International Journal of Robotics and Control Systems
Vol. 5, No. 3, 2025, pp. 1770-1789
ISSN 2775-2658
http://pubs2.ascee.org/index.php/ijrcs

# ESI-YOLO: Enhancing YOLOv8 with Efficient Multi-Scale Attention and Wise-IoU for X-Ray Security Inspection

Arinal Haq [a,1], Nanik Suciati [a,2,*], Ngoc Dung Bui [b,3]

[a] Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[b] Faculty of Information Technology, University of Transport and Communications, Hanoi 100000, Vietnam
[1] ananghaq99@gmail.com; [2] nanik@if.its.ac.id; [3] dnbui@utc.edu.vn
* Corresponding Author

ARTICLE INFO

**Article history**
Received June 02, 2025
Revised July 03, 2025
Accepted July 25, 2025

**Keywords**
YOLOv8;
X-Ray Security Inspection;
Efficient Multi-Scale Attention (EMA);
Wise-IoU;
Attention Mechanism

ABSTRACT

Security inspection is a priority for preventing threats and criminal activities in public places. X-ray imaging can help with the closed luggages checking process. However, interpreting X-ray images is challenging due to the complexity and diversity of prohibited items. This paper proposes ESI-YOLO, an enhanced YOLOv8-based model for prohibited item detection in X-ray security inspection. The model integrates Efficient Multi-Scale Attention (EMA) and Wise-IoU (WIoU) loss function to improve multi-scale feature representation and detection accuracy. EMA improves multi-scale feature representation, while WIoU enhances bounding box regression, particularly in cluttered and overlapping scenarios. Comprehensive experiments on the CLCXray and PIDray datasets validate the effectiveness of ESI-YOLO. A systematic exploration for the optimal placement of EMA integration on YOLOv8 architecture reveals that the scenario with direct integration in both backbone and neck sections emerges as the most effective configuration without introducing significant computational complexity. Ablation experiments demonstrate the synergistic effect of combining EMA and WIoU in ESI-YOLO, outperforming individual component additions. ESI-YOLO demonstrates notable advancements over the baseline YOLOv8 model, achieving mAP50 improvements of 0.9% on CLCXray and 3.5% on the challenging hidden subset of PIDray, with a computational cost of 8.4 GFLOPs. Compared to other nano-sized models, ESI-YOLO exhibits enhanced accuracy while maintaining computational efficiency, making it a promising solution for practical X-ray security inspection systems.



## 1. Introduction

Security inspection is a top priority for preventing threats and criminal activities in public places, such as airports, train stations, and government buildings [1], [2]. One of the most common methods used to ensure security in public spaces is the inspection of luggage through X-ray imaging. [3], [4]. These inspections allow security personnel to detect contraband items, including weapons, explosives, and other hazardous materials hidden in passenger luggage [5]-[7]. X-ray imaging facilitates the examination of items inside closed bags or suitcases without the need for manual unpacking [8], [9]. However, X-ray screening still poses several challenges. The resulting X-ray images are often complex and require accurate interpretation by trained personnel. Furthermore,

because hundreds or thousands of items must be scanned daily, X-ray security screening must be both fast and efficient. Manual screening, even when assisted by X-ray images, is often considered less effective and time-consuming. Therefore, the increasing pace and scale of security operations highlight the need for automation to support the screening process.

Deep learning-based object detection methods offer promising solutions to address the limitations of traditional X-ray security inspection. They offer powerful capabilities for feature extraction and pattern recognition. These models can make X-ray security inspection faster and more accurate [6], [10]. Several studies have been conducted to develop detection models for security inspection on X-ray images [11]-[16]. Several challenges must be considered when developing deep learning models, such as overlapping items in bags and the diversity of sizes, shapes, and variations in prohibited items. Objects in X-ray images have low contrast, textures, and edges that are less prominent than those in natural images [17]-[20], which can negatively impact detection accuracy [17], [19], [21]. Along with technological advancements, several studies have adopted large-scale deep neural networks and transformer-based object detection models to improve detection accuracy [8], [15], [18], [22]-[24]. These models have proven capable of recognizing complex patterns and subtle features in visual data, such as X-ray images. However, despite this approach is promising in terms of performance, significant challenges may arise due to the larger computational requirements. This poses limitations in terms of accessibility and scalability, particularly in resource-constrained environments.

In practical security inspection scenarios, detection models that strike a balance between accuracy, speed, and computational efficiency are essential. The YOLO-based model is advantageous in this context because it offers a compact design and rapid detection capabilities while ensuring reliable performance [25]-[30]. The efficiency of the YOLO architecture is attributed to its single-stage detection approach, which analyzes the entire image in a single forward pass [31]. This enables real-time object detection, making it ideal for scenarios with limited computational power or those that demand quick processing [32]-[37]. Several studies have adopted YOLO-based models to detect prohibited items in X-ray images [11], [12], [14], [38]-[40]. Ren et al. [12] proposed a lightweight object detection model based on the YOLOv4 architecture called LightRay. The proposed model achieved good performance with a small light size. Gan et al. [11] proposed the YOLO-CID method based on the YOLOv7 architecture. The experimental results obtained from the PIDray dataset indicate that modifications to the baseline, such as modifying the backbone and incorporating attention mechanisms, can improve the model performance by 4.9% compared with the base YOLOv7 model.

The Attention Mechanism (AM) is a method for diverting the attention of a deep-learning model to important parts while ignoring irrelevant parts [41]-[45]. In the context of computer vision, AM has proven to be effective in enhancing the model's ability to analyze and interpret complex visual data efficiently. Several AM modules have been proposed for computer vision tasks, including Squeeze-and-Excitation Networks (SENet) [46], Convolutional Block Attention Mechanism (CBAM) [47], and Coordinate Attention (CA) [48]. Previous studies have demonstrated that integrating attention mechanisms into object detection models can significantly improve their performance [49]-[51]. Recently, attention mechanisms have been incorporated into YOLO-based architectures to enhance the detection of prohibited items in X-ray images. For instance, Ren et al. [12] applied CBAM to YOLOv4 to improve the characteristics of small objects. Zhao et al. [7] introduced a Label-Aware mechanism aimed at improving the accuracy of detecting overlapping items. Although this approach enhances accuracy, it often necessitates more intricate network architectures and greater computational resources. Gan et al. [11] applied Shuffle Attention (SA) to the neck of the YOLOv7 architecture. S. Han, Jiang, and Wu [52] used CA on the backbone and neck of YOLOv5 to capture information and enable the model to identify targets more accurately.

Despite these advancements, the optimal placement of attention modules within YOLO-based architectures remains an open question. Based on related studies [52]-[57], various strategies for integrating AM modules at different locations within the architecture require further investigation to

determine their impact on the overall model performance. Although AMs have proven effective in enhancing object detection capabilities, there is a need for systematic exploration to determine the most effective placement strategies, particularly in the context of prohibited items detection. Moreover, challenges such as varying object scales and complex visual patterns in X-ray imagery highlight the need for further research and innovation.

In response to the issues, this paper proposes ESI-YOLO, an enhanced YOLOv8-based model, designed to improve multi-scale feature representation in X-ray security inspections. The model innovatively integrates Efficient Multi-Scale Attention (EMA) on optimal placement and the Wise-IoU (WIoU) loss function into the YOLOv8 architecture to enhance the detection of prohibited items. The integration of the EMA module with the YOLOv8 architecture aims to enhance its feature extraction capabilities by focusing more attention on significant features across different scales, thereby enhancing the model's detection performance. The contributions of this paper are as follows:

1) This study proposes ESI-YOLO, an enhanced YOLOv8-based model that innovatively integrates EMA on optimal placement and Wise-IoU loss function on the YOLOv8 architecture for prohibited items detection. To validate the effectiveness of the proposed ESI-YOLO model, comprehensive experiments were conducted using benchmark datasets. Experiments conducted on the CLCXray [7] and PIDray [6] datasets showed that the ESI-YOLO model achieved improved performance over the baseline model.
2) This study systematically explored various scenarios for integrating the Efficient Multi-Scale Attention (EMA) module into the YOLOv8 architecture, with a particular focus on the backbone and neck sections, to evaluate their impact on the detection performance. The direct integration scenario of the EMA module in both the backbone and neck sections of YOLOv8 emerged as the most effective configuration, consistently enhancing the overall performance.
3) To enhance the accuracy and robustness of bounding box regression in the presence of variations in the object scale and shape, this paper integrates the Wise-IoU (WIoU) loss function into the model development process. Its efficacy in managing complex backgrounds with overlapping challenges renders it suitable for detecting prohibited items. This loss function significantly improved the overall model performance.

## 2. Proposed Method

This paper proposes ESI-YOLO, an enhanced model based on YOLOv8, designed to improve the multi-scale feature representation in X-ray security inspections. The model innovatively integrates Efficient Multi-Scale Attention (EMA) on optimal placement and the Wise-IoU loss function into the YOLOv8 architecture to enhance the detection of prohibited items. The proposed method was derived from a series of research stages, including a literature review, data exploration, model design, model evaluation, and result analysis. The overall architecture of the proposed ESI-YOLO is shown in Fig. 1.

### 2.1. Baseline Method (YOLOv8)

YOLOv8 [58] is an iteration of the YOLO series developed by the same author as YOLOv5 [59]. Although it was not the latest iteration at the time this study was conducted, the model performance is still very good in recent studies and is still being optimized in various fields to date with good community support. YOLOv8 has an adjustable scaling factor. This implies that it can meet the requirements under different conditions. YOLOv8 incorporates cutting-edge backbone and neck designs that improve the feature extraction and object detection capabilities. In YOLOv8, the C2f module replaces the C3 module from the YOLOv5 architecture. The C2f module was developed by drawing inspiration from the C3 module of YOLOv5 and the ELAN concept [60]. With more residual connections, the C2f structure enables YOLOv8 to achieve a richer gradient flow while remaining lightweight. The SPP module from YOLOv5 is substituted with SPPF (Spatial Pyramid Pooling-Faster), which serves the same purpose of managing features at various scales but with

reduced computational demands. The neck architecture of YOLOv8 employs the PAN-FPN structure, which is derived from the PANet backbone network. The Detection Head in YOLOv8 utilizes the decoupling-head and anchor-free design from the YOLOv6 model [28].

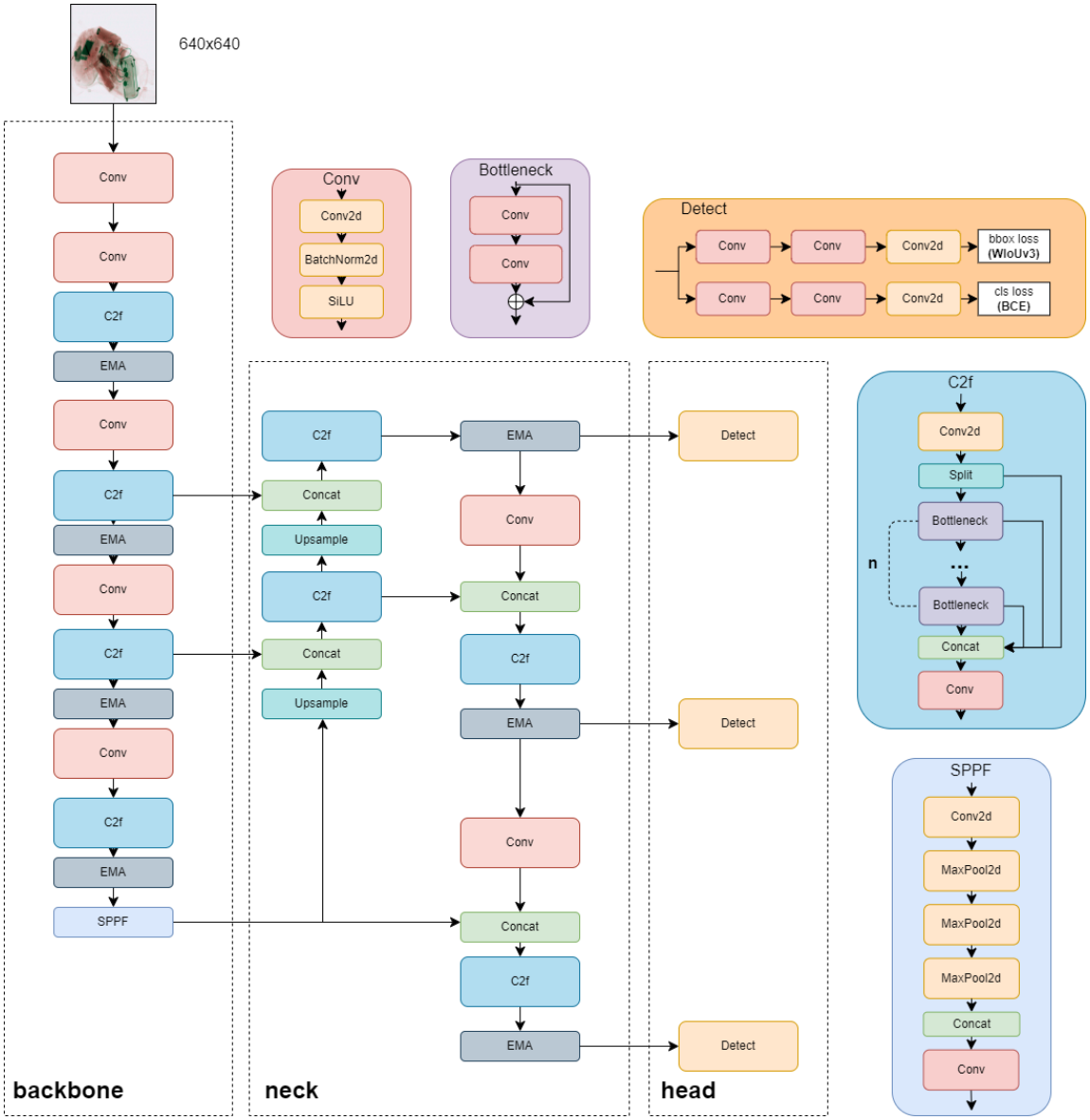



**Fig. 1.** ESI-YOLO overall architecture

In this study, the YOLOv8n (nano) variant was primarily employed as the baseline model owing to its advantages in computational efficiency and inference speed. YOLOv8n is crafted specifically for use in real-time scenarios on devices with limited resources, which is particularly pertinent to the context of detecting prohibited items during security inspections. In such scenarios, the system must process images swiftly and accurately without relying on high-powered hardware. Furthermore, the utilization of a lightweight model, such as YOLOv8n, facilitates the isolation and evaluation of the impact of integrating the EMA module on detection performance. Consequently, YOLOv8n serves as an ideal baseline for assessing the effectiveness of the EMA module placement in enhancing the accuracy of prohibited object detection.

## 2.2. Integration of Efficient Multi-Scale Attention

Efficient Multi-Scale Attention (EMA) [61] is an attention mechanism that focuses on important features at various scales with minimal computation. EMA can efficiently capture spatial and channel

information through a lightweight multi-scale mechanism. This capability is particularly crucial in the domain of prohibited item detection, where objects frequently exhibit variations in size, shape, and position. Unlike CBAM and CA that process channel and spatial separately, EMA integrates both in one stage, making it more efficient in the context of real-time object detection. EMA provides a better trade-off between accuracy and complexity than other attention modules. EMA uses parallel substructure in modules to avoid sequential processing and large network depth, preserving information while reducing computational cost. The overall structure of EMA module is shown in Fig. 2.

The main features of EMA include feature grouping, parallel subnetworks, and cross-spatial learning. The input features are divided into several channel groups to reduce complexity and enable parallel processing. EMA captures spatial context from two main directions, y using adaptive average pooling separately on the horizontal ($X$) and vertical ($Y$) dimensions as demonstrated in (1) and (2).

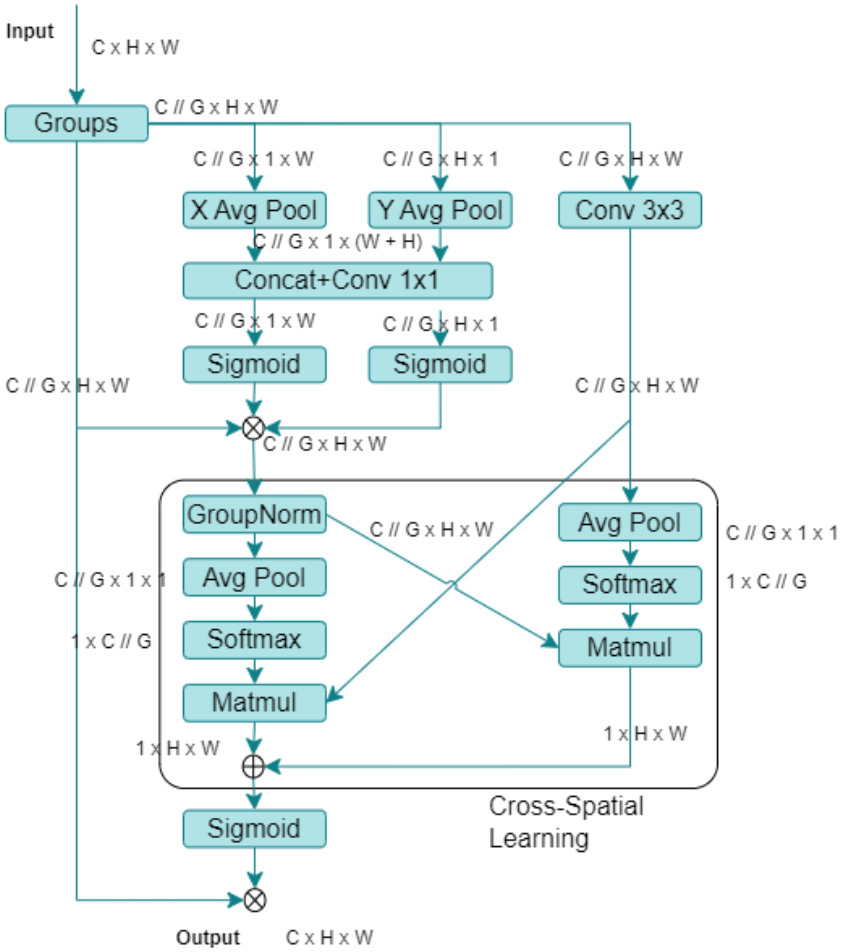



**Fig. 2.** EMA module architecture illustrating main feature such as feature grouping, parallel subnetworks, and cross-spatial learning

This allows the model to understand broader spatial structures without explicitly enlarging the receptive field.

$$Z_C^H = \frac{1}{W} \sum_{0 \le i \le W} x_C(H, i), \tag{1}$$

$$Z_C^W = \frac{1}{H} \sum_{0 \le j \le H} x_C(j, W), \tag{2}$$

where $C$ represents the count of input channels, while $H$ and $W$ refer to the spatial dimensions of the input features, respectively, and $x_C$ indicates the input feature. $Z_C^H$ represent the output of average pooling on $H$ (height) dimension for $C$ channel, $Z_C^W$ represent the output on average pooling on $W$ (width) dimension for $C$ channel.

The pooled features are concatenated and processed through 1×1 convolutions, then used as gating masks modulated with a sigmoid function. This gating selectively controls the flow of spatial information. While one path uses gating and normalization (GroupNorm), the other path uses 3×3 convolutions to preserve local information. These two paths are then combined through a global attention mechanism. EMA determines global attention weights to enhance the initial spatial features by utilizing 2D global adaptive average pooling with reconstruction, as illustrated in (3), along with the softmax normalization function. This process allows the model to adjust its focus on important areas in the spatial features. Through feature grouping and multi-scale structure, it builds short-term and long-term dependencies effectively.

$$Z_C = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} x_C(i,j) \tag{3}$$

where $C$ represents the count of input channels, while $H$ and $W$ refer to the spatial dimensions of the input features, respectively, and $x_C$ indicates the input feature. $Z_C$ represents the output of global average pooling for $C$ channel.

The experiments in this study are aimed at investigating the optimal placement of the EMA module within the YOLOv8 architecture for prohibited items detection. Strategic placement of the EMA module is expected to improve the quality of the features used for the final prediction, and to allow the model to focus more on important areas in the image, especially small or hidden objects. Several experimental scenarios for the module placement were considered in this study. In each scenario, the placement of the EMA module is evaluated by positioning it on the backbone, neck, or both sections of the YOLOv8 architecture. The scenarios explored in this paper are as follows;

(1) The EMA module is directly integrated into the YOLOv8 structure, either on the backbone, neck, or both. In the backbone section, The EMA module is placed after each C2f block to strengthen the feature representation in the early stages of extraction. In the neck section, The EMA module is placed before the transition to the head, aiming to enrich the multiscale features before the final prediction is performed.

(2) The integration of EMA internally into the C2f block, which is then referred to as C2f_EMA. The EMA module is placed before the last convolutional layer in the C2f block, resulting in more informative feature maps than the previous implementation. The original C2f block in YOLOv8 is replaced by C2f_EMA. The architectural design of C2f_EMA block is explained in Fig. 3 (a).

(3) The EMA module is integrated into the Bottleneck part of the C2f block, which then referred to as C2f_BtlEMA. This placement allows the processing of smaller features maps, thus reducing model complexity without sacrificing feature quality. The original C2f block is replaced by C2f_BtlEMA in the same position as in the second scenario. The changes in the bottleneck structure are shown in Fig. 3 (b), and the details of the C2f_BtlEMA block are shown in Fig. 3 (c).

### 2.3. Improve Loss Function (WIoU)

In this study, the Wise-IoU v3 (WIoU v3) [62] loss function is employed as part of an enhanced strategy to improve accuracy in the detection of prohibited items. This function replaces the default CIoU used in the YOLOv8 model, as CIoU has shown limitations in meeting the high demands for speed and accuracy in complex object detection scenarios. WIoU v3 introduces a dynamic gradient allocation mechanism that allows the model to focus more effectively on medium-quality samples,

which often reflect real-world conditions. WIoU introduces the Dynamic Non-Monotonic Focusing Mechanism (FM) approach. This mechanism aims to address the shortcomings of conventional loss functions that tend to produce zero gradients when there is no overlap between the prediction and ground truth. In addition, WIoU also takes into account the outlier degree of the anchor box to assess prediction quality, allowing for more prudent allocation of gradients. Additionally, this function enhances gradient distribution, enabling the model to detect objects at various scales more accurately and to accelerate convergence during training. By integrating non-monotonic behavior, WIoU v3 aids in minimizing overfitting to extreme samples while preserving the model's responsiveness to changes in object shape and size, a crucial feature for identifying prohibited items that are often hidden or have irregular structures. The WIOU loss function calculated using formula:

$$\mathcal{L}_{wiou} = rR_{wiou}\mathcal{L}_{iou}, \qquad r = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{4}$$

$$R_{wiou} = \left(\frac{\left(x - x_{gt}\right)^2 + \left(y - y_{gt}\right)^2}{\left(W_g^2 + H_g^2\right)^*}\right) \tag{5}$$

$$\mathcal{L}_{iou} = 1 - IOU \tag{6}$$

In the equations, $(x, y)$ represents the center coordinates of the anchor box, while $(x_{gt}, y_{gt})$ indicates the target box centroid coordinates. The dimensions of the smallest bounding box are denoted by $W_g$ and $H_g$. The factor of gradient enhancement $r$, which is influenced by the hyperparameters $\alpha$, $\delta$, and the nonmonotonic focusing factor $\beta$, is dynamically adjusted. The loss function is denoted by $\mathcal{L}_{iou}$ The hyperparameter $\alpha$ regulates the magnitude of the gradient applied to anchor boxes based on their quality. A higher $\alpha$ value results in a steeper gradient difference between anchor boxes of medium quality and those of high or low quality. The hyperparameter $\delta$ controls sensitivity to outliers by mitigating the impact of extremely poor-quality anchor boxes, thereby preventing the generation of harmful gradients. The hyperparameter $\beta$ defines the quality threshold in a non-monotonic focus mechanism, helping to identify anchor boxes that are considered medium quality to receive maximum gradient updates.

### 2.4. Experimental Dataset

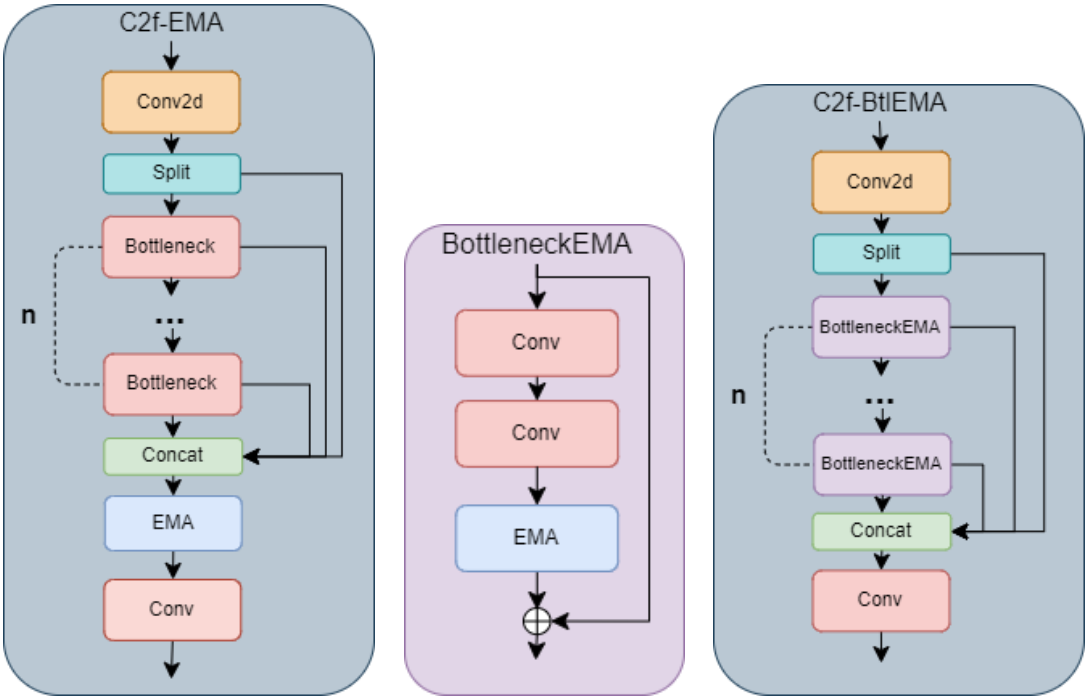
The main dataset used is the publicly available Cutters and Liquid Containers X-ray Dataset (CLCXray) [7]. The dataset contained 9059 X-ray images from real and simulated manual baggage scanning. The dataset includes 12 object categories with five types of cutters and seven types of liquid containers. This dataset has the largest amount of liquid container data compared to other security X-ray datasets. The data classes were imbalanced, particularly for the liquid container classes. The image resolution used in the experiment was 640×640. The dataset was divided into fixed training and testing sets with ratios of 80%:20%. To validate the best model, this study used the PIDray dataset [6]. The dataset covers various real-world cases of prohibited item detection, especially those involving deliberately hidden objects, thereby presenting significant challenges to conventional detection systems. PIDray consists of 124486 X-ray images covering 12 distinct categories of prohibited items. The dataset is one of the largest in the x-ray security images field and offers unique challenges related to intra-class variation, class imbalance, and occlusion issues. Both datasets were chosen because they contain prohibited items that are quite different from each other, with varying scenarios and complexity. For image preprocessing, each image in the dataset undergoes auto-orientation to ensure the image orientation matches its respective metadata, and the image is resized to stretch to 640×640.

### 2.5. Experimental Setup

In this paper, the hardware used in the experiment was a 4-Core CPU Processor with 30 GB RAM and an NVIDIA Tesla T4 GPU for training and testing in the model development experiments.

This paper uses Python 3.10 and PyTorch framework 2.3. In addition, in the training phase, each experiment did not use the pretrained YOLOv8 model. The decision to train the model from scratch in this study was based on technical considerations, particularly related to the modification of the YOLOv8 architecture. This modification directly affects the internal structure and learning mechanism of the model, so that the pre-trained weights of the standard architecture are no longer compatible or optimal for use. In addition, X-ray images have very different visual characteristics from conventional RGB images, such as object transparency and internal contrast, which are not covered in common datasets such as COCO. Therefore, training from scratch is necessary for the model to learn domain-specific features thoroughly and consistently.

The hyperparameter settings used for model training for each experiment were based on the YOLOv8 default settings. The number of epochs used is 100, with SGD optimizer and 0.01 learning rate. The momentum used was 0.937 with three warmup epochs. Employing these default configurations allows the research to maintain computational efficiency and time effectiveness, particularly in exploratory studies or under limited resource conditions. Moreover, the default settings provided by YOLOv8 have undergone extensive tuning and validation, demonstrating generally optimal performance across a wide range of datasets. Additionally, using default hyperparameters allows researchers to focus on the main aspects of the improvement that proposed in this study. This is important to ensure that experiments are compared fairly. As a baseline, default hyperparameters provide an initial representative overview of the model's capabilities.




(a) C2f_EMA (b) BottleneckEMA (c) C2f_BtlEMA

**Fig. 3.** EMA Module Integration for C2f_EMA and C2f_BtlEMA

## 2.6. Evaluation Metrics

The model evaluation used several metrics that are commonly used for object detection. In object detection, various metrics can measure how well a model detects objects in images. The model is measured by the precision of position and class of objects. Common evaluation metrics are precision, recall, and mean Average Precision (mAP). Besides performance metrics, measurements of model computation size, such as parameter count and FLOPs, are used for evaluation. The number of model parameters represents required memory resources. FLOPs represent the floating-point operations of the model, understood as calculation count. In deep learning, parameters and FLOPs are used to measure an algorithm's computational complexity.

In object detection, for a detection result to be considered correct, the model must accurately identify the object and its location. True Positive (TP) occurs when the model correctly predicts both the location and class according to the actual value and the object is still within the specified IoU threshold. False Positive (FP) occurs when the model correctly predicts the class according to the actual value, but the location falls below the IoU threshold, or the model predicts an object that does not exist in the actual values. False Negative (FN) occurs when the model fails to detect or does not detect an object that exists in the actual data. Precision, recall, and mAP are calculated using (7)-(10):

$$Precision = \frac{TP}{(TP + FP)} \tag{7}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{8}$$

$$AP_i = AP\ of\ class\ i \tag{9}$$

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \tag{10}$$

GFLOPs are calculated using (11):

$$GFLOPs = \frac{Floating\ Point\ per\ Operations}{10^9} \tag{11}$$

## 3. Results and Discussion

### 3.1. EMA Integration Experiments

The evaluation results for model development, particularly focusing on the optimal placement exploration of the EMA module within the YOLOv8 architecture, are presented in Table 1. The experiment adhered to the scenario outlined in the research method design. Notably, all scenarios demonstrated improved performance compared to the YOLOv8 base model across various performance metrics. Among the evaluation metrics utilized in the experimental dataset, the placement of the EMA module on both sections (EMA-both) of the YOLOv8 architecture exhibited the most promising results for mAP50 and mAP metrics, with an approximate 0.7% increase in mAP. However, this configuration also resulted in the largest parameters and GFLOPs across all scenarios, the increase in parameters and GFLOPs from the original model remains within acceptable limits. The YOLOv8 with integration of the C2f_BtlEMA block in both sections (C2f_BtlEMA-both), ranks as the second-best in performance while being lighter. This scenario delivers results similar to the EMA-both scenario but with fewer parameters and lower computational demands.

**Table 1.** EMA module placement experiments on CLCXray dataset

| Model | mAP50 | mAP | #param | GFLOPs |
|---|---|---|---|---|
| Base (YOLOv8n) | 81.6 | 69.5 | 3.013M | 8.21 |
| EMA-backbone | 82.1 | 69.9 | 3.014M | 8.29 |
| EMA-neck | 81.9 | 69.7 | 3.014M | 8.27 |
| EMA-both | 82.3 | 70.3 | 3.015M | 8.35 |
| C2f_EMA-backbone | 81.9 | 69.9 | 3.016M | 8.39 |
| C2f_EMA-neck | 81.9 | 69.9 | 3.015M | 8.33 |
| C2f_EMA-both | 82.0 | 69.8 | 3.017M | 8.52 |
| C2f_BtlEMA-backbone | 82.0 | 69.8 | 3.013M | 8.24 |
| C2f_BtlEMA-neck | 81.9 | 69.8 | 3.013M | 8.23 |
| C2f_BtlEMA-both | 82.1 | 70.2 | 3.014M | 8.26 |

To further evaluate the generalization capabilities of the two best scenario models, experiments were conducted utilizing multi-scale training by setting the multi-scale hyperparameter to True during model training. This approach was employed to assess the models' ability to manage the diversity and variation of objects within images, which presents a significant challenge in the detection of prohibited items. The evaluation outcomes of these multi-scale training experiments are presented in Table 2. From the results, the base model exhibits a significant decline from its previous default training outcomes, whereas the proposed model maintains a robust performance. The EMA scenario models demonstrate resilience in multi-scale training, suggesting an enhanced adaptability to varying object sizes and image variations. This improved generalization capability may be attributed to the EMA module's ability to capture and emphasize relevant features across different scales. The consistent performance of the EMA-both model under multi-scale conditions indicates its potential for real-world applications where object sizes may vary significantly.

**Table 2.** Evaluation results on multi-scale training settings on CLCXray dataset

| Model | mAP50 | mAP |
|---|---|---|
| Base (YOLOv8n) | 81.5 | 68.8 |
| EMA-both | 81.9 | 69.7 |
| C2f_BtlEMA-both | 81.8 | 69.6 |

To further validate the results obtained from the models, the two best scenario model from the experiment was evaluated on another prohibited items detection dataset, namely, PIDray. The evaluation results on the PIDray dataset, presented in Table 3. In the evaluation results on PIDray dataset, the EMA-both model produces better performance compared to the base model and C2f_BtlEMA-both model. The EMA-both model shows significant improvements over the base model, particularly in the mAP metrics for each test set. In hidden subset, the EMA-both model achieves an mAP50 of 59.7% and 44.9%. The EMA-both model can improve performance with an increase in mAP metrics of approximately 2.9% and mAP50 metrics of approximately 3% on hidden subset. The EMA-both model's enhanced significant performance in the hidden test set indicates its potential for addressing this challenge. The C2f_BtlEMA-both model performs comparably to the base model, with minor variations across test set.

**Table 3.** Evaluation results on PIDray dataset

| Model | Easy | | Hard | | Hidden | |
|---|---|---|---|---|---|---|
| | mAP50 | mAP | mAP50 | mAP | mAP50 | mAP |
| Base (YOLOv8n) | 81.1 | 68.5 | 83 | 65.3 | 56.6 | 42 |
| EMA-both | 81 | 69.3 | 83 | 65.6 | 59.7 | 44.9 |
| C2f_BtlEMA-both | 80.7 | 68.7 | 82.3 | 65.2 | 58.2 | 43.1 |

For further analysis, this experiment was compared with several other popular attention mechanisms for computer vision such as CBAM and CA. The data, architecture, and model training settings followed the same configurations used in the experimental settings. The results of the comparison of the experiments with several models with other AM modules are shown in Table 4. The placement of other AM modules follows the optimal placement that has been obtained previously, namely, by placing them in both parts of YOLOv8. The EMA module exhibited the best performance in the experiment for detecting prohibited items in this study. In terms of performance and number of parameters metrics, the EMA module yielded the best results in the experiment. Although, in terms of GFLOPs metrics, the CA metric has slightly better results, the configuration in the C2f_BtlEMA scenario has better results with slightly better performance. Based on the comprehensive evaluation results, the EMA-both scenario emerges as the optimal placement strategy in this study. The consistent performance enhancements observed across various datasets and training configurations indicate that the EMA, particularly EMA-both model effectively address the limitations of the baseline model without introducing significant computational complexity.

**Table 4.** Comparison with other AM module in optimal placement on CLCXray dataset

| Model | mAP50 | mAP | #Param | GFLOPs |
|---|---|---|---|---|
| Base | 81.6 | 69.5 | 3.01M | 8.21 |
| + CBAM | 81.9 | 69.7 | 3.19M | 8.34 |
| + CA | 82 | 70.2 | 3.04M | 8.27 |
| + EMA | 82.3 | 70.3 | 3.01M | 8.35 |
| + C2f_BtlEMA | 82.2 | 70.2 | 3.01M | 8.26 |

### 3.2. WIoU Loss Integration Experiments

Based on the evaluation results on the CLCXray dataset shown in Table 5, the model using the WIoU method showed the best performance compared to other IoU loss variants. The model using Wise IoU v3 (WIoU) [62] showed the best performance with an mAP50 value of 82.2% and an overall mAP of 69.7%. Compared to the baseline CIoU which obtained an mAP50 of 81.6% and PIoU [63] with an mAP50 of 82%, WIoU provided consistent improvements in both the high precision metric (mAP50) and the average overall precision (mAP). These results indicate that WIoU is able to produce more accurate and stable bounding box predictions, which are very important in detecting suspicious objects in X-ray images. Thus, the integration of WIoU in the YOLOv8-based Prohibited Items Detection system in this study has the potential to increase the effectiveness and reliability in identifying the items, making it a notable choice for security applications. These advantages confirm that the integration of WIoU in the detection system can improve the accuracy of suspicious object identification, thereby contributing significantly to the effectiveness of the developed security system.

**Table 5.** Comparison with other IoU loss on CLCXray dataset

| Model | mAP50 | mAP75 | mAP |
|---|---|---|---|
| CIoU (base) | 81.6 | 78.6 | 69.5 |
| PIoU | 82 | 78.8 | 69.7 |
| WIoU | 82.2 | 79.2 | 69.7 |

### 3.3. Ablation Experiments

The ablation experiments for the proposed ESI-YOLO model were conducted using the CLCXray and PIDray datasets. These experiments aimed to demonstrate the impact of various components on the performance of the proposed ESI-YOLO model. The study examined the effects of integrating EMA and WIoU loss separately into the baseline model. The experimental results, as presented in Table 6, illustrate the performance for the CLCXray dataset. Incorporating EMA led to improvements in both mAP50 and mAP, achieving values of 82.3% and 70.3%, respectively, with a slight increase in parameters and GFLOPs. The integration of WIoU alone enhanced mAP50 to 82.2%, although it showed only a slight increase in mAP compared to the baseline. The final proposed model, ESI-YOLO, which combines both EMA and WIoU, achieved the highest mAP50 of 82.5%, increase 0.9% approximately to the baseline, and the highest mAP75 with a score of 79.5%. Although the ESI-YOLO model shows improvement in the evaluation metrics compared to the baseline, the mAP value decreases slightly compared to the model with only EMA. This indicates that the model is better at detecting objects with a high threshold (50, 75), but the overall decrease in mAP may indicate a slight reduction from EMA only model in the model's ability in higher threshold within the CLCXray dataset.

Based on the ablation experiment results presented in Table 7, the findings on the PIDray dataset highlight the efficacy of the proposed components in ESI-YOLO. The integration of EMA into the baseline YOLOv8n model resulted in enhanced performance across all subsets, with significant improvements observed in the challenging hidden subset (mAP50 increased from 56.6% to 59.7%, mAP from 42% to 44.9%). The incorporation of WIoU loss alone yielded modest enhancements, particularly in the easy subset. The comprehensive ESI-YOLO model, which combines both EMA and WIoU, achieved the most favorable overall outcomes, with mAP50/mAP scores of 82.6%/69.8% on the easy subset, 84.1%/66.1% on the hard subset, and 60.1%/44.8% on the hidden subset. These

results represent substantial improvements over the baseline, especially in the hidden subset, where mAP50 increased by 3.5% and mAP by 2.8%. The findings demonstrate that the integration of EMA and WIoU significantly enhances the model's capability to detect prohibited items across varying levels of difficulty, with particular efficacy in identifying challenging hidden objects.

**Table 6.** Ablation experiment results on CLCXray dataset

| Model | EMA | WIoU | mAP50 | mAP75 | mAP | #Param | GFLOPs |
|---|---|---|---|---|---|---|---|
| YOLOv8n | | | 81.6 | 78.6 | 69.5 | 3.013M | 8.21 |
| +EMA | ✓ | | 82.3 | 79.3 | 70.3 | 3.014M | 8.35 |
| +WIoU | | ✓ | 82.2 | 79.2 | 69.7 | 3.013M | 8.21 |
| ESI-YOLO | ✓ | ✓ | 82.5 | 79.5 | 70.1 | 3.014M | 8.35 |

**Table 7.** Ablation experiment results on PIDray dataset

| Model | EMA | WIoU | Easy | | Hard | | Hidden | |
|---|---|---|---|---|---|---|---|---|
| | | | mAP50 | mAP | mAP50 | mAP | mAP50 | mAP |
| YOLOv8n | | | 81.1 | 68.5 | 83 | 65.3 | 56.6 | 42 |
| +EMA | ✓ | | 81 | 69.3 | 83 | 65.6 | 59.7 | 44.9 |
| +WIoU | | ✓ | 81.9 | 68.2 | 83 | 64.6 | 57.3 | 41.7 |
| ESI-YOLO | ✓ | ✓ | 82.6 | 69.8 | 84.1 | 66.1 | 60.1 | 44.8 |

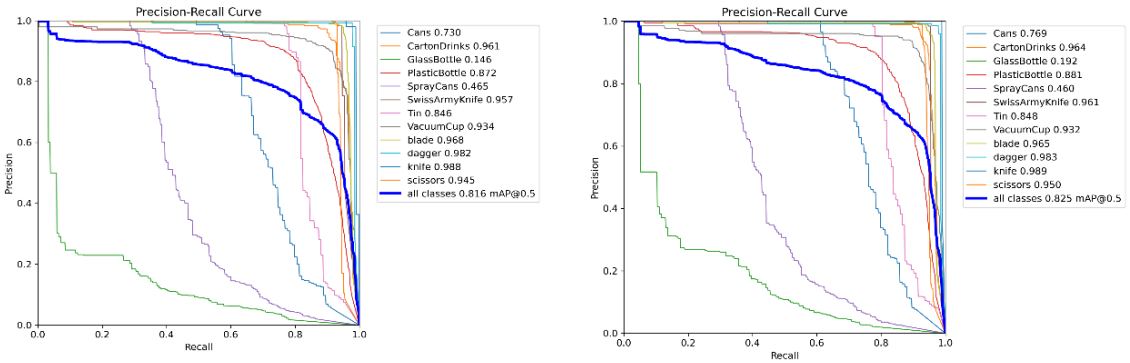
The ablation study highlights the synergistic effect of combining EMA and WIoU in proposed ESI-YOLO model, as the full model outperforms individual component additions. This comprehensive approach enables more robust detection across diverse scenarios, addressing the complexities of prohibited item identification in X-ray security screening. The substantial improvements on the hidden subset underscore ESI-YOLO potential to enhance security screening processes, particularly for concealed or obscured objects that pose significant challenges in real-world applications.

### 3.4. Visualization and Results Analysis

A comparison of the precision-recall curves of the baseline model and the ESI-YOLO model on the CLCXray and PIDray datasets is shown in Fig. 4 and Fig. 5. Based on the analysis performed on the Precision-Recall curves shown, it can be seen that the ESI-YOLO model shows superior performance compared to YOLOv8n as the baseline, especially in terms of object detection accuracy. This superiority is evident in the evaluation results on the CLCXray and PIDray (hidden test set) datasets. On the CLCXray dataset, the ESI-YOLO model has an mAP50 value of 82.5%, which is higher than that of YOLOv8n which only reaches 81.6%. This difference shows that ESI-YOLO is able to identify and classify objects more accurately than the baseline. In addition, the Precision-Recall distribution also shows better consistency across different object categories, especially on items that have high complexity in their identification. The superiority of ESI-YOLO was further strengthened when tested on the PIDray dataset (hidden test set), where this model recorded an mAP50 value of 60.1%, higher than YOLOv8n which only reached 56.6%. This difference indicates that ESI-YOLO has better generalization in detecting objects in more complex and varied environments. In addition, the analysis of specific objects such as guns, knives, and scissors shows that ESI-YOLO provides more precise detection, with a lower error rate compared to the baseline. This is due to the architectural optimization factor in ESI-YOLO with the integration of EMA and Wise-IoU that allows for improved model performance.

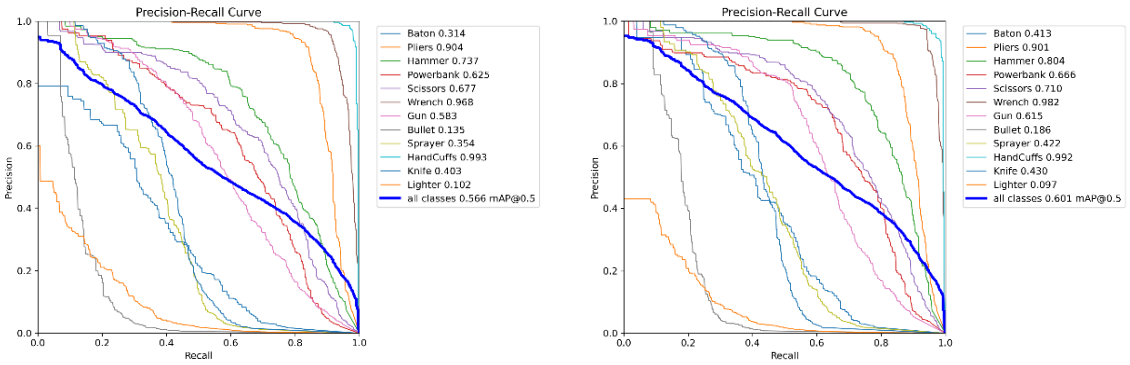
From the sample prediction results of the models shown in Fig. 6, ESI-YOLO consistently provides a higher confidence score for correctly detected ground-truth objects, indicating better detection accuracy and more reliability in identifying various objects. This not only ensures that each item examined is correctly identified but also reduces the risk of errors that could occur. In addition, ESI-YOLO is able to recognize more objects in a single overlapping or hidden image more accurately, allowing it to handle complex scenarios that often arise in security or baggage screening applications. The consistency exhibited by this model in various scenarios shows the advantage of a

more robust architecture in understanding object characteristics compared to the baseline model. Despite the model enhancing the baseline model performance, the model still encounters challenges in accurately detecting entire objects that exhibit a significant degree of overlap, as illustrated in the first sample in Fig. 6. Furthermore, in the last sample, a single object is erroneously identified as two distinct objects, although the confidence level for the correct object remains higher. However, comparison with ground truth shows ESI-YOLO results are more accurate and match actual labels, strengthening the claim that this model is optimal for object detection in challenging environments. With these advantages, ESI-YOLO becomes more efficient and reliable for improving deep-learning based prohibited items detection systems. However, although the model performs better, there are cases where object location is identified but the item type is misclassified. This may be due to subtle differences between item types in prohibited items detection.




(a) YOLOv8n (b) ESI-YOLO

**Fig. 4.** Comparison of Precision-Recall Curve on CLCXray




(a) YOLOv8n (b) ESI-YOLO

**Fig. 5.** Comparison of Precision-Recall Curve on PIDray (hidden test set)

### 3.5. Comparative Experiments with Other Models

Comparative results of various object detection models applied to the CLCXray dataset presented on Table 8. ESI-YOLO distinguishes itself among the models in several key findings. It achieves the highest mAP50 (82.5%) and mAP (70.1%) scores among the nano-sized models, surpassing other nano-sized YOLO models. It maintains an optimal balance between accuracy and efficiency, with only 3M parameters and 8.4 GFLOPs, comparable to YOLOv8n. In the category of small-sized models, ESI-YOLOs exhibits superior performance with 83.3% mAP50 and 71.8% mAP, outperforming YOLOv5s, YOLOv8s, and YOLO11s. When compared to larger two-stage detectors such as Faster R-CNN and RetinaNet, ESI-YOLO achieves significantly higher accuracy with considerably fewer parameters and reduced computational cost. Among the YOLO variants, ESI-YOLO consistently enhances the baseline YOLOv8 model in both nano and small sizes. The nano version of ESI-YOLO is particularly noteworthy, as it surpasses even some larger models while maintaining a highly lightweight architecture. In summary, ESI-YOLO excels in detection

performance for X-ray security inspection tasks while maintaining computational efficiency. Its ability to deliver high accuracy with a compact model makes it ideal for real-time security screening applications, where both speed and accuracy are crucial.

**Table 8.** Comparison with other models on CLCXray dataset

| Model | mAP50 | mAP | #Param | GFLOPs |
|---|---|---|---|---|
| Faster R-CNN [64] | 80.2 | 64.5 | 43.3M | 173.3 |
| RetinaNet [65] | 76.7 | 59.3 | 36.6M | 146.3 |
| YOLOv5n [59] | 81.4 | 68.1 | 2.5M | 7.2 |
| YOLOv6n [66] | 80.8 | 68.9 | 4.2M | 11.9 |
| YOLOv8n [58] | 81.6 | 69.5 | 3M | 8.2 |
| YOLO11n [67] | 81.4 | 69.1 | 2.6M | 6.5 |
| ESI-YOLO (n) | 82.5 | 70.1 | 3M | 8.4 |
| YOLOv5s [59] | 82.8 | 70.7 | 9.1M | 24.1 |
| YOLOv8s [58] | 82.5 | 71.5 | 11.1M | 28.7 |
| YOLO11s [67] | 82.4 | 71.4 | 9.4M | 21.6 |
| ESI-YOLO (s) | 83.3 | 71.8 | 11.1M | 29.2 |

### 3.6. Discussion

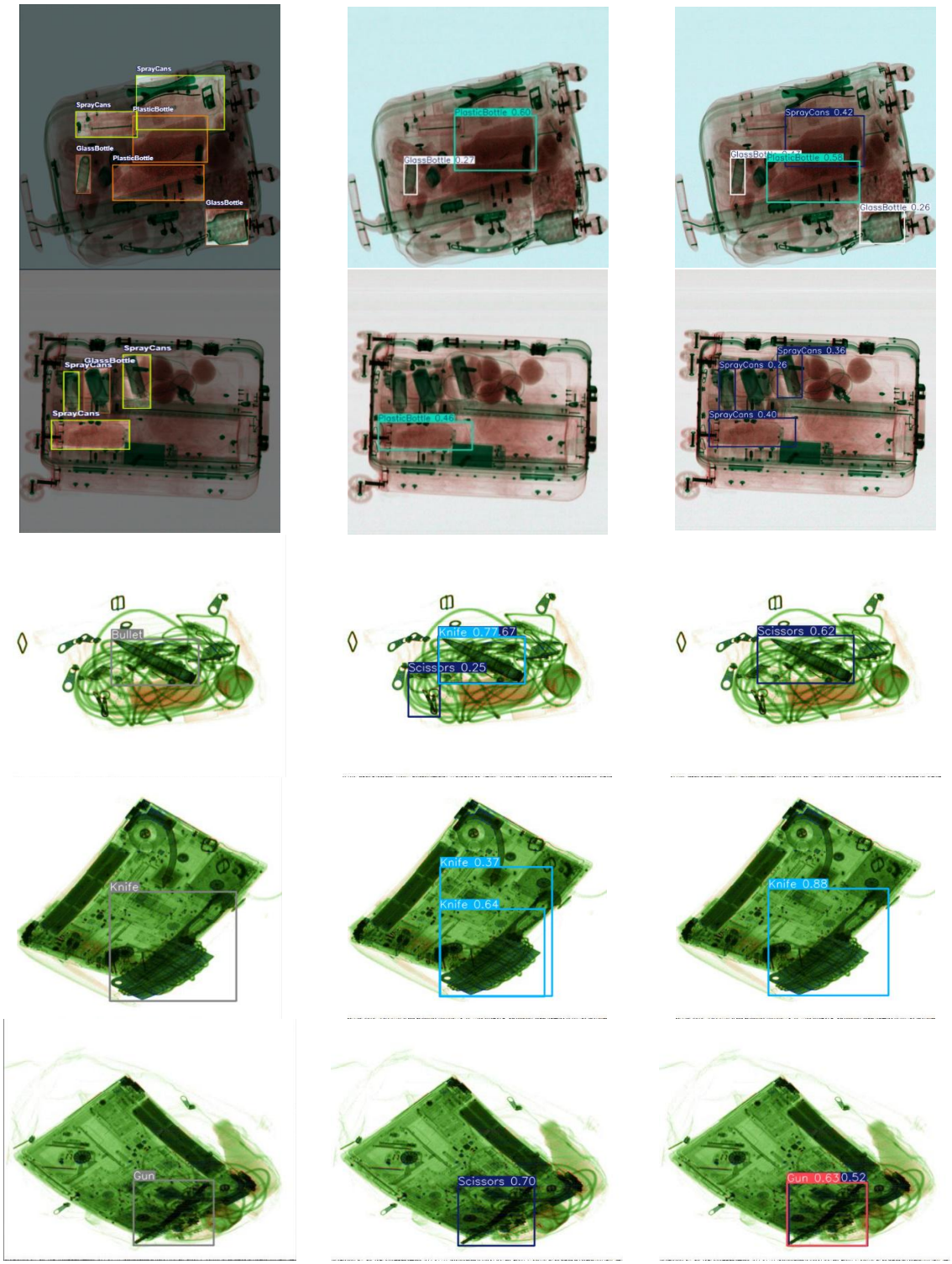
This paper proposes ESI-YOLO, an enhanced YOLOv8-based model designed to improve multi-scale feature representation in X-ray security inspections. The ESI-YOLO model consistently demonstrates superior performance compared to the baseline YOLOv8 model in detecting prohibited objects within the PIDray and CLCXray datasets. Evaluation results show that ESI-YOLO achieves mAP50 improvements of 0.9% on the CLCXray dataset and 3.5% on the PIDray hidden subset, respectively. This performance gain is primarily attributed to the integration of the EMA module, which adaptively enhances focus on critical features of small or obstructed objects. Simultaneously, the use of WIoU as a loss function improves consistency in object localization, resulting in more precise bounding boxes.

Systematic exploration reveals that direct integration of EMA into both the backbone and neck sections (EMA-both) of YOLOv8 yields the most effective results. These findings are supported by consistently strong performance across experimental trials. The EMA-both configuration delivers the best outcomes compared to other scenarios tested, without introducing significant computational complexity. By applying EMA to both the backbone and neck, the model effectively balances the extraction of low-level features with the synthesis of high-level semantic information, leading to improved detection accuracy across various scales. Consequently, the model becomes more adept at identifying prohibited items, which often pose challenges due to their diverse sizes, shapes, and contextual appearances within X-ray images.

Results analysis reveal that the combination of the EMA module and WIoU in ESI-YOLO significantly contributes to overcoming detection challenges in X-ray images, where objects frequently appear obscure or overlap. ESI-YOLO demonstrates the capability to accurately recognize more objects in a single image, even when they are overlapping or obscured, thereby enabling it to manage complex scenarios commonly encountered in security or baggage inspection applications. The model's consistency across various scenarios underscores the robustness of its architecture in understanding object characteristics compared to the baseline model. This is a positive indication of EMA integration, allowing the model to effectively concentrate attention on detected objects. The integration of WIoU further enhances detection outcomes, as evidenced by improved object localization over the baseline model. Comparisons with the ground truth indicate that ESI-YOLO's results are more accurate and aligned with the actual labels, reinforcing the assertion that this model is optimal for object detection in challenging environments.

Despite ESI-YOLO superior performance compared to the baseline model, several limitations warrant consideration. First, the model continues to encounter challenges in detecting objects at high thresholds, resulting in a suboptimal mAP metric, particularly in CLCXray images. This decline can be attributed to the difficulty in maintaining high precision at extreme IoU levels (above 0.85), which

is potentially expected as limitation of the WIoU loss function. Wise-IoU is a variant of IoU-based loss that seeks to balance the penalty for poor predictions and the reward for good predictions. The goal is to make the model more stable and accurate in bounding box regression. This trend is observed across both datasets, the increase in mAP50 and mAP75 still indicates relevant practical improvements for real-world applications, while the decrease in mAP reflects a trade-off in the loss function design that prioritizes generalization and prediction stability.




(a) Ground Truth (b) YOLOv8n (c) ESI-YOLO

**Fig. 6.** Sample of model prediction results

Second, although the use of EMA enhances focus on spatial features that lead to improve model performance, it increases the model's computational load compared to the baseline model. This could potentially reduce inference speed in real-time scenarios. However, the increase remains relatively small and manageable, especially when compared to the YOLOv8n baseline. Third, the quality of images within the datasets plays a crucial role in model performance. Class imbalance remains a persistent issue, with certain object categories appearing more frequently than others. This results in bias during model training and affects generalization to unseen data. Furthermore, minor differences between object classes in the images continue to pose challenges. Although data augmentation techniques from YOLOv8 have been applied to mitigate this issue, the improvements achieved remain limited.

## 4. Conclusion

This paper presents ESI-YOLO, an enhanced YOLOv8-based model designed to improve multi-scale feature representation for X-ray security inspections. ESI-YOLO integrates Efficient Multi-Scale Attention (EMA) at optimal placements and incorporates the Wise-IoU (WIoU) loss function into the YOLOv8 architecture to enhance the prohibited items detection. Experimental results on the CLCXray and PIDray datasets demonstrate mAP50 improvements of 0.9% and 3.5% (on the hidden subset), respectively, over the baseline YOLOv8 model. A systematic exploration of EMA integration scenarios reveals that applying EMA to both the backbone and neck sections yields the most effective performance. This configuration consistently enhances detection accuracy across various settings without introducing significant computational overhead. The integration of the WIoU loss function further improves bounding box regression accuracy and robustness. Ablation studies and comparative evaluations confirm the effectiveness of ESI-YOLO in balancing detection accuracy and computational efficiency, particularly among lightweight (nano-sized) models. The proposed model offers a promising solution for real-time X-ray security inspection systems, enabling more accurate detection of prohibited items while maintaining operational efficiency. Despite its advantages, this study is limited by the scope of the datasets used and the reliance on convolutional YOLOv8-based architectures. Future work may explore alternative model architectures, such as Transformer-based designs, or further optimize the model for edge deployment. Additionally, expanding the dataset to include a wider range of object categories, imaging conditions, and more challenging scenarios would enhance the model's generalizability and robustness in real-world applications.

**Author Contribution:** All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

**Acknowledgment:** The authors would like to thank the Lembaga Pengelola dana Pendidikan (LPDP), Indonesia, and Department of Informatics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia, for supporting this research.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

[1] D. Li, X. Hu, H. Zhang, and J. Yang, "A GAN based method for multiple prohibited items synthesis of X-ray security image," *Optoelectronics Letters*, vol. 17, no. 2, pp. 112-117, 2021, https://doi.org/10.1007/s11801-021-0032-7.

[2] J. Wu, X. Xu and J. Yang, "Object Detection and X-Ray Security Imaging: A Survey," *IEEE Access*, vol. 11, pp. 45416-45441, 2023, https://doi.org/10.1109/ACCESS.2023.3273736.

[3] M. Chouai, M. Merah, and M. Mimi, "CH-Net: Deep adversarial autoencoders for semantic segmentation in X-ray images of cabin baggage screening at airports," *Journal of Transportation Security*, vol. 13, no. 1-2, pp. 71-89, 2020, https://doi.org/10.1007/s12198-020-00211-5.

[4] X. Ji *et al*., "Filtered selective search and evenly distributed convolutional neural networks for casting defects recognition," *Journal of Materials Processing Technology*, vol. 292, p. 117064, 2021, https://doi.org/10.1016/j.jmatprotec.2021.117064.

[5] C. Miao *et al*., "SIXray: A Large-Scale Security Inspection X-Ray Benchmark for Prohibited Item Discovery in Overlapping Images," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2114-2123, 2019, https://doi.org/10.1109/CVPR.2019.00222.

[6] L. Zhang, L. Jiang, R. Ji, and H. Fan, "PIDray: A Large-Scale X-ray Benchmark for Real-World Prohibited Item Detection," *International Journal of Computer Vision*, vol. 131, no. 12, pp. 3170-3192, 2023, https://doi.org/10.1007/s11263-023-01855-1.

[7] C. Zhao, L. Zhu, S. Dou, W. Deng and L. Wang, "Detecting Overlapped Objects in X-Ray Security Imagery by a Label-Aware Mechanism," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 998-1009, 2022, https://doi.org/10.1109/TIFS.2022.3154287.

[8] T. Viriyasaranon, S.-H. Chae, and J.-H. Choi, "MFA-net: Object detection for complex X-ray cargo and baggage security imagery," *PLOS ONE*, vol. 17, no. 9, p. e0272961, 2022, https://doi.org/10.1371/journal.pone.0272961.

[9] M. Berger, Q. Yang, and A. Maier, "X-ray Imaging," *Medical Imaging Systems*, pp. 119-145, 2018, https://doi.org/10.1007/978-3-319-96520-8_7.

[10] X. Pei, C. Ma, J. Zhou, J. Yang, and Y. Xu, "Contraband detection algorithm for X-ray security inspection images based on global semantic enhancement," *IET Image Processing*, vol. 18, no. 13, pp. 4356-4367, 2024, https://doi.org/10.1049/ipr2.13256.

[11] N. Gan *et al*., "YOLO-CID: Improved YOLOv7 for X-ray Contraband Image Detection," *Electronics*, vol. 12, no. 17, p. 3636, 2023, https://doi.org/10.3390/electronics12173636.

[12] Y. Ren, H. Zhang, H. Sun, G. Ma, J. Ren, and J. Yang, "LightRay: Lightweight network for prohibited items detection in X-ray images during security inspection," *Computers & Electrical Engineering*, vol. 103, p. 108283, 2022, https://doi.org/10.1016/j.compeleceng.2022.108283.

[13] C. Liqun and J. Yaqin, "Improved x-ray prohibited items detection algorithm for YOLOv7," *2023 IEEE 6th International Conference on Automation, Electronics and Electrical Engineering (AUTEEE)*, pp. 505-510, 2023, https://doi.org/10.1109/AUTEEE60196.2023.10407853.

[14] W. Teng and H. Zhang, Y. Zhang, "X-ray Security Inspection Prohibited Items Detection Model based on Improved YOLOv7-tiny," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 7, pp. 1279-1287, 2024, https://www.iaeng.org/IJAM/issues_v54/issue_7/IJAM_54_7_05.pdf.

[15] W. Zhang, Q. Zhu, Y. Li, and H. Li, "MAM Faster R-CNN: Improved Faster R-CNN based on Malformed Attention Module for object detection on X-ray security inspection," *Digital Signal Processing*, vol. 139, p. 104072, 2023, https://doi.org/10.1016/j.dsp.2023.104072.

[16] J. Park, G. An, B.-N. Lee, and H. Seo, "Real-time CNN-based object detection of prohibited items for X-ray security screening," *Radiation Physics and Chemistry*, vol. 232, p. 112681, 2025, https://doi.org/10.1016/j.radphyschem.2025.112681.

[17] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging," *Pattern Recognition*, vol. 122, p. 108245, 2022, https://doi.org/10.1016/j.patcog.2021.108245.

[18] H. Sima, B. Chen, C. Tang, Y. Zhang, and J. Sun, "Multi-Scale Feature Attention-DEtection TRansformer: Multi-Scale Feature Attention for security check object detection," *IET Computer Vision*, vol. 18, no. 5, pp. 613-625, 2024, https://doi.org/10.1049/cvi2.12267.

[19] M. Rafiei, J. Raitoharju and A. Iosifidis, "Computer Vision on X-Ray Data in Industrial Production and Security Applications: A Comprehensive Survey," *IEEE Access*, vol. 11, pp. 2445-2477, 2023, https://doi.org/10.1109/ACCESS.2023.3234187.

[20] D. Pfeiffer, F. Pfeiffer, and E. Rummeny, "Advanced X-ray Imaging Technology," *Molecular Imaging in Oncology*, pp. 3-30, 2020, https://doi.org/10.1007/978-3-030-42618-7_1.

[21] D. Velayudhan, T. Hassan, E. Damiani, and N. Werghi, “Recent Advances in Baggage Threat Detection: A Comprehensive and Systematic Survey,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1-38, 2023, https://doi.org/10.1145/3549932.

[22] Y. Wei, Y. Liu, and H. Wang, “Cooperative distillation with X-ray images classifiers for prohibited items detection,” *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107276, 2024, https://doi.org/10.1016/j.engappai.2023.107276.

[23] M. S. H. Shovon, S. J. Mozumder, O. K. Pal, M. F. Mridha, N. Asai and J. Shin, "PlantDet: A Robust Multi-Model Ensemble Method Based on Deep Learning For Plant Disease Detection," *IEEE Access*, vol. 11, pp. 34846-34859, 2023, https://doi.org/10.1109/ACCESS.2023.3264835.

[24] W. Sarai, N. Monbut, N. Youngchoay, N. Phookriangkrai, T. Sattabun, and T. Siriborvornratanakul, “Enhancing baggage inspection through computer vision analysis of x-ray images,” *Journal of Transportation Security*, vol. 17, no. 1, p. 1, 2024, https://doi.org/10.1007/s12198-023-00270-4.

[25] Z. Cheng, L. Gao, Y. Wang, Z. Deng, and Y. Tao, “EC-YOLO: Effectual Detection Model for Steel Strip Surface Defects Based on YOLO-V5,” *IEEE Access*, vol. 12, pp. 62765-62778, 2024, https://doi.org/10.1109/ACCESS.2024.3391353.

[26] R. Gai, Y. Liu and G. Xu, "TL-YOLOv8: A Blueberry Fruit Detection Algorithm Based on Improved YOLOv8 and Transfer Learning," *IEEE Access*, vol. 12, pp. 86378-86390, 2024, https://doi.org/10.1109/ACCESS.2024.3416332.

[27] H. Ren, F. Jing and S. Li, "DCW-YOLO: Road Object Detection Algorithms for Autonomous Driving," *IEEE Access*, vol. 13, pp. 125676-125688, 2025, https://doi.org/10.1109/ACCESS.2024.3364681.

[28] Z. Chen, Q. Zhu, X. Zhou, J. Deng and W. Song, "Experimental Study on YOLO-Based Leather Surface Defect Detection," *IEEE Access*, vol. 12, pp. 32830-32848, 2024, https://doi.org/10.1109/ACCESS.2024.3369705.

[29] L. Chen, G. Li, S. Zhang, W. Mao, and M. Zhang, “YOLO-SAG: An improved wildlife object detection algorithm based on YOLOv8n,” *Ecological Informatics*, vol. 83, p. 102791, 2024, https://doi.org/10.1016/j.ecoinf.2024.102791.

[30] H. An, Z. Liang, M. Qin, Y. Huang, F. Xiong, and G. Zeng, “Wood defect detection based on the CWB-YOLOv8 algorithm,” *Journal of Wood Science*, vol. 70, no. 1, p. 26, 2024, https://doi.org/10.1186/s10086-024-02139-z.

[31] E. Casas, L. Ramos, E. Bendek and F. Rivas-Echeverría, "Assessing the Effectiveness of YOLO Architectures for Smoke and Wildfire Detection," *IEEE Access*, vol. 11, pp. 96554-96583, 2023, https://doi.org/10.1109/ACCESS.2023.3312217.

[32] Z. Liu, R. M. Rasika D. Abeyrathna, R. M. Sampurno, V. M. Nakaguchi, and T. Ahamed, “Faster-YOLO-AP: A lightweight apple detection algorithm based on improved YOLOv8 with a new efficient PDWConv in orchard,” *Computers and Electronics in Agriculture*, vol. 223, p. 109118, 2024, https://doi.org/10.1016/j.compag.2024.109118.

[33] S. R. Bakana, Y. Zhang, and B. Twala, “WildARe-YOLO: A lightweight and efficient wild animal recognition model,” *Ecological Informatics*, vol. 80, p. 102541, 2024, https://doi.org/10.1016/j.ecoinf.2024.102541.

[34] M. Cui, Y. Lou, Y. Ge, and K. Wang, “LES-YOLO: A lightweight pinecone detection algorithm based on improved YOLOv4-Tiny network,” *Computers and Electronics in Agriculture*, vol. 205, p. 107613, 2023, https://doi.org/10.1016/j.compag.2023.107613.

[35] M. Wei and W. Zhan, “YOLO_MRC: A fast and lightweight model for real-time detection and individual counting of Tephritidae pests,” *Ecological Informatics*, vol. 79, p. 102445, 2024, https://doi.org/10.1016/j.ecoinf.2023.102445.

[36] Z. Diao, X. Huang, H. Liu, and Z. Liu, “LE-YOLOv5: A Lightweight and Efficient Road Damage Detection Algorithm Based on Improved YOLOv5,” *International Journal of Intelligent Systems*, vol. 2023, no. 1, pp. 1-17, 2023, https://doi.org/10.1155/2023/8879622.

[37] J. Cao, W. Bao, H. Shang, M. Yuan, and Q. Cheng, "GCL-YOLO: A GhostConv-Based Lightweight YOLO Network for UAV Small Object Detection," *Remote Sensing*, vol. 15, no. 20, p. 4932, 2023, https://doi.org/10.3390/rs15204932.

[38] F. Guan, H. Zhang, and X. Wang, "An improved YOLOv8 model for prohibited item detection with deformable convolution and dynamic head," *Journal of Real-Time Image Processing*, vol. 22, no. 2, p. 84, 2025, https://doi.org/10.1007/s11554-025-01665-3.

[39] L. Han, C. Ma, Y. Liu, J. Jia, and J. Sun, "SC-YOLOv8: A Security Check Model for the Inspection of Prohibited Items in X-ray Images," *Electronics*, vol. 12, no. 20, p. 4208, 2023, https://doi.org/10.3390/electronics12204208.

[40] Q. Cheng, T. Lan, Z. Cai and J. Li, "X-YOLO: An Efficient Detection Network of Dangerous Objects in X-Ray Baggage Images," *IEEE Signal Processing Letters*, vol. 31, pp. 2270-2274, 2024, https://doi.org/10.1109/LSP.2024.3451311.

[41] M. -H. Guo *et al*., "Attention mechanisms in computer vision: A survey," *Computational Visual Media*, vol. 8, no. 3, pp. 331-368, 2022, https://doi.org/10.1007/s41095-022-0271-y.

[42] Y. Luo, M. Jiang and Q. Zhao, "Visual Attention in Multi-Label Image Classification," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 820-827, 2019, https://doi.org/10.1109/CVPRW.2019.00110.

[43] X. Yang, "An Overview of the Attention Mechanisms in Computer Vision," *Journal of Physics: Conference Series*, vol. 1693, no. 1, p. 012173, 2020, https://doi.org/10.1088/1742-6596/1693/1/012173.

[44] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48-62, 2021, https://doi.org/10.1016/j.neucom.2021.03.091.

[45] X. Zhu, D. Cheng, Z. Zhang, S. Lin and J. Dai, "An Empirical Study of Spatial Attention Mechanisms in Deep Networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6687-6696, 2019, https://doi.org/10.1109/ICCV.2019.00679.

[46] J. Hu, L. Shen, G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132-7141, 2018, https://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html.

[47] S. Woo, J. Park, J. Y. Lee, I. S. Kweon, "Cbam: Convolutional block attention module," *Proceedings of the European conference on computer vision (ECCV)*, pp. 3-19, 2018, https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html.

[48] Q. Hou, D. Zhou, J. Feng, "Coordinate attention for efficient mobile network design," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13713-13722, 2021, https://openaccess.thecvf.com/content/CVPR2021/html/Hou_Coordinate_Attention_for_Efficient_Mobile_Network_Design_CVPR_2021_paper.html.

[49] R. A. A. Saleh and H. M. Ertunç, "Attention-based deep learning for tire defect detection: Fusing local and global features in an industrial case study," *Expert Systems with Applications*, vol. 269, p. 126473, 2025, https://doi.org/10.1016/j.eswa.2025.126473.

[50] X. Nie, M. Duan, H. Ding, B. Hu and E. K. Wong, "Attention Mask R-CNN for Ship Detection and Segmentation From Remote Sensing Images," *IEEE Access*, vol. 8, pp. 9325-9334, 2020, https://doi.org/10.1109/ACCESS.2020.2964540.

[51] C. Peng, X. Li and Y. Wang, "TD-YOLOA: An Efficient YOLO Network With Attention Mechanism for Tire Defect Detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1-11, 2023, https://doi.org/10.1109/TIM.2023.3312753.

[52] S. Han, X. Jiang and Z. Wu, "An Improved YOLOv5 Algorithm for Wood Defect Detection Based on Attention," *IEEE Access*, vol. 11, pp. 71800-71810, 2023, https://doi.org/10.1109/ACCESS.2023.3293864.

[53] M. -A. Chung, Y. -J. Lin and C. -W. Lin, "YOLO-SLD: An Attention Mechanism-Improved YOLO for License Plate Detection," *IEEE Access*, vol. 12, pp. 89035-89045, 2024, https://doi.org/10.1109/ACCESS.2024.3419587.

[54] G. Mao *et al*., "SRS-YOLO: Improved YOLOv8-Based Smart Road Stud Detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 10092-10104, 2025, https://doi.org/10.1109/TITS.2025.3545942.

[55] L. Cao, Q. Wang, Y. Luo, Y. Hou, J. Cao, and W. Zheng, "YOLO-TSL: A lightweight target detection algorithm for UAV infrared images based on Triplet attention and Slim-neck," *Infrared Physics & Technology*, vol. 141, p. 105487, 2024, https://doi.org/10.1016/j.infrared.2024.105487.

[56] B. Huang, Y. Ding, G. Liu, G. Tian, and S. Wang, "ASD-YOLO: An aircraft surface defects detection method using deformable convolution and attention mechanism," *Measurement*, vol. 238, p. 115300, 2024, https://doi.org/10.1016/j.measurement.2024.115300.

[57] Y. Li, M. Zhang, C. Zhang, H. Liang, P. Li, and W. Zhang, "YOLO-CCS: Vehicle detection algorithm based on coordinate attention mechanism," *Digital Signal Processing*, vol. 153, p. 104632, 2024, https://doi.org/10.1016/j.dsp.2024.104632.

[58] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," *Github*, 2023, https://github.com/ultralytics/ultralytics.

[59] G. Jocher *et al*., "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," *Zenodo*, 2020, https://doi.org/10.5281/zenodo.3908559.

[60] C. -Y. Wang, A. Bochkovskiy and H. -Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7464-7475, 2023, https://doi.org/10.1109/CVPR52729.2023.00721.

[61] D. Ouyang *et al*., "Efficient Multi-Scale Attention Module with Cross-Spatial Learning," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2023, https://doi.org/10.1109/ICASSP49357.2023.10096516.

[62] Z. Tong, Y. Chen, Z. Xu, and R. Yu, "Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism," *arXiv*, 2023, https://doi.org/10.48550/arXiv.2301.10051.

[63] C. Liu, K. Wang, Q. Li, F. Zhao, K. Zhao, and H. Ma, "Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism," *Neural Networks*, vol. 170, pp. 276–284, 2024, https://doi.org/10.1016/j.neunet.2023.11.041.

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

[65] T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327, 2020, https://doi.org/10.1109/TPAMI.2018.2858826.

[66] C. Li *et al*., "YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications," *arXiv*, 2022, https://doi.org/10.48550/arXiv.2209.02976.

[67] G. Jocher and J. Qiu, "Ultralytics YOLO11," *Github*, 2024, https://github.com/ultralytics/ultralytics.