Article

Improved YOLOv8 for Dangerous Goods Detection in X-ray Security Images

Aili Wang¹, Pengfei Yuan¹, Haibin Wu¹,∗, Yuji Iwahori² and Yan Liu³

¹ Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Application, Harbin 150080, China; aili925@hrbust.edu.cn (A.W.);
² Computer Science, Chubu University, Kasugai 487-8501, Japan; iwahori@isc.chubu.ac.jp
³ College of Electron and Information, University of Electronic Science and Technology of China, Zhongshan Institute, Zhongshan 528402, China; yanliu@zsc.edu.cn

* Correspondence: woo@hrbust.edu.cn

Abstract: X-ray security images face significant challenges due to complex backgrounds, item overlap, and multi-scale target detection. Traditional methods often struggle to accurately identify objects, especially under cluttered conditions. This paper presents an advanced detection model, called YOLOv8n-GEMA, which incorporates several enhancements to address these issues. Firstly, the generalized efficient layer aggregation network (GELAN) module is employed to augment the feature fusion capabilities. Secondly, to tackle the problems of overlap and occlusion in X-ray images, the efficient multi-scale attention (EMA) module is utilized, effectively managing the feature capture and interdependencies among overlapping items, thereby boosting the model’s detection capability in such scenarios. Lastly, addressing the diverse sizes of items in X-ray images, the Inner-CIoU loss function uses auxiliary bounding boxes at varying scale ratios for loss calculation, ensuring faster and more effective bounding box predictions. The enhanced YOLOv8 model was tested on the public datasets SIXRay, HiXray, CLCXray, and PIDray, where the improved model’s mean average precision (mAP) reached 94.4%, 82.0%, 88.9%, and 85.9%, respectively, showing improvements of 3.6%, 1.6%, 0.9%, and 3.4% over the original YOLOv8. These results demonstrate the effectiveness and universality of the proposed method. Compared to current mainstream X-ray images of dangerous goods detection models, this model significantly reduces the false detection rate of dangerous goods in X-ray security images and achieves substantial improvements in the detection of overlapping and multi-scale targets, realizing higher accuracy in dangerous goods detection.

Keywords: X-ray images; dangerous goods detection; GELAN; EMA; inner-CIoU

1. Introduction

In recent years, with the increasingly complex global security situation, the demand for luggage screening of prohibited items at public places such as airports, rail transit, and customs has been continuously increasing. Traditional X-ray security checks mainly rely on manual screening, which is not only inefficient but also has its accuracy and stability limited by the experience and concentration of security personnel, leading to missed detections and false alarms. To improve the accuracy and efficiency of detection, deep learning-based X-ray image detection techniques for prohibited items have emerged.

The intelligent analysis of X-ray security images falls under the task of object detection in computer vision. The model needs to not only identify whether there are prohibited items in the X-ray images and classify them but also determine their locations. Generally speaking, X-ray security images have some unique characteristics, which increase the difficulty of implementing prohibited item detection using computer vision techniques:

1. Cluttered Goods Issue: Goods inside luggage are usually arranged chaotically. Due to the transmission effect of X-rays, different goods often overlap in security images.
Traditional image processing methods cannot effectively handle target detection affected by occlusion.

2. Imaging Angle Issue: Fixed perspective X-rays can easily result in unfavorable imaging angles, causing extreme and distorted visual displays of prohibited goods in X-ray images, making it difficult to extract features of prohibited items accurately and effectively.

3. Goods Size Issue: There are many different sizes of goods in luggage. The detection of prohibited items can be considered a multi-scale and multi-target detection problem. The size differences easily lead to the detection model overlooking smaller objects, resulting in missed detections.

The bag of visual words (BoVW) method has been very popular for object detection in X-ray images, which typically involves feature extraction through descriptors, followed by feature clustering using the k-means algorithm, and finally classification using support vector machines (SVMs) or sparse representation. In 2011, a BoVW model was employed on a relatively limited dataset for object recognition in X-ray images. The researchers combined various feature descriptors such as scale-invariant feature transform (SIFT), speeded-up robust features (SURF), and binary robust independent elementary features (BRIEF) with feature detectors like difference of Gaussians (DoGs), hessian Laplace, Harris, and so on, followed by SVM training on k-means clustered BoVW. The experimental results showed that the combination of the DoG feature detector and SIFT feature descriptor achieved better performance, obtaining a 65% accuracy on 200 security images [1]. In 2013, Turcsany et al. proposed a unique BoVW algorithm that used the SURF feature detector and descriptor to extract features and SVM to classify guns in security images, achieving an accuracy of 99.07% [2]. In 2015, Fliton et al. evaluated various feature descriptors combined with classification algorithms for detecting pistols and bottles in 3D CT images. They found that descriptors based on density histograms, specifically the density histogram (DH) and the density gradient histogram (DGH), outperformed other methods like SIFT and rotation invariant feature transform (RIFT) in classification tasks. This comparison highlighted the superior performance of DH and DGH in recognizing these objects [3]. In 2016, Mery et al. further employed the BoVW method to form a dictionary for each category, consisting of SIFT feature descriptors extracted from randomly cropped and stitched images. The resulting feature descriptors were classified using sparse representation, achieving a recognition accuracy of 95%, and 85% under occlusion conditions [4].

In another BoVW method, Zhang et al. used SVM to train local potential low-level image features extracted from 15 different categories of datasets, achieving an 80.1% recognition accuracy on 100 images per category [5]. Inspired by various research results of BoVW methods, Kundegorski et al. comprehensively evaluated various feature descriptors in image recognition tasks based on BoVW. The combination of FAST and SURF trained by SVM was the best-performing feature detector and descriptor in firearm recognition tasks on large datasets, achieving a recognition accuracy of 94% [6]. Similar work detailed the evaluation of multiple computer vision (CV) techniques, showing that clustering SIFT-extracted features with k-means and then classifying with a sparse representation-based k-nearest neighbor (KNN) algorithm performed well, ultimately achieving a 94.7% accuracy on GDXray [7]. However, these traditional image algorithms all rely on manual feature extraction, which is limited in the features extracted and only suitable for simple scenarios in security images, failing to effectively identify dangerous goods in real-world scenarios.

To address these challenges, deep-learning-based X-ray image detection techniques for prohibited items have gradually developed, providing effective solutions to enhance the accuracy and efficiency of security detection. In 2016, Akcay et al. first applied convolutional neural networks (CNNs) to security images. They studied the use of transfer learning with CNNs to evaluate its effectiveness in identifying dangerous goods in security images [8]. In 2017, Rogers et al. explored the use of dual-energy X-ray security images for identifying dangerous goods. They examined high-energy and low-energy X-ray images captured by dual-energy X-ray machines. Using the UCL TIP dataset, they generated
The datasets with different input channels were fed into the VGG-19 network for training, and the results indicated that the dual-channel and four-channel inputs achieved better recognition accuracy, reaching 95% [9]. In 2018, inspired by the limited X-ray security image datasets, Zhao et al. proposed a three-stage algorithm. The first stage classified and labeled the X-ray image dataset using angle information of foreground objects extracted from input images. The second stage generated new dangerous goods using a network model similar to generative adversarial networks (GANs) to improve the quality of generated images. Finally, a small classification network was used to verify whether the generated images belonged to the correct category [10]. Yang et al. further improved the GAN model to produce higher quality X-ray security images. The experimental results showed that the proposed method generated visually superior images [11]. In 2019, Miao et al. proposed a class-balanced hierarchical refinement algorithm for the SIXray dataset, which addressed class imbalance and clutter by extracting image features from three consecutive layers, with the later layers upsampled and connected to the previous layer. By refining each layer and removing redundant information, the recognition accuracy was improved [12]. Guo innovated the SSD model by replacing the base network and integrating multi-scale features, which significantly enhanced the model’s ability to recognize small-sized dangerous goods [13].

You only look once (YOLO), known for its streamlined structure and superior detection performance, has garnered extensive attention in the field of computer vision. Various improved YOLO models have been applied to the task of detecting prohibited items in X-ray security images. Lu Guanyou enhanced the YOLOv3 network by applying the k-means clustering algorithm to determine prior boxes that better match the size of the targets, thereby speeding up the detection process by minimizing the number of predicted bounding boxes [14]. In 2022, Yu et al. proposed replacing the multi-space pyramid pooling structure in YOLOv4 with the spatial pyramid pooling and fully connected spatial pyramid convolution (SPPFCSPC) module and introduced the convolutional block attention module (CBAM) to enhance the accuracy of detecting prohibited items. Wu optimized the SPP module in the YOLOv4 network by using dilated convolution techniques to increase the network’s receptive field. This optimization improved the detection capability for small objects and mitigated the issue of object overlap in security images [15]. Mu designed a dilated dense convolution module for the YOLOv4 algorithm and introduced an attention module, achieving an average precision of 80.16% on the SIXray dataset [16]. Dong proposed an enhanced version of the YOLOv5 network model by adding convolutional block attention modules to improve the model’s feature extraction capability and optimizing the prediction boxes with a weighted bounding box fusion algorithm during the testing phase [17]. Xiang et al. developed an improved prohibited items detection algorithm based on YOLOv5s. They utilized a re-parameterization concept to enhance the recurrent expanded pyramid structure (Rep), enabling the backbone network to extract more feature information and optimizing detection results with an attention mechanism excitation module [18]. Li introduced a Swin transformer module into the YOLOv5s model, improving the feature extraction capability and detection accuracy by incorporating the focal loss function [19]. Cheng improved the feature extraction capability of the YOLOv7 network using skip connections and a $1 \times 1$ convolution architecture to optimize a multi-branch stacking module [20]. Guo et al. proposed the YOLO-C model, which employs a compound backbone network and introduces a feature enhancement module to improve the non-linear expression capability of network features [21]. In 2024, Dong proposed an improved model, YOLOv8s-BiOG, which introduced a dynamic convolution module, a weighted bidirectional feature pyramid network (BiFPN), and a global attention mechanism. The dynamic convolution module replaces some convolution modules in the backbone and neck networks to refine local features of prohibited items and enhance feature extraction. The BiFPN improves the model’s ability to fuse features at different scales, while the global attention mechanism reduces feature loss and enhances detection performance. Experiment-
tal results on the SI2P-xray and OPIXray datasets showed that the average precision (mAP) for various prohibited items reached 93.4% and 91.8%, respectively [22].

In 2022, the Ultralytics team released the source code for the YOLOv8 model, offering five different scales and channel numbers (n, s, m, l, x). This model can better meet the needs of different scenarios, maximizing both real-time performance and accuracy, and holds significant potential for development in prohibited items detection. Additionally, by employing rotated box techniques, they enhanced the detection capability for contraband in the X-ray images of arbitrary orientations. The aforementioned methods have improved the accuracy of detecting hazardous materials in X-ray security images, providing a variety of approaches for applying deep learning in the field of X-ray detection of dangerous goods. However, due to the projective nature of X-rays, objects made of the same material exhibit similar characteristic information during X-ray transmission, blurring the boundaries between targets and background, making them difficult to distinguish, and causing interference in identification.

This paper introduces a new network based on YOLOv8n, called YOLOv8n-GEMA, which integrates a generalized efficient layer aggregation network (GELAN), the efficient multi-scale attention (EMA), and the inner-CIoU to implement a prohibited item detection algorithm for X-ray security imaging. The main contributions of the proposed method are summarized as follows:

1. GELAN Network Architecture: This architecture, which merges the strengths of CSPNet and ELAN, is particularly beneficial for X-ray image analysis where items are often obscured or overlapped. Its ability to optimize gradient paths significantly enhances learning efficiency, crucial for accurately identifying hidden or obscured objects in cluttered luggage scenarios. The incorporation of diverse computational blocks, like the RepNCSPELAN4 module, improves feature extraction capabilities, critical for detecting small or oddly shaped prohibited items that might otherwise be missed. The lightweight and scalable nature of GELAN ensures that the system can be efficiently deployed in real-time applications, such as airport security screenings, without sacrificing performance.

2. Efficient Multi-Scale Attention Module: The challenges posed by fixed imaging angles and the resulting distorted visual displays of items in X-ray images can be mitigated by the EMA module. This module’s use of parallel processing and advanced attention mechanisms allows for effective feature representation even at unfavorable angles. By employing multi-scale and cross-spatial learning techniques, EMA can adjust its focus across various scales, crucial for identifying both large and small items within a single scan, enhancing the accuracy and reliability of the detection process.

3. Inner-CIoU Loss Function: The inner-CIoU loss function introduces a transformative approach by incorporating a scale factor to adjust the size of auxiliary bounding boxes, enhancing the precision of bounding box regression. This method refines the traditional IoU calculations by allowing the model to handle various item sizes more accurately, particularly useful in cluttered X-ray images where small and overlapping items must be detected distinctly. The adaptation of the scale factor ratio not only maintains the simplicity and directness of IoU-based loss calculations but also significantly improves their effectiveness. By better addressing the challenges of generalization and convergence that are common with traditional IoU-based loss, inner-CIoU promotes quicker, more accurate model training and improved detection performance, making it particularly effective in the high-stakes environment of security screening.

2. Methods

In this methodology section, we introduce three main improvements based on the YOLOv8n detection framework, which significantly enhanced the performance of our model. First, the RepNCSPELAN4 module was designed and introduced into the backbone network to address the occlusion issues caused by item stacking, increasing the diversity of
learned features, reducing computational demands, and enhancing learning capabilities and inference speed. Second, the efficient multi-scale attention (EMA) module was modified and combined into the neck. EMA, through its parallel processing strategy, effectively handles the feature capture and interdependencies among overlapping items, thereby enhancing the model’s capability to detect in such scenarios. Lastly, the inner-CIoU loss function was adopted. Inner-CIoU uses auxiliary bounding boxes at varying scale ratios for loss calculation, improving the model’s generalization and convergence across different detection tasks. By calculating losses with these auxiliary bounding boxes, inner-IoU refines the regression process, ensuring faster and more effective bounding box predictions. This is particularly beneficial in multi-scale and multi-target detection scenarios, aiding the model in better recognizing items of varying sizes, preventing the overlooking of smaller objects, and thereby reducing missed detections. The YOLOv8 network was modified by GELAN combined with EMA in this paper, named YOLOv8n-GEMA, as shown in Figure 1.

![Figure 1. The structure of YOLOv8n-GEMA.](image)

### 2.1. GELAN

In the X-ray detection of dangerous goods, the GELAN architecture notably incorporates the RepNCSPELAN4 module, which features optimized gradient paths and efficient feature fusion capabilities, crucial for enhancing detection accuracy and speed [23]. By integrating the gradient path optimization of CSPNet with the layer aggregation efficiency of ELAN, the RepNCSPELAN4 module significantly boosts the feature extraction capabilities. This is particularly effective when processing X-ray images with overlapping and variably sized items. The module optimizes the gradient transmission and reduces the computational load, enabling the network to quickly and accurately identify and locate dangerous goods, especially in scenarios where the items are chaotically arranged and extensively overlapping. Additionally, its lightweight design ensures efficient real-time applications, adaptable to various model sizes and complexities. The following will provide a detailed introduction to the structure of this module.

By merging two neural network architectures, CSPNet [24] and ELAN [25], both designed with gradient path optimization, GELAN is a generalized efficient layer aggregation network that balances lightweight design, fast inference speed, and high accuracy,
as shown in Figure 2. The functionality of ELAN, which initially relied solely on stacking convolutional layers, has been extended to a new architecture capable of utilizing various computational blocks.

![Figure 2](image)

**Figure 2.** The structure of CPSNet, ELAN, and GELAN. (a) CPSNet. (b) ELAN. (c) GELAN.

GELAN combines the two powerful neural network architectures of CSPNet and ELAN. CSPNet enhances gradient combination by dividing the feature map of the base layer into two parts. One part goes through a computational block (such as Res block, ResX block, or Dense block), while the other part is directly combined with the processed feature map in the next stage. This cross-stage partial operation increases the diversity of learned features and reduces the overall computational load and memory usage. The gradient flow truncate operation, which introduces transition layers at the end of both paths, helps to increase the richness of gradient combinations, improving the learning capacity of the network. Consequently, CSPNet not only enhances the gradient combination but also improves inference speed without sacrificing accuracy, making it suitable for real-time applications.

ELAN, on the other hand, optimizes the gradient length by effectively stacking computational blocks. This reduces the shortest gradient path and ensures efficient gradient propagation. ELAN integrates VoVNet’s layer aggregation with CSPNet’s gradient optimization to improve the network’s overall efficiency. By analyzing and optimizing gradient paths, ELAN ensures better convergence, especially for deep networks, and facilitates efficient gradient flow, which makes the network both scalable and flexible, adaptable to various tasks and datasets without degrading performance.

In this paper, GELAN incorporates the RepNCSPELAN4 module within any block, enhancing its functionality and performance. The structure of the RepNCSPELAN4 module is designed to optimize the gradient path and improve computational efficiency, which is a key component of the GELAN network. The detailed architecture of the RepNCSPELAN4 module is illustrated in Figure 3, highlighting its configuration and contributions to the overall efficiency and accuracy of the GELAN network. Compared to the original YOLOv9 network, we improved YOLOv8n by introducing GELAN module and EMA module (called GEMA) and replacing the loss function with inner-CIoU in this paper.
In this paper, GELAN incorporates the RepNCSPELAN4 module within any block, enhancing its functionality and performance. The structure of the RepNCSPELAN4 module is designed to optimize the gradient path and improve computational efficiency, which is a key component of the GELAN network. The detailed architecture of the RepNCSPELAN4 module is illustrated in Figure 3, highlighting its configuration and contributions to the overall efficiency and accuracy of the GELAN network. Compared to the original YOLOv9 network, we improved YOLOv8n by introducing GELAN module and EMA module (called GEMA) and replacing the loss function with inner-CIoU in this paper.

Figure 3. The structure of RepNCSPELAN4.

RepNCSPELAN4 is a critical feature extraction and fusion module. It integrates Conv layers and RepNCSP to achieve efficient feature extraction and fusion. Structurally, the RepNCSP component includes Conv layers and multiple RepNBottleneck modules. These modules leverage the gradient path optimization of CSPNet and the efficient layer aggregation of ELAN. By combining CSPNet and ELAN, RepNCSPELAN4 significantly enhances feature extraction, thereby improving performance in object detection tasks. The integration of CSPNet ensures optimized gradient paths, facilitating better training and convergence. Its lightweight and fast design makes it suitable for real-time applications, while its modular structure allows scalability, adapting to various model sizes and complexities. The advanced design of RepNCSPELAN4 provides a robust and efficient solution for object detection, ensuring high accuracy and fast inference speeds.

2.2. EMA

The impressive effectiveness of channel or spatial attention mechanisms in generating more distinguishable feature representations has been demonstrated in various computer vision tasks. However, reducing channel dimensionality when modeling cross-channel relationships can negatively impact the extraction of deep visual representations. This paper adopts a novel efficient multi-scale attention (EMA) module [26].

The EMA module plays a crucial role in enhancing the effectiveness of X-ray detection systems for dangerous goods. Its design leverages both channel and spatial attention mechanisms to improve the necessary feature discrimination for accurate and efficient detection of prohibited items in X-ray security images. By maintaining full channel dimensionality and integrating multi-scale convolutional structures (1 × 1 and 3 × 3 convolutions), the EMA module effectively captures both detailed and broader spatial features. This capability is essential for distinguishing overlapping items within the typically chaotic arrangement of luggage in X-ray images. The parallel substructures of the EMA module enable it to handle items of varying sizes within luggage. The multi-scale design allows the module to
detect smaller items that might otherwise be overlooked, as well as accurately characterize larger items, addressing the challenges of multi-scale detection. By integrating enhanced attention mechanisms that focus on areas of interest within the X-ray images, the EMA module reduces the likelihood of false positives and missed detections. We will provide a detailed introduction to the structure and principles of the EMA module. The structure of the EMA module is shown in Figure 4.

![Figure 4. The structure of EMA module.](image)

Parallel substructures help networks avoid extensive sequential processing and significant depth. We applied this parallel processing strategy to our EMA module, as shown in Figure 4. In this section, we will discuss how the EMA module learns effective channel descriptions without reducing channel dimensionality in convolution operations, thereby producing better pixel-level attention for high-level feature maps. Specifically, we extracted the shared component of the $1 \times 1$ convolution from the CA module, referring to it as the $1 \times 1$ branch in our EMA. To aggregate multi-scale spatial structure information, we placed a $3 \times 3$ convolution kernel in parallel with the $1 \times 1$ branch for rapid responses, naming it the $3 \times 3$ branch. Considering feature grouping and multi-scale structures, this approach efficiently establishes both short- and long-range dependencies to enhance performance.

### 2.2.1. Feature Grouping

For any given input feature map $X \in \mathbb{R}^{C \times H \times W}$, EMA divides $X$ into $G$ sub-features along the channel dimension to learn different semantics. This grouping can be represented as $X = [X_0, X_1, \ldots, X_{G-1}]$, where $X_i \in \mathbb{R}^{C/G \times H \times W}$. Generally, we assume $G \ll C$ and that the learned attention weights will be used to enhance the feature representation of the regions of interest in each sub-feature.

### 2.2.2. Parallel Subnetworks

The large local receptive fields of neurons enable them to gather multi-scale spatial information. Consequently, EMA employs three parallel routes to extract attention weight descriptors from the grouped feature maps. Two of these routes are within the $1 \times 1$ branch, while the third is in the $3 \times 3$ branch. To capture dependencies across all channels and reduce computational costs, we modeled the cross-channel information interaction along the channel dimension. Specifically, the $1 \times 1$ branch uses two 1D global average pooling operations to encode the channels along two spatial directions, while the $3 \times 3$ branch uses a single $3 \times 3$ kernel to capture multi-scale feature representations.

Given that there is no batch coefficient in the dimension of the convolution function for normal convolutions, the number of convolution kernels is independent of the batch coefficients of the forward operational inputs. For example, in PyTorch 2.3.0, the parameter dimension of a normal 2D convolution kernel is $[out, inp, k, k]$, where $out$ represents the output planes, $inp$ represents the input planes, and $k$ denotes the kernel size. This does not involve the batch dimensions.
Accordingly, we reshaped and permuted the G groups into the batch dimension and redefined the input tensor with a shape of \([C/G, H, W]\). In the \(1 \times 1\) branch, similar to the CA module, we concatenated the two encoded features along the height dimension of the images, sharing the same \(1 \times 1\) convolution without dimensionality reduction. After factorizing the outputs of the \(1 \times 1\) convolution into two vectors, two non-linear Sigmoid functions were used to fit the 2D binomial distribution upon linear convolutions. To achieve different cross-channel interactive features between the two parallel routes in the \(1 \times 1\) branch, we aggregated the two channel-wise attention maps within each group via simple multiplication. On the other hand, the \(3 \times 3\) branch captures local cross-channel interaction via a \(3 \times 3\) convolution to enlarge the feature space. This way, EMA not only encodes inter-channel information to adjust the importance of different channels but also preserves precise spatial structure information within the channels.

### 2.2.3. Cross-Spatial Learning

Benefiting from the ability to establish interdependencies between channels and spatial locations, cross-spatial learning has been extensively researched and widely applied in various recent computer vision tasks. In this paper, a method was adopted for aggregating information across different spatial dimensions to achieve richer feature aggregation.

Here, we introduced two tensors: one is the output of the \(1 \times 1\) branch, and the other is the output of the \(3 \times 3\) branch. We then used 2D global average pooling to encode the global spatial information from the \(1 \times 1\) branch outputs, while the \(3 \times 3\) branch outputs are directly transformed to the corresponding dimension shapes before the joint activation mechanism of channel features. Specifically, the dimensions are reshaped to \(R_1 \times C/G \times R_3 \times C/G \times HW\). The 2D global pooling operation formula is shown in Formula (1):

\[
z_c = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} x_c(i, j)
\]  

(1)

This method is designed to encode the global information and model long-range dependencies. For efficient computation, the natural non-linear function Softmax for 2D Gaussian maps is employed at the outputs of 2D global average pooling to fit the linear transformations. By multiplying the outputs of the parallel processing with matrix dot-product operations, we derived our first spatial attention map. This approach collects different scale spatial information in the same processing stage. Furthermore, we similarly utilized 2D global average pooling to encode the global spatial information in the \(3 \times 3\) branch, while the \(1 \times 1\) branch is transformed to the corresponding dimension shape directly before the joint activation mechanism of channel features, i.e., \(R_1 \times C/G \times R_3 \times C/G \times HW\).

After that, the second spatial attention map, which preserves the precise spatial positional information, was derived. Finally, the output feature map within each group was calculated as the aggregation of the two generated spatial attention weight values followed by a Sigmoid function. This approach captured pixel-level pairwise relationships and highlighted the global context for all pixels. The final output of EMA is the same size as \(X\), making it efficient and effective to integrate into modern architectures.

As discussed above, the attention factors are guided solely by the similarities between the global and local feature descriptors within each group. By considering the cross-spatial information aggregation method, EMA models both long-range dependencies and embeds precise positional information. Fusing context information at different scales enables CNNs to produce better pixel-level attention for high-level feature maps. The parallelization of convolution kernels appears to be a more powerful structure for handling both short- and long-range dependencies using the cross-spatial learning method. Unlike the progressive behavior of limited receptive fields, utilizing \(3 \times 3\) and \(1 \times 1\) convolutions in parallel leverages more contextual information among the intermediate features.
2.3. Inner-CIoU

In X-ray detection of dangerous goods, utilizing the inner-IoU loss function can significantly enhance the precision and speed of the bounding box regression, which is crucial for detection accuracy. X-ray images often involve complex backgrounds and overlapping items, where traditional IoU loss functions might lead to gradient vanishing or slow convergence under these circumstances. Inner-IoU optimizes the loss calculation by introducing auxiliary bounding boxes and a scale factor, allowing the model to quickly adjust the predicted boxes to accurately capture dangerous goods of various sizes. Additionally, the implementation of inner-IoU helps reduce detection errors caused by significant variations in item sizes, thereby improving the overall efficiency and accuracy of detection. Below, we will provide a detailed explanation of the principles behind inner-CIoU [27].

With the swift advancement of detectors, the bounding box regression (BBR) loss function has undergone continuous updates and improvements. IoU, a crucial component of the prevalent bounding box regression loss functions, is defined in Formula (2), as follows:

\[
\text{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (2)
\]

B and B\(^{gt}\) denote the predicted box and the ground truth (GT) box, respectively. Once IoU is defined, the corresponding loss can be expressed as shown in Formula (3):

\[
L_{\text{IoU}} = 1 - \text{IoU} \quad (3)
\]

Currently, IoU-based loss functions have become mainstream and dominant. Most existing methods build upon IoU and incorporate additional loss terms. For instance, GloU was introduced to address the gradient vanishing issue when the overlap area between the anchor box and the GT box is zero. The GloU loss function is defined as shown in Formula (4), where C is the smallest box that covers both B and B\(^{gt}\):

\[
L_{\text{GloU}} = 1 - \text{IoU} + \frac{|C - B \cap B^{gt}|}{|C|} \quad (4)
\]

In comparison to GloU, the DIoU loss function introduces a new distance loss term to the IoU, primarily by minimizing the normalized distance between the center points of the two bounding boxes. This leads to faster convergence and improved performance.

\[
L_{\text{DIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} \quad (5)
\]

where \(b\) and \(b^{gt}\) are the center points of B and B\(^{gt}\), respectively, and \(\rho(\cdot)\) refers to the Euclidean distance, where \(c\) is the diagonal of the minimum bounding box.

The CIoU loss further considered the shape loss and added a shape loss term on the basis of DIoU loss.

\[
L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (6)
\]

where \(\alpha\) is a positive trade-off parameter, and its definition is as follows:

\[
\alpha = \frac{v}{(1 - \text{IoU}) + v} \quad (7)
\]

where \(v\) measures the consistency of aspect ratio:

\[
v = \frac{4}{\pi^2} \left( \arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (8)
\]
$w^c$ and $h^c$ denote the width and height of the target box and $w$ and $h$ denote the width and height of the predicted box. When the aspect ratio of the target box and predicted box is the same, CIoU will degrade to DIoU.

While the aforementioned bounding box regression loss functions can accelerate convergence and enhance detection performance by adding new geometric constraints to the IoU loss functions, they do not address the inherent limitations of the IoU loss itself, which significantly impacts detection quality. To address this issue, we adopted inner-IoU scaling factor, typically within the range of values $[0.5, 1.5]$.

The center points of the GT box and the inner GT box are represented by $(x^c, y^c)$, while $(x_c, y_c)$ represents the center points of the anchor and the inner anchor. The width and height of the GT box are denoted as $w^t$ and $h^t$, respectively, while the width and height of the anchor are represented by $w$ and $h$. The variable "ratio" corresponds to the scaling factor, typically within the range of values $[0.5, 1.5]$.

\begin{align*}
b^t_1 &= x^t - \frac{w^t \ast \text{ratio}}{2}, \quad b^t_2 = x^t + \frac{w^t \ast \text{ratio}}{2} \\
b^t_3 &= y^t - \frac{h^t \ast \text{ratio}}{2}, \quad b^t_4 = y^t + \frac{h^t \ast \text{ratio}}{2} \\
b_1 &= x_c - \frac{w \ast \text{ratio}}{2}, \quad b_r = x_c + \frac{w \ast \text{ratio}}{2} \\
b_t &= y_c - \frac{h \ast \text{ratio}}{2}, \quad b_b = y_c + \frac{h \ast \text{ratio}}{2}
\end{align*}

The center points of the GT box and the inner GT box are represented by $(x^t_c, y^t_c)$, while $(x_c, y_c)$ represents the center points of the anchor and the inner anchor. The width and height of the anchor are represented by $w$ and $h$. The variable "ratio" corresponds to the scaling factor, typically within the range of values $[0.5, 1.5]$.

$\pi \ast I_{\text{ratio}}(B, B^t) = \pi \ast I_{\text{ratio}}(B, B^t)$

Figure 5. Description of inner-IoU.
\begin{align*}
\text{inter} &= \left( \min(b^*_t, b_r) - \max(b^*_t, b_l) \right) \times \left( \min(b^*_b, b_b) - \max(b^*_b, b_b) \right) \quad (13) \\
\text{union} &= (w^*_t \times h^*_t) \times (\text{ratio})^2 + (w \times h) \times (\text{ratio})^2 - \text{inter} \quad (14) \\
\text{IoU}^{\text{inner}} &= \frac{\text{inter}}{\text{union}} \quad (15)
\end{align*}

Inner-IoU loss retains some characteristics of IoU loss while introducing its own unique features. Similar to IoU loss, the range of values for inner-IoU loss is \([0, 1]\). Since the difference between the auxiliary bounding boxes and the actual bounding boxes is merely in scale, the calculation method of the loss function remains the same, and the inner-IoU deviation curve closely resembles the IoU deviation curve. Applying inner-IoU loss to \(L_{\text{Inner} - \text{CIoU}}\) enhances the performance of the bounding box regression:

\[ L_{\text{Inner} - \text{CIoU}} = L_{\text{CIoU}} + \text{IoU} - \text{IoU}^{\text{inner}} \quad (16) \]

Through the previous discussion, we have understood the advantages and disadvantages of existing IoU-based bounding box regression loss functions. Although these methods have addressed some critical issues in detection to a certain extent, their limitations also impact the quality of detection results. To resolve these issues, we introduced a novel inner-IoU loss function and elaborated on its design concepts and advantages. Below, we will summarize the core features of inner-IoU loss and its effectiveness in practical applications.

### 3. Results

#### 3.1. Dataset Description and Experimental Environment Settings

The experimental environment is set up using the Pycharm 2023.2.3 and the Pytorch 2.3.0 deep learning framework. Model training is performed on an Intel i5-13490 CPU @ 2.5 GHz processor, with an NVIDIA GeForce RTX 4070 graphics card. The training is conducted over 300 epochs. MixUp and Mosaic are used as the primary data augmentation methods, with the Mosaic probability set to 1 and the MixUp probability set to 0.6. In addition, a horizontal flip is adopted and the parameter is set to 0.5, which allows the model to learn the mirror representation of the object, enhancing robustness to changes in direction; data enhancement was performed with scale transformation and the parameter was set to 0.5, which allowed the model to adapt to objects of different sizes, which is crucial for recognizing the same objects at different distances.

To evaluate the performance of our proposed method, we conducted experiments on four X-ray image datasets: SIXray [12], CLCXray [28], PIDray [29], and HiXray [30]. Each dataset had the following distinct characteristics:

1. SIXray contains annotations for six types of dangerous items and focuses primarily on the issue of overlapping objects.
2. HiXray contains eight different types of contraband items not covered in SIXray. This dataset features a greater variance in the size of dangerous items, ranging from laptops to lighters.
3. CLCXray emphasizes the overlap between the target and similar backgrounds, as well as between multiple targets. It includes annotations for various liquid containers and consists of 12 categories.
4. PIDray is the largest dataset with annotated images of dangerous items, covering a wide range of real-world contraband detection scenarios, especially intentionally hidden items. It includes 12 categories of contraband items with high-quality annotated segmentation masks and bounding boxes.

These datasets contain images of various hazardous materials and provide a comprehensive test for assessing detection accuracy and robustness. The categories and the number of instances for each category in the four datasets are shown in Table 1. Some sample images from each dataset are shown in Figure 6.
Table 1. Class distribution of SIXray, HiXray, CLCXray, and PIDray datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SIXray</th>
<th>HiXray</th>
<th>CLCXray</th>
<th>PIDray</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gun</td>
<td>Portable_Charger_1</td>
<td>12,421</td>
<td>Blade</td>
</tr>
<tr>
<td>2</td>
<td>Knife</td>
<td>Portable_Charger_2</td>
<td>7788</td>
<td>Dagger</td>
</tr>
<tr>
<td>3</td>
<td>Wrench</td>
<td>Water</td>
<td>3092</td>
<td>Knife</td>
</tr>
<tr>
<td>4</td>
<td>Pliers</td>
<td>Laptop</td>
<td>10,042</td>
<td>Scissors</td>
</tr>
<tr>
<td>5</td>
<td>Scissor</td>
<td>Mobile_Phone</td>
<td>53,835</td>
<td>Swiss Army Knife</td>
</tr>
<tr>
<td>6</td>
<td>Tablet</td>
<td>4918</td>
<td>Cans</td>
<td>789</td>
</tr>
<tr>
<td>7</td>
<td>Cosmetic</td>
<td>9949</td>
<td>Carton Drinks</td>
<td>1926</td>
</tr>
<tr>
<td>8</td>
<td>Nonmetallic_Lighter</td>
<td>883</td>
<td>Glass Bottle</td>
<td>540</td>
</tr>
<tr>
<td>9</td>
<td>Plastic Bottle</td>
<td>5998</td>
<td>Sprayer</td>
<td>4227</td>
</tr>
<tr>
<td>10</td>
<td>Vacuum Cup</td>
<td>2166</td>
<td>Handcuffs</td>
<td>3388</td>
</tr>
<tr>
<td>11</td>
<td>Spray Cans</td>
<td>1077</td>
<td>Knife</td>
<td>5549</td>
</tr>
<tr>
<td>12</td>
<td>Tin</td>
<td>856</td>
<td>Lighter</td>
<td>6157</td>
</tr>
</tbody>
</table>

Figure 6. Sample images from the SIXray, HiXray, CLCXray, and PIDray datasets.

In evaluating the performance of object detection systems, this paper employs four key metrics, namely, recall, F1 score, precision, and mean average precision (mAP). These metrics collectively provide a thorough assessment of the system’s accuracy and robustness. By applying these metrics, we can comprehensively evaluate the performance of target detection systems in terms of accuracy, sensitivity to detecting positives, and the
overall quality of detection, providing insights into areas for potential improvement in detection systems.

### 3.2. Experimental Results and Analysis

In order to verify the detection performance of the improved YOLOv8n-GEMA network on X-ray images of dangerous goods, this paper compares the proposed improved algorithm with current mainstream object detection algorithms on PIDray: YOLOv4, YOLOv5s, YOLOv7, and YOLOv8. A comparative experiment was conducted using mAP as a measure of detection performance, and the results are shown in Table 2. As shown in Table 2, our model achieved an mAP of 85.9% on the PIDray dataset, outperforming YOLOv4, YOLOv5s, YOLOv7, and YOLOv8n by 9.5%, 2.3%, 2.7%, and 4.1%, respectively. Significant improvements were observed in the detection of “Sprayer” and “Knife” objects. The improvement in detecting “Knife” is attributed to the integration of the GELAN module, which enhances feature fusion, thereby increasing the accuracy for “Knife”, which may be difficult to recognize due to its placement. Additionally, “Sprayer” is prone to occlusion, and the EMA module effectively captures detailed and extensive spatial features, resulting in a higher accuracy in recognizing “Sprayer”.

<table>
<thead>
<tr>
<th>Methods</th>
<th>YOLOv4</th>
<th>YOLOv5s</th>
<th>YOLOv7</th>
<th>YOLOv8n</th>
<th>YOLOv10</th>
<th>RT-DETR</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baton</td>
<td>89.6</td>
<td>99.3</td>
<td>99.6</td>
<td>99.4</td>
<td>99.3</td>
<td>99.5</td>
<td>99.4</td>
</tr>
<tr>
<td>Pliers</td>
<td>90.3</td>
<td>99.4</td>
<td>99.5</td>
<td>99.4</td>
<td>99.4</td>
<td>99.5</td>
<td>99.4</td>
</tr>
<tr>
<td>Hammer</td>
<td>88.7</td>
<td>96.4</td>
<td>95.2</td>
<td>96.5</td>
<td>95.8</td>
<td>99.3</td>
<td>97.0</td>
</tr>
<tr>
<td>Powerbank</td>
<td>89.1</td>
<td>96.5</td>
<td>95.4</td>
<td>96.4</td>
<td>96.5</td>
<td>98.2</td>
<td>97.7</td>
</tr>
<tr>
<td>Scissors</td>
<td>84.5</td>
<td>92.9</td>
<td>90.2</td>
<td>87.5</td>
<td>92.6</td>
<td>94.8</td>
<td>90.1</td>
</tr>
<tr>
<td>Wrench</td>
<td>85.0</td>
<td>96.1</td>
<td>92.4</td>
<td>95.8</td>
<td>96.5</td>
<td>96.6</td>
<td>93.8</td>
</tr>
<tr>
<td>Gun</td>
<td>19.8</td>
<td>25.2</td>
<td>28.8</td>
<td>20.3</td>
<td>36.1</td>
<td>14.6</td>
<td>28.1</td>
</tr>
<tr>
<td>Bullet</td>
<td>86.7</td>
<td>95.9</td>
<td>96.2</td>
<td>94.0</td>
<td>94.1</td>
<td>98.1</td>
<td>95.3</td>
</tr>
<tr>
<td>Sprayer</td>
<td>74.5</td>
<td>63.7</td>
<td>76.5</td>
<td>61.3</td>
<td>48.9</td>
<td>73.1</td>
<td>79.2</td>
</tr>
<tr>
<td>Handcuffs</td>
<td>92.4</td>
<td>99.3</td>
<td>99.5</td>
<td>98.9</td>
<td>99.2</td>
<td>99.3</td>
<td>99.3</td>
</tr>
<tr>
<td>Knife</td>
<td>38.6</td>
<td>52.0</td>
<td>40.6</td>
<td>47.4</td>
<td>47.3</td>
<td>62.6</td>
<td>64.2</td>
</tr>
<tr>
<td>Lighter</td>
<td>78.0</td>
<td>86.4</td>
<td>84.8</td>
<td>84.1</td>
<td>84.5</td>
<td>87.6</td>
<td>85.9</td>
</tr>
<tr>
<td>mAP50</td>
<td>76.4</td>
<td>83.6</td>
<td>83.2</td>
<td>81.8</td>
<td>82.5</td>
<td>85.3</td>
<td>85.9</td>
</tr>
</tbody>
</table>

The proposed YOLOv8n-GEMA model’s detection performance on PIDray, SIXray, CLCXray, and HiXray datasets is illustrated in Figure 7, which includes the following: (a) a single sample image with obstructions, (b) a single sample image with cluttered item placement, (c) a single sample image with special imaging angles, (d) a multi-sample image, (e) an overlapping sample image, (f) a single sample image, (g) a multi-sample image with obstructions, (h) an overlapping sample image, and (i) a single sample image. Specifically, Figure 7, (a), (b), (c) are from the PIDray dataset, (d) and (e) from the SIXray dataset, (f) and (g) from the CLCXray dataset, and (h) and (i) from the HiXray dataset.

In Figure 7, (g) is the detection result of occlusion. Before the improvement, only four occlusion targets were identified, all of which were PlasticBottle. After the improvement, five occlusion targets were identified, and CartonDrinks were identified. In addition, (h) refers to overlapping detection results. The detected categories are Portable_Charger_1 and Mobile_Phone. The accuracy of the model before improvement is 0.74 and 0.78, and the accuracy of the model after improvement is 0.81 and 0.82, respectively. From the X-ray security images, it is evident that our proposed method can accurately detect hazardous substance targets. In Figure 7c,e,g, the inclusion of the GELAN module enhanced feature fusion capabilities. The EMA module, with its multi-scale attention mechanism, is able to capture and integrate features at different levels, enabling the model to effectively identify and locate targets even in the presence of obstructions. Specifically, the parallel $1 \times 1$ and $3 \times 3$ convolutional structures of the EMA module allow it to simultaneously handle local...
details and broader regional features, maintaining sensitivity and responsiveness to crucial features in complex scenarios such as overlapping or partially obstructed items.

Figure 7. Dangerous goods’ detection results by YOLOv8 and improved YOLOv8n-GEMA. (a) image with obstructions (PIDray). (b) image with cluttered item placement (PIDray). (c) image with special imaging angles (PIDray). (d) A multi-object image (SIXray). (e) An overlapping sample image (SIXray). (f) A single object image (CLCXray). (g) obstructions (CLCXray). (h) overlapping (HiXray). (i) single object (HiXray).
Because X-ray images of dangerous goods have the characteristics of clutter, occlusion, and different scales, case segmentation technology shows its unique advantages in these cases. Different from traditional object detection, instance segmentation can not only identify the object category in the image, but also accurately segment the independent pixel level region for each object instance in the image. This means that even if the objects overlap or are partially obscured, instance segmentation can accurately delineate the specific outline of each object, providing more detailed information to help judge the nature of the object. The use of case segmentation can more accurately detect the type and location of prohibited items, making security further improved. Therefore, this paper applies the above three improvements to segmentation, and adopts the PiDray dataset for training, in which each image contains at least one item of dangerous goods with bounding box and mask annotation. The detection results after training are shown in Figure 8.

![YOLOv8n vs YOLOv8n-GEMA](image)

**Figure 8.** The instance segmentation and comparison of X-ray detection of dangerous goods. (a) A multi-object image. (b) overlapping. (c) single object.

From Figure 8, it can be seen that in (a), there are two hazardous targets, both detected by YOLOv8 and YOLOv8n-GEMA, but YOLOv8n-GEMA achieved a higher confidence level than YOLOv8. In (b), there are two occluded hazardous items; both models correctly detected the scissors, but YOLOv8 incorrectly segmented the area of the knife, whereas YOLO-GEMA correctly framed it. In (c), which is a single hazardous item image, YOLOv8 failed to detect it, while YOLO-GEMA correctly detected it with a high confidence level.

In order to further determine the detection situation of different types of hazardous materials, we analyzed the AP, accuracy precision, recall, and F1 scores of the improved method proposed in this paper for each hazardous material category in the SIXray dataset, as shown in Table 3. From Table 3, it can be seen that the precision, accuracy, recall, and F1 score of the gun are the highest among all categories, because the gun has the largest number of samples in the SIXray dataset. The accuracy, recall rate, and F1 score of the knife are relatively low, which is due to the flat and slender structure of the knife itself, which is prone to stacking with other items, resulting in missed detections.

In the PiDray dataset, using the average values of various dangerous goods as a comprehensive evaluation metric, our method was compared with the original YOLOv8n, as shown in Figure 9.
Table 3. Improved model-performance analysis for each category.

<table>
<thead>
<tr>
<th>Categories</th>
<th>AP (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1 Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baton</td>
<td>97.5</td>
<td>97.5</td>
<td>99.2</td>
<td>94.2</td>
</tr>
<tr>
<td>Pliers</td>
<td>96.2</td>
<td>96.2</td>
<td>99.8</td>
<td>96.8</td>
</tr>
<tr>
<td>Hammer</td>
<td>95.1</td>
<td>95.1</td>
<td>93.0</td>
<td>89.7</td>
</tr>
<tr>
<td>Powerbank</td>
<td>89.3</td>
<td>89.3</td>
<td>95.9</td>
<td>88.5</td>
</tr>
<tr>
<td>Scissors</td>
<td>81.8</td>
<td>81.8</td>
<td>90.3</td>
<td>70.5</td>
</tr>
<tr>
<td>Wrench</td>
<td>66.4</td>
<td>66.4</td>
<td>97.5</td>
<td>90.1</td>
</tr>
<tr>
<td>Gun</td>
<td>52.5</td>
<td>52.5</td>
<td>12.5</td>
<td>19.7</td>
</tr>
<tr>
<td>Bullet</td>
<td>80.1</td>
<td>80.1</td>
<td>94.9</td>
<td>85.5</td>
</tr>
<tr>
<td>Sprayer</td>
<td>95.2</td>
<td>95.2</td>
<td>56.5</td>
<td>60.9</td>
</tr>
<tr>
<td>Handcuffs</td>
<td>98.1</td>
<td>98.1</td>
<td>99.5</td>
<td>96.3</td>
</tr>
<tr>
<td>Knife</td>
<td>67.9</td>
<td>67.9</td>
<td>60.6</td>
<td>45.5</td>
</tr>
<tr>
<td>Lighter</td>
<td>86.2</td>
<td>86.2</td>
<td>82.4</td>
<td>73.1</td>
</tr>
</tbody>
</table>

Figure 9. Performance comparison between YOLOv8n and YOLOv8n-GEMA: (a) mAP50; (b) mAP50-95; (c) train/box-loss; (d) precision; (e) recall; (f) val/box-loss.
It is evident that during 300 training cycles, the improved YOLOv8n-GEMA algorithm performed well in terms of precision, recall, mAP50, and mAP50-95. These metrics showed a noticeable improvement over YOLOv8n, indicating that our method is more efficient at detecting dangerous goods and is better suited for X-ray security inspection tasks. Furthermore, both training and validation loss values were lower than those of YOLOv8n, due to the inner-CIoU loss function which is specifically designed to address the issues of occlusion and overlapping in X-ray images. This loss function enhances the model’s ability to precisely locate targets in complex scenes by adding considerations for the distance between bounding box centers and shape differences on top of the IoU calculation. The design of inner-CIoU not only reduces the error between predicted and actual bounding boxes but also optimizes the model performance in environments with high occlusion and layered items, thus achieving lower loss values and faster convergence during training, effectively enhancing the detection performance and robustness of the model in X-ray security images.

4. Discussion

To validate the robustness and versatility of the proposed enhanced model, experiments were conducted on four X-ray security image datasets: SIXray, HiXray, CLCXray, and PiDray. The model’s performance was benchmarked against leading target detection algorithms including YOLOv4, YOLOv5s, YOLOv7, YOLOv8n, YOLOv10n, and RT-DETR. Among them was the RT-DETR model, which requires a computational capacity of 108 GFLOPs, compared to only 16.2 GFLOPs for our model. Despite the significant difference in computational demand—nearly seven times less—our algorithm still surpassed RT-DETR in performance. As demonstrated in Table 4, the proposed method outperformed these mainstream detection algorithms across all datasets. Compared to the baseline model YOLOv8n, mAP improved by 3.6%, 1.6%, 0.9%, and 3.4% on each dataset, confirming the effectiveness of the improved approach.

<table>
<thead>
<tr>
<th>Models</th>
<th>F</th>
<th>SIXray</th>
<th>HiXray</th>
<th>CLCXray</th>
<th>PiDray</th>
<th>SIXray</th>
<th>HiXray</th>
<th>CLCXray</th>
<th>PiDray</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv4</td>
<td>-</td>
<td>85.3</td>
<td>75.6</td>
<td>72.6</td>
<td>76.4</td>
<td>306.6</td>
<td>306.5</td>
<td>305.8</td>
<td>305.0</td>
</tr>
<tr>
<td>YOLOv5s</td>
<td>-</td>
<td>93.1</td>
<td>81.0</td>
<td>88.3</td>
<td>83.6</td>
<td>355.0</td>
<td>348.2</td>
<td>348.0</td>
<td>348.1</td>
</tr>
<tr>
<td>YOLOv7</td>
<td>-</td>
<td>91.1</td>
<td>80.5</td>
<td>86.5</td>
<td>83.2</td>
<td>398.2</td>
<td>397.5</td>
<td>397.2</td>
<td>397.2</td>
</tr>
<tr>
<td>YOLOv8n</td>
<td>-</td>
<td>90.8</td>
<td>80.4</td>
<td>88.0</td>
<td>82.5</td>
<td>787.3</td>
<td>778.7</td>
<td>783.4</td>
<td>789.0</td>
</tr>
<tr>
<td>YOLOv8s</td>
<td>-</td>
<td>94.0</td>
<td>82.3</td>
<td>88.6</td>
<td>86.1</td>
<td>211.5</td>
<td>211.0</td>
<td>211.1</td>
<td>211.1</td>
</tr>
<tr>
<td>YOLOv9</td>
<td>-</td>
<td>91.4</td>
<td>80.9</td>
<td>89.0</td>
<td>83.5</td>
<td>756.3</td>
<td>756.3</td>
<td>756.6</td>
<td>756.5</td>
</tr>
<tr>
<td>YOLOv10n</td>
<td>-</td>
<td>91.6</td>
<td>81.1</td>
<td>88.0</td>
<td>82.5</td>
<td>801.1</td>
<td>801.3</td>
<td>802.0</td>
<td>802.0</td>
</tr>
<tr>
<td>RT-DETR</td>
<td>-</td>
<td>93.8</td>
<td>82.3</td>
<td>86.8</td>
<td>85.3</td>
<td>91.9</td>
<td>91.9</td>
<td>92.0</td>
<td>91.9</td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
<td>94.4</td>
<td>82.0</td>
<td>88.9</td>
<td>85.9</td>
<td>377.7</td>
<td>376.9</td>
<td>377.0</td>
<td>377.0</td>
</tr>
</tbody>
</table>

Ablation experiments were conducted to verify the effectiveness of integrating various modules into the network, as presented in this paper. The experimental outcomes are shown in Table 5. Method A utilizes the GELAN structure, Method B incorporates the EMA attention mechanism, and Method C involves replacing the loss function with inner-CIoU. Furthermore, ablation studies highlight the contribution of each component to the overall performance increase. These studies clearly indicate that each individual enhancement plays a crucial role in the robustness and accuracy of the model. Not only does the proposed model reduce false positives and false negatives, but it also shows remarkable stability in performance across different types of X-ray security images, which often vary in quality and complexity. These experiments demonstrate the validity of the proposed enhancements in the network configuration.
Table 5. Results of ablation experiment.

<table>
<thead>
<tr>
<th>Models</th>
<th>SIXray mAP50 (%)</th>
<th>SIXray 50–95 (%)</th>
<th>HiXray mAP50 (%)</th>
<th>HiXray 50–95 (%)</th>
<th>CLCXray mAP50 (%)</th>
<th>CLCXray 50–95 (%)</th>
<th>PIDray mAP50 (%)</th>
<th>PIDray 50–95 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOLOv8n</td>
<td>90.8</td>
<td>65.2</td>
<td>80.4</td>
<td>50.6</td>
<td>88.0</td>
<td>76.4</td>
<td>81.8</td>
<td>70.7</td>
</tr>
<tr>
<td>YOLOv8n+A</td>
<td>93.5</td>
<td>70.8</td>
<td>81.4</td>
<td>52.2</td>
<td>88.8</td>
<td>78.8</td>
<td>83.8</td>
<td>74.6</td>
</tr>
<tr>
<td>YOLOv8n+B</td>
<td>93.7</td>
<td>71.2</td>
<td>81.9</td>
<td>52.2</td>
<td>88.2</td>
<td>79.0</td>
<td>85.0</td>
<td>74.7</td>
</tr>
<tr>
<td>YOLOv8n+C</td>
<td>92.0</td>
<td>67.6</td>
<td>80.4</td>
<td>50.7</td>
<td>88.0</td>
<td>77.5</td>
<td>81.4</td>
<td>69.2</td>
</tr>
<tr>
<td>YOLOv8n+A+B</td>
<td>93.4</td>
<td>71.3</td>
<td>81.6</td>
<td>52.3</td>
<td>88.7</td>
<td>79.4</td>
<td>84.6</td>
<td>74.9</td>
</tr>
<tr>
<td>YOLOv8n+A+C</td>
<td>94.3</td>
<td>72.2</td>
<td>82.3</td>
<td>52.4</td>
<td>88.7</td>
<td>78.9</td>
<td>84.6</td>
<td>75.4</td>
</tr>
<tr>
<td>YOLOv8n+B+C</td>
<td>94.0</td>
<td>71.3</td>
<td>82.5</td>
<td>52.4</td>
<td>88.6</td>
<td>79.3</td>
<td>85.7</td>
<td>75.4</td>
</tr>
<tr>
<td>YOLOv8n+A+B+C</td>
<td>94.4</td>
<td>72.4</td>
<td>82.0</td>
<td>52.3</td>
<td>88.9</td>
<td>78.9</td>
<td>85.9</td>
<td>75.9</td>
</tr>
</tbody>
</table>

Figure 10 illustrates the training process of the ablation experiment, which reveals that the enhanced techniques introduced in this study effectively enhance the model’s accuracy in detecting dangerous goods, thereby confirming the performance superiority of the proposed method. YOLOv8n has a computational load of 8.9 GFLOPs, while the GELAN module, based on YOLOv8n, requires 16.1 GFLOPs, adding an additional 7.2 GFLOPs.

We conducted an experiment to evaluate the improvement in detecting overlapping items using the EMA module compared to YOLOv8n, as shown in Figure 11.

It can be seen from the ablation experiment that the inner-CIoU loss function is effective in improving the precision. By comparing YOLOv8n+A+B with YOLOv8n+A+B+C, the inner-CIoU loss function is adopted in SIXray, HiXray, CLCXray, and PIDray increased by 1.0%, 0.4%, 0.2% and 1.3%, respectively.

Figure 11. Visualization results of ablation experiment training process.
Figure 10. Visualization results of ablation experiment training process.

We conducted an experiment to evaluate the improvement in detecting overlapping items using the EMA module compared to YOLOv8n, as shown in Figure 11.

![YOLOv8n](image1.png)  ![YOLOv8n + EMA](image2.png)

(a) (b) (c)

Figure 11. A comparison of the test results with and without EMA. (a) Single target; (b) Multi targets; (c) Targets of similar shapes.

From Figure 11, it is evident that the EMA module enhances the detection of overlapping items compared to YOLOv8n. In (a), which shows single-object occlusion detection, the YOLOv8n model failed to detect the scissors, whereas YOLOv8n with the EMA module successfully detected them. In (b), which involves multi-object overlap detection, YOLOv8n detected only the baton and missed the handcuffs; however, with the EMA module, both items were detected. In (c), for another multi-object overlap scenario, YOLOv8n detected only the scissors and missed the knife, whereas the EMA module enabled the detection of both items.

To better demonstrate the state of X-ray images of dangerous goods within the network, Figures 12 and 13 show the feature maps after the first convolution layer and after passing through the EMA module, respectively. In Figure 11, basic edge detection is displayed along with deeper features of the backpack’s contents. As the layers of the network deepen, the features extracted by each layer become increasingly abstract. Higher layers contain less information about the specific input and more information about the targets.

![Feature Maps](image3.png)

Figure 12. Feature maps obtained after the input image passes through the first layer.
Electronics 2024, 13, x FOR PEER REVIEW 22 of 25 successfully detected them. In (b), which involves multi-object overlap detection, YOLOv8n detected only the baton and missed the handcuffs; however, with the EMA module, both items were detected. In (c), for another multi-object overlap scenario, YOLOv8n detected only the scissors and missed the knife, whereas the EMA module enabled the detection of both items.

To better demonstrate the state of X-ray images of dangerous goods within the network, Figures 12 and 13 show the feature maps after the first convolution layer and after passing through the EMA module, respectively. In Figure 11, basic edge detection is displayed along with deeper features of the backpack’s contents. As the layers of the network deepen, the features extracted by each layer become increasingly abstract. Higher layers contain less information about the specific input and more information about the targets.

Figure 12. Feature maps obtained after the input image passes through the first layer.

Figure 13. Feature maps obtained after the input image passes through the EMA module.

5. Conclusions

In this paper, we have modified and evaluated an enhanced detection model, YOLOv8n-GEMA, which significantly advances the field of X-ray security images for detecting dangerous goods. The proposed model incorporates three novel components: the generalized efficient layer aggregation network, the efficient multi-scale attention module, and the inner-CIoU loss function, each tailored to address specific challenges in the detection process.

The GELAN architecture optimizes feature extraction and learning efficiency, which is crucial for the complex visual scenarios typically encountered in X-ray imagery. This network structure facilitates better gradient flow and reduces computational load, making it ideal for real-time application in security checkpoints. The EMA module enhances the model’s ability to discern relevant features from cluttered and overlapping items, a common issue in the X-ray images of luggage. By focusing on effective channel and spatial descriptions, EMA allows the model to maintain high accuracy even under varying image conditions. Lastly, the inner-CIoU loss function introduces adjustments in bounding box predictions that are more sensitive to the unique sizes and shapes of different objects, improving the precision and reliability of the model.

Experimental results from testing on four distinct X-ray image datasets—SIXray, HiXray, CLCXray, and PIDray—demonstrate the superiority of the YOLOv8n-GEMA model over existing state-of-the-art methods like YOLOv4, YOLOv5s, YOLOv7, and YOLOv8n. Significant improvements in mAP were observed across all four datasets, confirming the effectiveness of the integrated enhancements. In future, the use of pruning techniques to remove redundant network parameters and connections will reduce the computational resource requirements of the model, which allows for an increase in detection speed without sacrificing accuracy.
Author Contributions: Conceptualization, P.Y., Y.I., H.W., Y.L. and A.W.; methodology, P.Y.; software P.Y.; validation Y.I., Y.L. and P.Y.; writing—review and editing Y.L., P.Y., Y.I., H.W. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Key Research and Development Plan Project of Heilongjiang (JD2023SJ19), the Natural Science Foundation of Heilongjiang Province (LH2023F034), the Science and Technology Project of Heilongjiang Provincial Department of Transportation (HJK2024B002) and the high-end foreign expert introduction program (G2022012010L).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.


Conflicts of Interest: The authors declare no conflicts of interest.

References

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.