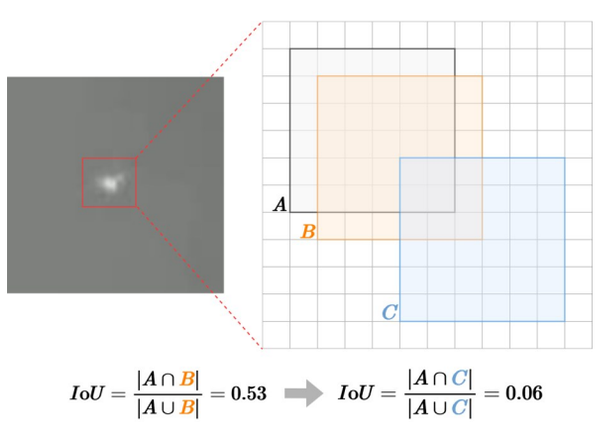

Fig. 4 The sensitivity analysis of IoU on infrared small UAV
图4 IoU在红外小型无人机上灵敏度分析

viation of the Gaussian distribution. This approach has the advantage of better describing the spatial distribution of the objects, not solely relying on their geometric shapes. NWD takes into account the geometric features and spatial distributions between two bounding boxes, even when they do not overlap, by considering their geometric relationships[10]. This measurement approach avoids the sensitive issues of IoU with small objects, making it more suitable for quantifying small object matches.

$$\mathrm{NWD}(N_A,N_B)=\exp\left(-\frac{W_2^2(N_A,N_B)}{C}\right) \quad , \quad (2)$$

$$W_2^2(N_A,N_B)=\left\|\left(\left[cx_A,cy_A,\frac{w_A}{2},\frac{h_A}{2}\right]^T,\left[cx_B,cy_B,\frac{w_B}{2},\frac{h_B}{2}\right]^T\right)\right\|_2^2 \quad , \quad (3)$$

where N_A and N_B are Gaussian distributions modeled by $A=(cx_A, cy_A, w_A, h_A)$ and $B=(cx_B, cy_B, w_B, h_B)$, $W_2^2(N_A, N_B)$ is the distance measurement, C is the constant closely related to the dataset.

2 Experiment and deployments

2.1 Dataset introduction

This article is based on the experimental analysis of selected subsets from the 3rd Anti-UAV Workshop & Challenge[11]. The dataset consists of a total of 9841 small target drone images captured under thermal infrared, involving complex environmental factors such as dynamic backgrounds, complex motions, scale changes, and small targets. The labels in this dataset only include the category of drones. It is noteworthy that the Signal-to-Noise Ratio (SNR) of this dataset is 12.615 dB. The dataset is divided into training and testing sets in a 7:3 ratio. The distribution of the dataset is shown in Fig. 5.

2.2 Evaluation index

To verify the performance of the model, this article selects Average Precision (AP) to evaluate the model performance, where the AP@0.5 and the AP@0.5:0.95 represent the detection accuracy of two models under different IoU thresholds. The AP@0.5 is the average accuracy of a certain category when IoU is 0.5. The average accuracy is based on different confidence levels, including the curve area of Precision (P) and Recall (R). And AP@0.5:0.95 is the average accuracy of a certain category when IoU is taken every 0.05 from 0.5 to 0.95. This indicator requires a higher degree of overlap in the target box. The False Alarm Rate (FAR) is depicted in Eq. (6), the Miss Rate (MR) is represented by Eq. (7), with the Average Precision (AP) illustrated in Eq. (8).

$$\mathrm{P}=\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right) \quad , \quad (4)$$

$$\mathrm{R}=\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\right) \quad , \quad (5)$$

$$\mathrm{FAR}=\left(\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}}\right) \quad , \quad (6)$$

$$\mathrm{MR}=1-\mathrm{R} \quad , \quad (7)$$

$$\mathrm{AP}=\int_0^1 \mathrm{P}(\mathrm{R})d\mathrm{R} \quad , \quad (8)$$

where TP means true positive, TN means true negative, FP means false positive, and FN means false negative.

2.3 Experimental environment and parameter settings

In this work, all experiments were conducted on the Ubuntu 22.04 operating system, with 128 GB of RAM and an Intel i9-13900K processor. The system was equipped with an NVIDIA RTX 3090 Ti graphics card with 24 GB of video memory; the deep learning framework used was Pytorch 1.12.1; and the programming language was Python 3.10.

The optimization algorithm used for model training was Stochastic Gradient Descent (SGD). The initial learning rate was 0.01, the momentum was 0.937, and the weight decay coefficient was 0.0005. In addition, the model was trained for 300 epochs, the batch size of the dataset was set to 64.

2.4 Experimental results

The experimental results regarding the allocation of different loss function weights for different feature layers are shown in Table 1. From the ablation experimental results in Table 2, it can be observed that adding the NWD

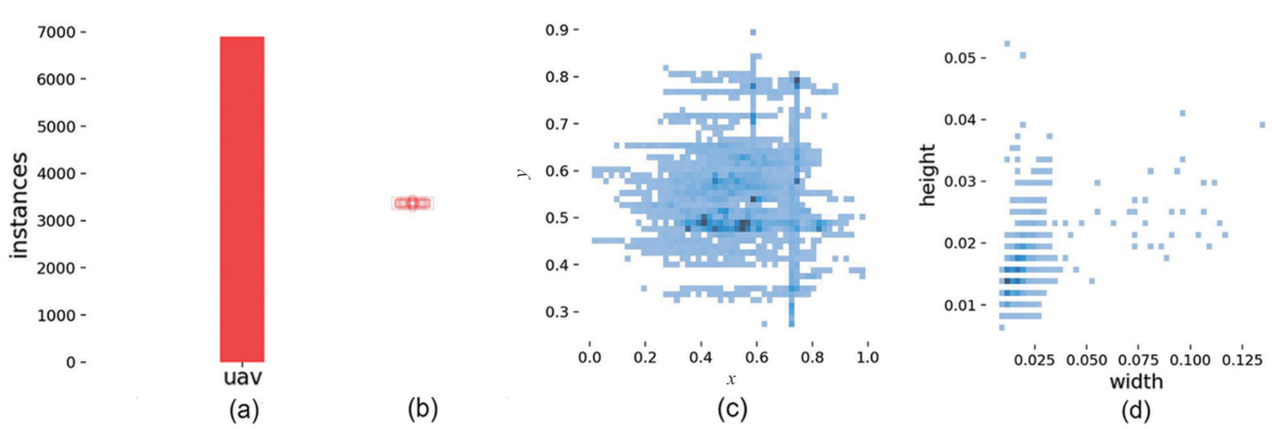

Fig. 5 Dataset analysis: (a) the label category distribution; (b) the bounding box size distribution; (c) the label center position distribution; (d) the label size distribution

图5 数据集分析:(a)标签类别分布;(b)边界框尺寸分布;(c)标签中心位置分布;(d)标签尺寸分布

metric with a coefficient of 0.5 results in a 2.7% improvement in the AP@0.5 compared to the original model. Furthermore, when using the complete NWD metric, the AP@0.5 is improved by 5.2%. Moreover, introducing a small object detection head with a resolution of 160×160 pixels from the P2 layer also leads to a 3.7% increase in the AP@0.5 compared to the baseline results. Finally, by combining the complete NWD loss function with the small object detection head, the overall model achieves a 7.2% improvement. The improved model also shows a 3.9% increase in the AP@0.5:0.95 compared to the original model.

The performance comparison chart of *AP* is shown in Fig. 6. The partial experimental result is shown in Fig. 7.

Table 1 Comparison of the different weighting coefficient results
表1 不同权重系数的结果对比

Weighting Coefficients	*AP*@0.5 (%)	*AP*@0.5:0.95 (%)	*FAR* (%)	*MR* (%)
$1\cdot L_{P2}+1\cdot L_{P3}+1\cdot L_{P4}+0.4\cdot L_{P5}$	81.2	44.9	6.4	28.4
$1\cdot L_{P2}+4\cdot L_{P3}+1\cdot L_{P4}+0.4\cdot L_{P5}$	86.8	47.6	4.7	19.4
$4\cdot L_{P2}+1\cdot L_{P3}+1\cdot L_{P4}+0.4\cdot L_{P5}$	84.2	46.0	7.8	22.8
$4\cdot L_{P2}+4\cdot L_{P3}+1\cdot L_{P4}+0.4\cdot L_{P5}$	88.4	48.6	4.0	17.4

Table 2 Comparison of ablation experiments of improved methods
表2 改进方法的消融实验对比

Models	*AP*@0.5 (%)	*AP*@0.5:0.95 (%)	*FAR* (%)	*MR* (%)
YOLOv5s	84.7	46.1	3.1	23.6
YOLOv5s+0.5NWD	87.4	47.9	4.0	17.4
YOLOv5s+NWD	89.9	48.1	4.4	14.6
YOLOv5s+P2	88.4	48.6	4.0	17.4
YOLOv5s+NWD+P2	91.9	50.0	4.2	12.9

To validate the effectiveness of the improved YOLOv5s in this paper, a comparison was made with various mainstream detection networks using the official sub-dataset from the 3rd Anti-UAV Workshop & Challenge. The comparison results are shown in Table 3.

2.5 Deployment

Our algorithm is deployed on the BM1684X TPU, and the specific flow is shown in Fig. 8. Generally speaking, deep learning-based algorithms have two main results in the quantization process: int8 and FP16. According to the characteristics of infrared small targets, learning-based target detection algorithms will have a significant accuracy decline in the quantization process.

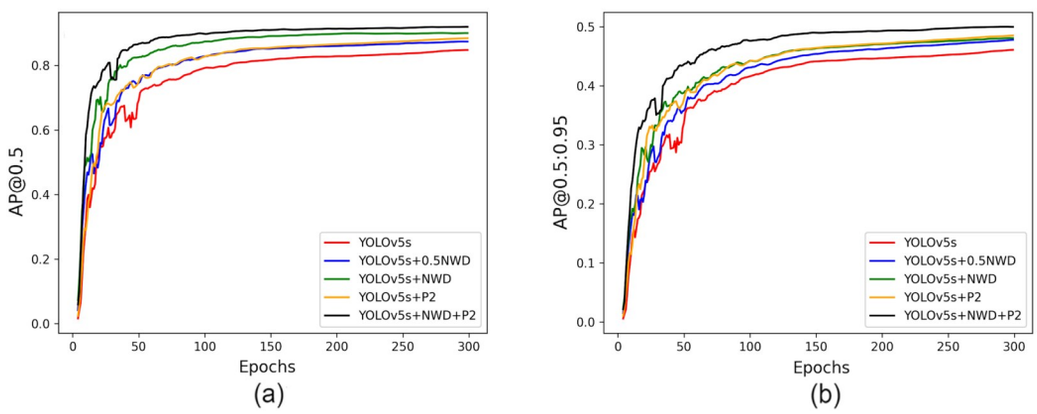

Fig. 6 Performance comparison of the AP: (a) AP@0.5; (b) AP@0.5:0.95
图6 模型在AP值上的性能表现:(a)AP@0.5;(b)AP@0.5:0.95

Table 3 Comparison of improved YOLOv5s with other methods
表3 改进的模型与其他同类算法的比较

Models	AP@0.5 (%)	AP@0.5:0.95 (%)	FAR (%)	MR (%)	Parameter (M)	GFLOPs	Speed (FPS)	Weights (MB)
SSD-ResNet50	60.4	22.8	29.7	46.2	13.1	15.0	200	105.1
Faster-RCNN-ResNet50	78.3	30.9	21.2	40.0	41.1	134.5	50	330.3
RetinaNet-ResNet50	82.1	33.8	18.5	33.2	32.0	127.5	43	257.3
YOLOv3	83.3	45.2	8.7	22.4	9.3	23.1	526	18.9
YOLOv5s	84.7	46.1	3.1	23.6	7.0	15.8	625	14.4
YOLOv5m	86.6	48.8	3.1	20.4	20.8	47.9	303	42.2
YOLOv5l	87.7	49.1	3.8	17.6	46.1	107.6	196	92.8
YOLOv8s	89.5	48.9	6.3	17.3	11.1	28.4	435	22.5
YOLOv5s+NWD+P2	91.9	50.0	4.2	12.9	7.7	26.8	400	16.3

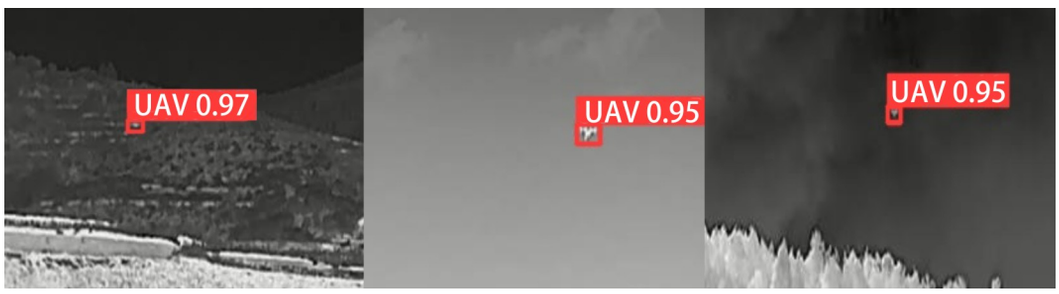

Fig. 7 Some examples of the detection result on the improved model
图7 改进的模型的案例表现

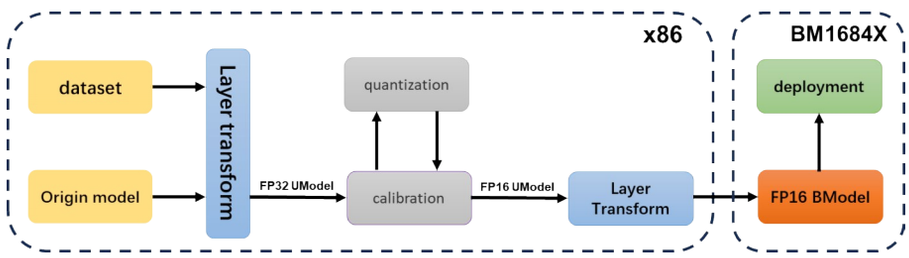

Fig. 8 Framework of the deployment
图8 模型部署的架构流程

The experimental results show that when the algorithm accuracy is quantized to FP16, the target detection efficiency is the highest, while the quantization to int8 has the weakest effect to the extent that no valid data can be counted. The deployment experimental results are shown in Table 4.

Table 4 Result of the deployment
表4 部署的实验结果

BM1684X	FPS	AP@0.5 (%)	*AP*@0.5:0.95(%)
FP32	12	91.9	50.0
FP16	95	87.8	49.8
INT8	163	–	–

3 Conclusion

To address the challenge of small drone detection in infrared devices, this paper proposes a light weight detection model. The model introduces a small object feature extraction layer at the P2 level of the backbone network, connecting it to a high-resolution detection head, thereby enhancing the network's capability to perceive small objects. Additionally, the paper adopts the NWD metric to replace the original IoU-based metric, as the NWD metric provides better measurements for small object instances and improves the model's detection accuracy. The paper conducts several comparative experiments on the partial sub-dataset provided by the 3rd Anti-UAV Workshop & Challenge. The results demonstrate that the proposed model outperforms other mainstream detection models in terms of both *AP*@0.5 and *AP*@0.5:0.95 evaluation metrics, validating the effectiveness of the proposed approach. Furthermore, the proposed method maintains high performance levels concerning parameter count, computational complexity, and model weight file size.

References

[1] Rawat S S, Alghamdi S, Kumar G, *et al*. Infrared small target detection based on partial sum minimization and total variation [J]. *Mathematics*, 2022, **10**(4): 671.
[2] Fang H, Xia M, Zhou G, *et al*. Infrared small UAV target detection based on residual image prediction via global and local dilated residual networks [J]. *IEEE Geoscience Remote and Sensing Letters*, 2021, **19**: 1-5.
[3] Ren S, He K, Girshick R, *et al*. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, **39**(6): 1137-1149.
[4] Lin T Y, Goyal P, Girshick R, *et al*. Focal Loss for Dense Object Detection [J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2020, **42**(2):318-327.
[5] Liu W, Anguelov D, Erhan D, *et al*. Ssd: Single shot multibox detector [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2016, Part I 14 (pp. **21**-37).
[6] Redmon J, Divvala S, Girshick R, *et al*. You Only Look Once: Unified, Real-Time Object Detection [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,(pp. **779**-788).
[7] Guo K, He C, Yang M, *et al*. A pavement distresses identification method optimized for YOLOv5s [J]. *Scientific Reports*, 2022, **12**(1): 3542.
[8] Wang J, Xu C, Yang W, *et al*. A normalized Gaussian Wasserstein distance for tiny object detection [J]. *arXiv preprint*, 2021, 2110.13389.
[9] Liu S, Qi L, Qin H, *et al*. Path Aggregation Network for Instance Segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018,(pp. **8759**-8768).
[10] Wang J, Yang W, Li H-C, *et al*. Learning center probability map for detecting objects in aerial images [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, **59**(5): 4307-4323.
[11] Zhao J, Li J, Jin L, *et al*. The 3rd anti-uav workshop & challenge: Methods and results [J]. *arXiv preprint*, 2023, 2305.07290.