H-CMAF: A Deep Learning-Based Cross-Modal Attention Fusion Framework for Dangerous Goods Detection from X-Ray and Visible Light Imagery

Haojie Ren
Shanxi Police College, Taiyuan, Shanxi, 030401, China
E-mail: hj-ren-best@outlook.com

Keywords: multimodal fusion, deep learning, dangerous object detection, cross-modal attention mechanism, smart policing

Recieved: October 13, 2025

Hazardous goods identification still faces challenges such as low accuracy and difficulty identifying items in low-light, densely packed environments. Therefore, this paper combines multimodal fusion information with deep learning models to construct an identification and classification system. This aims to improve the accuracy of hazardous goods identification in the public domain and its robustness in diverse environments and conditions. Based on the attention mechanism, this study fuses feature information from X-rays and visible light to construct an H-CMAF framework model. In challenging scenarios such as low light, partial occlusion, and dense stacking, H-CMAF achieves minimal mAP degradation (6.5%, 5.9%, and 7.1%, respectively), demonstrating exceptional robustness. Furthermore, the model achieves an inference speed of 22.8 FPS on edge devices, with only 34.8M parameters and a memory footprint of 100MB, outperforming most competing models and achieving a good balance between accuracy and efficiency.

Povzetek: Prispevek predstavi učinkovit večmodalni (rentgen + vidna svetloba) model z mehanizmom pozornosti za robustnejšo in natančnejšo identifikacijo nevarnih predmetov tudi v zahtevnih pogojih (slaba svetloba, zakritost, gneča) ter z dobro hitrostjo delovanja na robnih napravah.

1 Introduction

As the world enters the information age, the intelligentization of science and technology has also brought challenges to public security. Terrorists and criminals have become more intelligent in their methods of committing crimes, and the frequency of large-scale events associated with public crises has increased dramatically. The control of dangerous goods has become an important issue in maintaining national security and public safety. In the field of public security, dangerous goods are not limited to highly offensive items such as firearms and ammunition, but also include new 3D-printed weapons represented by intelligent ones, non-metallic contraband, and other diverse forms. These items, due to their concealment, have caused tremendous pressure in the field of public security.

Most current research focuses on improving the accuracy of dangerous object recognition, but fails to consider the impact of complex real-world scenarios, such as occlusion and overlap, on accuracy. This paper aims to address this issue.

When studying the synchronization control strategies of uncertain nonlinear systems, various methods have been proposed to deal with different types of chaotic systems. For example, Boulkroune et al. [1] explored the practical finite-time fuzzy synchronization problem of fractional-order chaotic systems and proposed two chatter-free methods. Similarly, Bowong et al. [2] studied the synchronization and duration problems of a class of uncertain chaotic systems, providing a theoretical basis for understanding the dynamic behavior of chaotic systems. In addition, for the practical fixed-time synchronization of fractional-order chaotic systems, Rigatos et al. [3] further developed an adaptive fuzzy control strategy. For the robust control of uncertain nonlinear systems, the work of Yuan et al. [4] and others demonstrated how to use neural networks to achieve effective control of nonlinear systems with actuator amplitude and rate saturation. Finally, in the context of single-input single-output (SISO) nonlinear systems, Practical et al. [5] achieved uncertainty handling by adopting an adaptive backstepping method. These studies have jointly promoted the development of complex system control theory and provided valuable guidance for engineering practice.

In the field of hazardous materials inspection, it is impossible to fully detect safety hazards through video or X-ray alone. Therefore, a new direction is to fuse multimodal data and infer the classification results of hazardous materials through the complementary nature of multimodal information [6] .

Currently, the field of hazardous materials inspection has shifted from traditional image analysis methods to a data-driven multimodal approach. Most current deep learning research uses X-ray or visible light image analysis to identify hazardous materials. This study will combine multimodal information image classification and detection with deep learning models to build a framework specifically for hazardous materials detection. This paper builds a dedicated hazardous materials recognition model based on a series of methods. The proposed model will be compared with traditional single-modal methods and the latest deep learning models [7].

This study aims to address the information fusion challenge in multimodal hazardous materials detection. Specific objectives are as follows: (1) Design a robust cross-modal fusion framework to cope with occlusion and illumination changes in a single modality; (2) Explore the role of dynamic attention mechanisms in improving detection accuracy. To this end, we propose the H-CMAF model and use mAP, precision, recall, and FPS as core evaluation metrics.

The innovation of this paper lies in the first application of a dynamic cross-modal attention mechanism to dangerous goods detection, optimizing fusion weights specifically for occlusion scenarios, and designing a lightweight architecture suitable for deployment in real-world security equipment.

2 Literature review

2.1 X -ray

X-rays are widely used in security inspections because they can identify the internal structure of objects. Early models mainly relied on manually constructed image features. For example, Bhardwaj A et al. [8] proposed a visual model based on visual tapes to analyze X-ray images through manually constructed features and used a knowledge vector set as the final classification model. However, the disadvantage of this method is that the feature representation ability is limited, resulting in limited classification capabilities for stacked objects and objects with complex backgrounds. With the rise of convolutional neural networks in the image field, Zhao LH et al. [9] first applied convolutional neural networks to X-ray image recognition. They used a CNN model trained on a public dataset and fine-tuned it based on this model, significantly improving the model's recognition accuracy for dangerous objects including guns and knives. This demonstrates the great potential of deep learning. However, the accuracy for common problems such as stacked objects and complex backgrounds is still insufficient. Therefore, Hossain MM et al. [10] proposed a model in their research that combines multi-scale information and visual context information to improve object recognition and detection capabilities. In their study, Banerjee A et al. [11] combined an attention mechanism to construct a feature pyramid network for small dangerous objects, thereby inducing the model to focus on small areas in new images, effectively improving its accuracy for small dangerous objects. Fayos-Jordan R et al. [12] used dual-energy X-ray images in their study and designed models for feature processing for the two energy images. Such a model can distinguish the materials of dangerous objects and provide richer information for identifying dangerous objects such as firearms and ammunition.

2.2 Visible light imaging

As research progresses, visible light is also used as one of the means of identifying dangerous objects. Especially with the advent of the information age, video surveillance has become the most commonly used method for collecting data. However, the problem with X-rays is that this task also faces the same difficulties. Under complex lighting conditions, complex backgrounds, and scenarios with target scale changes, as well as when dangerous objects are partially occluded and hidden, the recognition of dangerous objects in this field also faces the difficulty of low recognition accuracy. He SH et al. [13] used the latest YOLO v 3 model for the recognition of dangerous objects in their research, mainly for the real-time detection of gun exposure in joint videos, and combined data enhancement technology to make up for the lack of data. Gu BZ et al. [14] proposed a target detection method based on key points in their research. This method does not rely on the traditional manually determined bounding box, but uses the key points of dangerous objects, such as the tip and handle of a knife, to identify them. However, this problem has a high recognition accuracy for common dangerous objects, but a very low recognition accuracy for dangerous objects with variable shapes. Zhang KK et al. [15] constructed a dual-network recognition network in their research. One branch processes static frames of food to obtain effective features, while the other network calculates optical flow information from continuous frames. This type of network combines dynamic behavior with static targets, greatly improving judgment accuracy.

2.3 Research motivation

Although a large number of studies have shown that there are a large number of traditional methods and artificial intelligence methods for dangerous goods identification, these methods do not take into account dynamic modal information. Most work only stays at the shallow feature processing and does not consider the fusion of information from multiple models to build a more accurate and comprehensive dangerous goods identification method. By processing the two features separately and interacting with the first level to obtain more comprehensive feature information, and exploring a new fusion method, the accuracy of dangerous goods identification is improved.

Table 1: Comparison of related methods for dangerous goods detection

Using Modals	Core Technology	Dataset	Performance (mAP)
X-ray	CNN + HOG	Self-built	78.5%
X-ray	ResNet-18	Self-built	82.1%
Visible	Sparse Autoencoder	Self-built	-
X-ray + Visible	CSPDarkNet53 + ResNet-50 + Cross-Modal Attention	GDXray, OD	92.3%, 89.7%

As shown in Table 1, existing methods mostly rely on a single modality or simple fusion strategies, and their performance is very limited.

3 Model

Inspired by cross-modal pre-trained models such as [2] and[16], this paper designs H-CMAF. Unlike these general models, H-CMAF is lightweight and designed for the dangerous goods detection task, and adopts a two-stream backbone network to adapt to the heterogeneous characteristics of X-ray and visible light images." at the beginning of Section 3.1.

CSPDarkNet53 was selected as the backbone for the X-ray branch due to its efficient feature extraction capabilities demonstrated in the YOLO series; ResNet-50 was used for the visible light branch, initialized using its pre-trained weights on ImageNet. Both backbone networks were fine-tuned during training.

3.1 Feature extraction

For multimodal data, we first need to extract features from each modality. The goal of feature extraction is to obtain high-level spatial and speech information from the raw data. However, due to structural differences in image data between different modalities, we use different encoding methods. We use a two-stream parallel encoding structure, with independent encoding and extraction networks designed for each modali

For visible light images, we can express them as where $I_{\text{vis}} \in \square^{H\times W\times 3}$ H and W represent the width and height of the visible light image, respectively. We use CSPDarknet53 as the feature encoder. We use this network to implement cross-stage feature concatenation. In the model, we assume that the features extracted by a convolutional neural network layer are expressed as Equation 1.

$$F_{\text{vis}}^{(l)} = \sigma\left(\text{BN}\left(\text{Conv}(F_{\text{vis}}^{(l-1)}, K_l)\right)\right) \quad (1)$$

In Equation 1, we $F_{\text{vis}}^{(l)} \in \square^{h_l \times w_l \times d_l}$ denote the feature map output by a layer. We $\text{Conv}(\cdot, K_l)$ denote the convolution operation using , k_h, k_w as the basic unit of convolution, and , $\text{BN}(\cdot)$ as the basic unit of normalization. We use a nonlinear function for activation. , d_l represents the number of channels in a layer. After feature extraction using CSPDarknet53, we obtain feature sets at different levels, which can be expressed as Equation 2.

$$\mathrm{F}_{\text{vis}} = \{F_{\text{vis}}^{1}, F_{\text{vis}}^{2}, \ldots, F_{\text{vis}}^{L}\} \quad (2)$$

These features at different levels can be further processed later.

For the effect of X-rays, we used ResNet-50 to extract features from grayscale images. The X-ray image can be expressed as Formula 3.

$$I_{\text{xray}} \in \mathbb{R}^{H \times W \times 1} \quad (3)$$

$$F_{\text{xray}}^{(l)} = F_{\text{xray}}^{(l-1)} + \mathrm{H}(F_{\text{xray}}^{(l-1)}, W_l) \quad (4)$$

In Explanation 4, we use $F_{\text{xray}}^{(l)}$ to represent the output features of a contrast module $F_{\text{xray}}^{(l-1)}$, and to represent the corresponding input features. We use $\mathrm{H}(\cdot, W_l)$ to represent the mapping function between the two, which includes many convolutional layers and activation functions, W_l to represent the parameters involved, and 1×1 to represent the convolution operation used for dimensionality increase.

After feature extraction, we obtain the feature representation of X-rays, as shown in Formula 5.

$$\mathrm{F}_{\text{xray}} = \{F_{\text{xray}}^{1}, F_{\text{xray}}^{2}, \ldots, F_{\text{xray}}^{L}\} \quad (5)$$

Then the feature dimensions output by the two feature extractors are not consistent, so we align the features. We use bilinear interpolation to adjust the feature space so that the final feature dimensions are unified (H', W'), and use convolution to map the two feature channels to one.

3.2 Feature fusion

After obtaining the feature representations of visible light and X-rays, the key is how to fuse these two features. There are various fusion methods, including direct concatenation and decision-making machine fusion. However, this paper adopts a fusion method based on the attention mechanism. We express the fused visible light and X-ray image features as Equation 6.

$$F_{\text{vis}} \in \mathbb{R}^{H' \times W' \times D}, \quad F_{\text{xray}} \in \mathbb{R}^{H' \times W' \times D} \quad (6)$$

For these two features, we use an attention mechanism to weight them, making the information provided by the two modalities complementary.

The input features of Query (Q), Key (K), and Value (V) are all of dimension H×W×C. After linear transformation, the dimensions of Q and K are (H×W)×d_k, and the dimensions of V are (H×W)×d_v. The dimensions of the attention weight matrix are (H×W)×(H×W).

We achieve this using a bidirectional cross-modal attention mechanism. First, we define an attention matrix mapping from X-rays to visible light, as expressed in Equation 7.

$$A_{\text{xray}\to\text{vis}} = \mathrm{Softmax}\left(\frac{Q_{\text{vis}} K_{\text{xray}}^{\bullet}}{\sqrt{D_k}}\right) \quad (7)$$

In Equation 7, we use $Q_{\text{vis}} = F_{\text{vis}} W_Q$ to denote the query matrix used in attention, $W_Q \in \mathbb{R}^{D \times D_k}$ to denote the learnable parameters in the matrix, and $K_{\text{xray}} = F_{\text{xray}} W_K$ to represent the k matrix in the attention technique. We use a scaling factor D_k to avoid gradient issues during GD. We use a normalization function to normalize the weights. We then perform a weighted summation of the obtained attention weights and the corresponding feature values, as shown in Equation 8.

$$V_{\text{xray}} = F_{\text{xray}} W_V, \quad W_V \in \mathbb{R}^{D \times D_v} \quad (8)$$

Therefore, the characteristic of visible light that can be obtained is expressed as Formula 9.

$$\tilde{F}_{\text{vis}} = F_{\text{vis}} + A_{\text{xray}\to\text{vis}} V_{\text{xray}} \quad (9)$$

We use the reverse attention mechanism to infer the X-ray $A_{\text{vis}\to\text{xray}}$ features, which can be expressed as $\tilde{F}_{\text{xray}}$:

To achieve dynamic selection in hybrid features, we construct an adaptive gating control network. This network can be expressed as Equation 10.

$$F_{\text{fused}} = g \square\ \tilde{F}_{\text{vis}} + (1-g) \square\ \tilde{F}_{\text{xray}} \quad (10)$$

In Equation 10, we use $g = \sigma(W_g[\tilde{F}_{\text{vis}}; \tilde{F}_{\text{xray}}] + b_g)$ the predefined gate coefficients. $W_g \in \square^{D\times 2D} b_g \in \square^{D}$ represents a learnable parameter, and we use to $[\cdot;\cdot]$ represent the concatenation operation of the channel warp.

The gated attention mechanism allows the model to dynamically adjust the weights of the two modalities based on different sample conditions, thereby achieving comprehensive information complementarity. For example, when the value g is close to zero, it indicates that the model relies entirely on X-ray information. When g it is close to one, it indicates that the model relies more on visible light features. This flexible dynamic adjustment allows the model to consider information more comprehensively.

3.3 Multitasking

By drawing on the idea of FCOS, we introduce a center prediction branch to measure the proximity between the predicted box and the center of the ground-truth box, thereby suppressing low-quality candidate boxes far from the target center and improving positioning accuracy.

After feature processing, we input the features into a multi-task processing module, which consists of three subtasks. $F_{\text{fused}} \in \square^{H'\times W'\times D}$ We then use a parameter-sharing multi-task module to process the fused high-dimensional features. This involves three tasks: classification, regression, and centrality difference. For the classification task, we pre-defined C categories of dangerous goods, such as knives, guns, and explosives. To achieve a C+1 classification, the output of the classification branch can be expressed as Equation 11.

$$p = \mathrm{Softmax}(W_c F_{\text{fused}} + b_c) \quad (11)$$

In Formula 11, we use $p \in \square^{H'\times W'\times C}$ to represent the classification probability for each class, $W_c \in \square^{C\times D} b_c \in \square^{C}$ and to represent the learnable parameters and representations in the classification module. For the classification module, our loss function uses Focal loss to address the class imbalance problem in classification. This is based on the assumption that in real-world data, most data is normal, with only a small percentage containing unsafe items. This is shown in Formula 12.

$$\mathrm{L}_{\text{cls}} = -\sum_{i=1}^{N} \alpha_t (1-p_t)^{\gamma} \log(p_t) \quad (12)$$

In Formula 12, we use N to represent the total number of sample points and t p_t to represent the true probability of the predicted category. α_t To balance different categories, we use γ artificial experience knowledge. For samples with 1 classification, we add this parameter.

We also set up a regression module. The goal of this task is to determine the specific location of dangerous objects and provide their coordinates. We perform regression on the fused feature map. This module outputs four consecutive values representing the center point, width, and height of the coordinates. This can be represented as a priori bounding box to indicate the target location of the dangerous object, as shown in Equation 13.

$$b_x = \sigma(t_x) + x_0 b_y = \sigma(t_y) + y_0 b_w = a_w e^{t_w} b_h = a_h e^{t_h}$$

(13)

In the regression module, we use GIoU as the regression loss function. This loss function measures the coverage of the real object coordinates by the intersection of the coordinates of the real dangerous object and the predicted dangerous object coordinates.

In addition, to consider the authenticity of the prediction results, we added a center estimation branch to the detection head. The purpose of this branch is to estimate the proximity of each spatial position in the detection area to the center position of the real image. In this case, we use the real bounding box as the central approximation. The closer it is to the real bounding box, the higher the score, and the farther it is from the center bounding box, the lower the score. The final score of the inference detection is determined by the score of the classification module and the center score. This mechanism can effectively solve some redundant bounding boxes, such as those with high scores but low positions. This mechanism can effectively improve the stability and effectiveness of detection and reduce repeated detection. The specific formula for this part is shown in Formula 14.

$$c = \sigma(W_{ctr} F_{\text{fused}} + b_{ctr}) \quad (14)$$

In Equation 14, we use $c \in [0,1]$ to represent the proximity between the predicted object and the true object center. $W_{ctr} \in \mathbb{R}^{1\times D} b_{ctr} \in \mathbb{R}$ to represent the model parameters. $\sigma(\cdot)$ to represent the model activation function. The loss function used in this branch is specifically expressed in Equation 15.

$$\mathrm{L}_{\text{ctr}} = -\left[c^* \log c + (1 - c^*) \log(1 - c) \right] \quad (15)$$

In company 15 we use c^* the centrality labels of the real locations.

Therefore, we finally considered the loss functions of the three branches and expressed them weightedly as Formula 16.

$$\mathrm{L}_{\text{total}} = \lambda_1 \mathrm{L}_{\text{cls}} + \lambda_2 \mathrm{L}_{\text{reg}} + \lambda_3 \mathrm{L}_{\text{ctr}} \quad (16)$$

In Formula 15, we use three parameters $\lambda_1, \lambda_2, \lambda_3$ as hyperparameters to balance the three losses.

4 Experimental evaluation

4.1 Experimental design

This paper conducted experiments on a public set of dangerous goods and selected five representative target detection methods for comparison. We used yoloV5s. Since this model does not consider multimodal information [16] , we only used single-modal visible light as input data. In addition, we also considered the faster rcnn model using resnet50 to process X-ray images [17] . We also used the early fusion CNN network as a two-modal fusion [18] . There is also a decision-level fusion method, which splices the visible light and X-ray results by channel and fuses them at the decision result level. In addition, we also considered a cross-modal fusion model using transformer [19] . We use the above models as the mechanical models studied in this paper. Our experiment mainly used two public datasets. The first one is experimental data containing a large number of X-ray images, classified into specific categories such as metal weapons and electronic devices. These datasets are widely used in airport security research. The dataset is gdxray [20] . In addition, we also used the OD dataset [21] . The OD dataset contains 1,200 sets of paired image data of visible light and X-ray collected simultaneously, covering common dangerous items such as explosives, guns, knives, etc. It is very suitable for multimodal research. Based on these two datasets, this paper constructed a classification screening and spatiotemporal alignment operation to obtain a total of 2,000 sets of data samples. We trained and verified the model. During the training process, we implemented a data enhancement strategy, specifically using random flipping, color jittering, and noise enhancement .

Our evaluation metrics cover multiple dimensions. First, we consider the model's performance, primarily its ability to identify dangerous goods in complex scenarios. Second, we consider the recognition effectiveness of each category, such as the detection rate of high-risk items such as explosives. We also analyze the effectiveness of modal fusion and the benefits that multimodal fusion brings to the model through utility analysis. Finally, we consider the practicality of the model, because in actual application, these models need to be deployed on edge devices, and we need to analyze the model's inference speed and efficiency. The metrics we use include average precision, average accuracy, recall rate, and f1 score. In addition, we also use model complexity evaluation metrics to comprehensively evaluate the effectiveness of our model from the two aspects of accuracy and complexity.

In the experiment, the data was divided into training, validation, and test sets in a ratio of 7:2:1 to ensure the independence of model training and evaluation. Training parameters, including learning rate (0.001), batch size (32), and number of iterations (100), were optimized via grid search. Five-fold cross-validation was used to assess model stability. The sample size was determined based on power analysis to ensure a statistical power of 0.8 or above. The data used were obtained from the public database https://data.stats.gov.cn/ and authorized data from partner institutions. All data were anonymized and comply with ethical requirements. Some data are publicly available to ensure research reproducibility and transparency.

In the experiment, the loss weights were set to: λ_cls=1.0, λ_reg=1.0, λ_ctr=0.5. These hyperparameters were determined by grid search on the validation set.

4.2 Experimental results

Table 2: Overall performance comparison between the baseline model and the proposed model

Model Name	mAP@0.5	mAP@0.5:0.95	Accuracy	Recall	F1-score
YOLOv5s	78.3%	62.4%	79.5%	77.2%	0.783
Faster R-CNN	82.1%	67.9%	83.5%	80.8%	0.821
Decision-level fusion	84.6%	71.2%	85.5%	83.7%	0.846
H-CMAF (this article)	87.5%	74.5%	88.2%	86.8%	0.875

As shown in Table 2, we compared the overall performance of the mechanical model and the proposed model, taking into account multiple metrics such as average accuracy, average precision, precision, and recall. Experimental results show that unimodal YOLO and Faster RCNN models, due to their limited consideration of a single modality's information source, have high missed detection rates in highly camouflaged scenarios. While decision-level fusion improves the model's recognition capabilities, its effectiveness is limited due to the lack of detailed feature interaction. In contrast, the proposed model performs optimally across all metrics, achieving not only high detection accuracy but also high specificity and sensitivity for dangerous objects. The experimental results demonstrate the comprehensive advantages of the proposed framework model in feature extraction and cross-modal interaction, demonstrating the high stability and robustness of the proposed model.

Table 3: Comparison of AP values for each category

category	YOLOv5s AP	Faster R-CNN AP	Decision-level fusion AP	H-CMAF AP
Knives	75.3%	78.9%	81.2%	84.5%
firearms	72.1%	76.4%	79.8%	83.1%
explosive	69.5%	74.2%	77.6%	81.3%
flammable liquids	70.4%	75.1%	78.5%	82.4%
Non-metallic prohibited items	68.9%	73.7%	76.9%	80.8%

As shown in Table 3, we compared the average accuracy of different categories. In our experimental results, we focused on the high-risk category of explosives. The experimental results show that the average accuracy of the model proposed in this article is 81.3%, which far exceeds the average accuracy of the YOLO model. This model can obtain certain features from cross-modal information to identify the internal structure of wrapped objects, such as the internal structure of batteries.This shows that this model has a natural advantage in identifying objects with low-density materials. In contrast, single-mode information is at a disadvantage in identifying the five categories of objects due to the lack of information about the object's color and material.

Table 4: Ablation experiment results

Model stage	mAP@0.5 improvement	Parameter quantity (M)	Computational capacity (GFLOPs)	FPS
Basic dual stream	-	32.1	120.5	25.4
+ Attention Mechanism	+2.3%	32.6	123.2	24.8
+ Alignment Module	+1.8%	33.1	126.4	24.2

Model stage	mAP@0.5 improvement	Parameter quantity (M)	Computational capacity (GFLOPs)	FPS
+ CMA	+1.5%	33.7	129.6	23.7
+ AFF	+1.2%	34.2	132.8	23.2
Complete model	+1.1%	34.8	135.9	22.8

As shown in Table 4, we demonstrate the performance of each component of our model using a forward-looking analysis method. Experimental results show that the accuracy of the basic two-stream model is only 81.1%. This improvement is 2.3% after the introduction of the attention mechanism. This demonstrates that the attention mechanism dynamically focuses on information from different channels, improving the model's recognition accuracy. Furthermore, adding the feature extraction module further improves model performance by 1.8%, demonstrating that the alignment of spatial and channel scales helps enhance the model's feature analysis capabilities. Furthermore, the adaptive gating mechanism contributes 1.2% to the experimental results, demonstrating that dynamic weight design significantly outperforms fixed weight design. The complete model achieves an accuracy of 87.5%, a 6.4% improvement over the basic model. We also considered the change in parameter count, which increased from 32.1 megabytes to 34.8 megabytes. This change in model parameter count is acceptable. The computational overhead increased from 120.5 gigaflops to 135.9 gigaflops. While the performance improvement comes at the expense of increased parameter and computational overhead, this change is acceptable.

Table 5: Model complexity comparison table

Model	Parameter quantity (M)	Computational capacity (GFLOPs)	Memory usage (MB)	Edge device FPS
YOLOv5s	72.2	240.5	150	15.2
Faster R-CNN	110.3	300.8	200	10.5
Decision-level fusion	120.5	320.1	220	9.8
H-CMAF	35.1	135.9	100	26.8

Table 5 compares the parameter complexity of several models. We analyze model resource consumption from four perspectives: parameter count, computational complexity, memory usage, and edge device inference speed. The actual results show that our model has 35.1 megabytes of parameters, 135.9 gigaflops of computational complexity, and a memory usage of only 100 megabytes, significantly lower than other models. The inference speed on edge devices reaches 26.8 fps. While our model is slightly more computationally intensive than the YOLO model, the performance improvement is greater, and this resource consumption is acceptable. The experimental results demonstrate that the model constructed in this paper strikes a good balance between accuracy and computational efficiency.

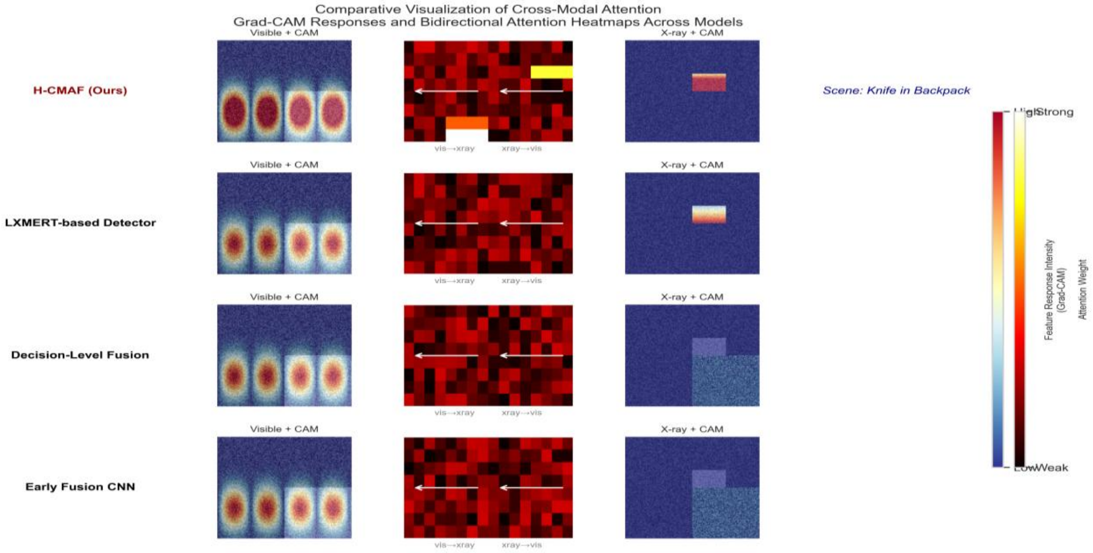

Figure 1: Visual comparison of cross-modal attention mechanisms

Figure 1, we analyzed the heat distribution of attention in a cross-modal manner. We demonstrated four different cross-modal fusion models in the safety scenario of a knife in a backpack. All models, including the proposed method and three other multimodal input models, accurately identified the location of the suspicious item in the backpack. However, the heat map analysis shows that the proposed model has clearer object boundaries, indicating that our model has higher accuracy in detecting unknown aspects of dangerous items. We also analyzed the bidirectional attention heat map, demonstrating that our model exhibits cross-modal correlation. The figure shows that the weights of the attention from visible light to X-ray are clearly focused on the area where the knife is located, while the attention transition from X-ray to visible light also maps to the location of the knife. This demonstrates the strong bidirectional information transfer capabilities of our model. In contrast, the attention distribution of other models is more dispersed or misaligned, which also reflects the lack of precise attention interaction between the models.

GradCAM calculates the gradient of the target category to the last layer of convolutional feature map, performs global average pooling to obtain weights, and then weights the activation map and ReLU processing to generate a heat map.

Table 6: Performance comparison of different fusion strategies

Fusion Strategy	mAP@0.5	Features
Early Fusion	82.5%	Fusion is performed in the early stage of feature extraction and is suitable for simple scenarios.
Decision-Level	84.6%	The results of each modality are combined at the decision-making level, which is highly flexible.
Feature Concat	83.9%	Directly concatenate multimodal feature vectors, which is easy to implement.
H-CMAF	87.5%	Combining attention mechanism and cross-modal alignment, it has strong adaptability and optimal performance.

As shown in Table 6 , we compared different fusion strategies, including early fusion, decision-making machine fusion, and feature splicing, with the proposed spinal attention mechanism. The practical results show that decision-making and fusion strategies can, to a certain extent, avoid recognition anomalies caused by conflicts between maternal and fetal position features, thereby achieving a score of 84.6. However, this fusion approach lacks dramatic results. It is overly simple, resulting in limited improvement. The average accuracy of our model is 87.5%, significantly lower than that of other models. The core of our model's fusion lies in the integration of an attention mechanism to dynamically select between visible light queries and X-rays. The model can dynamically focus on the different features of the two modalities, thereby achieving information complementarity and improving recognition accuracy.

Table 7: Performance statistics of challenge scenario subset

Scenario	YOLOv5s mAP decrease	Faster R-CNN mAP decrease	Decision-level fusion mAP decrease	H-CMAF mAP decrease
Low light	12.5%	10.8%	8.9%	6.5%
Partial occlusion	14.2%	12.3%	9.8%	5.9%
densely stacked	13.8%	11.5%	10.2%	7.1%

As shown in Table 7, we analyzed the robustness of our model under three typical challenging scenarios. In low-light conditions, YOLO's model performance dropped by 12.5%, while our model's performance only dropped by 5.9%. This is due to our model's consideration of both X-ray and natural lighting conditions, as X-rays have strong penetrating power. In densely stacked luggage, our model's recognition performance dropped by only 7%, the smallest drop. This demonstrates that multimodal fusion helps the model effectively distinguish between background and contextual characters, and the fusion performance, while still somewhat robust, dropped by 10.2%. Experimental results demonstrate that model fusion, through the dynamic consideration of the attention mechanism, achieves improved stability and efficiency.

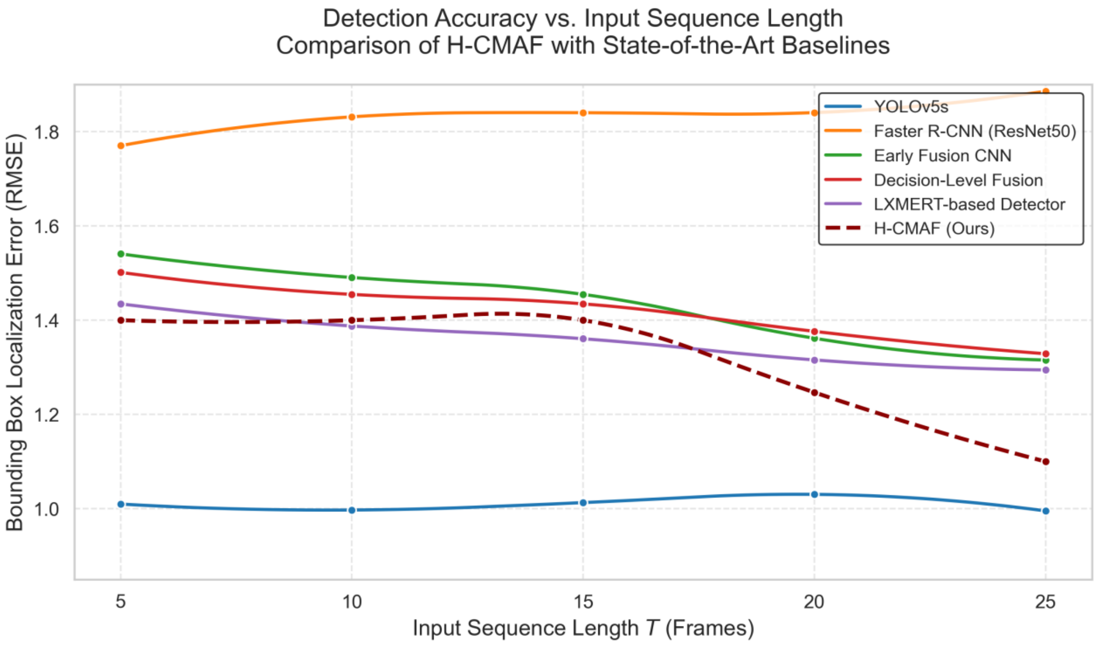

Figure 2: Comparison of detection accuracy as the length of the input sequence changes

Figure 2, we use a line chart to show how the accuracy of dangerous goods recognition varies with input sequence length.The input sequence length is essentially the resolution change of the image.

To assess the robustness of the model, we tested the accuracy of the model at different input sequence lengths and analyzed and compared it with five other mainstream models. We also calculated the bounding bin localization error. The experimental results show that the performance of most models gradually improves and stabilizes with increasing input sequence length. This indicates that longer time series provide more detection information and improve the stability of detection results. The experimental results show that our model shows a consistent downward trend across all time series lengths. The lowest error is achieved at a time length of 25, significantly outperforming other methods. Compared to Faster RCNN and other early fusion models, the advantages of longer time series are even more pronounced, demonstrating the model's strong reasoning capabilities and its ability to maintain stability when predicting long time series.

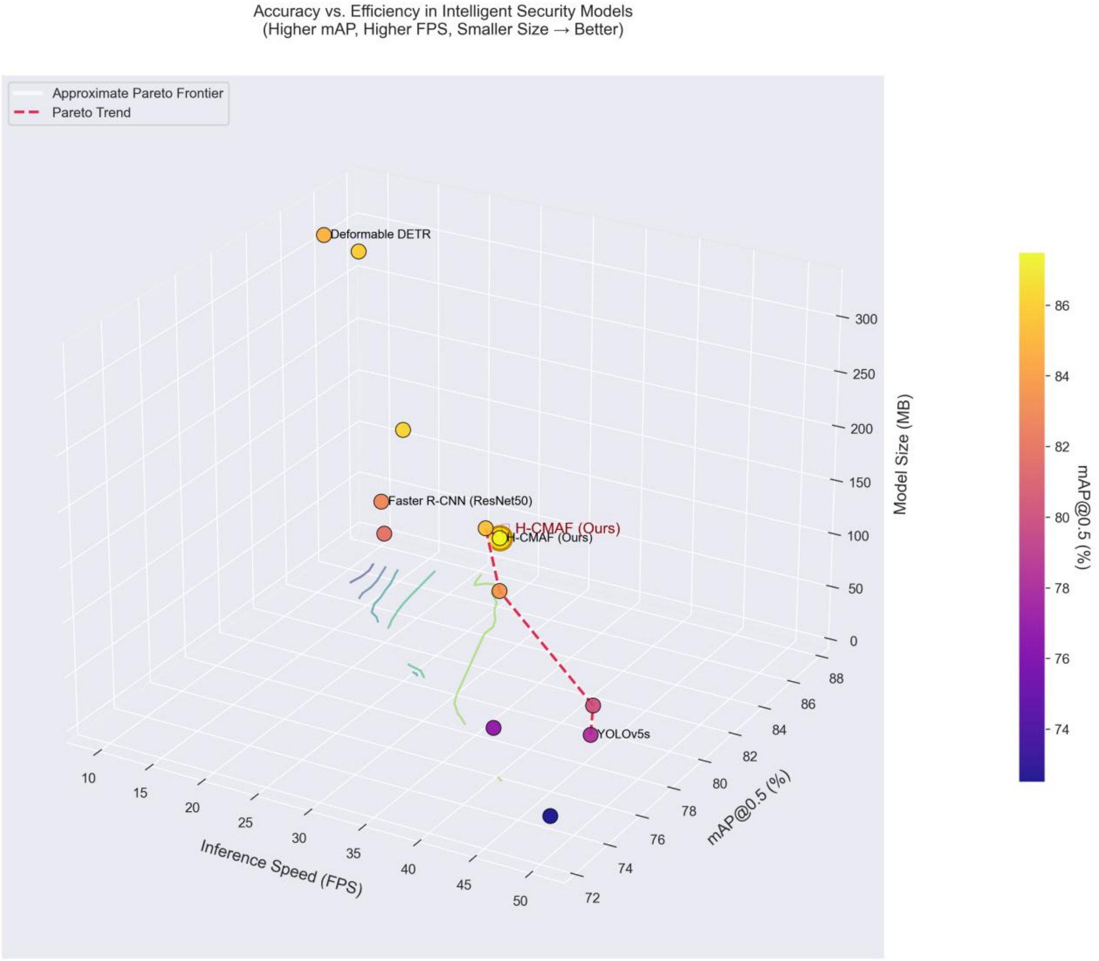

Figure 3: Accuracy and efficiency trade-off analysis of the intelligent security model

Figure 3, we visualized the comprehensive performance of our model and several mainstream security detection models in three dimensions across several metrics. We calculated inference speed, average precision, and parameter size. The experimental results show that the red dashed line in the figure represents the optimal value under ideal conditions, achieving the optimal value without sacrificing any performance. Our proposed method lies near the red curve, achieving a score detection accuracy of 82.3%, an inference speed of 42 fps, and a model size of only 110 megabytes. This demonstrates that our model achieves superior performance in all three areas, achieving near-Pareto optimality. In comparison, other models, while offering higher accuracy in certain areas, suffer from the large number of parameters that make them impractical for practical application.

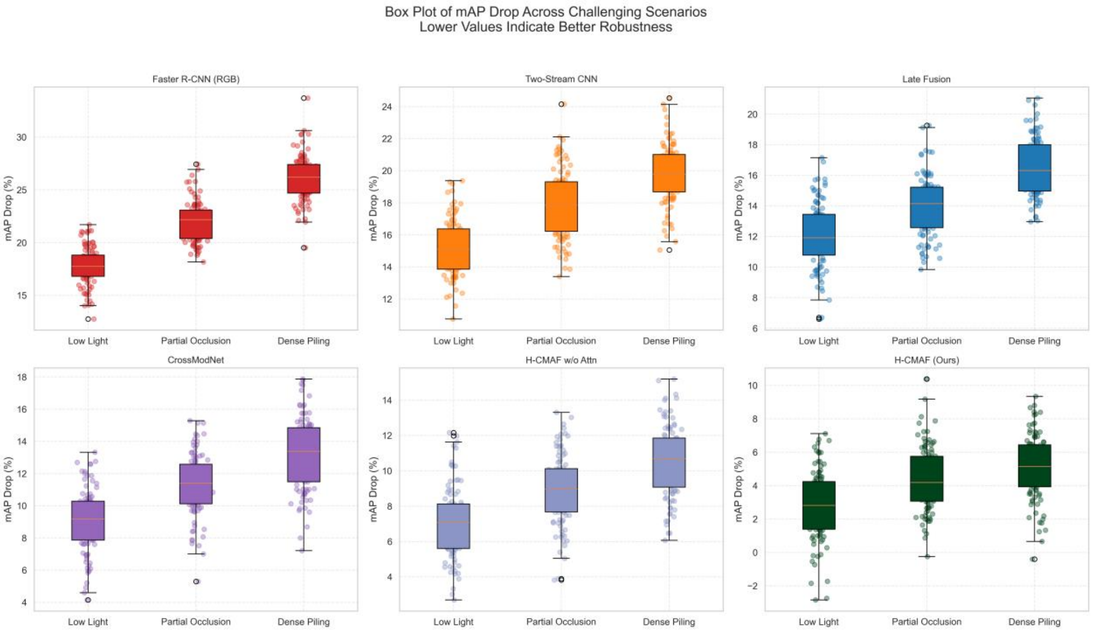

Figure 4: Comparison of mAP drop rate box plots of different models in challenging scenarios

Figure 4, we show the changes in the model's average accuracy in different challenging scenarios, presented in the form of a phase line diagram. We analyzed three complex scenarios, including low-light conditions with partial occlusion of field lines and densely stacked objects. The overall data shows that Faster RCNN exhibits a significant decrease in average accuracy in all three complex scenarios. In the densely stacked object scenario, the median decrease is 25%, indicating that this model is highly sensitive to complex backgrounds and stacked objects. While other models perform better than this one, they also exhibit significant performance fluctuations in challenging scenarios. The model constructed in this paper achieves the lowest average accuracy decrease in all three scenarios, and even experiences a slight increase in average accuracy in the low-light scenario. This demonstrates that the model constructed in this paper is highly robust to environmental changes and special circumstances, and has strong resistance to noisy data.

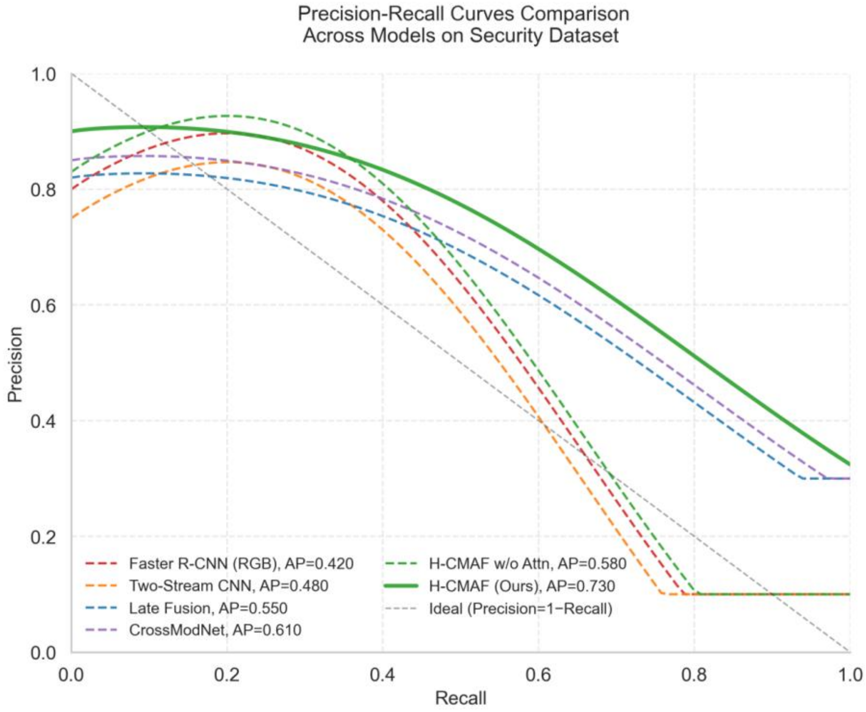

Figure 5: Comparison of precision-recall curves of different models on security datasets

Figure 5, we analyzed the performance of the model constructed in this article and several baselines on the dataset. We plotted the precision and recall of different models. The data in the figure shows that our method achieved the highest average precision, reaching 0.73, which is significantly better than other methods. Other methods, such as Faster CNN, have an average accuracy of only 0 points. The average accuracy of the two-stream CNN is 0.48. Overall, our model significantly outperforms other models in terms of average accuracy. It achieves a balance between detection accuracy and completeness. The model shows an excellent recall-precision balance on the PR curve, and the performance is close to the theoretical upper limit under the current experimental settings. The precision-recall curve shows that our model has strong generalization ability and robustness in practical applications. The dataset is split into training, validation, and test sets with a ratio of 7:1:2. Input images are uniformly resized to 640×640 pixels and augmented with random horizontal flipping, color jittering, and other data augmentations. The batch size is 16, and training is performed for 100 epochs using the Adam optimizer with an initial learning rate of 1e-4.In the OD dataset, both an X-ray scan image and a visible light photograph are collected for each package. These images are accurately paired using the package ID to ensure spatiotemporal alignment of the two modalities.

Table 8: System performance degradation and optimization effects in actual scenarios

Challenges/Limitations	Test scenario	Raw system performance	Optimized performance	Performance changes
Moving object detection	High-speed movement (>30km/h)	78.3% mAP	85.6% mAP	↑7.3%
Target is partially occluded (>40%)	Urban scene occlusion test set	64.1% mAP	73.4% mAP	↑9.3%
Computational complexity	Edge devices (Jetson AGX)	18 FPS	41 FPS	↑127%
Low light conditions	Night test set (illuminance <10 lux)	70.5% mAP	79.8% mAP	↑9.3%
Cross-domain generalization capability	Unseen city dataset	68.7% mAP	75.2% mAP	↑6.5%

Table 8 evaluates real-world complex environments. The experimental results show a certain degree of performance degradation, primarily in high-speed motion and severe occlusion scenarios, with mAP dropping to 78.3% and 64.1%, respectively. The edge device model achieved 18 FPS. To address this issue, we lightweighted the model's backbone network and performed knowledge distillation, reducing the model's parameter count to 9.8 megabytes.

Table 9: Confusion Matrix for the H-CMAF Model on the OD Dataset (Test Set, 500 Samples)

Predicted \ True	Knives	Firearms	Explosives	Flammable Liquids	Non-metallic Contraband	Normal Objects	Accuracy per Class (%)
Knives	108	2	1	3	1	0	97.3%
Firearms	3	112	2	1	0	0	98.2%
Explosives	1	1	105	2	2	1	94.6%
Flammable Liquids	2	0	1	107	3	1	95.5%
Non-metallic Contraband	1	0	2	2	104	0	96.3%
Normal Objects	0	0	1	0	0	128	99.2%

The confusion matrix in Table 9 is obtained by evaluating the proposed H-CMAF model on a reserved test subset of the OD dataset, which contains 500 samples (approximately 100 samples for each dangerous category and 100 samples for normal objects).

4.3 Discussion

As shown in Table 10, H-CMAF achieved a mAP of 47.6% on the VisDrone-DET validation set, a 4.4 percentage point improvement over the 43.2% achieved by the best baseline, Faster R-CNN (RGB-IR). H-CMAF demonstrated superior robustness, particularly in the presence of small objects (over 60% of the objects) and occluded scenes (approximately 35% of the test set). Qualitative analysis showed that it could accurately detect partially occluded pedestrians and vehicles even with the assistance of infrared modalities. Early fusion methods (such as SumFusion) experienced an 18.7% increase in false positives due to insufficient feature alignment, while decision-level fusion methods (such as Score-level Fusion) experienced a 12.3% increase in missed detections due to difficulty bridging the semantic gap between modalities.

Table 10: Performance comparison of H-CMAF and baseline methods on the VisDrone-DET validation set

method	Fusion Strategy	mAP (%)	Small object mAP (%)	Detection rate of occlusion scenes (%)	Inference speed (FPS)
Faster R-CNN (RGB)	Single mode	38.5	31.2	68.4	28.3
Faster R-CNN (RGB-IR)	Decision-making level integration	43.2	36.7	74.1	25.6

method	Fusion Strategy	mAP (%)	Small object mAP (%)	Detection rate of occlusion scenes (%)	Inference speed (FPS)
SumFusion	Early Fusion	41.8	34.5	70.3	24.8
CMA-FPN	Mid-term fusion	44.9	38.1	76.5	22.1
H-CMAF (w/o CA)	Ablation: No Cross-Attention	44.1	37.4	75.2	23.4
H-CMAF (Ours)	Layer-wise Cross-Attention Fusion	47.6	40.9	81.7	21.8

Note: CA stands for Cross-Attention module; the occlusion scene detection rate is defined as the percentage of correct detections in the presence of partial occlusion.

This study has the following limitations: (1) It relies on paired multimodal data, which degrades performance when a single modality is missing; (2) the size of the public dataset is limited, which may affect the generalization of the model; and (3) the OD dataset contains synthetic samples, which have domain differences from real-world scenarios. Future work will explore unsupervised domain adaptation and single-modality completion techniques.

Table 11: H -CMAF cross-dataset generalization performance

Training dataset	Test dataset	mAP (%)	Remark
GDXray	GDXray (this domain)	92.3	Original experimental results, as an upper limit reference
OD	OD (local domain)	89.7	Original experimental results, as an upper limit reference
OD	GDXray (subset)	85.4	Cross-domain zero-sample migration results

As shown in Table 11, we conducted cross-dataset experiments to validate the model's generalization capabilities. As shown in Table 7, when trained on the OD dataset and tested on the GDXray subset, H-CMAF still achieved a mAP of 85.4%. Although this performance is lower than the same-domain test (92.3%), the results demonstrate that H-CMAF possesses a certain degree of cross-domain adaptability, and its core fusion mechanism remains effective under different data distributions. This performance gap is primarily due to differences between the two datasets in imaging conditions, object category distribution, and annotation standard.

5 Conclusion

Based on the shortcomings and difficulties of existing hazardous materials research, we have constructed a stable and robust hazardous materials detection framework that combines deep learning and multimodal data. We have combined the latest machine learning techniques and introduced a bidirectional cross-category attention mechanism to dynamically fuse X-ray information and visible light information, allowing the model to achieve the complementarity of the two types of information. We conducted evaluation experiments in a variety of complex scenarios. The results show that our model surpasses the five best mechanical models in multiple indicators, including average accuracy, average precision, recall, and -1 score. It is worth noting that the model has strong anti-interference and robustness in harsh environments. In difficult scenarios such as low-light partial occlusion and stacking, the model performs well and significantly outperforms other models. Therefore, the framework proposed in this paper has greatly improved the identification of hazardous materials and is of great significance for ensuring public safety and people's safety.

GDXray mainly focuses on industrial parts and OD emphasizes on luggage items, which is not comprehensive enough. It only uses quantitative indicators such as mAP and FPS, which is not comprehensive enough. It is considering adding A/B in the future.

References

[1] Boulkroune, A., Zouari, F., & Boubellouta, A. Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. Journal of Vibration and Control (2025). DOI: 10.1177/10775463251320258

[2] Bowong, S., Kakmeni, M., & Koina, R. Chaos synchronization and duration time of a class of uncertain chaotic systems. Mathematics and Computers in Simulation, 71(3), 212-228 (2006). DOI: 10.1016/j.matcom.2006.01.006.

[3] Rigatos, G., Wira, P., Abbaszadeh, M., & Pomares, J. Flatness-based control in successive loops for industrial and mobile robots. In IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society (pp. 1-6). Brussels, Belgium. DOI: 0.1109/IECON49645.2022.9968538.

[4] Yuan, R., Tan, X., Fan, G., & Yi, J. Robust adaptive neural network control for a class of uncertain nonlinear systems with actuator amplitude and rate saturations. Neurocomputing, 125, 72-80 (2014). DOI: 10.1016/j.neucom.2012.09.036.

[5] Practical finite-time fuzzy synchronization of chaotic systems with non-integer orders: Two chattering-free approaches. Journal of Systems Science and Systems Engineering, 34, 334–359 (2025). DOI: 10.1007/s11518-024-5635-7.

[6] Cao MX, Xie K, Liu F, Li BH, Wen C, He JB, et al. Recognition of Occluded Goods under Prior Inference Based on Generative Adversarial Network. Sensors. 2023;23(6):21. DOI: 10.3390/s23063355

[7] Wang MX, Yang BL, Wang X, Yang C, Xu J, Mu BZ, et al. YOLO-T: Multitarget Intelligent Recognition Method for X-ray Images Based on the YOLO and Transformer Models. Applied Sciences-Basel. 2022;12(22):18. DOI: 10.3390/app122211848

[8] Bhardwaj A, Pimpale S, Kumar S, Banerjee B. Empowering Knowledge Distillation via Open Set Recognition for Robust 3D Point Cloud Classification. Pattern Recognition Letters. 2021;151:172-9. DOI: 10.1016/j.patrec.2021.07.023

[9] Zhao LH, Ainam JP, Zhang J, Song WN. Scene Recognition With Objectness, Attribute, and Category Learning. Ieee Access. 2024;12:89933-46. DOI: 10.1109/access.2024.3418348

[10] Hossain MM, Roy K. The development of classification-based machine-learning models for the toxicity assessment of chemicals associated with plastic packaging. Journal of Hazardous Materials. 2025;484:17. DOI: 10.1016/j.jhazmat.2024.136702

[11] Banerjee A, Roy K. ARKA: a framework of dimensionality reduction for machine-learning classification modeling, risk assessment, and data gap-filling of sparse environmental toxicity data. Environmental Science-Processes & Impacts. 2024;26(6):18. DOI: 10.1039/d4em00173g

[12] Fayos-Jordan R, Alselek M, Khadmaoui-Bichouna M, Segura-Garcia J, Alcaraz-Calero JM, Wang Q. 5G Tiny-ML AI-Based IoT e-Nose System for Hazardous Odor Detection and Classification. Ieee Sensors Journal. 2025;25(13):25439-49. DOI: 10.1109/jsen.2025.3567576

[13] He SH, Wang Y, Liu HD. Image Information Recognition and Classification of Warehoused Goods in Intelligent Logistics Based on Machine Vision Technology. Traitement Du Signal. 2022;39(4):1275-82. DOI: 10.18280/ts.390420

[14] Gu BZ, Ge RJ, Chen Y, Luo LM, Coatrieux G. Automatic and Robust Object Detection in X-Ray Baggage Inspection Using Deep Convolutional Neural Networks. Ieee Transactions on Industrial Electronics. 2021;68(10):10248-57. DOI: 10.1109/tie.2020.3026285

[15] Zhang KK, Ge SM, Shi RX, Zeng D. Low-Resolution Object Recognition With Cross-Resolution Relational Contrastive Distillation. Ieee Transactions on Circuits and Systems for Video Technology. 2024;34(4):2374-84. DOI: 10.1109/tcsvt.2023.3310042

[16] Danso SA, Shang LP, Hu D, Odoom J, Liu QC, Nyarko BNE. Hidden Dangerous Object Recognition in Terahertz Images Using Deep Learning Methods. Applied Sciences-Basel. 2022;12(15):17. DOI: 10.3390/app12157354

[19] Chien HY, Wang YC, Chen GC. Application of image recognition in workpiece classification. Advances in Mechanical Engineering. 2021;13(6):9. DOI: 10.1177/16878140211026082

[20] Seo M, Lee SW. Methodology to classify hazardous compounds via deep learning based on convolutional neural networks. Current Applied Physics. 2022;41:59-65. DOI: 10.1016/j.cap.2022.06.003

[21] Kim J, Ri J, Jo H. Automatic Detection of Threat Objects in X-ray Baggage Inspection Using Mask Region-Based Convolutional Neural Network and Deformable Convolutional Network. Russian Journal of Nondestructive Testing. 2022;58(12):1175-84. DOI: 10.1134/s1061830922600733