Enhancing YOLOv5 for Small Target Detection in UAV Aerial Images Via Multi-Scale Feature Fusion

Yuejuan Jing*, Zhouzhou Liu, Han Peng
[1]School of Computer Science, Xihang University, Xi'an, Shaanxi 710077, China
E-mail: jingyjyuej@hotmail.com
*Corresponding author

Keywords: high-altitude image, target detection, YOLOv5, multi-scale feature

Received: February 11, 2025

This paper briefly introduces unmanned aerial vehicle (UAV) aerial photography technology and the you only look once version 5 (YOLOv5) model for image target detection. In order to enhance the performance of the YOLOv5 model, modifications were implemented by increasing one feature fusion network layer in its neck network and an anchor box of a smaller scale in its head network. Subsequently, simulation experiments were conducted using images captured by UAVs to compare the improved model with the region-based convolutional neural network (R-CNN) and traditional YOLOv5 models. The findings indicated that the enhanced YOLOv5 model achieved faster convergence to stability during the training process. The R-CNN model required 130 iterations, the traditional YOLOv5 model required 120 iterations, and the improved YOLOv5 model only needed 80 iterations. Moreover, compared to the R-CNN and traditional YOLOv5 models, the improved YOLOv5 model demonstrated superior accuracy in identifying and localizing small targets in UAV aerial images. The improved YOLOv5 model had a precision of 0.989, a recall rate of 0.988, an F value of 0.988, and an intersection over union of 0.987 for small target recognition in images.

Povzetek: V članku je predstavljen izboljšani YOLOv5 z večnivojsko fuzijo značilk, ki učinkovito zaznava majhne tarče na UAV posnetkih s hitrejšo konvergenco in boljšo lokalizacijo v kompleksnih ozadjih.

1 Introduction

In the fields of environmental monitoring, disaster assessment, urban planning, and agricultural management [1, 2], unmanned aerial vehicle (UAV) aerial images offer a broad perspective and data support. However, the target objects required in the aforementioned application areas are often small targets within the images, and aerial images typically contain numerous complex backgrounds and small-sized targets, which significantly increases the challenge of rapidly and accurately detecting the small targets of interest. Traditional target detection methods are constrained by their feature expression capabilities and computational efficiency [3, 4], making it difficult to meet the high precision and real-time requirements for detecting small targets in UAV aerial images. In recent years, deep learning technology, a robust machine learning method, has been successfully utilized in image classification, target detection, and other projects [5, 6]. Its strong feature extraction capabilities and generalization characteristics offer a new approach to small target detection in UAV aerial images [7, 8]. This paper briefly introduces the UAV aerial photography technology and the you only look once version 5 (YOLOv5) model used for image target detection. In order to improve the performance of the YOLOv5 model, its neck network was improved by increasing one feature fusion network layer, and anchor boxes of corresponding scales were added in the head network. Finally, simulation experiments were carried out using the images captured by UAVs. The final experimental results verified that the improved YOLOv5 model can more accurately identify small targets in UAV aerial images and locate them in the images compared with the region-based convolutional neural network (R-CNN) and traditional YOLOv5 models.

2 Related works

Table 1: Related works

Author	Research content
Li et al. [9]	They developed an approach for detecting small targets in infrared remote sensing images, which addresses the issues of missing inspections and false alarms better than current advanced data-driven detection methods.
Zou et al. [10]	They introduced a defect detection method, YOLOv7-EAS, that enhances YOLOv7's ability to detect camera module images with complex backgrounds and small target defects.
Qiu et al. [11]	They developed a pixel-level local contrast measurement method that achieves a higher detection rate and lower false alarm rate compared to other methods while also achieving a high speed.

The research related to the recognition and detection of small targets in images is shown in Table 1. Some of the research focuses on infrared remote sensing images, attempting to detect small targets in infrared images. The color of infrared images is relatively single, and the targets to be detected often have higher heat than the surrounding environment and are generally more prominent in infrared images. Therefore, the detection of small targets in infrared images is relatively easier. Some of the research is focused on the improvement of the small target detection model, and the detection accuracy is improved by changing the structural parameters of the detection model. Some research is focused on improving the detection accuracy of small targets in local images, and the accuracy of the small target detection algorithm is improved by processing local images. The research of this paper attempts to improve the image small target detection model YOLOv5 to enhance the performance of the detection model. The reason for choosing YOLOv5 is that compared with YOLOv1-YOLOv4, the structural optimization of YOLOv5 has greatly improved its detection speed and accuracy. YOLOv6-YOLOv8, as the successors of YOLOv5, have improved the overall accuracy of the algorithm through structural adjustments, but they also lead to longer training times, affecting the iteration and optimization of the algorithm model. In some application environments that require rapid iteration and deployment, it can instead become a limiting factor. Secondly, the application scenarios of small target detection algorithms usually have sufficient detection accuracy, and there is no need to pursue the ultimate accuracy. The complex structures of YOLOv60-YOLOv8 that can bring higher accuracy may not lead to better performance but will result in higher costs. The detection accuracy and speed of YOLOv5 are balanced, and its classic model architecture has good potential for improvement. Therefore, this paper finally made improvements based on YOLOv5. The improvement method was to add another image feature fusion layer in the neck network and add an anchor box of the corresponding scale in the head network.

3 Target detection methods for UAV images

In image target recognition, a convolutional neural network (CNN) is a frequently used deep learning algorithm. However, while the algorithm is effective in recognizing images, its performance is poor when faced with images captured by UAV aerial photography. This is due to the complex background in aerial images and the small size of targets to be recognized [12]. For image target detection, the R-CNN series is a commonly utilized method. The principle of these algorithms for target identification in images is generating candidate regions through a neural network, classifying and recognizing these regions using another neural network, and adjusting their positions. Although the accuracy of this type of target recognition algorithm is generally high, a two-stage calculation results in low-efficiency recognition [13].

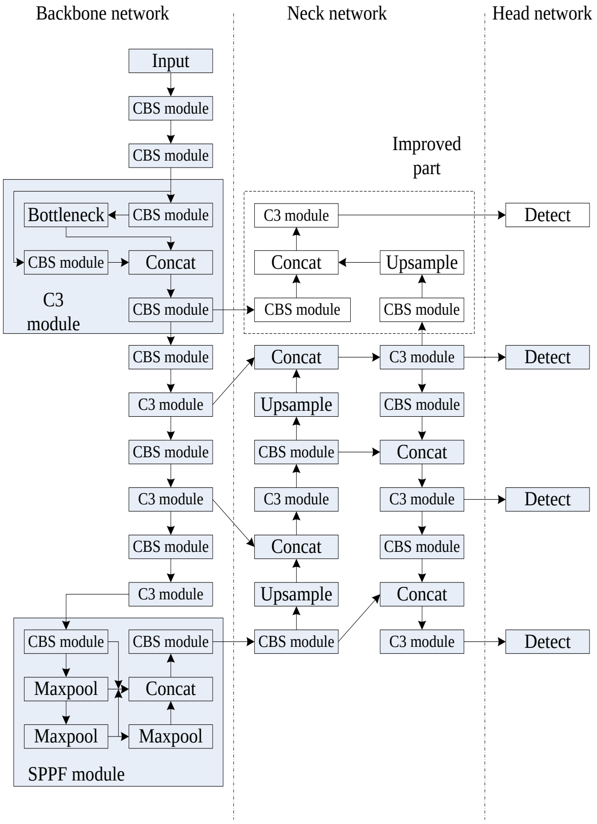


Figure 1: Basic structure of the improved YOLOv5 model

The various YOLO series models differ in the detailed construction within the three-part structure. In this study, the YOLOv5 model is utilized, and improvements are made to its neck network and head network structures to further enhance its performance in detecting small targets [14]. In the neck network, the number of layers in the feature fusion network is increased to enable the combination of smaller-sized image features with surface features. In the head network, very small anchor boxes are added to match with the increased number of feature fusion network layers. The basic structure of the enhanced YOLOv5 model is illustrated in Figure 1, and the structure in the dashed box of the neck network is the improved structure [15].

In the enhanced YOLOv5 model, features are extracted in the backbone network after an image is input. After processing by multiple CBS (conv + batch normalization + sigmoid linear unit) modules, C3 modules, and a spatial pyramid pooling fast (SPPF) module, image features at various scales are obtained [16]. Subsequently, the extracted features at different scales are integrated in the neck network. CBS modules and an up-sampling layer individually process the features extracted at different scales from the backbone network. Then, they are concated in the concat layer to generate the fused image features at various scales. Finally, the fused image features are recognized in the head network using anchor boxes of corresponding scale sizes.

The small target recognition process for UAV aerial images using the improved YOLOv5 model mentioned above is as follows.

① The UAV aerial photography flight route is set, then the UAV flies along the set route to collect images using the camera carried.

② The image is processed by denoising and enhancement. In this paper, the denoising is done by the bilateral filtering method [17], and the filtering formula used for enhancement is:

$$f(x,y) = (G(x,y) - m_g)\frac{cs_f}{cs_g + (1-c)cs_f} + bm_f + (1-b)m_g \text{ , (1)}$$

where $G(x,y)$ is the grey value of original pixel (x,y), $f(x,y)$ is the grey value of pixel (x,y) after the enhancement operation, m_g is the mean grey value of the original pixel, s_g is the standard deviation of the grey level of the image pixel, s_f is the variance of the grey level of the original pixel, c is the contrast expansion constant, 0.3 here, and b is the luminance coefficient, 0.2 here.

③ The image is input into the enhanced YOLOv5 algorithm for computation. The convolution formula [18] for the CBS module is:

$$H_i = \sigma(H_{i-1} \otimes \omega_i + b_i) \text{ , (2)}$$

where H_i and H_{i-1} are the feature maps of the output from the i - and $i-1$-th layers, ω_i is the weights in the structure of the i -th layer, b_i is the bias in the structure at the i -th layer, $\sigma(\cdot)$ is the activation function, and $\otimes$ is cyclic multiplication. In addition to the CBS module in the SPPF module, there are three pooling layers. This paper adopts maximum pooling for the sake of facilitating calculation, which utilizes a pooling box to move on the feature map, compresses the feature data within the box in the process of moving, and takes the maximum value in the box as the compression value.

④ After the forward computation, anchor boxes of different sizes are used in the head network part to recognize the targets inside the boxes. At the same time, regressive calculation is performed on the grid points of the anchor boxes to obtain the coordinates of the anchor boxes in the original image. Simply speaking, based on the anchor box coordinates in the feature fusion graph, the coordinates of the anchor box in the original image are calculated inversely according to the scale of the feature fusion graph.

4 Simulation experiments

4.1 Experimental data

The simulation experiments utilized independently collected images captured by a DJI Mavic 3 Pro UAV equipped with a Hasselblad camera, medium telephoto camera, and telephoto camera. In the process of collecting images using the camera carried by the UAV, the flight route was first designed according to the terrain of the area to be photographed. On sunny days, an image was taken every 30 seconds during the flight along the route. A total of 3,000 shots were taken, and some of the UAV images are displayed in Figure 2. Manual annotation was employed to delineate the small targets in the images and label their categories. During the labeling process, three people formed a group to ensure consistency.

Figure 2: Partial images

4.2 Experimental settings

The relevant parameter settings of the improved YOLOv5 model determined by orthogonal tests are outlined in Table 2. To evaluate the effectiveness of the algorithm, it was compared with the R-CNN algorithm and the traditional YOLOv5 model. The traditional YOLOv5 model was structured by excluding the enhanced part from the improved YOLOv5 model while keeping the remaining parameters the same. The R-CNN model was structured as follows: two conv layers - one maxpool layer - two conv layers - one maxpool layer - three conv layers - one maxpool layer - three conv layers - one maxpool layer - three conv layers - one region-of-interest (ROI) pooling layer. In the convolutional layer, there were 32 convolution kernels with a specification of 3 × 3, and the activation function was set to sigmoid. All three models adopted the same training dataset and test dataset.

Table 2: Relevant parameter settings for the improved YOLOv5 model

Parameter	Setting	Parameter	Setting
Resolution of the input image	640 × 640	Number of trainings	200
Batch	8	Optimizer	Adam [15]
Learning rate	0.01	Learning rate momentum	0.945

4.3 Evaluation criteria

The evaluation criteria of the detection algorithm's small target detection performance can be divided into two aspects: one is the recognition performance of the small target category, and the other is the localization precision of the small target frame. For the identification performance of small target categories, the confusion matrix was used to calculate precision (P), recall rate (R) and F values:

$$\begin{cases} P = \dfrac{TP}{TP + FP} \\ R = \dfrac{TP}{TP + FN} \\ F = \dfrac{2PR}{P + R} \end{cases}, (3)$$

where TP is the number of true positive samples, FP is the number of false positive samples, and FN is the number of false negative samples.

The localization accuracy for small target frames was measured using the degree of target frame overlap:

$$IOU = \frac{DR \bigcap GT}{DR \bigcup GT}, (4)$$

where IOU is the degree of target frame overlap [15], DR denotes the target frame forecasted by the algorithm, and GT denotes the actual target frame in the image.

4.4 Experimental results

As illustrated in Figure 3, the loss functions of the three algorithms during training gradually converged as the number of training iterations increased, and they stabilized after a certain number of training. The R-CNN model required 130 iterations, the traditional YOLOv5 model required 120 iterations, and the improved YOLOv5 model only needed 80 iterations. The improved algorithm demonstrated the fastest stabilization of convergence, followed by the traditional YOLOv5 model, and the R-CNN algorithm showed the slowest convergence. Furthermore, at the point of convergence stabilization, the improved algorithm exhibited the lowest training loss.

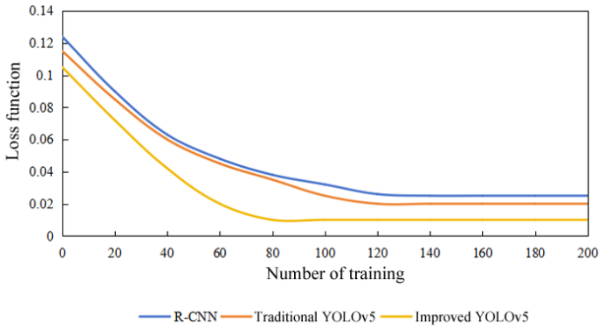
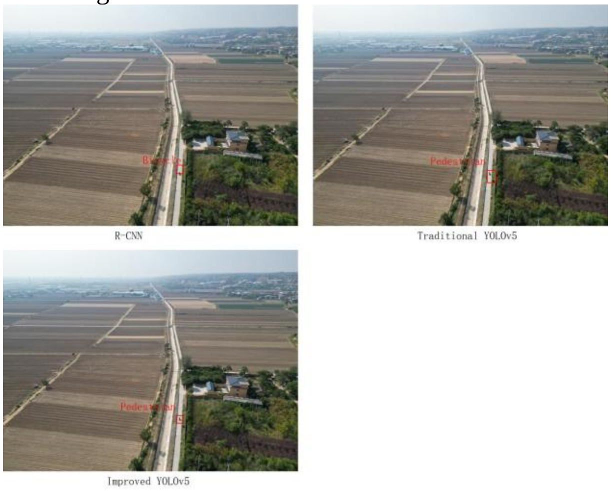


Figure 3: Convergence curve

In Figure 4, it is evident that when presented with the same aerial image, the improved YOLOv5 model accurately localized the small targets and correctly identified their types. Although the traditional YOLOv5 model identified the types of small targets, it exhibited a bias in localizing small targets. The R-CNN algorithm not only demonstrated localization bias for small targets but also exhibited bias in the identification of the categories of small targets.



R-CNN

Traditional YOLOv5

Improved YOLOv5

Figure 4: Partial detection results of three small target detections

The recognition and localization performance of the three small target detection algorithms are summarized in Table 3. The improved YOLOv5 model exhibited the best category recognition and target localization performance in detecting small targets in UAV aerial images, followed by the traditional YOLOv5 model, and the R-CNN algorithm demonstrated the lowest performance. The category recognition performance of the improved YOLOv5 model can be attributed to the increased layers of feature fusion and the utilization of smaller-scale image features. Additionally, the reduction in anchor box size due to the smaller scale allowed for better fitting of the localization frame to the small targets.

Table 3: Performance comparison

	R-CNN	Traditional YOLOv5	Improved YOLOv5
Precision	0.754	0.852^{*}	0.989^{+-}
Recall rate	0.749	0.843^{*}	0.988^{+-}
F	0.751	0.847^{*}	0.988^{+-}
Intersection over union	0.654	0.789^{*}	0.987^{+-}

Note: * indicates $p < 0.05$ in the comparison between the R-CNN and traditional YOLOv5 models, i.e., the difference is significant; + indicates $p < 0.05$ in the comparison between the R-CNN and improved YOLOv5 models, i.e., the difference is significant; - indicates $p < 0.05$ in the comparison between the traditional YOLOv5 and improved YOLOv5 models, i.e., the difference is significant.

Furthermore, during the testing process, the improved YOLOv5 model had false detections and positioning deviations when detecting small targets in aerial images. Analyzing the deviation cases, it was found that the appearance of the small targets falsely detected in the aerial images was indeed easily confused with other types from the planar perspective, while the error of the positioning box comes from the error brought by the restoration of anchor box coordinates in the feature fusion map.

5 Discussion

With the rapid development of UAV technology and its wide application in various fields, the demand for effective detection and recognition of targets in images captured by UAVs is increasing day by day. However, in these application scenarios, due to factors such as long shooting distance and wide viewing angle, the targets in these images are often small in size and sparsely distributed, which brings great challenges to the target detection algorithms. Traditional target detection algorithms based on deep learning perform well in general scenarios but often face the problem of insufficient accuracy when dealing with small targets. YOLOv5, as a small target detection model, achieves a good balance in detection speed and accuracy, but it still has room for improvement. In this paper, multi-scale feature fusion was adopted to improve the YOLOv5 model. Through multi-scale feature fusion, feature graphs of different levels can be combined, thereby capturing more abundant context information and enhancing the model's recognition ability for small targets, in order to handle targets of different sizes and shapes. Specifically, low-level feature graphs contain more detailed information, which is helpful for locating small targets, while high-level feature graphs provide higher-level semantic information, which is helpful for the recognition and classification of targets.

In this paper, based on the traditional YOLOv5 model, an additional feature fusion layer was added to the neck network part, and an anchor box of a corresponding scale was added to the head network part. Then, the images captured by a UAV were used for simulation experiments. It was found that the improved YOLOv5 model converged faster in the training process; compared with the R-CNN and traditional YOLOv5 models, the improved YOLOv5 model could more accurately identify the types and locations of small targets. The improved YOLOv5 model added a new feature fusion layer to the structure of the traditional YOLOv5 model, which can obtain feature graphs of smaller scales, thereby obtaining more abundant detail information of small targets in the image; in addition, due to the addition of a feature fusion layer of a smaller scale, the corresponding small-scale target box was also added in the head network, and the smaller target box can be more fitting when locating small targets.

The limitation of this paper lies in those only minor modifications were made to the YOLOv5 model, with no significant breakthrough in principle. Moreover, self-built UAV aerial images were used for model training, resulting in insufficient generalization. Therefore, the future research direction is to further improve the model and simultaneously expand the image range used in model training to enhance its generalization.

6 Conclusion

This paper briefly introduces UAV aerial photography technology and the YOLOv5 model for detecting image targets. To enhance the performance of the YOLOv5 model, improvements were made to its neck network by increasing the number of layers in the feature fusion network and incorporating anchor boxes of corresponding scales in the head network. Simulation experiments were conducted using UAV-captured images to compare the improved model with the R-CNN and traditional YOLOv5 models. The key findings are as follows. The improved YOLOv5 model exhibited the fastest convergence and the smallest loss function when stabilized. When detecting aerial images, the improved YOLOv5 model accurately located small targets and identified their types. The traditional YOLOv5 model identified the types of small targets but showed a deviation in the localization of small targets. The R-CNN algorithm exhibited deviations not only for small targets but also for the categories of small targets. The improved YOLOv5 model demonstrated the best category recognition and target localization performance in detecting small targets in UAV aerial images. The traditional YOLOv5 model was the second best, while the R-CNN algorithm performed the worst in recognizing and localizing small targets.

References

[1] Wu W, Liu H, Li L, Long Y, Wang X, Wang Z, Li J, Chang Y (2021). The application of a local fully convolutional neural network combined with the YOLO v5 algorithm in small target detection of remote sensing images. *PloS one*, 16(10), pp. e0259283. http://doi.org/10.1371/journal.pone.0259283

[2] Kıraç E, Özbek S (2024). Deep learning-based object detection with unmanned aerial vehicle equipped with embedded system. *Journal of Aviation*, 8(1), pp. 15-25. http://doi.org/10.30518/jav.1356997

[3] Ma M, Wang D, Sun H, Zhang T (2019). Infrared Dim-Small Target Detection Based on Robust Principal Component Analysis and Multi-Point Constant False Alarm. *Acta Optica Sinica*, 39(8), pp. 0810001. http://doi.org/10.3788/AOS201939.0810001

[4] Zhu H, Huang Y, Xu Y, Zhou J, Deng F, Zhai Y. (2024). Unmanned aerial vehicle (UAV) object detection algorithm based on keypoints representation and rotated distance-IoU loss. *Journal of Real-Time Image Processing*, 21(2), pp. 1-14. http://doi.org/10.1007/s11554-024-01444-6

[5] Chen H, Liu D (2019). Dim Moving Small Target Detection by Local and Global Variance Filtering on Temporal Profiles in Infrared Sequences. *Aerial Weapon*, 26(6), pp. 43-49. http://doi.org/CNKI:SUN:HKBQ.0.2019-06-008

[6] Tyagi S (2024). Deep reinforcement learning based framework for tactical drone deployment in rigorous terrains: From modeling to real-world implementation. *Web 3.0-Deep reinforcement*

learning based framework for tactical drone deployment in rigorous terrains: From modeling to real-world implementation, pp. 39-53.

[7] Wang L, Yu J (2024). Infrared Weak and Small Target Detection Algorithm Based on Deep Learning. *International Journal of Advanced Network, Monitoring and Controls*, 9(3), pp. 47-55. http://doi.org/10.2478/ijanmc-2024-0026

[8] Yu Q, Liu A, Yang X, Diao W (2024). An Improved Lightweight Deep Learning Model and Implementation for Track Fastener Defect Detection with Unmanned Aerial Vehicles. *Electronics*, 13(9), pp. 1781. http://doi.org/10.3390/electronics13091781

[9] Li R, Shen Y (2023). YOLOSR-IST: A deep learning method for small target detection in infrared remote sensing images based on super-resolution and YOLO. *Signal Processing: the Official Publication of the European Association for Signal Processing (EURASIP)*, 208, pp. 108962-1-108962-12. http://doi.org/10.1016/j.sigpro.2023.108962

[10] Zou H, He G, Yao Y, Zhu F, Zhou Y, Chen X (2023). YOLOv7-EAS: A Small Target Detection of Camera Module Surface Based on Improved YOLOv7. *Advanced Theory and Simulations*, 6(11), pp. 2300397.1-2300397.13. http://doi.org/10.1002/adts.202300397

[11] Qiu Z B, Ma Y, Fan F, Huang J, Wu MH, Mei XG (2022). A pixel-level local contrast measure for infrared small target detection. *Defense Technology*, 18(9), pp. 1589-1601. http://doi.org/10.1016/j.dt.2021.07.002

[12] Choi L, Chung W Y, Park C G (2024). CSI-Net: CNN Swin Transformer Integrated Network for Infrared Small Target Detection. *International Journal of Control, Automation and Systems*, 22(9), pp. 2899-2908. http://doi.org/10.1007/s12555-024-0089-8

[13] Guo T, Zhou B, Luo F, Zhang L, Gao X (2024). DMFNet: Dual-Encoder Multistage Feature Fusion Network for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62, pp. 1-14.

[14] Deng L, Song J, Zhu X H (2022). When Infrared Small Target Detection Meets Tensor Ring Decomposition: a Multiscale Morphological Framework. *IEEE Transactions on Aerospace and Electronic Systems*, 58(4), pp. 3162-3176. http://doi.org/10.1109/taes.2022.3147435

[15] Chen D, Sun S, Lei Z, Shao H, Wang Y (2021). Ship Target Detection Algorithm Based on Improved YOLOv3 for Maritime Image. *Journal of Advanced Transportation*, 2021(10), pp. 1-11. http://doi.org/10.1155/2021/9440212

[16] Pan C, Zhao H, Sun M (2024). Real-time target detection system in scenic landscape based on improved yolov4 algorithm. *Informatica*, 48(8), pp. 35-48. http://doi.org/10.31449/inf.v48i8.5700

[17] Gao F, Huang T, Sun J, Wang J, Hussain A, Yang E (2018). A New Algorithm of SAR Image Target Recognition Based on Improved Deep Convolutional Neural Network. *Cognitive Computation*, 2018(5), pp. 1-16. http://doi.org/10.1007/s12559-018-9563-z

[18] Kim S (2017). Infrared variation reduction by simultaneous background suppression and target contrast enhancement for deep convolutional neural network-based automatic target recognition. *Optical Engineering*, 56(6), pp. 063108.

[19] Tang H, Han Y, Zheng J, Wang Z, Wang L (2025). An optimized yolov5s-rd framework for efficient target detection in remote sensing images. *Informatica*, 49(18), pp. 1-28. http://doi.org/10.31449/inf.v49i18.7849