OBJECT DETECTION IN SECURITY SCENE BASED ON IMPROVED YOLOv5

Kunwei LV[1], Ruobing WU[2], Zhiren XIAO[3], Ping LAN[4,*]

In this study, traditional manual security detection methods suffer from significant drawbacks, especially low efficiency and high cost. To overcome these challenges, this paper introduces a new approach: First, a parallel convolutional module is designed and enhanced by a hybrid attention mechanism, which significantly improves the network's ability to process complex image data. Second, a decoupled detection header is devised, aiming to enhance the neural network's performance in classification tasks and regressions. Lastly, a hybrid data enhancement strategy and an anchor frame adaptive matching technique are integrated, enhancing the network's robustness. These innovations aim to significantly boost object detection efficiency and capability, improve detection accuracy, and extend method applicability to diverse scenarios. The approach surpasses the benchmark YOLOv5m by 6.10%, demonstrating its effectiveness.

Keywords: anchored frame matching approach, decoupled header, hybrid attention, target identification, and data augmentation

1. Introduction

Air terminals, rapid transit train depots, swift courier hubs, and additional mass transit junctions extensively utilize radiographic safety inspection systems to ensure the security of passengers by examining bags and parcels. Nevertheless, the conventional manual scrutiny approaches are becoming increasingly impractical due to their high subjectivity, low efficiency, substantial expenses, and susceptibility to inaccuracies and misdirections. With the swift progression of machine vision technology, deep learning-based item recognition has gained prominence in various domains, including video monitoring, healthcare imagery evaluation, smart manufacturing, self-driving vehicles, and human-machine interaction. Item recognition, a fundamental and challenging undertaking in

[1] School of Information Science and Technology, Tibet University, Lhasa, China. lkw872473172@163.com

[2] School of Information Science and Technology, Tibet University, Lhasa, China. w1233737@qq.com

[3] School of Information Science and Technology, Tibet University, Lhasa, China. xiaozhiren0222@163.com

[4] School of Information Science and Technology, Tibet University, Lhasa, China. Lanping@utibet.edu.cn, *Corresponding author

machine vision, underpins advanced applications like instance segmentation, image analysis, and video tracking. These tasks not only necessitate locating items within an image but also categorizing them. In the context of security screening, item recognition technology can revolutionize the process by enabling real-time identification of prohibited items. This application can significantly decrease inspector workloads and boost operational efficacy, playing a pivotal role in the realms of intelligent transportation, logistics, and public safety.

The study of deep convolutional neural networks (DCNN) underwent a paradigm shift following the seminal work of Hinton et al. [1], who utilized AlexNet, a profound convolutional neural network, for extensive image categorization, clinching victory in the 2012 ImageNet competition. This milestone marked a new direction in DCNN research, particularly within the realm of object detection, where models with one and two phases are now dominant.

The two-stage detection models involve three primary steps: (1) Girshick et al.'s method combines semantic segmentation with object detection, significantly enhancing detection accuracy through a comprehensive, multilevel feature representation [2]. (2) The introduction of ROI pooling layers and RPN by Shaoqing Ren et al. led to the development of Faster R-CNN [3]. (3) Kaiming He et al. proposed a groundbreaking neural network architecture, the Residual Network (ResNet), introducing the concept of residual learning [4].

In contrast, models with one phase in detection, exemplified by the YOLO [5] and SSD [6] families, are based on regression analysis. These models, along with the R-CNN family, represent a candidate region-based approach to object detection.

Originally, deep convolutional neural networks were employed by R-CNN as an alternative to traditional object detection methods. This adoption has since spurred a substantial increase in the use of DCNNs for target identification, leading to the development of numerous effective models that leverage DCNNs to address various challenges in object detection.

The YOLO (You Only Look Once) object detection framework, introduced by Redmon et al. [5], represented a significant breakthrough as the inaugural neural network framework capable of real-time object detection. To achieve an optimal trade-off between detection accuracy and processing speed, subsequent versions and improvements were developed, drawing inspiration from related research. These include YOLOv4 [7], YOLOX [8], and YOLOv7 [9], each integrating additional modular structures and enhancement techniques.

Xu et al. [10] proposed an attention mechanism grounded in cognitive science theory, presenting a novel approach to managing computational resources in deep learning. This deep learning attention mechanism addresses the challenge of information overload by focusing limited computational capacity on select critical tasks. Selvaraju et al. [11] developed Grad-CAM, a technique employing a heatmap to visualize the network's prediction process, thus providing partial insights into the functioning of neural networks.

Sun et al. [12] contributed to this field by developing a multiscale self-attention module. This module, which synthesizes self-attention in both spatial and channel dimensions, enables the network to extract information across multiple scales by grouping convolved feature data. However, a specific challenge in object detection within security screening scenarios is the prevalence of mutual occlusion and overlapping among targets. In such contexts, the channel information of feature maps assumes greater importance than spatial data.

In object detection networks, the detection head plays a crucial role in processing fused feature maps to generate final detection frames and labels. Originally, the YOLO family adopted a coupled detection head, where both localization and classification branches were integrated and shared. However, this approach can lead to conflicts between localization and classification tasks due to their differing feature representation requirements, potentially impeding network performance [13].

Song et al. [14] conducted experimental research on the localization and classification subtasks within object detection tasks. Their findings suggest that a convolutional head is better suited for localization, whereas a dense head (dense-head) is more suitable for classification tasks. This insight underscores the importance of designing detection heads that align with the specific demands of each subtask.

Furthermore, the design of a priori frames significantly influences model performance. Anchor frames, or sets of predefined a priori frames, are employed in object detection models to fine-tune the network’ s final output and provide a more nuanced detection mechanism. Literature reveals the emergence of object detection models that employ an anchor frame-free strategy [15]. Comparative analyses between anchor frame-free and anchor frame-based methods, under identical network structures, reveal that anchor frame-free approaches exhibit superior performance in hazardous material detection tasks.

Nevertheless, a middle ground between anchor frame-free and anchor frame methods can be achieved by generating dataset-specific groups of anchor frames using clustering-based methods [16, 17]. However, it is crucial to consider that for tasks with potential targets exhibiting varied aspect ratios, the anchor frame-free approach may negatively impact the model's performance. This emphasizes the importance of customizing anchor frame strategies to cater to the unique needs of each object detection task. A range of routine changes to the training samples is known as data augmentation, and it assists in teaching the model more fundamental characteristics of the dataset and improves its ability to adjust to small changes in the samples (thereby decreasing sensitivity to change). Two popular techniques for enhancing data are mosaic [17] and mix up [16].

In choosing YOLOv5[18] as the base model, we consider its excellent performance in various target detection tasks. YOLOv5 achieves a good balance between accuracy and speed and is especially suitable for security scenario detection with high real-time requirements. In addition, YOLOv5 has relatively low computational resource requirements compared to other models, making it more suitable for deployment in resource-limited environments. Specifically, the performance of the YOLOv5m version on multiple benchmark datasets shows that it possesses high detection accuracy and robustness, making it an ideal choice for this study.

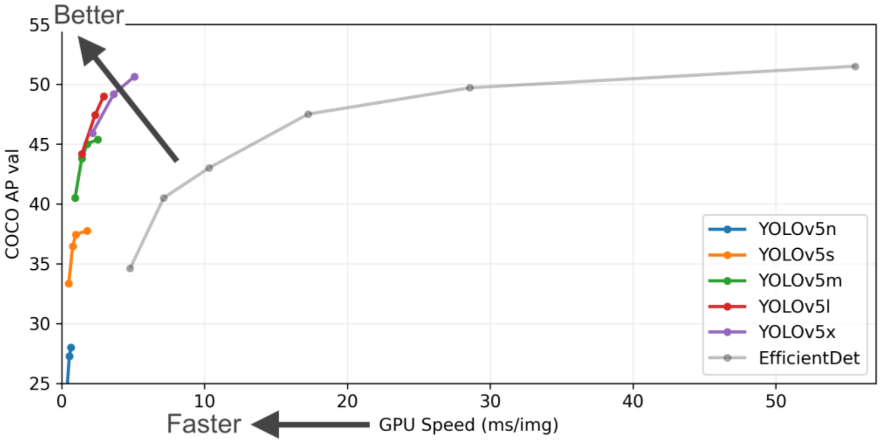


Fig. 1. YOLOv5 model performance

In this study, YOLOv5 serves as the benchmark model, and our primary contributions include:

(1) In this paper, a parallel convolution module that integrates hybrid attention mechanisms is introduced. This module leverages the synergistic effects of various attention mechanisms to enhance the network's focus on specific tasks, particularly in security screening scenarios. The parallel structure facilitates the acquisition of richer gradient flow information, thereby enhancing the network's analytical capabilities.

(2) To resolve the discrepancy between classification and regression tasks in object detection, this paper introduces a decoupled detection head as a replacement for the coupled detection head employed in the original model. This alteration seeks to enhance the network's performance tailored to specific tasks.

(3) In this paper, two new strategies are introduced to enhance the efficacy of networks in security screening: hybrid data enhancement methods and anchor frame adaptive matching techniques.

The structure of the paper unfolds as follows: Section 2 outlines the methodology and enhancements introduced in this study. Section 3 encompasses the experimental setup, dataset description, model assessment metrics, analysis of experimental results, ablation studies, and comparative experiments. Finally, Section 4 presents the conclusions drawn from this research.

2. Techniques

2.1 Module RCES

The RCES (Residual Convolutional Efficient Squeeze-and-Excitation) module in our study is specifically designed to amplify salient features while concurrently suppressing less relevant ones. This is achieved by learning the significance of each feature channel and accordingly assigning variable weights across different regions of the feature map. A key component of this mechanism is the ECA-SENet hybrid attention module, as illustrated in Fig.. 2. This module integrates two distinct networks: the Efficient Channel Attention (ECA) and the Squeeze-and-Excitation (SE) networks [19, 20].

The ECA-SENet module enhances the ability of the convolutional neural network to prioritize specific channels by assigning them greater weights. This selective weighting of feature maps is contingent upon the nature of the task at hand. In the context of security screening scenarios, such a mechanism significantly boosts the network’s performance by focusing on the most pertinent features for analysis.

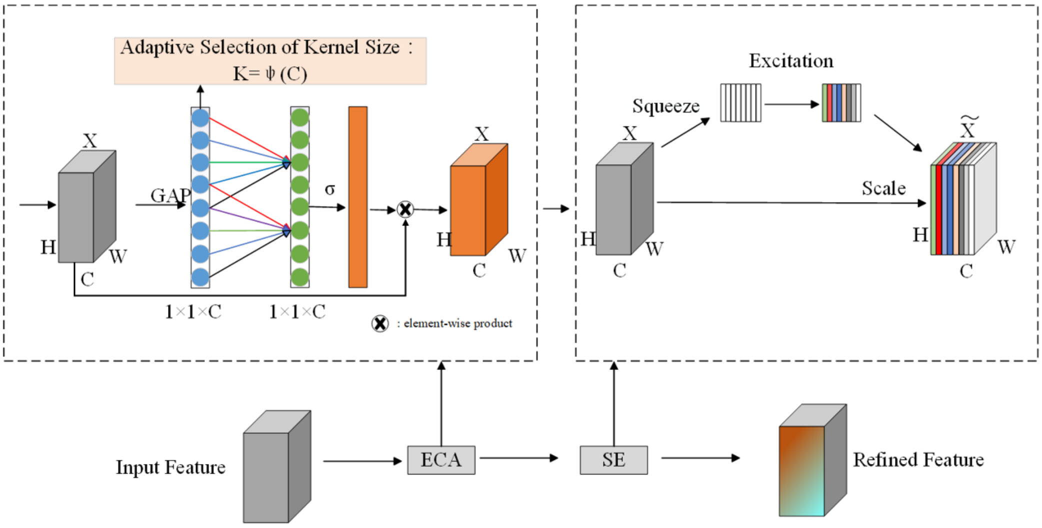


Fig. 2. ECA-SENet

Fig. 1 in the paper illustrates the dimensions of the feature map, indicating 'H', 'W', and 'C' as the height, width, and channel count, respectively. The Efficient

Channel Attention (ECA) module, a critical component of our model, effectively reduces computational complexity compared to a dense layer. This reduction is achieved by integrating a 1 × 1 convolutional layer subsequent to the global average pooling layer, thereby facilitating efficient cross-channel interaction.

To maximize the utilization of channel information, this paper introduces equation (1). This equation is pivotal in enabling the network to adaptively select the extent of the convolutional kernel, which is contingent on the channel count present. This adaptability allows for more effective processing and integration of channel-specific information, thereby enhancing the overall efficacy of the network in handling diverse channel quantities:

$$\varphi(C) = \left|\frac{log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{1}$$

In the context of the equation, 'C' denotes the channel count in the feature map. The parameters 'γ' and 'b' are constants, assigned values of 2 and 1, respectively. These constants are pivotal in determining the extent of the convolution kernel, denoted as Φ(C). An important aspect of this formula is the condition |t| odd, which signifies that if the computed value of t is not an odd number, it should be rounded to the closest odd number. This adjustment is critical to ensure the symmetry and effectiveness of the convolution kernel, enabling it to adapt to the varying channel dimensions more effectively.

To construct a channel descriptor, the Squeeze Excitation (SE) module initially executes a compression operation on the feature maps, which are spatially dimensioned as H×W. This operation consolidates the feature mappings across these spatial dimensions to transform a feature map of dimensions H×W×C into a channel descriptor of dimensions 1×1×C. This transformation effectively condenses the global spatial information of the feature maps into the channel descriptors, ensuring that the input layer can utilize this condensed form of data effectively. Equation (2) in the paper mathematically delineates this compression operation, providing a clear representation of how the channel descriptors are derived from the feature maps.

$$F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{2}$$

The global spatial information within the network is encapsulated in a collection of local descriptors, denoted as uc. Following this, a channel-dependent self-selecting gate mechanism is employed. This mechanism is pivotal in enabling each channel to selectively highlight informative features while attenuating less significant ones, a process guided by learning sample-specific activations.

To empower the network with the capability to extract finer features from complex images, this paper innovatively designed a parallel convolution module. This design draws inspiration from the deep residual network and innovates upon the original C3 module of YOLOv5 by introducing parallel branching. This addition allows the network to access a more comprehensive gradient flow information, while maintaining its lightweight architecture.

Building upon this foundation, we integrate the ECA-SENet module, as previously discussed. The resultant module, combining the parallel branching strategy with ECA-SENet, is designated as the RCES module. Fig. 3 in our paper visually contrasts this newly developed RCES module (b) with the original C3 module (a), illustrating the enhancements and modifications made for this research.

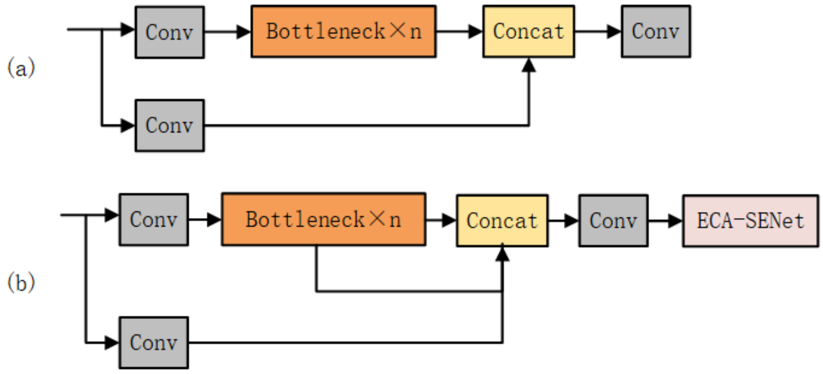


Fig. 3. Comparing Modules

2.2 Decoupled Head for Detection

To mitigate conflicts between localization and classification tasks in object recognition, our recognition head comprises two distinct components: a convolutional network dedicated to regression tasks for the target frame, and a dense network focusing on classification tasks. Fig. 4 in our paper illustrates the structures of both the coupled and decoupled recognition heads.

In the decoupled head, feature maps are bifurcated into two branching networks. The first branch, a convolutional network, is tasked with the localization job. It extracts features using a 3×3 downsampled convolutional layer, specifically tuned for this purpose. The second branch, a dense network, is designed for the classification task. It adjusts the channel dimensions of the feature map to correspond with the count of classes of the predicted target.

Subsequently, the feature maps are processed by two separate networks. The first network is responsible for predicting the anchor frame’ s dimensions – its height, width, and center coordinates. The second network focuses on calculating the intersection between the predicted and actual frames. This bifurcated approach enhances the robustness of the model and its generalization ability. The overarching objective of this network structure design is to independently extract and learn the target location and category information through different network branches, before eventually fusing these features for final output.

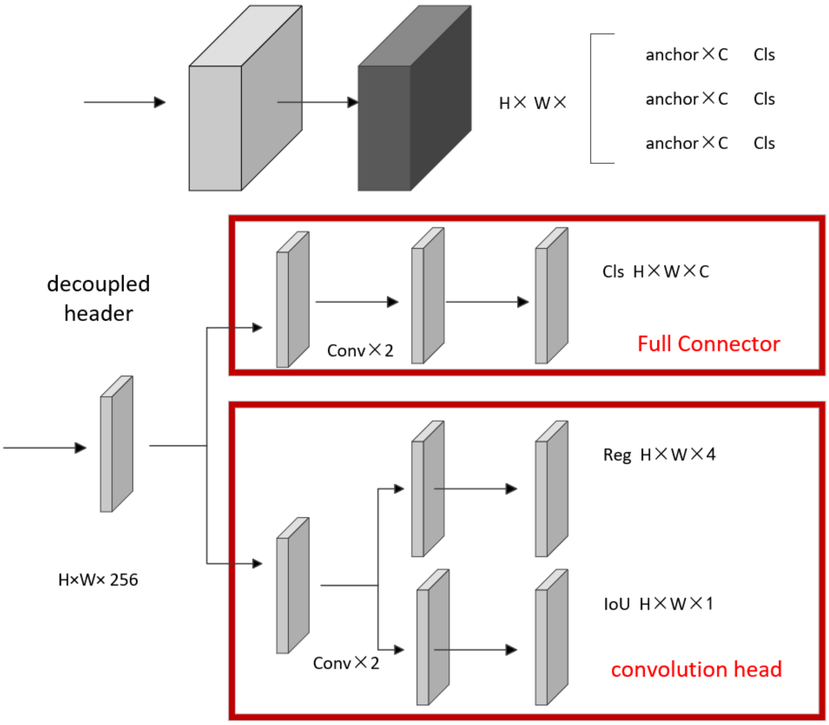


Fig. 4. Decoupling header and header

2.3 Mix up data enhancement

The mix-up data enhancement method, although simple in concept, has proven to be a highly effective technique for data augmentation. This method involves selecting two random samples from the training dataset and performing a basic random weighted summation of these samples. Essentially, this process blends two images together, creating a new composite image that retains elements from both original images. This technique is particularly beneficial in scenarios where diversity in training data can lead to more robust models.

As illustrated in Fig. 5 of our paper, the mix-up data enhancement method has been determined to be particularly well-suited for the task scenario addressed in this research. By employing this approach, the training dataset can be augmented in a way that introduces variability and complexity, thereby enhancing the network's capability to generalize from the training dataset to new, unseen samples.

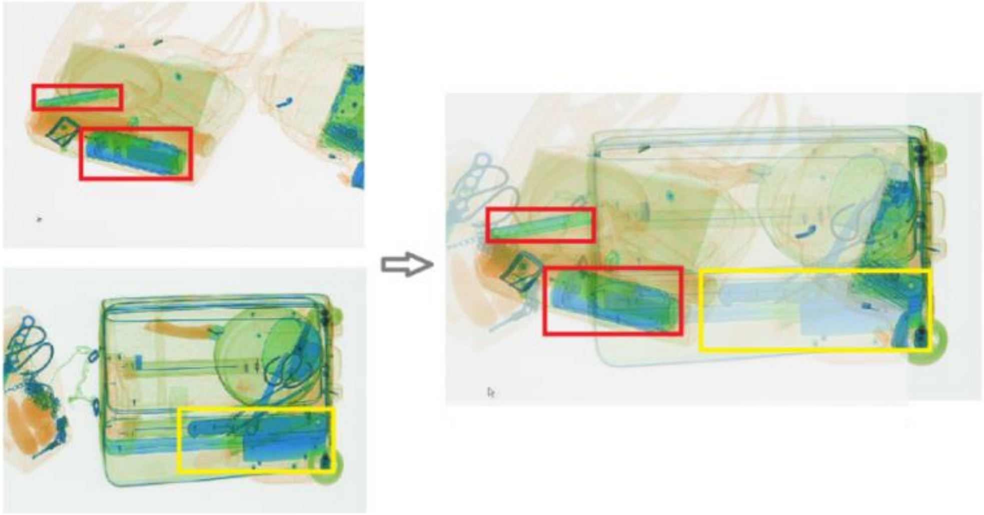


Fig. 5. Mix up data enhancement

2.4 Anchor frame alignment technique

This study develops a method to generate a customized set of anchor framesets specifically for a given dataset. This process involves analyzing and computing the characteristics of the anchor frames within the dataset. To achieve this, a new combination of K-means clustering and genetic algorithms is used in this paper. This methodology allows for the efficient identification and selection of the most representative anchor frames, ensuring optimal compatibility with the specific features of the dataset.

The algorithm's workflow is comprehensively depicted in Fig. 6 of our paper. This illustration provides a step-by-step visual representation of the process, from the initial data input through to the final generation of the anchor frame groups. By combining K-means clustering with genetic algorithms, the accuracy and effectiveness of the anchor frame selection process are improved, thus enhancing the overall effectiveness of the target detection model.

Our method starts with initializing the {k} cluster centroids, which are chosen empirically to best fit the features of the dataset, to represent the initial positions of the anchor frames in K-means clustering. After that, each bounding box in the dataset is grouped with other comparable bounding boxes by assigning it to the closest cluster centroid using the Euclidean distance. Iteratively averaging the bounding boxes allocated to each cluster, the centroids are recalculated until they stabilize.

We use a genetic approach to refine these clusters after the K-means initialization to make sure they are best suited to the dataset. Each member of the population is represented by a set of anchor frames in this algorithm, and the average Intersection over Union (IoU) between the anchor frames and the ground truth bounding boxes is used to determine each member's fitness. Based on their fitness scores, a roulette wheel selection mechanism chooses people for reproduction in a probabilistic manner. To introduce diversity, a single-point crossover method permits parent individuals to trade portions of their anchor frame sets to form offspring. To avoid premature convergence, mutation is administered with a probability of 0.1 and modifies randomly chosen anchor frames. Until a predetermined number of generations or a plateau in fitness improvement is reached, this process iterates across several generations, possibly creating better-suited anchor frames through selection, crossover, and mutation. In order to ensure well-matched anchor frames for more accurate and dependable object detection, this integrated technique handles the issue of various aspect ratios and scales within the dataset. It is especially well-suited to the intricate requirements of security scene analysis.

3. Investigation and evaluation

3.1 Setting for experimentation and metrics for assessment

Table 1 provides a comprehensive overview of the experimental environment configuration employed in this research. Our study utilized the EDS dataset [21], which encompasses a total of 31,655 instances of target objects. This dataset is composed of 14,219 images captured using three different scanning devices, featuring 10 different categories of objects. Each image within the dataset is meticulously annotated by professionals.

For training the model, this paper constructs the training and test sets by aggregating and randomly dividing the data collection. The training set comprises 1743 samples, while the test set comprises 12476 samples. This partitioning resulted in a training-to-test ratio of approximately 7:1, guaranteeing a thorough assessment of the model's effectiveness.

Table 1

Experimental environment configuration

parameters	configure
CPU	Intel(R) Xeon(R) Platinum 8255C
GPU	NVIDIA GeForce RTX 3090
system environment	Ubuntu 18.04
multilingualism	Python 3.8
Accelerated environment	Cuda 11.1
PyTorch version	1.8.0

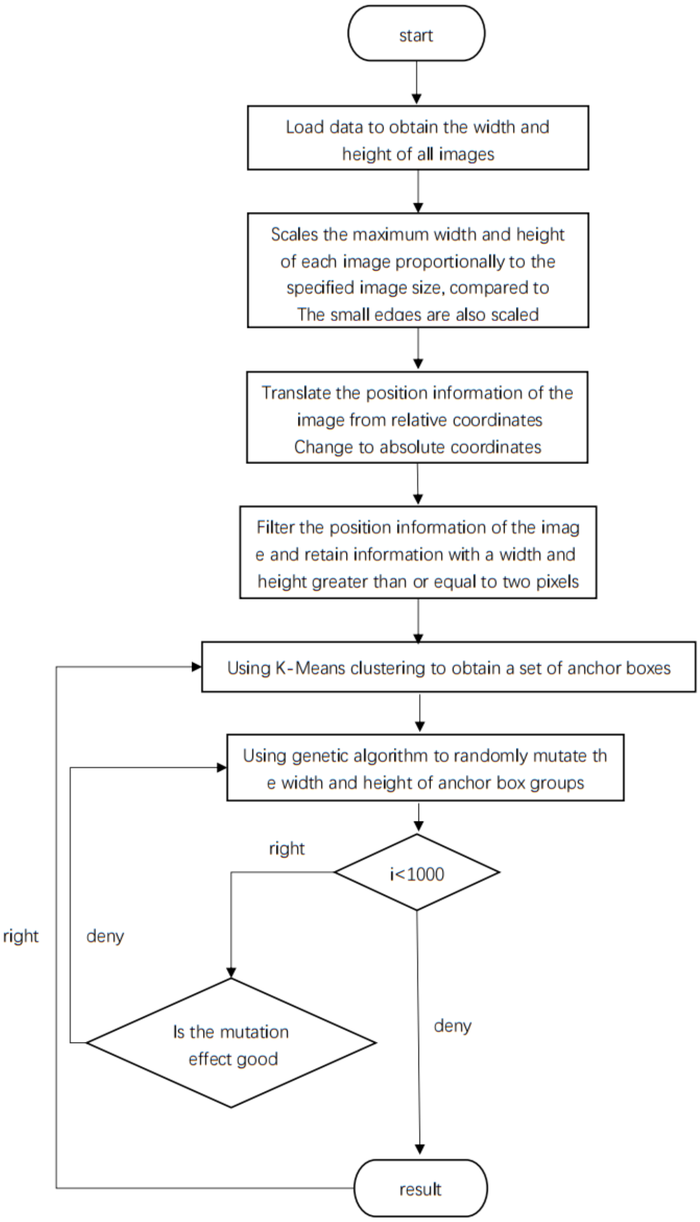

start
Load data to obtain the width and height of all images
Scales the maximum width and height of each image proportionally to the specified image size, compared to The small edges are also scaled
Translate the position information of the image from relative coordinates Change to absolute coordinates
Filter the position information of the imag e and retain information with a width and height greater than or equal to two pixels
Using K-Means clustering to obtain a set of anchor boxes
Using genetic algorithm to randomly mutate th e width and height of anchor box groups
right
i<1000
right
deny
Is the mutation effect good
deny
result

Fig. 6. Matching strategy for anchor boxes.

The evaluation of our model is based on key metrics, including mean average precision (mAP), recall (Recall), and precision (Precision). These metrics offer crucial perspectives on the model's performance, ensuring a comprehensive assessment of its capabilities.

The computation of these metrics is formally illustrated by the following equation:

$$Precision = \frac{TP}{TP+FP} \tag{9}$$

$$Recall = \frac{TP}{TP+FN} \tag{10}$$

$$AP = \sum_{i=1}^{n-1}(r_{i+1} - r_i)P_{inter}(r_i + 1) \tag{11}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{12}$$

Precisely, the model's precision is quantified as the percentage of correctly identified positive categories out of all positively labeled samples. This calculation involves considering false positives (FP), which represent negatively assessed samples that the model erroneously classified as positive.

On the other hand, recall evaluates the effectiveness of the model in accurately detecting positive categories. This metric is determined by comparing the true positives (TP), which are the accurately classified positive samples by the model, to the false negatives (FN), which are samples incorrectly classified as negative.

To thoroughly evaluate the performance of the model, this paper employs the mean accuracy (mAP) metric. The mAP metric combines precision and recall, offering a holistic measure of the model's efficiency.

Notably, the mAP was calculated for different intersections over the union (IoU) threshold. The metric was denoted as mAP 0.5 when a threshold of 0.5 was assumed. Additionally, model performance was evaluated for a series of IoU thresholds between 0.5 and 0.95, increasing by 0.05, denoted as mAP 0.5:0.95.3.2. experimental analysis.

Table 2 presents the experimental results pertaining to the RCES module. In our dataset, compared to the benchmark model, the introduction of the RCES module resulted in notable improvements across various metrics. Specifically, there was a substantial increase in mean average precision at an IoU of 0.5 (mAP 0.5) by 1.40%, a significant enhancement in mean average precision across the range of IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95) by 2.50%, a marginal increase in accuracy by 0.10%, and a substantial boost in checking completeness by 2.00%. These findings underscore the efficacy of the RCES module in enhancing the model's performance across multiple evaluation criteria.

Table 2

Experimental results of RCES module

Model	mAP 0.5	mAP 0.5:0.95	accurate	recall rate
YOLOv5m	0.781	0.559	0.836	0.706
YOLOv5m_RCES module	0.795	0.584	0.837	0.726

To enhance accuracy and provide some insight into the neural network's functionality, this paper employs a Grad-CAM [22] heatmap visualization to gain a clearer understanding of the attentional mechanisms of the network. In this visualization, the network's focus is represented through a color gradient, with warmer colors indicating higher attention or 'heat' at specific locations within the image.

Fig. 7 in our paper showcases the detection results on the original image, which initially contained four items. Grad-CAM visualization allows us to compare the outcomes of the YOLOv5s network before and after the integration of the ECA_SE hybrid attention module. Fig. 8 provides this comparison, with (a) depicting the focus of the enhanced network and (b) showing the focus of the pre-enhanced network. Upon examination, it becomes evident that the improved network exhibits more focused attention on the objects of interest, while the pre-improved network demonstrates a more dispersed attention pattern.

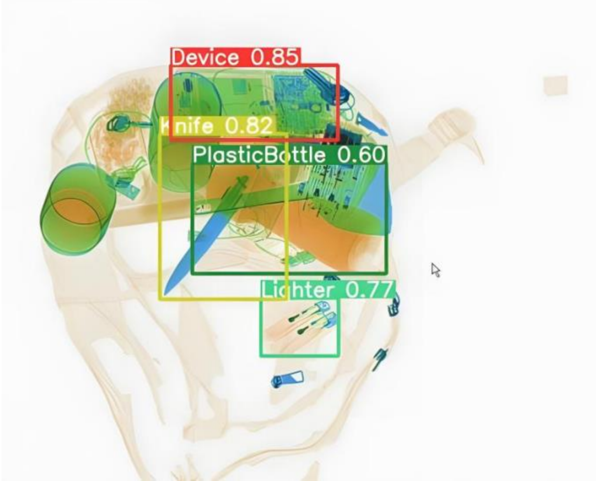
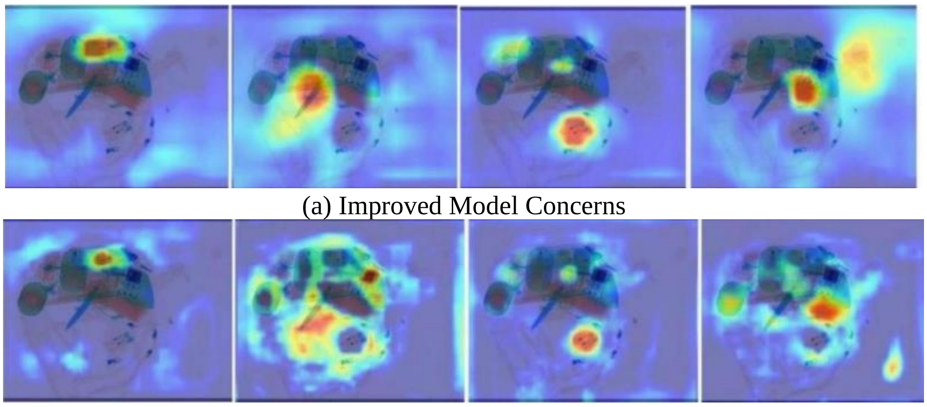


Fig. 7. Detection result of original image



(a) Improved Model Concerns

(b) Pre-improvement model concerns

Fig. 8. Grad-CAM heat map visualization

Table 3 provides a comprehensive presentation of the experimental outcomes concerning the anchor frame adaptive matching strategy. After implementing this strategy in our dataset, significant improvements were observed in several key metrics. Specifically, there was a notable increase in mean average precision at an IoU of 0.5 (mAP 0.5) by 0.7%, a substantial enhancement in mean average precision across the range of IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95) by 0.8%, and a significant boost in recall by 1.1%. These findings underscore the effectiveness of the anchor frame adaptive matching strategy in enhancing the model's performance across various evaluation criteria.

However, it is worth noting that there was a marginal decrease in precision by 0.7% when comparing the results to the baseline model. This trade-off between recall and precision warrants further consideration in the context of specific application requirements.

Table 3

Experimental results of the anchor frame adaptive matching strategy

Model	mAP 0.5	mAP 0.5:0.95	accurate	recall rate
YOLOv5m	0.781	0.559	0.836	0.706
YOLOv5m_autoanchor	0.788	0.567	0.829	0.717

Table 4 presents the outcomes of the Mix up data enhancement technique experiment. With the introduction of this technique, our dataset experienced significant improvements in multiple key metrics. Specifically, there was a substantial increase in mean average precision at an IoU of 0.5 (mAP 0.5) by 1.60%, a noteworthy enhancement in mean average precision across the range of IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95) by 1.60%, a substantial boost in accuracy by 1.70%, and a notable improvement in checking completeness by 1.00%. These findings demonstrate the efficacy of the Mix up data enhancement technique in enhancing the model's performance across various evaluation criteria when compared to the baseline model.

Table 4

Experimental results of hybrid data enhancement strategies

Model	mAP 0.5	mAP 0.5:0.95	accurate	recall rate
YOLOv5m	0.781	0.559	0.836	0.706
YOLOv5m-mixup	0.797	0.575	0.853	0.716

Table 5 provides a comprehensive display of the experimental findings associated with the decoupled detection head. The integration of this approach yielded notable improvements in several key metrics. Specifically, there was a significant increase in mean average precision at an IoU of 0.5 (mAP 0.5) by 0.5%, a substantial enhancement in mean average precision across the range of IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95) by 1.3%, and a notable improvement in checking completeness by 1.2%. These findings highlight the effectiveness of the decoupled detection head in improving the model's performance across various evaluation criteria.

However, it is worth noting that there was a marginal decrease in accuracy by 0.3% when comparing the results to the benchmark model. This trade-off between accuracy and other metrics warrants further consideration in the context of specific application requirements.

Table 5

Experimental results of decoupled detection head

Model	mAP 0.5	mAP 0.5:0.95	accurate	recall rate
YOLOv5m	0.781	0.559	0.836	0.706
YOLOv5m_decoupled head	0.786	0.572	0.833	0.718

3.3 Ablation experiment

Table 6

Ablation experiments

	RCES module	Auto anchor	mixup	Decouple head	mAP 0.5	mAP 0.5:0.95	P	R
1					0.781	0.559	0.836	0.706
2	√				0.791	0.574	0.842	0.724
3	√	√			0.801	0.582	0.834	0.742
4	√	√	√		0.811	0.603	0.847	0.745
5	√	√	√	√	0.823	0.620	0.859	0.753

After implementing all the enhanced strategies, our model demonstrated significant overall improvements in key performance metrics. Notably, there was a substantial increase in mean average precision at an IoU of 0.5 (mAP 0.5) by 4.20%, a remarkable enhancement in mean average precision across the range of IoU thresholds from 0.5 to 0.95 (mAP 0.5:0.95) by 6.10%, a notable improvement in precision by 2.30%, and a substantial boost in recall by 4.70%. These comprehensive improvements underscore the effectiveness of our combined strategies in enhancing the model's overall performance. The results of the ablation experiment are shown in Table 6..

Recognizing that while the introduction of some strategies led to slight performance degradation in specific aspects, the net gain in overall performance demonstrates the successful integration and synergy of these strategies. 3.4 comparative experiment.

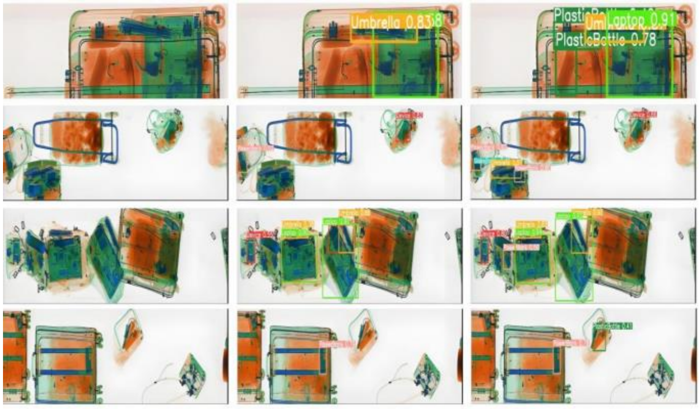
In the comparative experimental section, several prominent target detection models underwent comprehensive evaluation. The results, as presented in Table 7, unequivocally demonstrate the distinct advantages of our enhanced model over the other network models in every aspect assessed. Our model exhibits superior performance across a range of metrics, reaffirming its effectiveness and competitiveness.

Table 7

Performance of different models on EDS dataset

Model	mAP 0.5	mAP 0.5:0.95	P	R
SSD	0.794	0.561	0.845	0.731
Faster R-CNN	0.819	0.618	0.853	0.751
YOLOv5m	0.781	0.559	0.836	0.706
YOLOv7	0.787	0.569	0.851	0.742
YOLOv8m	0.797	0.605	0.828	0.724
YOLOv5m-improved	0.823	0.620	0.859	0.753

For a more intuitive evaluation, this paper provides a direct comparison of the actual detection results achieved by YOLOv5m before and after the implementation of our improvement strategies. As depicted in Fig. 9, our strategies effectively address the challenges associated with error detection and missing detection, particularly in the case of small targets and occlusions within complex backgrounds. These visual results provide compelling evidence of the significant enhancements brought about by our strategies in real-world detection scenarios.



(a) Lnitial image (b) YOLOv5m (c) Improved YOLOv5

Fig. 9. Comparison of detection effect

4. Conclusions

Traditional manual security detection is not only slow but also less accurate. To address these issues, a parallel convolutional module, RCES, is proposed in this paper, utilizing a hybrid attention mechanism. Built upon YOLOv5m as a baseline model, the design aims to improve the network's detection performance in complex scenes. To enhance the network's performance in classification and regression tasks, a decoupled detection head tailored to our design is introduced, surpassing the effectiveness of the original decoupled detection head. The incorporation of a hybrid data enhancement strategy and an anchor frame adaptive matching strategy significantly contributes to the network's robustness. When compared to other mainstream target detection models, the method presented in this paper demonstrates notable advantages across various aspects.

Acknowledgements

Lhasa science and technology plan project under Grant LSKJ202405

REFERENCES

[1]. *A. Krizhevsky, I. Sutskever, G.E. Hinton*, ImageNet classification with deep convolutional neural networks, Commun. ACM, 60, 2017, 84–90.

[2]. *R.B. Girshick, J. Donahue, T. Darrell, J. Malik*, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2013, 580-587.

[3]. *S. Ren, K. He, R. Girshick and J. Sun*, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, Doi: 10.1109/TPAMI.2016.2577031.

[4]. *K. He, X. Zhang, S. Ren, J. Sun*, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 770-778.

[5]. *J. Redmon, S.K. Divvala, R.B. Girshick, A. Farhadi*, You Only Look Once: Unified, Real-Time Object Detection, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, 779-788.

[6]. *W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.E. Reed, C.-Y. Fu, A.C. Berg*, SSD: Single Shot MultiBox Detector, European Conference on Computer Vision, 2015.

[7] *Li, Chaofeng, Wang, Baoping*, A YOLOv4 Model with FPN for Service Plates Detection, Journal of Electrical Engineering & Technology, DOI 10.1007/s42835-021-00993-1(2022).

[8] *L. Han, F. Li, H. Yu, K. Xia, Q. Xin, X. Zou*, BiRPN-YOLOvX: A weighted bidirectional recursive feature pyramid algorithm for lung nodule detection, Journal of X-Ray Science and Technology, DOI 10.3233/xst-221310(2023).

[9] *S. Hussein, A.-A. Sadam, L. Hamzah, A. Motaz*, Contrastive-based YOLOv7 for personal protective equipment detection, Neural Computing and Applications, DOI 10.1007/s00521-023-09212-6(2023).

[10]. *K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R.S. Zemel, Y. Bengio*, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, International Conference on Machine Learning, 2015.

[11]. *R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra*, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision, vol. 128, 2020, 336-359.

[12]. *F. Sun, X. Zhang, Y. Liu, H. Jiang*, Multi-Object Detection in Security Screening Scene Based on Convolutional Neural Network, Sensors, vol. 22, 2022, 7836.

[13]. *Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, Y.R. Fu*, Rethinking Classification and Localization for Object Detection, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 10183-10192.

[14]. *G. Song, Y. Liu, X. Wang*, Revisiting the Sibling Head in Object Detector, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, 11560-11569.

[15]. *Y. Huang, X. Fu and Y. Zeng*, Anchor-Free Weapon Detection for X-Ray Baggage Security Images, in IEEE Access, vol. 10, pp. 97843-97855, 2022, Doi: 10.1109/ACCESS.2022.3205593.

[16] *B. Jae Soon, Y. In Young, C. Jun Won*, DBN-Mix: Training dual branch network using bilateral mixup augmentation for long-tailed visual recognition, Pattern Recognition, DOI 10.1016/j.patcog. 2023. 110107(2023).

[17]. *E. R. Q. Fernandes, A. C. P. L. F. de Carvalho and X. Yao*, Ensemble of Classifiers Based on Multiobjective Genetic Sampling for Imbalanced Data, in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1104-1115, 1 June 2020, Doi: 10.1109/TKDE.2019.2898861.

[18] *Jocher, G*. (2020). YOLOv5 by Ultralytics (Version 7.0) [Computer software]. https://doi.org/10.5281/zenodo.3908559

[19]. *Q. Wang, B. Wu, P.F. Zhu, P. Li, W. Zuo, Q. Hu*, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, 11531-11539.

[20]. *J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu*, "Squeeze-and-Excitation Networks", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011-2023, 1 Aug. 2020, Doi: 10.1109/TPAMI.2019.2913372.

[21]. *R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, X. Liu*, Towards Real-world X-ray Security Inspection: A High-Quality Benchmark and Lateral Inhibition Module For Prohibited Items Detection, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 10903-10912.

[22] *Y. Liu, L. Tang, C. Liao, C. Zhang, Y. Guo, Y. Xia, Y. Zhang, S. Yao*, Optimized Dropkey-Based Grad-CAM: Toward Accurate Image Feature Localization, Sensors, DOI 10.3390/s23208351(2023).