EAI Endorsed Transactions
on AI and Robotics
Research Article EAI.EU

FDD-YOLO: A Lightweight Multi-scale Prohibited Items Detection Model

Zilong Xue[1], Bo Wang[1,*], Yuanwei Xie[1], Zhibin Li[2,3,*], Xiaozheng Fan[1] , Chenyoukang Lin[1] , Peiyang Wei[2,3], Linlin Chen[2,3], Xun Deng[2,3], Jianhong Gan[2,3]

[1] Department School of Software, Xinjiang University, Urumqi 830091, China;
[2] School of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China
[3] Dazhou Key Laboratory of Government Data Security, Sichuan University of Arts and Science, Dazhou, Sichuan 635000, China

Abstract

X-ray security inspection faces challenges such as severe occlusion, scale variation, and complex background when detecting prohibited items, requiring real-time and accurate detection. Although the YOLO series of models has high inference efficiency, they suffer from problems such as feature redundancy, insufficient fine-grained feature extraction, and limited adaptability to overlapping objects. To overcome these limitations, we propose FDD-YOLO and design three novel modules: (1) The Frequency Domain Decomposition Network (FDDN) in the backbone network enhances the edges of metal objects and the contours of liquid containers by decomposing high-frequency and low-frequency features while reducing computational redundancy; (2) The Deformable Elastic Fusion Pyramid (DEFP) in the neck network adopts dynamic channel allocation and multi-scale deformable convolution to handle the geometric changes of folded and overlapping objects; (3) The lightweight Dual-channel Convolution (DualConv) improves multi-scale feature capture through grouping and point-by-point convolution, thereby reducing the number of parameters while improving the accuracy of small object detection. Tests on the SIXray, HIXray, and private GIX datasets show that FDD-YOLO achieves 2.6%, 3.2%, and 8.6% higher mAP than YOLOv11n, respectively, achieving accuracies of 94.8%, 84%, and 71.8%, respectively. This framework also reduces the number of parameters by 30.6% and the number of FLOPs by 26.9%, achieving an optimal balance between accuracy and efficiency, setting a new technical benchmark for real-time security inspections.

Keywords: Frequency Domain Decomposition Network (FDDN); Deformable Elastic Fusion Pyramid (DEFP); Dual-channel Convolution (DualConv); Prohibited Items; X-ray Image

Received on 15 September 2025, accepted on 05 November 2025, published on 11 November 2025

doi: 10.4108/airo.10277

1. Introduction

Public safety relies on efficient baggage checks in key locations like airports and subways. X-ray imaging non-invasively reveals internal object structures and has become a core security tool worldwide [1]. However, manual inspection is prone to inefficiency and missed detection due to subjective bias [2], leading to congestion and risks to public security.

A promising new paradigm for automated X-ray security inspection has been made possible by the revolutionary advancements in deep learning, particularly the remarkable success of convolutional neural networks (CNNs) in general object detection [3]. Researchers have attempted to apply sophisticated algorithms such as Faster R-CNN [4], YOLO [5], and SSD to the challenge of X-ray contraband detection due to their exceptional performance in natural picture

[1]Corresponding author. Email: LiZhibin111@outlook.com

identification [6]. However, X-ray images have unique intrinsic properties, which make this direct transfer face significant obstacles, resulting in model performance far from meeting practical requirements [7]. The severe overlap and occlusion of items in luggage result in blurred contours and unclear boundaries of target objects, making it difficult to separate them, as well as high class differences and inter-class similarities. The same type of contraband (such as knives made of different materials) may show completely different textures and shapes, while everyday items with similar appearances (such as laptops and explosive blocks) may have similar grayscale distributions under X-rays [8]. Contraband also has extremely multi-scale characteristics. The size range of the objects to be inspected is extremely large, from tiny lighters to large laptops, requiring the model to have strong multi-scale perception capabilities.

Existing research addresses these challenges through two main approaches. One involves designing specialized network architectures, such as feature pyramid networks (FPN) and their variants, to improve multi-scale feature fusion [9]. For instance, Wang M et al. proposed a weight-guided dual-direction-fusion feature pyramid network (WDFPN) to handle scale variations in crowded scenes [10], while others introduced attention mechanisms to highlight suspicious regions [11]. The second approach utilizes physical priors in X-ray imaging, such as employing dual-energy data to distinguish organic and inorganic materials based on atomic number [12].

Although the aforementioned studies have made progress in the singular application of frequency-domain analysis or deformable convolutions, limitations persist. Firstly, most frequency-domain methods are computationally complex and lack co-design with lightweight network architectures, making them difficult to deploy on edge devices. Secondly, existing deformable convolution modules often operate in isolation, lacking a dynamic and elastic fusion mechanism with multi-scale feature pyramids at both channel and spatial dimensions. More importantly, few works can simultaneously and cooperatively address the dual challenges in X-ray images: the loss of high-frequency (detailed textures) and low-frequency (structural contours) information, and the geometric variations caused by occlusion and deformation.

This paper presents FDD-YOLO, a novel model for contraband detection in X-ray images that employs multi-feature adaptive enhancement through an efficient multi-dimensional collaborative framework. The proposed approach significantly improves feature extraction and fusion capabilities for handling complex X-ray image characteristics, with key contributions in the following four aspects.

The Frequency Domain Decomposition Network (FDDN) is designed to decompose the input feature map into high- and low-frequency components. The high-frequency path uses a fused depthwise separable convolution (DSIB) module with a reverse bottleneck design and residual connections to enhance local details. The low-frequency path uses a multi-scale edge enhancement (MSEE) module to sharpen the global outline of objects. Finally, the Squeeze and Excitation attention mechanism adaptively fuses high- and low-frequency information, effectively addressing the dual issues of blurred details and missing outlines.

The Deformable Elastic Fusion Pyramid (DEFP) module is designed specifically for processing deformable and multi-scale targets in X-ray images. It adaptively captures target features with different geometries through dynamic channel allocation and multi-branch deformable convolution, and utilizes lightweight attention maps to adaptively fuse forward and backward features from the FPN and backbone network, thereby significantly improving the flexibility and effectiveness of feature fusion.

To improve model efficiency and accuracy, we replaced standard convolutions in Backbone and the neck with our DualConv. This module utilizes a parallel structure of Group Convolution and Pointwise Convolution to achieve an equivalent multi-scale receptive field while significantly reducing computation and parameter count. This paves the way for model deployment on edge devices and meets the real-time requirements of contraband detection.

In order to further improve the model's ability to detect small objects, a self-built dataset called GIXray is also constructed in this paper.

2. Related work

Convolutional neural networks (CNNs) have greatly enhanced contraband detection accuracy and efficiency, leading to broader adoption in recent years. Deep learning-based methods are mainly divided into two-stage and one-stage detectors. Two-stage detectors, such as R-CNN and its successors (e.g., Fast R-CNN), first generate region proposals and then perform classification and regression [13]. These models are known for high accuracy and have been extensively studied in X-ray security inspection — for example, Zhang W et al. improved Faster R-CNN with ResNet-50 to detect overlapping items [14], while Sagar et al. introduced MSA R-CNN with multi-scale feature extraction to mitigate FPN information loss [15]. However, such methods suffer from high computational cost, structural complexity, and slow inference, making them less suitable for real-time applications like embedded scanners and limiting their practical use.

To overcome the limitations of two-stage detectors, research has shifted toward efficient single-stage models like SSD, YOLO, and DETR, which unify localization and classification within a single network to achieve faster inference. Improved YOLO variants are particularly prominent in X-ray contraband detection. For instance, Zhang H et al. enhanced YOLOv7-tiny with a FasterNet backbone, a PConv-ELAN neck module, and coordinate attention to improve small object detection [16]. Guan F et al. incorporated the ADown sampling module and DCNv2 into YOLOv8 for efficient feature extraction and higher accuracy, along with Fast SPPF_RE for better feature fusion [17]. Zhao C et al. introduced a label-aware (LA) method using gradient-based channel weighting to handle overlapping objects robustly [18]. Ding et al. proposed FE-DETR, integrating split-attention, CBAM, DCN, and a transformer prediction

head to enhance performance in detecting obscured objects [19]. Zhou Y et al. developed EI-YOLO using Normalized Wasserstein Distance (NWD) for improved bounding box regression [20].

Current research emphasizes lightweight design for edge device deployment. Zhou Y T et al. proposed a Low-Parameter Feature Aggregation (LPFA) structure that utilizes max pooling and NWD loss to enhance feature integration and small object detection [21]. Similarly, Jia L et al. employed MobileNetV3 as a lightweight backbone and combined SIoU loss with coordinate attention to improve YOLOv7 performance [22].

Despite significant progress, current research still faces common challenges: severe deformation in X-ray images due to angles and occlusions, and a lack of adaptive fusion strategies that effectively complement deep/shallow and high/low-frequency features while maintaining a lightweight design. We propose a Frequency Domain Decomposition Network to address these gaps to enhance detailed texture and structural information, effectively handling blur and occlusion. We employ deformable convolution to adaptively capture deformed objects through dynamic receptive fields, and introduce adaptive weight maps for improved feature fusion. Additionally, DualConv provides multi-scale receptive fields with minimal computational cost, facilitating lightweight deployment. We aim to achieve an optimal balance between speed and accuracy, meeting the stringent demands of real-world security inspection applications.

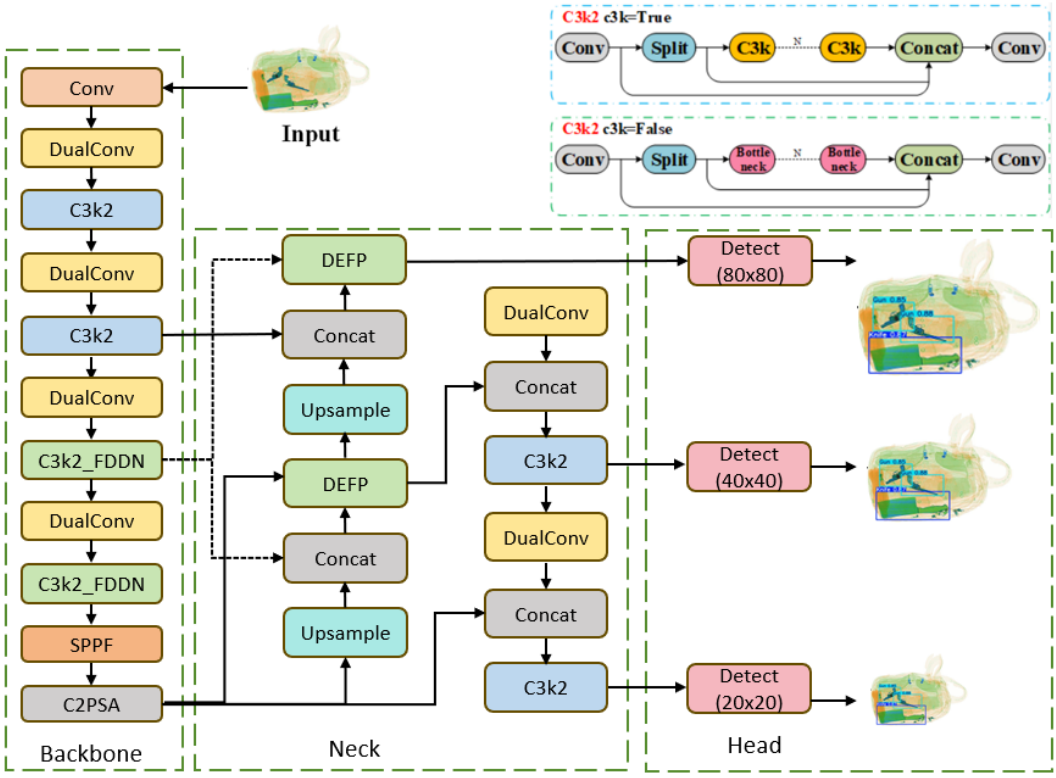


Figure 1. The structure of FDD-YOLO

for X-ray security inspection. However, it faces challenges in complex scenarios: continuous convolution in the backbone introduces feature redundancy and computational cost, standard convolutions struggle to capture fine-grained features such as metal edges or container contours, and overlapping objects often lead to feature confusion and false positives. To address these issues, we propose the FDD-YOLO model, which enhances detection accuracy while maintaining real-time capability.

FDD-YOLO integrates the Frequency Domain Decomposition Network (FDDN) with C3K to replace the C3K module in C3K2, forming C3K2_FDD. FDDN significantly reduces redundant feature generation through the decomposition of high and low frequency features and an adaptive fusion mechanism, thereby improving computational efficiency and enhancing the extraction ability of fine-grained features such as the edges of metal cutting tools and the contours of liquid containers. Additionally, through the designed Dynamic Elastic Fusion Pyramid (DEFP) to optimize the neck structure, an innovative multi-scale deformation adaptation mechanism is introduced to strengthen the ability to capture multi-scale features, effectively solving the feature confusion problem of

3. Proposed method

3.1. FDD-YOLO

The YOLO series is widely recognized in object detection for its strong real-time performance, making it highly suitable

overlapping items. Subsequently, DualConv replaces the standard convolution in the network structure [23], adopting dual-path convolution and group compression techniques to reduce parameters and lower computational load. The overall structure of FDD-YOLO 11 is shown in Figure 1.

3.2. Frequency Domain Decomposition Network

In complex visual scenes, high-frequency information (local details such as edges and textures) and low-frequency information (global features such as overall structure and background) in contraband images are complementary to

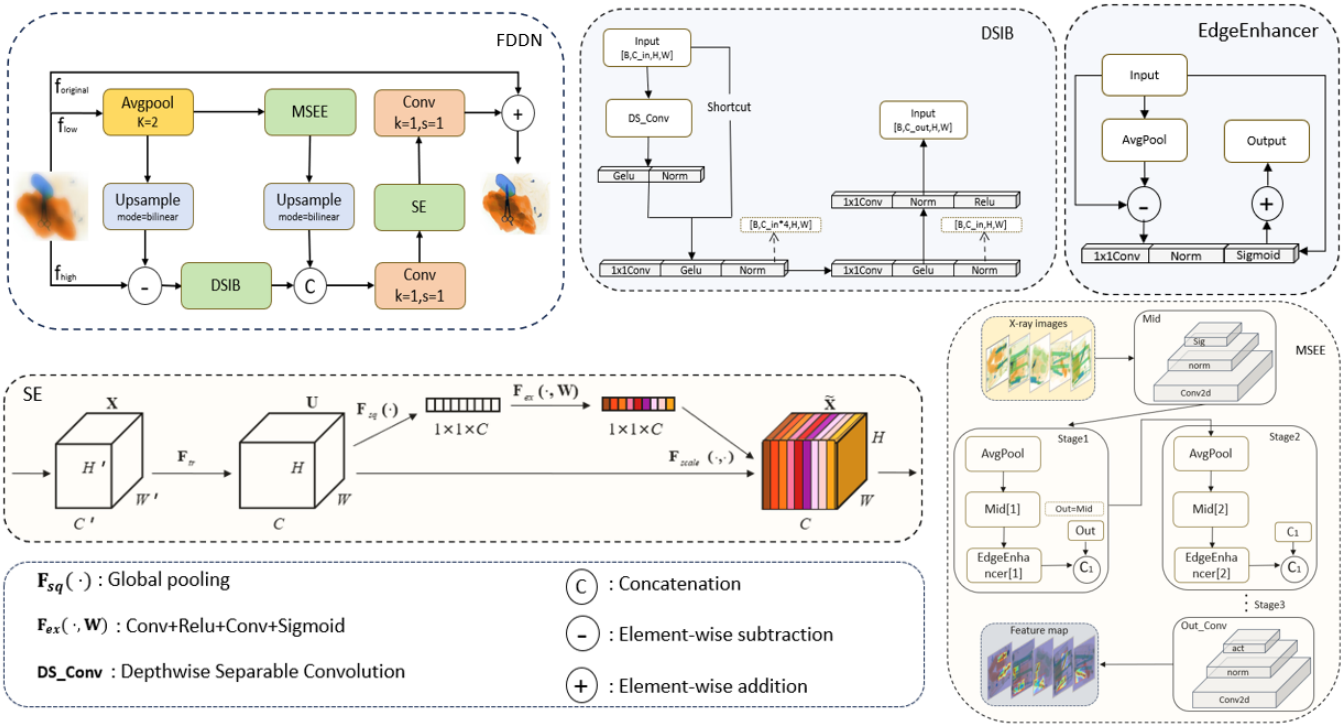


Figure 2. Detailed structure of the Frequency Domain Decomposition Networks

each other for visual tasks. Existing methods often fail to fully utilize these two types of information in a synergistic manner, and traditional frequency-domain processing methods are often accompanied by high computational overhead. To this end, we propose a lightweight Frequency Domain Decomposition Network (FDDN). This network first separates feature maps into high-frequency and low-frequency components using a lightweight frequency-domain decomposition layer. It then enhances high-frequency details and low-frequency structural features through an efficient two-way processing mechanism. This significantly improves the model's feature representation capabilities in scenes with complex backgrounds and translucent objects while maintaining extremely low computational complexity. The overall structure of the FDDN is shown in Figure 2. The FDDN consists of three core components: a frequency-domain decomposition layer, two-way feature enhancement, and an adaptive fusion mechanism.

The frequency domain decomposition layer uses average pooling downsampling to extract low-frequency components, based on the theoretical principle that average pooling acts as a low-pass filter in the frequency domain and can effectively capture fundamental low-frequency information of the image. The high-frequency components are obtained by subtracting the upsampled low-frequency components from the original image. Bilinear interpolation is chosen here instead of transposed convolution for upsampling to avoid the risk of overfitting caused by introducing additional learnable parameters. Meanwhile, the fixed kernel of bilinear interpolation ensures the purity of frequency components and effectively reduces computational complexity.

The high-frequency path is processed using our designed DSIB module, which integrates depthwise separable convolution [24], an inverted bottleneck structure, and residual connections. First, local details are extracted through grouped convolution, where the number of parameters is only 1/G of standard convolution (with G being the number of groups). This approach avoids redundant inter-channel computations. A 1×1 convolution is then used to expand the channels by a factor of four, forming an inverted bottleneck. This design breaks the limitations of the traditional bottleneck structure by performing nonlinear transformations in the expanded layer, significantly enhancing feature representation capability.

The GELU activation function is employed to provide smoother nonlinear mapping. Finally, a channel compression

layer restores the original number of channels via 1×1 convolution, forcing the network to focus on key features. The entire process substantially reduces computational load while preserving high-frequency detail sharpness. The corresponding computational method is defined in equations (1) and (2).

$$Y = GELU\left(BN\left(DepthwiseConv_{k\times k}(X)\right)\right) + X \quad (1)$$

$$Z = GELU\left(BN(Conv_{1\times1}(Y))\right) \quad (2)$$

Where $X \in R^{C\times H\times W}$ represents the input feature map, and C, H and W represent the number of channels, height, and width, respectively. $Y \in R^{C\times H\times W}$ represents the local features after grouped convolution, $Z \in R^{4C\times H\times W}$ represents the high-dimensional features after channel expansion, and $Conv_{k\times k}$ represents a $k\times k$ convolution. The low-frequency path is processed by our proposed Multi-Scale Edge Enhanced (MSEE) module. This structure constructs a multi-scale feature pyramid. The initial scale is generated by 1×1 convolution and sigmoid gating. Subsequent scales are gradually downsampled by cascaded average pooling (kernel=3, stride=1, padding=1) and the same 1 × 1 convolution gating. Each scale feature is enhanced by Edge Enhancer. Edge Enhancer is an edge enhancement module designed for complex security inspection scenarios. This module first smoothes the input features using 3×3 average pooling, which is equivalent to extracting local background information from the image. Then, by calculating the difference between the original and smoothed features, it accurately captures the intensity mutation of the local area, which is the essential characteristic of the object edge. Its calculation method is defined in Equation (3).

$$F_{low} = \bigoplus_{k=1}^{3}\left(AvgPool_{3\times3}^{(k)} \circ EdgeEnhancer(X)\right) \quad (3)$$

Where F represents the feature, F_{low} represents the low-level feature, $k \in \{1,2,3\}$, and $\bigoplus_{k=1}^{3}$ represents cascaded three-way average pooling and edge enhancement. The initial input $F_{\text{low}}^{(0)} = X$, $X \in R^{C\times H\times W}$ represents the input feature.

During the adaptive fusion stage, the varying importance of different frequency components for detecting specific contraband items—such as metallic objects relying on high-frequency signals and liquids depending on low-frequency information—motivates the use of a Squeeze-and-Excitation (SE) attention mechanism for dynamic frequency-domain calibration. First, low-frequency features are upsampled to the original resolution and concatenated with high-frequency features along the channel dimension. These fused features are then weighted using the SE mechanism, which employs global average pooling to compress spatial information into a channel-wise feature vector. This vector captures the average response intensity of each feature channel across the entire image. The feature refinement is further performed through a bottleneck structure (with a compression ratio of 16) consisting of two fully connected layers: the first uses ReLU activation to introduce nonlinearity during dimensionality reduction, while the second applies a sigmoid function to generate normalized attention weights between 0 and 1, effectively forming a parameterized feature selection filter.

The weighted dual-path features are subsequently fused spatially using a 3×3 convolution layer. This operation helps restore spatial correlations that might have been weakened during the frequency decomposition and enhancement stages. A residual connection is also incorporated to facilitate identity mapping, mitigate gradient vanishing, and preserve the integrity of the original features. The entire fusion process remains computationally efficient due to the extensive use of channel compression and 1×1 convolutions.

In the YOLOv11 architecture, we integrate FDDN with C3k2 to form C3k2_FDDN, strategically deployed at backbone layers P4 and P5. This design employs channel splitting to reduce computational load while maintaining training stability through residual connections, thereby enhancing detection robustness in security scenarios without compromising real-time performance.

3.3. Deformable Elastic Fusion Pyramid

DEFP (Deformable Elastic Fusion Pyramid) is a lightweight feature fusion architecture designed to address the key challenge of cross-scale object detection in X-ray security images. Its structure is shown in Figure 3.

First, a dynamic channel allocation mechanism is used to concatenate the input dual-path features y_1 and y_2 after bilinear interpolation and size alignment, resulting in f_1.

$$f_1 = Concat(y_1, y_2) \quad (4)$$

Then the high-level features X are obtained through the conversion layer.

$$x = Transform(f_1) \quad (5)$$

Where Transform(·) represents a feature transformation layer consisting of 1×1 convolution, batch normalization, and ReLU activation. Based on x, the channel importance weights are generated by the lightweight channel attention unit:

$$w = Softmax(W_2 \cdot ReLU(W_1 \cdot GAP(x))) \quad (6)$$

Where $GAP(\cdot)$ represents global average pooling, $W_1 \in R^{(c/4)\times 2c}$ and $W_2 \in R^{2c\times(c/4)}$ are the weights of the two linear transformation layers (compression ratio = 8).

After softmax normalization, the feature channels are dynamically sorted and reorganized by importance to obtain x_{sorted}. The reorganized features are intelligently divided into four groups.

$$\{x_1, x_2, x_3, x_4\} = Split(x_{sorted}) \tag{7}$$

The first three groups of inputs are multi-scale elastic deformation units. Each unit contains an offset generator (predicting the sampling point position offset through 3×3 convolution) and an adaptable convolution kernel (supporting three scales: 3 × 3, 5 × 5, and 7 × 7). This deformation perception capability can accurately fit non-rigid objects such

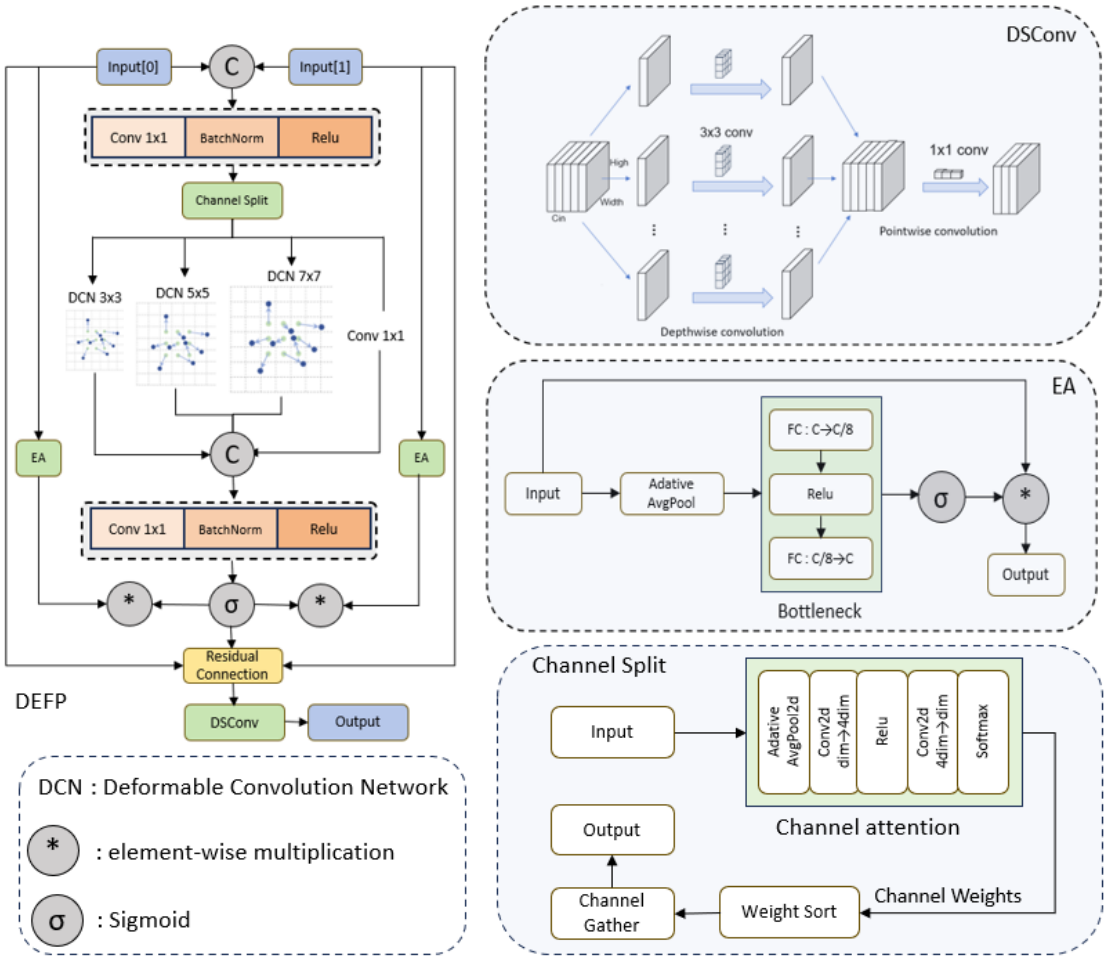


Figure 3. Detailed structure of the Deformable Elastic Fusion Pyramid

as folding knives and curved containers in X-ray images. The fourth group retains the original features to maintain information integrity. The computational load of processing the four sets of features is reduced through a channel averaging strategy. The deformation convolution process can be expressed as:

$$y(p) = \sum_{k=1}^{K} W_k * x(P + P_k + \Delta P_k) \tag{8}$$

$$\Delta p_k = W_{offset} * x_k \tag{9}$$

Where $y(p)$ represents the final calculated value at position p on the output feature map, K represents the total

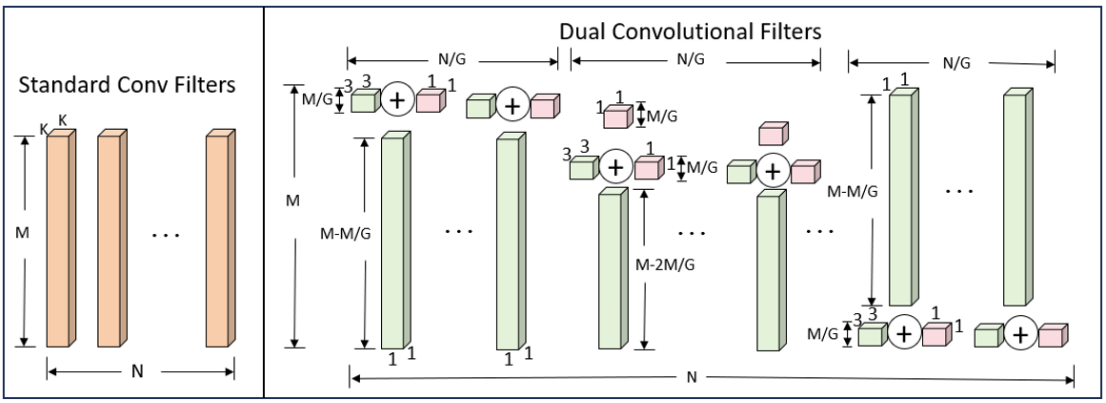
number of convolution kernel sampling points ($K = 9$ for a 3 ×3 convolution kernel), k is the index of the sampling point currently being calculated, W_k represents the convolution weight corresponding to the kth sampling point, x represents the input feature map, and P_k represents the fixed offset coordinate of the kth sampling point relative to the center point p in a standard convolution. ΔP_k is the learned offset. Essentially, it uses a small convolutional network W_{offset}, guided by the input feature x_k, to predict the optimal direction and distance Δp_k for each sampling point of the main convolution kernel. This allows the entire convolution operation to adaptively focus on irregular target geometries.

In the feature conversion stage, multi-scale outputs are integrated through 1×1 convolution to generate a spatial attention map.

$$a = \sigma(W_a * Concat(x_1, x_2, x_3, x_4)) \quad (10)$$

x_1, x_2, and x_3 are features processed by deformable convolutions of different scales (3x3, 5x5, and 7x7). The * symbol represents convolution, and W_a is the weight of the convolution kernel.

This is innovatively weighted fused with the dual-path efficient attention-enhanced features (features processed using an efficient attention module with an 8:1 compression ratio):



(a) Standard Convolution (b) Dual Convolution

Figure 4. Detailed structure of the (a)Standard Convolution and (b)Dual Convolution

$$out = a \cdot Attn(y_1) + (1-a) \cdot Attn(y_2) + y_1 + y_2 \quad (11)$$

The learned coefficients dynamically balance the contribution of features at different scales (such as detailed information at the P3 layer and semantic information at the P5 layer), and ultimately refine the features through depthwise separable convolution (3×3 grouped convolution and 1×1 point convolution):

$$Output = W_{1\times1} \cdot (W_{3\times3} * out) \quad (12)$$

In the YOLOv11 architecture, the DEFP module is deployed at the key fusion nodes P3 and P4 of the feature pyramid. This design is based on deep feature characteristics; layer P4 (downsampled by 1/16) carries the structural information of medium-sized contraband, while layer P3 (downsampled by 1/8) retains key details for small target detection. DEFP achieves breakthroughs through several lightweight design features: channel attention uses 1×1 convolutions instead of fully connected layers to reduce parameters; it combines Softmax to achieve adaptive channel feature selection; the deformable convolution's offset generator shares the computation path with the underlying feature extraction, avoiding additional overhead; and the multi-scale separable convolution architecture decouples spatial convolution from channel projection, significantly reducing the computational cost of standard convolutions in layer P4. This design provides an excellent balance between accuracy and efficiency for security inspection systems, significantly improving the detection rate of contraband in complex scenarios.

3.4. DualConv

In X-ray security image processing, the single convolution kernel of traditional standard convolution struggles to balance local texture details with global semantic expression. This makes optimizing the complex structures of metal objects and the low-contrast features of liquid contraband challenging. Furthermore, as the number of channels increases, the computational complexity of standard convolution increases significantly, resulting in excessive computational load in deep networks (such as the P5 layer). We introduced the DualConv dual-path convolution module to overcome these

limitations, which restructures the feature extraction paradigm through a dual-path parallel architecture. Its structural comparison with standard convolution is shown in Figure 4.

The DualConv module is deeply integrated into the P2, P3, P4, P5, and Neck feature fusion paths of the backbone network to construct a hierarchical optimization system. In its dual-path collaborative design, the grouped convolutional path decomposes the input channels through a spatial separation strategy and uses 3×3 convolutions to efficiently extract local texture features, while the point convolutional path establishes global channel interaction through 1×1 convolutions to accurately convey material density information. The dual outputs are fused element-wise, which enhances semantic feature expression while maintaining the integrity of the spatial structure, significantly improving the recognition ability of complex metal structures and low-contrast materials.

In system-level collaboration, DualConv is deeply coupled with higher-order modules. At the P4 layer, its optimized feature representation provides structured low-frequency input to the FDDN frequency-domain decomposition network, collaboratively improving liquid container detection accuracy. In the Neck path, its spatial compression features provide key input to the DEFP elastic fusion module, jointly improving the recognition accuracy of small-sized cutting tools. The module employs an adaptive grouping strategy, enhancing edge feature extraction in the shallow P2 layer and semantic compression in the deep P5 layer. It optimizes memory access efficiency through group convolution and

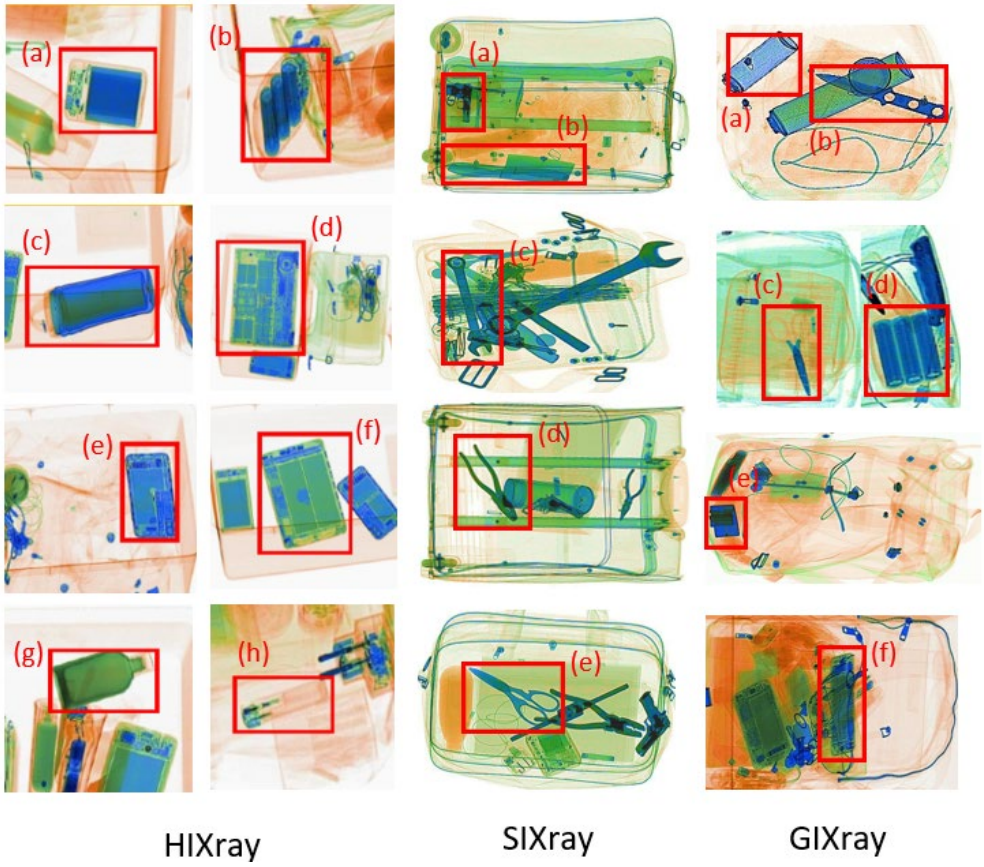


Figure 5. Examples of three datasets, HIXray contains 8 categories, SIXray contains 5 categories, and GIXray contains 6 categories.

eliminates channel computational redundancy through point convolution.

This module demonstrates exceptional performance in complex security inspection scenarios, significantly improving the accuracy of occluded object detection and significantly increasing the recognition rate of low-contrast contraband. This establishes an efficient, reliable feature extraction architecture foundation for real-time X-ray security inspection systems.

4. Experimental results and analysis

We conduct several tests on the SIXray [25], HIXray [26], and GIXray datasets to evaluate the effectiveness of the proposed method. We present the experimental results after introducing the datasets and performance evaluation criteria. Next, we conduct an ablation study to demonstrate the effectiveness of each module in the proposed method.

Finally, we compare the experimental results with some state-of-the-art techniques. Figure 5 shows several sample examples from these three datasets.

4.1. Datasets

The SIXray dataset is the largest prohibited item detection dataset released to date, containing 1,059,321 X-ray images, but only 8,929 images contain prohibited items, which are manually labeled into six categories. There are six common categories of prohibited items: gun, knife, wrench, pliers, scissors and hammer. However, this category is not used in the official labeling due to the low number of hammers. The training set included 7496 images, and the test set included 1433 images.

The HIXray dataset is a larger dataset of X-ray images compared to the SIXray dataset. HIXray contains 45,364 high-quality X-ray images of eight prohibited items. There are eight common categories: portable charger 1, portable charger 2, water, laptop, mobile phone, tablet, cosmetic, and non-metallic lighter. The training set includes 36,295 images, and the test set consists of 9,069 images.

Regarding the construction details of the GIXray dataset, all X-ray images in this dataset were acquired from real security inspection scenarios using a Smiths Detection HI-SCAN 6040aTiX dual-energy X-ray scanner. The acquisition process followed standard airport baggage security procedures. To enrich the diversity of small-sized prohibited items, we designed various placement protocols, including placing the target item at different depths in the baggage, overlapping it with other everyday items, and placing it at different angles. All images were independently labeled by three professionally trained annotators using LabelImg. The final annotation results underwent consistency testing, and inter-evaluator consistency was guaranteed by calculating the mean intersection-union ratio (mIoU), achieving a mIoU of 0.92, ensuring high-quality and consistent bounding boxes.

All baggage data collected for GIXray strictly adhered to data privacy protection guidelines. Due to the sensitive data contained in this dataset from real security inspection scenarios, the GIXray dataset cannot be publicly released at this time for security and privacy protection reasons. However, to ensure the rigor and verifiability of the research, we have provided the complete model architecture and training details in this paper. We believe that the FDD-YOLO method proposed in this study, as a general framework, is also applicable and effective on other publicly available X-ray security inspection datasets (e.g., SIXray, HIXray).

4.2. Evaluation criteria

Three datasets, SIXray, HIXray, and ZLXray, are used in this paper's experiments to assess FDD-YOLO's performance thoroughly. The primary metrics used to compare the experiments with the state-of-the-art detection techniques are Precision (P), Recall (R), Mean Average Precision (mAP), Parameters (Params), and Giga Floating Point Operations per Second (GFLOPs).

With TP (True Positives) representing successfully recognized object instances and FP (False Positives) representing background instances incorrectly categorized as objects, precision is the percentage of true positives among all samples anticipated as positive. Meanwhile, FN (False Negatives) denotes cases that are real objects but go unnoticed, and recall quantifies the percentage of genuine positives that were accurately predicted out of all actual positive samples. A standard measure for comparing the overlap of two bounding boxes is the Intersection over Union (IoU), which is usually employed to gauge how closely the anticipated and ground-truth boxes match. The detection is usually considered legitimate when the IoU is above a certain threshold (such as 0.5 or 0.7). One of the most popular comprehensive assessment measures in object identification is Mean Average Precision (mAP), which shows how well the model performs throughout the dataset. It is calculated by averaging all category-specific Average Precision (AP) values after calculating each category's AP. A model's memory footprint and computational complexity are directly impacted by its parameters, representing the total number of trainable parameters in the model; larger parameter counts often correspond to a more complicated model. A crucial measure of computational effectiveness and inference speed, GFLOPs (Giga Floating-point Operations) count the number of floating-point operations needed for a single forward run through the model. The following equations are used to represent them.

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$AP = \int_0^1 p(x)dx \tag{15}$$

$$mAP = \frac{1}{c}\sum_{i=1}^{c} AP_i \tag{16}$$

Where mAP is the average of all AP categories, C is the number of categories for the item, and AP may be computed using the area of the Precision-Recall curve.

4.3. Evaluation criteria

In this paper, we use the PyTorch 2.6.0 framework to build the network model, CUDA 12.1, Python 3.11, Intel(R) Core(TM) i9-14900K for the hardware CPU, and NVIDIA GeForce RTX 4090 D with 24GB memory for the graphics card. To be fair to all methods, we use the training set for training. For training, the test set is evaluated using SGD optimizer with configured momentum of 0.9, batch size of 16, epoch of 300, and initial learning rate set to 0.01. All the experiments are unified in terms of input resolution (640× 640) and training strategy (SGD optimizer, 500 epochs) to ensure fairness.

4.4. Ablation studies

To systematically verify the effectiveness of each innovative module, we conducted comprehensive ablation experiments on the most challenging GIXray dataset (focused on small object detection) and the most representative SIXray dataset. The baseline model was the original YOLOv11n model without any improvements.

We conducted comprehensive ablation studies to evaluate the contributions of each module in FDD-YOLO.

The introduction of the Frequency Domain Decomposition Network (FDDN) into the backbone improved mAP50 on SIXray and GIXray by 1.5% and 3.4%, respectively, while reducing computation by 6.3% and parameters by 9.3%. This confirms FDDN's effectiveness in enhancing feature extraction for blurred, small, and occluded objects through explicit decoupling of high- and low-frequency information.

The Deformable Elastic Fusion Pyramid (DEFP) in the neck further increased mAP50 by 2.1% on SIXray and 5.4% on GIXray, while reducing computation and parameters by 9.5% and 12.8%, respectively. Its dynamic channel allocation and multi-scale deformable convolution allowed the model to adapt more accurately to objects of varying shapes and complex backgrounds.

The lightweight DualConv module improved mAP50 by 1.0% and 1.5% on the two datasets while reducing computation and parameters by 15.9% and 16.7%.

Table 1. Ablation experiments on the SIXray dataset

FDDN	DEFP	DualConv	P	R	mAP50	mAP50:95	GFLOPs	Params
			93.6	86.8	92.2	72.0	6.3	2.58
✓			94.5	87.7	93.4	73.5	5.9	2.34
	✓		94.8	88.1	93.8	74.1	5.7	2.25
		✓	94.2	87.4	93.2	73.3	5.3	2.15
✓	✓		95.2	88.4	94.3	74.4	5.4	2.16
	✓	✓	95.0	88.2	93.9	74.2	4.9	1.97
✓		✓	94.9	88.2	94.1	74.2	5.1	1.95
✓	✓	✓	96.1	88.3	94.8	75.2	4.6	1.79

Table 2. Ablation experiments on the GIXray dataset

FDDN	DEFP	DualConv	P	R	mAP50	mAP50:95	GFLOPs	Params
			73.0	55.1	63.2	39.2	6.3	2.58
✓			73.1	61.5	66.6	42.3	5.9	2.34
	✓		73.8	58.0	68.6	42.8	5.7	2.25
		✓	74.1	55.6	64.7	41.7	5.3	2.15
✓	✓		73.7	56.4	70.7	43.3	5.4	2.16
	✓	✓	73.8	58.1	69.5	43.0	4.9	1.97
✓		✓	75.6	59.5	67.1	42.6	5.1	1.95
✓	✓	✓	79.9	62.5	71.8	44.1	4.6	1.79

Table 3. Comparison results with advanced methods on the SIXray dataset

Methods	P	R	mAP50	mAP50:95	GFLOPs	Params	Year
Faster R-CNN	84.5	77.5	82.8	65.5	210.4	41.09	2016
Mask R-CNN	86.1	76.7	83.9	65.4	283.5	60.04	2017
Grid R-CNN	85.6	77.8	83.3	66.0	328.8	64.32	2019
POD-Y	92.3	85.1	90.4	70.5	108.1	47.19	2023
AO-DETR	85.4	76.2	83.3	65.2	268.9	58.38	2024
MLSA-YOLO	92.5	86.5	90.8	71.0	32.5	12.14	2024
YOLOv10n	94.4	86.3	92.5	72.2	8.2	2.69	2024

YOLOv11n	93.6	86.8	92.2	72.0	6.3	2.58	2024
YOLOv12n	95.0	86.6	92.6	73.1	6.3	2.56	2025
RT-DETR	94.2	87.1	93.1	73.8	107.21	32.1	2025
D-FINE	93.8	86.9	92.8	73.2	91.35	31.5	2025
DEIM	94.5	87.3	93.5	74.1	91.42	31.62	2025
FDD-YOLO	**96.1**	**88.3**	**94.8**	**75.2**	**4.6**	**1.79**	**2025**

Its grouped and pointwise convolution design maintained feature extraction capacity with minimal computational cost.

Combining all three modules yielded the highest gains: mAP50 improved by 2.6% on SIXray and 8.6% on GIXray, with a total reduction of 26.9% in computation and 30.6% in parameters. The complete FDD-YOLO model achieves an optimal balance of accuracy, efficiency, and model size, making it well-suited for high-performance real-world security inspection systems.

4.5. Comparison Experiments

To comprehensively evaluate the advancedness and comprehensive advantages of the model proposed in this paper, we selected several of the most representative advanced object detection models for comparison, including the two-stage Faster R-CNN [4], Mask R-CNN [27], Grid R-CNN and the single-stage POD-Y [29], AO-DETR [1], MLSA-YOLO [30], YOLOv10n [31], YOLOv11n [32], and YOLOv12n [33], RT-DETR, D-FINE, DEIM [34]. To maintain fairness, all comparative tests were conducted using the same hardware setup and dataset splitting.

1) Comparative Experiments on the SIXray Dataset:

The SIXray dataset contains many simple negative samples, challenging the model's ability to localize rare objects in complex scenes. Table 3 compares the performance of various models on this dataset.

As shown in Table 3, our model achieves the best performance among all compared models on the SIXray dataset, achieving 94.8% mAP50 and 75.2% mAP50-95. Compared to YOLOv12n, a strong competitor in the lightweight field, our model achieves a 2.1% improvement in mAP50 while reducing computational overhead and parameter count by 1.7 GFLOPs and 0.77 MB, achieving an optimal trade-off between accuracy and efficiency. More importantly, compared to the baseline model, YOLOv11n, our approach achieves a 2.6% improvement in mAP50 and reduces computational overhead and parameter count by 1.7 GFLOPs and 0.79 MB, fully demonstrating the effectiveness of our improved solution.

2) Comparative Experiments on the HIXray Dataset:

The HIXray dataset offers high-quality images and detailed annotations, making it ideal for evaluating the model's overall performance in realistic scenarios. The comparative results are shown in Table 4.

Table 4 Comparison results with advanced methods on the HIXray dataset

Methods	P	R	mAP50	mAP50:95	GFLOPs	Params	Year
Faster R-CNN	78.5	70.2	73.1	44.8	210.4	41.09	2016
Mask R-CNN	79.8	69.6	72.9	44.3	283.5	60.04	2017
Grid R-CNN	81.5	70.1	73.8	45.4	328.8	64.32	2019
POD-Y	84.3	72.8	76.4	47.5	108.1	47.19	2023
AO-DETR	84.2	73.5	75.8	47.0	268.9	58.38	2024
MLSA-YOLO	87.2	78.3	80.1	50.3	32.5	12.14	2024
YOLOv10n	87.9	78.6	78.9	49.8	8.2	2.69	2024
YOLOv11n	86.5	78.5	80.8	51.1	6.3	2.58	2024
YOLOv12n	87.1	78.7	80.6	50.7	6.3	2.56	2025
RT-DETR	88.5	79.2	81.5	51.5	107.21	32.1	2025
D-FINE	87.8	79.0	81.0	51.0	91.35	31.5	2025
DEIM	88.3	79.5	81.8	51.8	91.42	31.62	2025
FDD-YOLO	**89.8**	**82.1**	**84.0**	**52.7**	**4.6**	**1.79**	**2025**

Table 5. Comparison results with advanced methods on the GIXray dataset

Methods	P	R	mAP50	mAP50:95	GFLOPs	Params	Year

Faster R-CNN	68.0	52.0	58.5	36.3	210.4	41.09	2016
Mask R-CNN	67.5	51.5	57.1	35.4	283.5	60.04	2017
Grid R-CNN	69.3	52.7	59.3	36.9	328.8	64.32	2019
POD-Y	70.2	53.1	59.8	37.1	108.1	47.19	2023
AO-DETR	69.5	52.8	59.3	36.8	268.9	58.38	2024
MLSA-YOLO	69.3	55.2	58.5	36.3	32.5	12.14	2024
YOLOv10n	69.9	55.3	61.6	37.6	8.2	2.69	2024
YOLOv11n	73.0	55.1	63.2	39.2	6.3	2.58	2024
YOLOv12n	68.0	58.5	62.9	39.2	6.3	2.56	2025
RT-DETR	75.5	58.8	67.5	42.5	107.21	32.1	2025
D-FINE	74.2	57.9	66.2	41.8	91.35	31.5	2025
DEIM	75.8	58.5	67.9	42.7	91.42	31.62	2025
FDD-YOLO	**79.9**	**62.5**	**71.8**	**44.1**	**4.6**	**1.79**	**2025**

Table 6. Performance comparison results of different categories on the GIXray dataset

Methods	mAP50	Lighter	Pressure	Knife	Scissors	Powerbank	Zippo
Faster R-CNN	58.5	64.6	67.3	51.4	35.7	76.3	55.7
Mask R-CNN	57.1	63.1	65.7	50.1	34.8	74.4	54.3
Grid R-CNN	58.7	64.9	68.1	51.0	36.2	76.5	55.9
POD-Y	59.8	66.1	68.8	52.5	36.5	77.9	56.9
AO-DETR	59.3	65.5	68.2	52.1	36.2	77.3	56.4
MLSA-YOLO	58.5	64.6	67.3	51.4	35.7	76.3	55.7
YOLOv10n	61.6	68.1	70.9	54.1	37.6	80.3	58.6
YOLOv11n	63.2	72.7	73.6	58.0	39.1	80.8	55.0
YOLOv12n	62.9	71.6	74.1	58.6	35.4	81.6	56.0
RT-DETR	67.5	75.2	76.8	63.1	42.8	84.5	63.1
D-FINE	66.2	73.8	75.3	61.7	41.9	83.1	61.8
DEIM	67.9	75.5	77.1	63.4	43.1	84.8	63.4
FDD-YOLO	**71.8**	**80.8**	**81.4**	**65.5**	**45.2**	**89.9**	**67.2**

Table 7. Performance Comparison on Edge Devices

Methods	Devices	FPS	Size(MB)	Inference(ms)
YOLOv11n	Jetson Nano	12.3	12.45	81.3
FDD-YOLO	**Jetson Nano**	**17.8**	**8.95**	**56.2**
YOLOv11n	Jetson Xavier NX	35.1	12.68	28.5
FDD-YOLO	**Jetson Xavier NX**	**47.6**	**9.05**	**21.0**

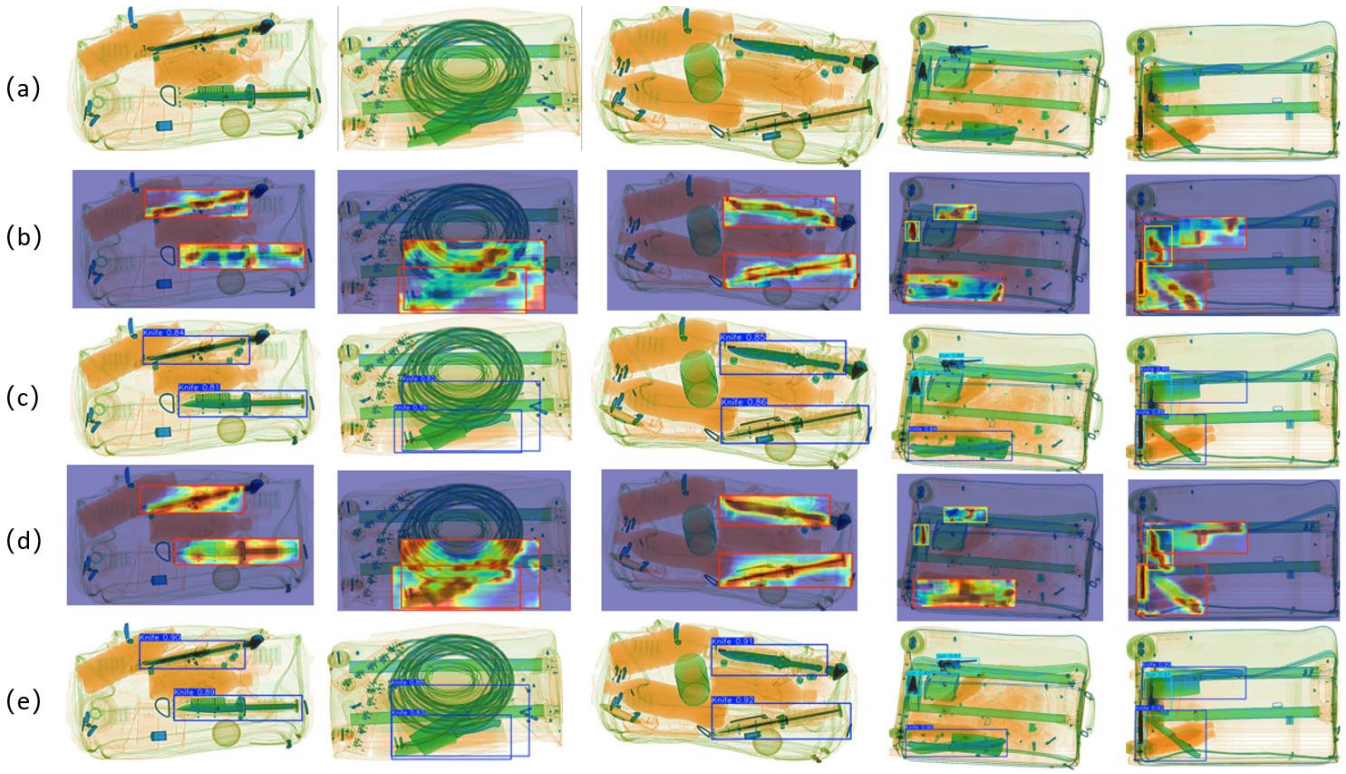


Figure 6. (a) Original image,(b) YOLOv11n baseline heatmap,(c) Baseline detection result,(d) FDD-YOLO heatmap, (e) FDD-YOLO detection result.

On the HIXray dataset, our model also demonstrated dominant performance, achieving a mAP50 score of 84.0%, surpassing all other compared models. Compared to our baseline model, YOLOv11n, we achieved a 3.2% lead in accuracy. Compared to YOLOv12n, we achieved a 3.4% lead in accuracy. Our proposed model outperformed all other two-stage models in accuracy, significantly reducing computational complexity and parameter count. This result demonstrates that our model is not overfit to a specific dataset but possesses strong generalization capabilities, enabling widespread improvement in detection accuracy on high-quality X-ray images.

3) Comparative Experiments on the GIXray Dataset

The GIXray dataset focuses on small contraband and is specifically designed to validate the model's robust detection capabilities for small objects, a core challenge in real-world security inspections. The comparative results are shown in Table 5.

Our model's advantages are most evident on the challenging GIXray small object dataset. Its mAP50 of 71.8% and Precision of 79.9% significantly surpass all compared models. Compared to YOLOv10n, its mAP50 is 10.2% higher. Compared to the baseline model YOLOv11-n, its mAP50 significantly improves by 8.6%. Table 6 shows that our model achieves significant improvements in mAP50 for the small object categories Lighter and Zippooil, demonstrating the exceptional effectiveness of our core innovations, the Frequency Domain Decomposition Network and the Deformable Elastic Fusion Pyramid, in enhancing detail perception and handling deformed and occluded small objects.

Extensive comparisons on these three authoritative datasets with varying focus confirm that our proposed improved model comprehensively surpasses existing detectors in detection accuracy while maintaining exceptionally low model complexity (number of parameters and computational effort), surpassing state-of-the-art detection models. Our work successfully achieved the goal of finding the optimal balance between accuracy and efficiency, providing a powerful solution for deploying high-performance X-ray contraband detection algorithms in embedded security inspection equipment with limited computing resources.

4.6. Deployment of edge devices

To validate the practical deployment efficacy of FDD-YOLO, we supplemented our evaluation with performance tests on typical edge devices. The testing platforms included the NVIDIA Jetson Nano (4GB) and Jetson Xavier NX, which are widely used in embedded AI applications. All models were optimized and accelerated using TensorRT, with an input resolution of 640×640 and a batch size of 1 to simulate real-time video stream processing. Key metrics including frames per second (FPS), peak memory consumption, and average latency are reported in Table 7.

4.7. Visualization Analysis

To gain a deeper understanding of the nature of FDD-YOLO's performance improvement over the baseline model YOLOv11n, this study conducted a detailed visual analysis. Based on samples from the test set, we compared the results across two dimensions: feature response and prediction.

The class activation heatmaps generated by Grad-CAM (Figure 6) demonstrate that our proposed model more accurately and densely focuses on the target, particularly in complex backgrounds and small target areas. While the baseline model's response is diffuse and easily distracted by background noise, FDD-YOLO's heatmaps are highly focused and fully cover the target area, demonstrating the FDDN module's ability to enhance high-frequency detail. In heavily occluded scenes, the baseline model exhibits only a weak response to visible portions, while our proposed model demonstrates strong anti-occlusion reasoning capabilities. The heatmap significantly covers occluded areas, demonstrating the role of the deformable convolutions in the DEFP module in adaptively adjusting the receptive field and inferring the complete structure. Even in a cluttered background with multiple metal objects, our proposed model clearly focuses on the subject and effectively suppresses background noise, demonstrating enhanced robustness to interference.

Visual analysis proves that FDD-YOLO can accurately allocate attention to the prohibited items area, which is a direct reflection of its high recognition rate. At the same time, it shows better stability and accuracy under occlusion, overlap and complex background conditions.

5. Conclusions and prospective research

This paper proposes a lightweight object detection model for X-ray security inspection images. It aims to address the inadequate performance of current detectors when dealing with challenges such as small objects, severe occlusion, and object deformation. Our core contribution is first introducing a Frequency Domain Decomposition Network that decouples and adaptively enhances the image's high-frequency details and low-frequency structural information, significantly improving the model's perception of blurred and small objects. Second, the Deformable Elastic Fusion Pyramid utilizes dynamic channel allocation and multi-scale deformable convolutions to fuse complex deformable objects and multi-scale features adaptively. Finally, the DualConv module is introduced to achieve efficient feature extraction using a parallel architecture, providing the model with excellent lightweight properties. Extensive experiments on three authoritative datasets, SIXray, HIXray, and GIXray, demonstrate that this scheme achieves state-of-the-art performance in balancing detection accuracy (mean Average Precision Index) and model efficiency (parameter count and computational effort). This approach provides a practical solution for deploying high-performance deep learning models in resource-constrained real-time security inspection systems.

Despite the encouraging results achieved in our current work, several directions remain worthy of future exploration. We have outlined a clear technological development roadmap. First, in terms of multimodal fusion, we will explore the deep fusion of material atomic number information and visual features based on dual-energy X-ray diffraction. Specific technical paths include: First, in early fusion, using material property maps as the fourth input channel of the network, establishing a correlation between physical properties and visual appearance from the very beginning of feature extraction. Then, employing a cross-modal attention mechanism, we will design an interactive attention module that allows RGB features and material features to guide and complement each other, enhancing each other to accurately distinguish between items that look similar but have vastly different materials (such as plastic toy guns and real iron). Second, addressing the core challenge of scarce labeled data, we will systematically study weak supervision and few-shot learning strategies, establishing image-level label-based detection. Using only weak labels such as "knife present in image" to train the detection model will significantly reduce labeling costs. Subsequently, we will establish a meta-learning framework to build a detector capable of quickly adapting to new categories of contraband, effectively detecting newly emerging threatening items with only a small number of samples (e.g., 1-5 images). This technical approach aims to fundamentally address the core pain points of data annotation difficulties and the rapid emergence of new threats in security inspection scenarios, and will greatly enhance the practicality and scalability of the model.

Acknowledgements.

This work was partially funded by the Xinjiang Uygur Autonomous Region Major Science and Technology Special Project (No. 2023A03001).

Natural Science Foundation Program of Xinjiang. Uygur Autonomous Region 2024D01A141, Tianchi Talents Program of Xinjiang Uygur Autonomous Region , the open project of Dazhou Key Laboratory of Government Data Security under grants ZSAQ202501, ZSAQ202502, and ZSAQ202507.

References

[1] Li M, Jia T, Wang H, et al. Ao-detr: Anti-overlapping detr for x-ray prohibited items detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024.

[2] Hassan T, Bettayeb M, Akçay S, et al. Detecting prohibited items in X-ray images: A contour proposal learning approach[C]//2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020: 2016-2020.

[3] Hättenschwiler N, Sterchi Y, Mendes M, et al. Automation in airport security X-ray screening of cabin baggage: Examining benefits and possible implementations of automated explosives detection[J]. Applied ergonomics, 2018, 72: 58-68.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-1149.
[5] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
[6] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
[7] Ou X, Chen X, Xu X, et al. Recent development in x-ray imaging technology: Future and challenges[J]. Research, 2021.
[8] Wang H, Jia T, Ma B, et al. Delving into cluttered prohibited item detection for security inspection system[J]. IEEE Transactions on Industrial Informatics, 2024, 20(10): 11825-11834.
[9] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
[10] Wang M, Du H, Mei W, et al. Weight-guided dual-direction-fusion feature pyramid network for prohibited item detection in x-ray images[J]. Journal of Electronic Imaging, 2022, 31(3): 033032-033032.
[11] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational visual media, 2022, 8(3): 331-368.
[12] Abbasi S, Mohammadzadeh M, Zamzamian M. A novel dual high-energy X-ray imaging method for materials discrimination[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2019, 930: 82-86.
[13] Xie X, Cheng G, Wang J, et al. Oriented R-CNN and beyond[J]. International Journal of Computer Vision, 2024, 132(7): 2420-2442.
[14] Zhang W, Zhu Q, Li Y, et al. MAM Faster R-CNN: Improved Faster R-CNN based on Malformed Attention Module for object detection on X-ray security inspection[J]. Digital Signal Processing, 2023, 139: 104072.
[15] Sagar A S M S, Chen Y, Xie Y K, et al. MSA R-CNN: A comprehensive approach to remote sensing object detection and scene understanding[J]. Expert Systems with Applications, 2024, 241: 122788.
[16] Zhang H, Teng W, He X, et al. Lightweight prohibited items detection model in X-ray images based on improved YOLOv7-tiny[J]. Journal of the Franklin Institute, 2025, 362(1): 107421.
[17] Guan F, Zhang H, Wang X. An improved YOLOv8 model for prohibited item detection with deformable convolution and dynamic head[J]. Journal of Real-Time Image Processing, 2025, 22(2): 84.
[18] Zhao C, Zhu L, Dou S, et al. Detecting overlapped objects in X-ray security imagery by a label-aware mechanism[J]. IEEE transactions on information forensics and security, 2022, 17: 998-1009.
[19] Ding J, Ye C, Wang H, et al. Foreign bodies detector based on detr for high-resolution x-ray images of textiles[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.
[20] Zhou Y, Xu X, Wang R. EI-YOLO: Efficiently Improved YOLO on Detection of Prohibited Items During Security Inspections[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Singapore: Springer Nature Singapore, 2024: 330-343.
[21] Zhou Y T, Cao K Y, Li D, et al. Fine-YOLO: a simplified X-ray prohibited object detection network based on feature aggregation and normalized Wasserstein distance[J]. Sensors, 2024, 24(11): 3588.
[22] Jia L, Wang T, Chen Y, et al. MobileNet-CA-YOLO: An improved YOLOv7 based on the MobileNetV3 and attention mechanism for Rice pests and diseases detection[J]. Agriculture, 2023, 13(7): 1285.
[23] Zhong J, Chen J, Mian A. DualConv: Dual convolutional kernels for lightweight deep neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 34(11): 9528-9535.
[24] Srivastava H, Sarawadekar K. A depthwise separable convolution architecture for CNN accelerator[C]//2020 IEEE Applied Signal Processing Conference (ASPCON). IEEE, 2020: 1-5.
[25] Miao C, Xie L, Wan F, et al. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2119-2128.
[26] Tao R, Wei Y, Jiang X, et al. Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10923-10932.
[27] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
[28] X. Lu, B. Li, Y. Yue, Q. Li, J. Yan, Grid R-CNN, in: Proc. IEEE Conf. Comput. Vis.Pattern Recog, IEEE, Long Beach, CA, USA, 2019, pp. 7355 – 7364, https://doi.org/10.1109/CVPR.2019.00754.
[29] Ma C, Zhuo L, Li J, et al. Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation[J]. Neurocomputing, 2023, 519: 1-16.
[30] Peng J, Lv K, Wang G, et al. MLSA-YOLO: a multi-level feature fusion and scale-adaptive framework for small object detection[J]. The Journal of Supercomputing, 2025, 81(4): 528.
[31] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems, 2024, 37: 107984-108011.
[32] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arXiv preprint arXiv:2410.17725, 2024.
[33] Tian Y, Ye Q, Doermann D. Yolov12: Attention-centric real-time object detectors[J]. arXiv preprint arXiv:2502.12524, 2025.
[34] Huang S, Lu Z, Cun X, et al. Deim: Detr with improved matching for fast convergence[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 15162-15171.