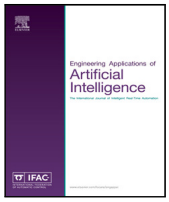
Engineering Applications of Artificial Intelligence 133 (2024) 108076

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

FDTNet: Enhancing frequency-aware representation for prohibited object detection from X-ray images via dual-stream transformers

Ziming Zhu, Yu Zhu *, Haoran Wang, Nan Wang, Jiongyao Ye, Xiaofeng Ling

School of Information Science and Engineering, East China University of Science and Technology, Shanghai, 200237, China

ARTICLE INFO

Keywords:
Prohibited object detection
X-ray image
Frequency feature perception
Dual-stream network
Attention module

ABSTRACT

With the extensive application of object detection in intelligent security, the demand for detecting prohibited items in X-ray images has become increasingly stringent. Unlike natural images, X-ray images present unique challenges such as complex backgrounds and mutual occlusion between prohibited and normal items. Consequently, applying traditional detection methods to X-ray images remains a significant challenge. To tackle these challenges, we have developed a unique frequency-aware dual-stream transformers (FDTNet) that is specifically designed for analyzing X-ray images. The FDTNet consists of two streams: one handles the original image, while the other deals with an image that has been enhanced with frequency domain features. In order to achieve precise detection of prohibited items, we introduce a frequency-aware module (FAM) that enhances the representation of prohibited items by utilizing information from the frequency domain. This FAM can be easily integrated into other backbones or detectors as it is a plug-and-play module. Additionally, to enhance the fusion of feature maps from both streams, we utilize a global and channel attention module (GCA) that aggregates texture representations for spatial feature streams. Our evaluation of the proposed FDTNet on the OPIXray datasets and PIDray datasets demonstrates that our detection mAP achieves 88.02 and 68.2, respectively. Extensive experiments conducted on publicly available datasets provide substantial evidence that our proposed network significantly improves the detection of prohibited items compared to state-of-the-art methods.

1. Introduction

Safety inspections are crucial for upholding social and public safety during activities like using public transportation or accessing sensitive departments. Security personnel rely on X-ray images from inspection machines to detect prohibited items concealed within backpacks. However, individuals often intentionally hide such items within non-prohibited objects to evade detection, posing a significant threat to public safety. Thus, there is an urgent need for a fast and effective method to help security inspectors accurately identify prohibited items in X-ray images.

X-rays are high-frequency electromagnetic waves with short wavelengths (ranging from 0.01 nm to 10 nm) and strong penetration capabilities. When an X-ray source irradiates an object during inspection, its absorption varies due to differences in density among objects. The resulting X-ray images reflect this information, such as metal objects appearing blue. Security inspectors can thus assess whether prohibited items are present without needing to unpack bags or containers. However, deliberate attempts to conceal prohibited items present challenges during inspections. Objects stacked within boxes or backpacks can cause overlapping in X-ray images, resulting in disorderly appearances that complicate item identification. Consequently, detecting prohibited items solely through X-ray imaging remains a challenging task due to the unique characteristics of these images compared to traditional natural ones. Existing methods for analyzing X-ray images fall into two main categories: one focuses on enhancing features by extracting edge information while the other enhances low-level and high-level features using different approaches aimed at improving classification and localization abilities. While significant progress has been made in applying object detection techniques to X-ray imaging tasks, there is still ample room for improvement—particularly regarding small-sized prohibited item identification and addressing overlap between objects within these specialized types of imagery.

Compared to objects in natural visible light images, objects within airtight packages in X-ray images possess distinct characteristics. In X-ray images, these objects are randomly overlapped and placed together (Mery et al., 2016; Ma et al., 2023; Miao et al., 2019). Consequently, object detection becomes challenging.

* Corresponding author.
E-mail address: zhuyu@ecust.edu.cn (Y. Zhu).

https://doi.org/10.1016/j.engappai.2024.108076
Received 13 October 2023; Received in revised form 8 January 2024; Accepted 8 February 2024
Available online 16 February 2024

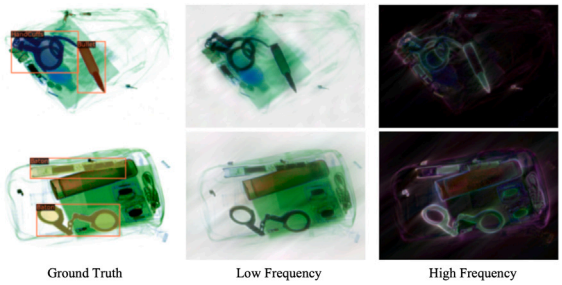


Fig. 1. From left to right, it represents the real annotation of the image object, the low frequency component and the high frequency component of the image.

X-ray security scanners display objects in different colors to help workers identify them, especially in images with complex backgrounds. The low-frequency component of an image represents areas where the intensity shift is minimal, while the high-frequency components refer to the edges and contours. To extract these frequency information, we utilize Fourier transform along with low-pass and high-pass filters, which are then visualized through inverse Fourier transform (Fig. 1). By extracting the high-frequency content, prohibited items can be effectively separated from the background. Hence, image frequency information serves as a valuable supplement to RGB data. We aim to leverage the inherent features outlined by high-frequency information for detecting prohibited items. Specifically, we focus on enhancing edge details by extracting features from high-frequency images. Various techniques exist for obtaining different frequency components within an image. For instance, Chen et al. (2021) utilized Discrete Cosine Transform (DCT) to convert RGB-domain images into frequency-domain representations and proposed an RGB-Frequency Attention Module (RFAM) for comprehensive feature representation by fusing RGB and frequency domains. Similarly, F^3-Net (Qian et al., 2020) employed Frequency-aware Decomposition (FAD) and Local Frequency Statistics (LFS) to detect fake elements in human faces. To emphasize key characteristics of targeted prohibited items during detection processes, enhancement via frequency angle is being considered.

Due to the specific requirements of security inspection tasks, the current public datasets for prohibited items in X-ray images primarily include OPIXray (Wei et al., 2020), PIDray (Wang et al., 2021c), GDXray (Mery et al., 2015), and SIXray (Miao et al., 2019). However, GDXray has a limited number of samples, featuring only three categories of prohibited items and lacking complex backgrounds and occlusions. As a result, object detection within this dataset does not pose significant challenges and offers limited support for contraband object detection research. On the contrary, although SIXray consists of an extensive collection of 1,059,231 images, the actual instances depicting prohibited items are relatively low, with only 8929 instances, representing merely 0.84% of all images. Additionally, both GDXray and SIXray datasets primarily focus on classification tasks and do not provide bounding box annotations necessary for precise object localization. Therefore, to address these limitations and ensure better representation in terms of category size and sample size for evaluating contraband object detection research, we have chosen the OPIXray and PIDray datasets as more suitable evaluation datasets.

In this paper, we introduce a novel dual-stream frequency-aware detection network that leverages both RGB information and frequency domain information from X-ray images. Our approach employs two distinct backbones to extract RGB features and frequency domain features separately, without parameter sharing. These feature maps are then fused to enhance the overall image representation. To specifically enhance the features in the frequency domain perspective, we propose a simple and flexible frequency-aware module (FAM). This module aggregates the features in high-frequency images, improving their discriminative power. Moreover, to better integrate the original RGB features with the frequency domain features, we propose a plug-and-play global and channel attention module (GCA). GCA combines spatial global representation abilities inspired by transformer with attention mechanisms in the channel dimension. By adopting this complementary approach, our method effectively promotes feature map representation learning. The main contributions of this paper can be summarized as follows:

- We propose a dual-stream frequency-aware detection network. One branch of the network focuses on extracting RGB features from the image, while the other branch extracts features from a frequency-enhanced feature map. To effectively fuse RGB and frequency information for prohibited item detection, we propose a global and channel attention module (GCA). This module is designed to enhance the feature map generated by the RGB branch through attention mechanisms in both spatial and channel dimensions.
- In light of the distinctive frequency information exhibited by prohibited items in X-ray images compared to the background, we have developed a frequency-aware module (FAM) to extract this specific characteristic. Within the FAM, we employ a frequency feature enhancement (FFE) operation that enhances the image features from a frequency perspective. This enables us to effectively capture and leverage the unique frequency information associated with prohibited items for improved detection performance.
- Extensive experiments on the OPIXray (Wei et al., 2020) and PIDray (Wang et al., 2021c) public datasets confirm the superiority of our proposed FDTNet over existing methods in detecting prohibited items. Our method achieves a detection mAP of 88.02 and 68.2 on PIDray and OPIXray datasets, respectively. Additionally, during testing runs on an NVIDIA TITAN RTX GPU, our method achieves a frame rate of 13.3 fps. This ensures that our method can be practically applied in security scanning scenarios without experiencing significant delays or disruptions when objects pass through the scanners.

2. Related work

2.1. X-ray object detection

The De-occlusion Attention Module (DOAM) (Wei et al., 2020) enhances image features by extracting edge and material information from prohibited items. These enhanced features are then fed into subsequent feature extraction and detection networks. EAOD-Net (Ma et al., 2022) incorporates a learnable Gabor revolution layer to capture the edge information of prohibited items, while CFPA-Net (Wei et al., 2021) utilizes the Cross-Layer Feature Extraction Fusion Module (CEF) to enhance semantics and localization information across high-level and low-level features. The Parallel Attention module (PA) in CFPA-Net captures long-range contextual information, resulting in more detailed features. Chen et al. (2023) introduce a mixed samples-driven methodology with DDPM to overcome small sample size limitations in x-ray image analysis for CFCS damage identification, incorporating synthesis of new samples through DDPM, integration with authentic measurements, and employing a DenseNet-based module within a mixed samples-driven architecture for diagnosis. IEFPN, proposed by Wang et al. (2021a), builds upon FPN by re-weighting different layers' features using a layer-based recalibration module (LRM). This promotes better exchange of information between feature layers. Additionally, IEFPN enhances location information in low-level features through the Channel-Attention-based Skip connection path (CSP). To address overlapping prohibited items, Zhao et al. (2022) introduce a new tag assignment method based on ATSS as an effective solution. Liu et al. (2022) propose a consistent multiscale feature mapping method with a combination of multiscale feature mapping, consistency strategy, and feature fusion model to improve the recognition

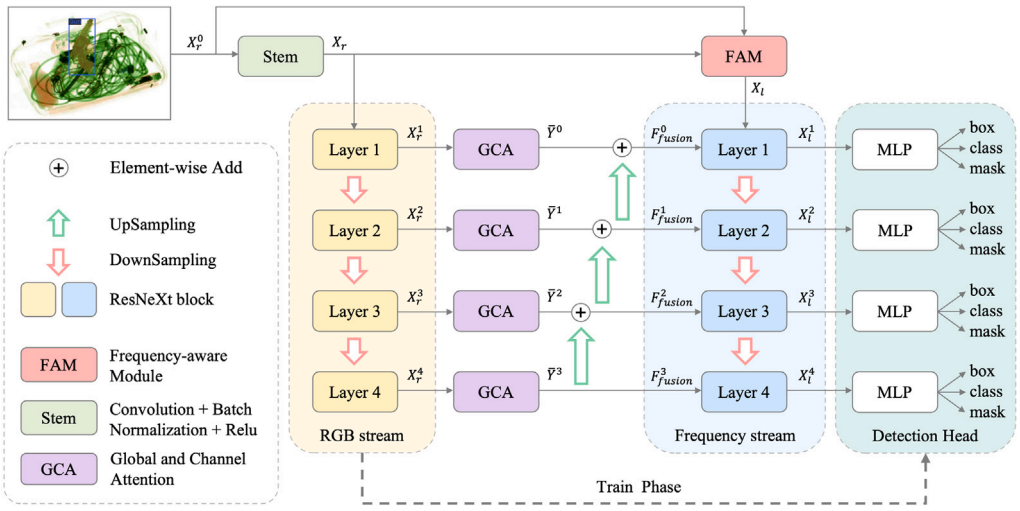


Fig. 2. Overall structure of the proposed FDTNet. Our approach employs a X-ray image as input for prohibited object detection. To enable frequency feature encoding and interaction, we introduce a dual-stream fusion model. The yellow branch represents the network's image RGB feature stream, and the blue branch represents the network's image frequency feature stream.

of defects in X-ray images. Ding et al. (2023) propose FE-DETR, a transformer-based object detection framework that improves the performance of anchor-based detectors for detecting foreign bodies (FBs) through split-attention, CBAM and DCN integration, MSFE module for feature dispersion handling, transformer as prediction head, and optimized training strategies.

These methods collectively aim to improve object detection performance by leveraging various attention mechanisms, fusion techniques, recalibration modules, and innovative tag assignment strategies.

2.2. Dual-stream network

The primary objective of the dual-stream network is to acquire more comprehensive image features from different perspectives. Conformer (Peng et al., 2021) achieves this by combining local convolutional features and global transformer-based features in a parallel and interactive manner. This fusion process results in richer image features. DS-Net (Mao et al., 2021), built upon the traditional ResNet block (He et al., 2016), organizes feature maps in the channel dimension and extracts local and global features using both convolutional and transformer operations. These extracted features are then fused through cross-attention mechanisms and concatenation. CBNet (Liu et al., 2020) employs multiple backbones with identical structures but independent parameters to extract image features. Each stage of the assistant backbone transmits its output to subsequent backbones via composite connections. The multi-scale feature maps generated by the final backbone are utilized for detection and segmentation tasks. PEL (Gu et al., 2021) utilizes a dual-stream network with EfficientNet (Tan and Le, 2019) as its backbone architecture. It takes RGB images as well as fine-grained frequency components as input, enhancing feature representations between streams through mutual enhancement modules.

In summary, these approaches employ various techniques such as parallel fusion of local–global information, composite connections among multiple backbones, or mutual enhancement between streams to obtain richer image features within their respective dual-stream networks.

2.3. Attention mechanism and transformer

The attention mechanism's primary goal is to mimic the human visual system, prioritizing important objects in an image rather than irrelevant backgrounds. SENet (Hu et al., 2018) utilizes global average pooling to represent overall information and a squeeze-and-excitation module to learn channel connections. CBAM (Woo et al., 2018) further enhances feature representation through both channel attention and spatial attention. Transformer (Vaswani et al., 2017) excel at extracting global features and have demonstrated outstanding performance in NLP tasks. In image classification, ViT (Dosovitskiy et al., 2020) introduced multi-head self-attention (MHSA). This concept inspired the development of various transformer-based backbone networks like PVT (Wang et al., 2021b), Swin (Liu et al., 2021), and MPViT (Lee et al., 2022) for image dense prediction tasks by dividing images into patches.

In order to learn the relationship between channels and global pixel locations of feature maps, we introduce channel attention into the traditional transformer block. This enhancement enables us to better understand how each channel contributes to the overall information across different areas of the image.

3. The proposed method

The framework of the proposed FDTNet is shown in Fig. 2. Frequency-aware module (FAM) obtains the frequency information of the input RGB image through SRM filter and the frequency feature enhancement operation in Section 3.2. The original image and the frequency image from FAM as two separate inputs are fed into a dual-stream network with ResNeXt101 as the backbone. To better fuse the features from RGB image branches, we proposed global and channel attention module (GCA) for extracting global features and channel features in Section 3.3.

3.1. Network architecture

We feed RGB X-ray image into the first backbone, which we denote as X_r. In the yellow branch shown in Fig. 2, the various stages of the first branch extract feature maps of different scales, referred to as X_r^s, where $0 \leq s \leq 4$. Simultaneously, the frequency information of the input image is extracted and fused with the feature image X_r^0 obtained from the previous backbone. The resulting fused feature map, denoted as X_l, is then input into another branch. This process enhances the detection performance of the network by extracting and enhancing the image's frequency information. Our detection heads are constructed using a simple MLP layer, eliminating the need for complex designs.

During the training phase, the feature maps X_r^s and X_l^s obtained from the various branches are utilized for detection tasks. Specifically,

the feature maps corresponding to $1 \le s \le 4$ are used and supervised in the detection heads.

We designate the feature map pixel closest to the center of each labeled bounding box as a positive sample and supervise it using focal loss (Lin et al., 2017), denoted as L_{pos}. We directly regress the class scores, bounding box coordinates and masks from the pixel features of the multi-scale feature map. The classification loss and mask loss, represented by L_{cls} and L_{seg}, are calculated using cross-entropy loss. For each bounding box, we predict the position offset $(\delta x, \delta y) \in R^{1\times2}$ and the two-dimensional sizes $(l, w) \in R^{1\times2}$. The regression loss, denoted as L_{reg}, is computed using L1 loss. To further enhance performance, we incorporate the IoU loss (L_{IoU}) between the predicted box and the ground truth box (Zhou et al., 2019). The total loss function is defined by assigning weights to the positive, classification, regression, IoU and mask components:

$$L_{total} = 0.85L_{pos} + 0.95L_{cls} + 0.25(L_{reg} + L_{IoU}) + 0.5L_{seg} \tag{1}$$

The overall loss of the network is as follows:

$$L = L_{total}^{lead} + \lambda L_{total}^{assist} \tag{2}$$

Where L_{total}^{lead} represents the total loss of the frequency stream, and L_{total}^{assist} represents the total loss of the original RGB stream. It is important to note that during the test phase, X_r^s will not be fed into the detection heads.

3.2. Frequency-aware module

The frequency-aware module is mainly composed of two parts: frequency information extraction and frequency feature enhancement, as shown in Fig. 3. To obtain the frequency information of the X-ray image, for an input RGB image X_r with size $H \times W \times 3$, use SRM filter (Fridrich and Kodovsky, 2012) to convert it into frequency image X_h with size $H\times W\times3$. SRM filter consists of three fixed filter operators (f_1, f_2, f_3).

$$f_1 = \frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$f_2 = \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & -8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & -8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \tag{3}$$

$$f_3 = \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$X_h = SRM(X_r, [f_1, f_2, f_3]) \tag{4}$$

In order to use the feature map X_r^0 from the lowest layer of RGB branch, we use the convolutional kernel with kernel size of 7, step size of 2, and padding of 3 to increase the number of image channels passing through the SRM filter from 3 to 64. Secondly, max pooling with a kernel size of 3×3 is used to reduce the width and height of the frequency image X_h to $\frac{1}{4}$ of the original image. The specific formula is as follows:

$$\overline{X_h} = MaxPooling_{3\times3}(Conv_{7\times7}(X_h)) \tag{5}$$

Inspired by CBAM (Woo et al., 2018), average pooling and maximum pooling are performed on the frequency image $\overline{X_h}$ along the channel dimension, and spatial attention is used to obtain edge features

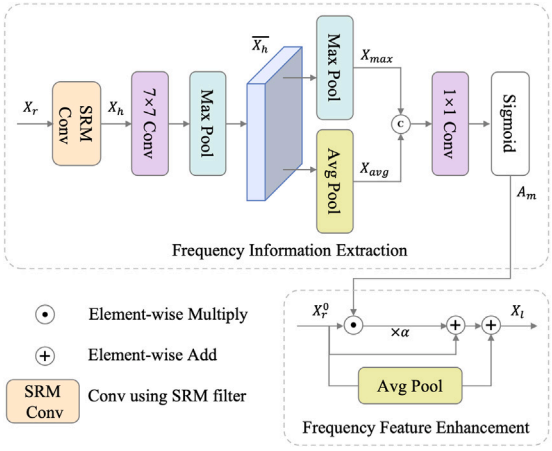


Fig. 3. Framework of frequency-aware Module.

in the frequency image $\overline{X_h}$ to generate feature maps $X_{avg} \in R^{\frac{H}{4}\times\frac{W}{4}\times1}$ and $X_{max} \in R^{\frac{H}{4}\times\frac{W}{4}\times1}$. After concat X_{avg} and X_{max}, use a 1×1 convolution to generate an attention map, denoted as $A_m(\frac{H}{4} \times \frac{W}{4} \times 1)$.

$$A_m = \sigma(Conv_{1\times1}(Concat(MaxPooling(\overline{X_h}), AvgPooling(\overline{X_h})))) \tag{6}$$

Where σ represents the sigmoid activation function. To enable the model to learn adaptively based on the frequency information of different images, we introduce a learnable hyper parameter α. In order to highlight the edge high-frequency feature of prohibited items, multiply X_r^0 with attention map A_m and enhance it with parameter α. To supplement the low-frequency information in X_r^0, average pooling is a simple and effective way to extract the low-frequency information of the feature map. Therefore, the calculation formula of feature map X_l that input into the second backbone is as follows:

$$X_l = (1 + \alpha \times A_m) \odot X_r^0 + AvgPooling_{3\times3}(X_r^0) \tag{7}$$

3.3. Global and channel attention

Due to the global representation capability of self-attention, we propose the global and channel attention module (GCA), as shown in Fig. 4. To reduce the number of parameters and computation of the model, the feature maps X_r^3 and X_r^4 from the RGB branch use a 3×3 convolution, batch normalization(BN) and ReLu to reduce the channel dimension. The number of channels of feature maps of X_r^1 and X_r^2 remains unchanged, and the obtained feature map is marked as F_r^s:

$$F_r^s = \delta(BN(Conv_{3\times3}(X_r^s))) \tag{8}$$

Where δ represents the Relu activation function. The feature map F_r^s is operated by three branches, as shown in Fig. 4. In order to obtain global features F_{gl}^s of different scale feature maps, the structure of transformer is used to calculate global attention:

$$\overline{F_r^s} = SRAttention(L_q(LN(F_r^s)), L_k(LN(F_r^s)), L_v(LN(F_r^s))) + F_r^s \tag{9}$$

$$\begin{aligned} L_q(F) &= FW^Q \in R^{hw\times C} \\ L_k(F) &= DownSampling(F, p)W^K \in R^{\frac{hw}{p^2}\times C} \\ L_v(F) &= DownSampling(F, p)W^V \in R^{\frac{hw}{p^2}\times C} \end{aligned} \tag{10}$$

$$SRAttention(Q, K, V) = Softmax(\frac{Q * K^T}{d_{head}}) * V \tag{11}$$

$$F_{gl}^s = MLP(LN(\overline{F_r^s})) + \overline{F_r^s} \tag{12}$$

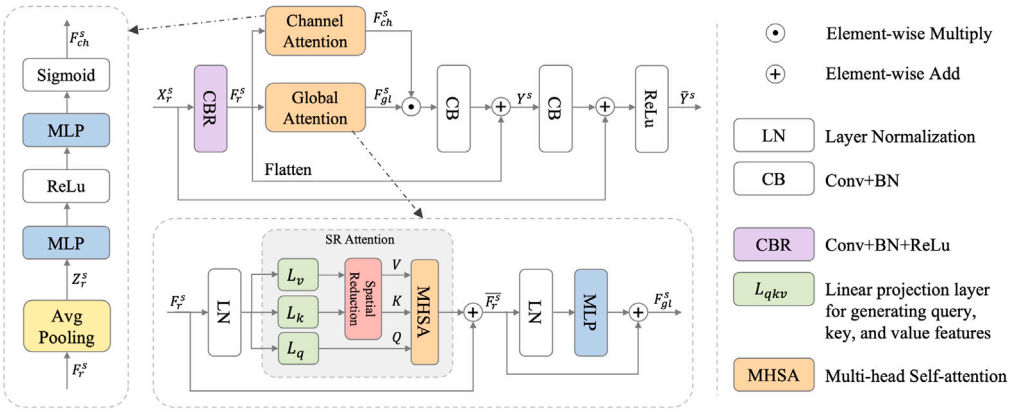


Fig. 4. Global and channel attention module.

Where LN stands for layer normalization, d_{head} means the dimension of each head in the multi-head attention mechanism is set to 64 in the experiment. $W^{Q/K/V}$ represent three learnable matrices for obtaining query, key and value with different weights. During the calculation process, the key and value are down-sampled to reduce the amount of calculation.

In order to obtain the relationship F_{ch}^{s} on the channel dimension of the feature map, an additional branch is added to the original transformer structure for calculating channel attention:

$$Z_r^s = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_r^s(i,j) \tag{13}$$

$$F_{ch}^{s} = \sigma(W_2 \delta(W_1 Z_r^s)) \tag{14}$$

$$Y^s = BN(Conv_{1\times1}(F_{ch}^{s} \odot F_{gl}^{s})) + F_r^s \tag{15}$$

Where W_1 and W_2 represent the two parameters of the fully connected layer. F_{gl}^{s} and F_{ch}^{s} perform feature fusion through element-wise multiplication and 1×1 convolution, and add skip connections at the same time.

When the GCA module acts on the third and fourth layer feature map X_r^3 and X_r^4, an additional 3×3 convolutional layer (as shown in the dashed box) and batch normalization are used to change the channel dimension of the feature map Y^s to make it equal to the number of channels at the time of input for subsequent multi-scale feature fusion, and finally use the Relu activation function to obtain the final output:

$$\overline{Y^s} = \delta(BN(Conv_{3\times3}(Y^s)) + X_r^s) \tag{16}$$

In the second branch, the feature maps from the RGB branch are fused before extracting features at different stages. To fuse the feature maps of different scales, $\overline{Y^s}$ will be up-sampled so that they have to the same scale. The fusion feature F_{fusion}^{k} is expressed as following:

$$F_{fusion}^{k} = \sum_{s=k+1}^{4} UpSampling((BN(Conv_{1\times1}(\overline{Y^s})))) \qquad 0 \le k \le 3 \tag{17}$$

4. Experiments

We conducted extensive experiments on the OPIXray (Wei et al., 2020) dataset and PIDray (Wang et al., 2021c) dataset, compared the proposed method with several state-of-the-art methods, and then demonstrated the effectiveness of the proposed module in the method by ablation study.

4.1. Datasets

We evaluate the proposed network on two prohibited items detection datasets. One is OPIXray Dataset. It contains 5 categories (Folding Knife (FO), Scissor (SC), Straight Knife (ST), Multi-tool Knife (MU) and Utility Knife (UT)), including 7109 train images and 1776 test images.

The other is perfermed on PIDray Dataset. It contains 12 categories (Baton, Pliers, Hammer, Power-bank, Scissors, Wrench, Gun, Bullet, Sprayer, Hand-Cuffs, Knife, Lighter), including a training set of 29,457 images and a test sets of 18,220 images, where the test sets is divided into easy, hard and hidden according to the difficulty of prohibited items detection. Hidden mode indicates that the prohibited item in the image is intentionally hidden. The results of each test set and the average of the three sets are recorded.

4.2. Implementation details

We employ the MMDetection (Chen et al., 2019) toolkit to implement our method, which is executed on a machine with NVIDIA TITAN RTX 24 GB. For the sake of fairness, all methods are trained with the train set, and the test set is used to evaluate. The proposed FDTNet uses ResNeXt101 as the backbone. Resize image to 512×512, and the entire network is trained with the stochastic gradient descent (SGD) algorithm with a momentum of 0.9 and a weight decay of 0.001. We train detectors for 12 epochs. The initial learning rate is set to 0.001 and the batch size is set to 4. Other parameters are the same as the default settings in MMDetection. After adjusting the input size of the image, the training settings used remain unchanged.

4.3. Evaluation metrics

For OPIXray, we adopt the evaluation metric in the article of OPIXray. For PIDray, we use COCO (Lin et al., 2014) evaluation metrics and calculate the mean precision (mAP). The bounding boxes are first ranked according to the confidence scores. The Intersection over Union (IoU) of each bounding box is calculated. The mAP score is calculated according to the different IoU thresholds set. mAP represents the mean average precision computed over the 10 IoU thresholds of 0.5:0.05:0.95, which is the primary challenge metric. mAP_{50} represents the mean average precision computed at a single IoU threshold of 0.5.

4.4. Result and analysis

In order to verify the effectiveness of the method proposed in this paper, we first conducts experiments on the OPIXray dataset, and compares it with several existing detection methods for prohibited items in X-ray images, as shown in Table 1. Among them, the size of all experimental input images is 512×512. Meanwhile, since the dataset

Table 1
Results of different detectors on the OPIXray dataset.

Method	FO	ST	SC	UT	MU	mAP_{50}
SSD (Liu et al., 2016)	76.91	35.02	93.41	65.87	83.27	70.90
SSD+DOAM (Wei et al., 2020)	81.37	41.50	95.12	68.21	83.83	74.00
SSD+LIM (Tao et al., 2021)	81.40	42.40	95.90	71.20	82.10	74.60
TST (Hassan and Werghi, 2020)	80.24	56.13	89.34	72.89	78.02	75.32
Cascade R-CNN+IEFPN (Wang et al., 2021a)	86.00	**70.22**	89.90	78.09	75.20	79.88
Faster R-CNN+IEFPN (Wang et al., 2021a)	86.39	64.18	88.74	**82.20**	**87.16**	81.73
FCOS (Tian et al., 2019)	86.41	68.47	90.22	78.39	86.60	82.02
FDTNet(ours)	**87.90**	60.20	**96.10**	78.90	87.10	**82.04**
Tensor pooling-driven (Hassan et al., 2022)[a]	85.28	**76.49**	88.03	80.62	89.41	83.96
FDTNet(ours)[a]	**89.10**	70.10	**96.60**	**84.80**	**90.50**	**86.20**
EAOD-Net (Ma et al., 2022)[b]	89.60	76.10	90.70	83.20	89.20	85.76
MCIA-FPN (Wang et al., 2022)[b]	89.08	74.48	89.99	**86.13**	89.75	85.89
POD-F-X (Ma et al., 2023)[b]	89.40	**78.70**	90.60	83.30	88.70	86.10
FDTNet(ours)[b]	**91.50**	74.60	**97.60**	85.20	**91.20**	**88.02**

[a] Indicates that the input size of the experiment is adjusted from 512 × 512 to 576 × 768.
[b] Indicates that the input size of the experiment is adjusted from 512 × 512 to 1333 × 800.

does not have annotations for semantic segmentation, the Cascade R-CNN detection framework that does not include the Mask branch is used. Due to the different sizes of the input images set in the experiments of different papers, for the sake of fairness, the methods marked with * and ** in the experimental results indicate that the size of the training image is 576 × 768 and 1333 × 800 respectively, and the aspect ratio of the image is maintained during the adjustment process. As the size of the input image becomes larger, the detection effect is improved to a certain extent. It can be seen from

Table 1 that under the input condition of 512 × 512, the method in this chapter has achieved the best performance in the two categories of folding knife (FO) and scissors (SC) and the average of five categories mAP_{50}, and is 8.04% higher than the DOAM method proposed on the OPIXray dataset. When the size of the input image is expanded to 1333 × 800, the detected mAP_{50} is improved by 5.98%. The proposed method achieves state-of-the-art performance for the scissors (SC) category under three input sizes. The Tensor pooling-driven algorithm (Hassan et al., 2022) highlights the contour features of prohibited items by using the tensor pooling module to generate multi-scale tensor maps. The MCIA-FPN network (Wang et al., 2022) uses average pooling and standard pooling to obtain the material features of prohibited items in the image, uses convolution to establish local cross-channel interactions, and converts material weights into channel weights. Compared with MCIA-FPN, the method in this paper achieves the best results in five results.

It is worth noting that in certain scenarios, FDTNet outperforms certain models with larger input image sizes of 1333 × 800 when the input image size is set to 576 × 768. This superiority can be attributed to our focus on not only extracting RGB features from the image but also leveraging the dual-stream architecture to extract features from frequency-enhanced feature maps. Additionally, the introduction of GCA enhances the frequency domain features that are more relevant to X-ray images through the attention mechanism in the spatial and channel dimensions. This means that while expanding the receptive field, FDTNet also preserves the fine-grained feature representation of frequency features.

In the OPIXray dataset, the test set is divided into OL1 (no occlusion or slight occlusion), OL2 (partial occlusion) and OL3 (severe occlusion or complete occlusion) according to the degree of occlusion. Table 2 records the test results of the method proposed in this chapter in three test sets with different occlusion levels, indicating that the method proposed in this chapter can achieve a certain degree of improvement regardless of the occlusion level.

The data in Table 3 are experiments conducted on the PIDray dataset. Since there are few studies on the PIDray dataset, the methods in this chapter are first compared with some of the most common target detectors. The data in the experiment are retrained owned. It

Table 2
Detection results of OPIXray dataset under different occlusion levels.

Method	OL1	OL2	OL3
SSD (Liu et al., 2016)	75.45	69.54	66.30
SSD+DOAM (Wei et al., 2020)	77.87	72.45	70.78
Tensor pooling-driven (Hassan et al., 2022)	79.46	73.82	72.91
MCIA-FPN (Wang et al., 2022)	82.24	81.71	79.58
Faster R-CNN+IEFPN (Wang et al., 2021a)	82.49	80.82	80.49
FDTNet(ours)	**82.60**	**82.30**	**80.60**

can be seen that the FDTNet detection network designed in this paper achieves the best detection performance and segmentation performance on all test sets. Compared with the SDANet network proposed on the PIDray dataset, the method in this chapter improves the average results of the three test sets by 6.6%, where the experimental data comes from Ref. Wang et al. (2021c). Both the method proposed in this chapter and CBNetV2 use two backbone networks without shared parameters (such as ResNeXt101), and both use Cascade Mask R-CNN as the detection framework. The superior performance of the CBNetV2 backbone network is illustrated by comparison. Meanwhile, the method proposed in this chapter improves the detection mAP on the simple test set by 1.2% compared with the original CBNetV2, and the average mAP on the three test sets increases by 0.7%. Relevant metrics for semantic segmentation also show significant improvements, from 55.4% mAP to 56.1%. Experimental results demonstrate the effectiveness of the proposed method. The parameters of the model are 198.58M, and when the input image size is 512 × 512, the detection speed can reach 13.3 frames per second. During the security check process, it can meet the needs of real-time detection.

Since the PIDray dataset was proposed in SDANet, we chose SDANet for the comparison of detection results. Fig. 5 shows the comparison between our method and SDANet in terms of detection results and segmentation results, as well as the visualization of feature maps in the final stages of the two branches. The first column in the figure shows the ground truth annotations of prohibited items, the second column shows the prediction results of SDANet, and the third column shows the prediction results of our method. By comparing the detection results, it can be seen that our method has higher accuracy without generating false positive samples.

The fourth column in Fig. 5 displays the visualization of the RGB branch feature maps in our proposed method, and the last column shows the visualization of the feature maps after frequency enhancement. By visualizing the feature maps extracted by the two branches in the final stages, it can be seen that after frequency domain feature enhancement, the network's attention can be more significantly focused on the region where prohibited items are located, indicating that the importance of frequency domain features is stronger than RGB

Table 3
Results of different detectors on the PIDray dataset.

Method	Detection *mAP*				Segmentation *mAP*			
	Easy	Hard	Hidden	Overall	Easy	Hard	Hidden	Overall
FSAF (Zhu et al., 2019)	65.2	59.0	48.0	56.7	–	–	–	–
RetinaNet (Lin et al., 2017)	66.4	58.1	45.8	56.8	–	–	–	–
ATSS (Zhang et al., 2020)	69.0	61.8	48.0	59.6	–	–	–	–
Faster R-CNN (Ren et al., 2015)	69.4	62.0	48.1	59.8	–	–	–	–
TOOD (Feng et al., 2021)	68.5	63.8	48.9	60.4	–	–	–	–
VFNet (Zhang et al., 2021)	70.3	62.8	48.6	60.6	–	–	–	–
Mask R-CNN (He et al., 2017)	70.7	63.5	49.6	61.3	61.4	53.5	39.0	51.3
SDANet (Wang et al., 2021c)	71.2	64.2	49.5	61.6	59.9	52.0	37.4	49.8
Co. Dist.+TOOD (Wei et al., 2024)	72.7	64.8	49.5	62.3	–	–	–	–
Cascade R-CNN (Cai and Vasconcelos, 2018)	72.3	65.4	50.1	62.6	–	–	–	–
Cascade Mask R-CNN (Cai and Vasconcelos, 2018)	74.3	67.3	53.4	65.0	62.4	54.7	40.8	52.6
CBNetV2 (Liang et al., 2021)	76.0	69.3	57.2	67.5	64.3	57.0	44.9	55.4
FDTNet(ours)	**77.2**	**69.6**	**57.9**	**68.2**	**65.2**	**57.3**	**45.7**	**56.1**

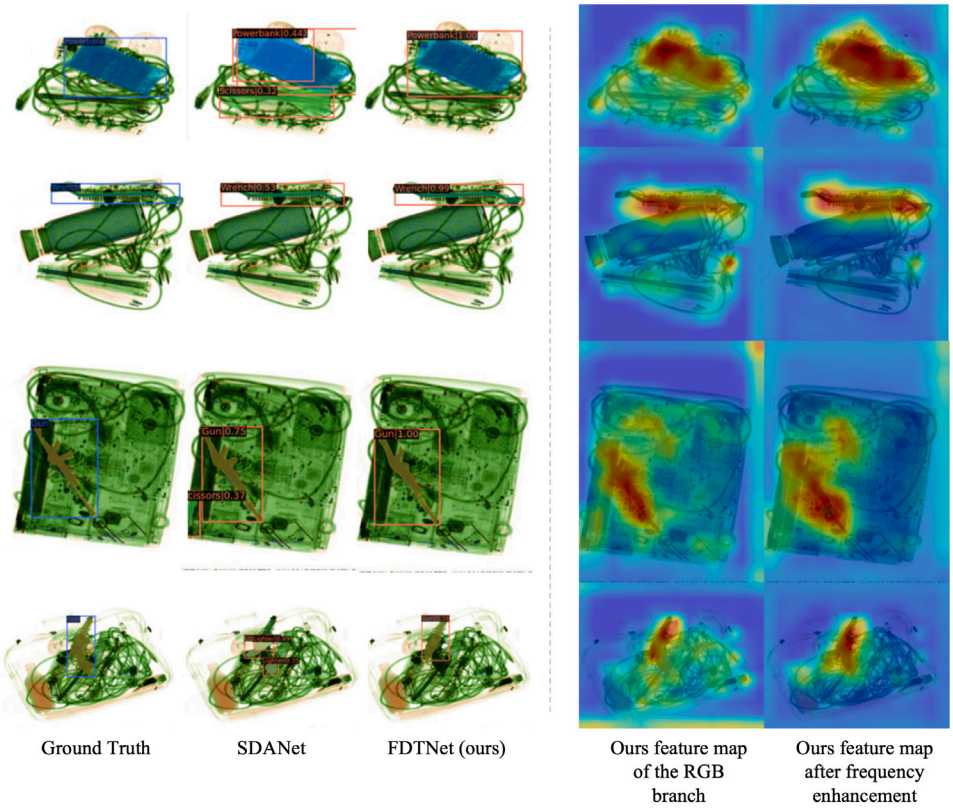


Fig. 5. Display of prohibited items detection results and visualization of feature maps.

features for X-ray images. In other words, by integrating the multi-scale features of the RGB branch enhanced by the GCA module, the model performance is improved and it is clear that the frequency branch pays more attention to the position information and features of the detected objects.

Fig. 6 presents detailed visualization results of our model on eighteen samples from the hard partition of the PIDray test set. For each sample, we display the predicted bounding box, class, and score results for prohibited object detection. Furthermore, we use yellow dashed boxes to highlight the prediction of samples with severe object occlusion and complex background information. It is worth noting that our model successfully predicts even those objects that are barely visible in the X-ray images.

4.5. Ablation study

We explore the effect of different modules on the proposed network performance through three ablation experiments.

Effects of GCA module and FAM module: To evaluate the effectiveness of our proposed GCA module and FAM module, we conducted various design experiments. The first row of Table 4 demonstrates that when both modules are used together, the frequency information from the FAM module can effectively combine with the features from the RGB stream through the GCA module, resulting in the best detection performance. In the second row of Table 4, we removed the FAM module, which led to repetitive or redundant features extracted from the two streams. This resulted in a decrease of 0.7 PIDray overall mAP and 1.2 OPIXray OL3 mAP compared to the first row. By replacing the GCA module with a direct up-sampling and summing operation (as observed in the third row of Table 4), we observed that the proposed GCA module contributed to a gain of 1.2 PIDray overall mAP and 2.2 OPIXray OL3 mAP compared to the first row. In the fourth row of Table 4, we removed both the GCA module and FAM module simultaneously, resulting in a decrease of 1.9 PIDray overall mAP and 4.3 OPIXray OL3 mAP compared to the first row.

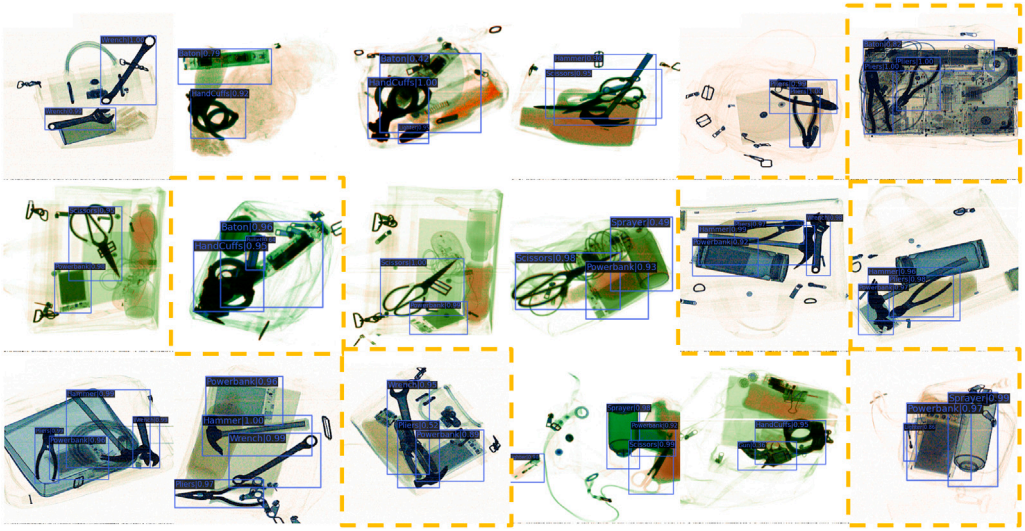


Fig. 6. More visualizations of the proposed FDTNet for prohibited object detection.

Table 4
Effectiveness of different modules.

GCA	FAM	PIDray (mAP)				OPIXray (mAP_{50})		
		Easy	Hard	Hidden	Overall	OL1	OL2	OL3
✓	✓	**77.2**	**69.6**	**57.9**	**68.2**	82.6	**82.3**	**80.6**
✓	×	76.2 (−1.0)	69.0 (−0.6)	57.3 (−0.6)	67.5 (−0.7)	**82.8** (+0.2)	78.8 (−3.5)	79.4 (−1.2)
×	✓	75.7 (−1.5)	69.0 (−0.6)	56.3 (−1.6)	67.0 (−1.2)	82.7 (+0.1)	80.8 (−1.5)	78.4 (−2.2)
×	×	74.4 (−1.8)	68.1 (−1.5)	55.4 (−2.5)	66.0 (−1.9)	80.6 (−2.0)	77.1 (−5.2)	76.3 (−4.3)

Table 5
Effectiveness of different branches of GCA.

Global attention	Channel attention	PIDray (mAP)				OPIXray (mAP_{50})		
		Easy	Hard	Hidden	Overall	OL1	OL2	OL3
✓	✓	**77.2**	69.6	**57.9**	**68.2**	82.6	**82.3**	**80.6**
✓	×	76.0 (−1.2)	69.0 (−0.6)	57.6 (−0.3)	67.5 (−0.7)	82.5 (−0.1)	81.8 (−0.5)	79.6 (−1.0)
×	✓	76.6 (−0.6)	**69.7** (+0.1)	57.1 (−0.8)	67.8 (−0.4)	82.5 (−0.1)	81.0 (−1.3)	78.9 (−1.7)
×	×	75.7 (−1.5)	69.0 (−0.6)	56.3 (−1.6)	67.0 (−1.2)	**82.7** (+0.1)	80.8 (−1.5)	78.4 (−2.2)

Table 6
Effectiveness of GCA position.

	PIDray (mAP)				OPIXray (mAP_{50})		
	Easy	Hard	Hidden	Overall	OL1	OL2	OL3
(a)	**77.2**	**69.6**	57.9	**68.2**	**82.6**	82.3	**80.6**
(b)	76.4 (−0.8)	68.9 (−0.7)	**58.4** (+0.5)	67.9 (−0.3)	82.4 (−0.2)	**82.6** (+0.3)	80.1 (−0.5)

Effects of different branches of GCA: By comparing the results of the first and second rows in Table 5, it is evident that there is a significant decrease in both PIDray overall mAP (−0.7) and OPIXray OL3 mAP (−1.0) when the channel attention module is removed. This highlights the importance of inter-channel interactions in achieving accurate detection. The model's ability to integrate information from multiple channels in the entire 2D space is crucial for optimal performance. Furthermore, upon removing the global attention module and comparing the results of the first and third rows in Table 5, we observed that the proposed global attention contributed to a gain of 0.4 PIDray overall mAP and 1.7 OPIXray OL3 mAP. These findings further emphasize the significance of both the channel attention module and the global attention module in improving the model's performance by facilitating effective interactions and information integration across different channels and spaces.

Effects of GCA position: As shown in Fig. 7(a), the feature map X_r^s from the first backbone will be used in two branches. One is used for feature fusion with feature maps of other scales and transmitted to the second backbone (blue branch). To improve the ability of RGB branch feature extraction, we input X_r^s into subsequent detection network (pink branch). Subsequent detection network includes FPN, RPN and detection head. During the experiment, we consider using the GCA module to enhance the feature map input to the subsequent detection network, so we changed the position of the GCA module, as shown in Fig. 6(b). As shown in Table 6, although method (b) enhances the feature maps of both branches, it results in a decrease in mAP for both the easy and hard sets of PIDray, as well as OL1 and OL2 of OPIXray, compared to method (a). In conclusion, method (a) performs better, which is why we chose to use it in our experiment.

Effects of loss coefficient: Table 7 presents the ablation experimental results for a single loss weighting coefficient in the detection performance of FDTNet. Our observations reveal that an excessively large coefficient for L_{total}^{assist} can lead to convergence issues in the network. However, within a reasonable range, the detection performance is not significantly affected by the coefficient.

Table 7
Effects of loss coefficient.

λ	PIDray (mAP)				OPIXray (mAP_{50})		
	Easy	Hard	Hidden	Overall	OL1	OL2	OL3
0.5	**77.2**	**69.6**	**57.9**	**68.2**	**82.6**	**82.3**	**80.6**
0.25	76.9 (−0.3)	69.1 (−0.5)	57.1 (−0.8)	67.7 (−0.5)	82.5 (−0.1)	82.0 (−0.3)	80.1 (−0.5)
0.85	76.5 (−0.7)	68.8 (−0.8)	57.3 (−0.6)	67.5 (−0.7)	82.4 (−0.2)	82.1 (−0.2)	80.0 (−0.6)
1.2	76.0 (−1.2)	68.1 (−1.5)	56.8 (−1.1)	67.0 (−1.2)	81.1 (−1.5)	81.0 (−1.3)	79.3 (−1.3)
2.0	/	/	/	/	/	/	/

Table 8
Comparison of inference complexity and parameters.

Method	Parameters (MB)	GFLOPs
Faster R-CNN (Ren et al., 2015)	43.51	218.91
Sparse R-CNN (Sun et al., 2021)	128.70	232.18
POD-F-X (Ma et al., 2023)	119.67	337.44
FDTNet(ours)	66.17	207.94

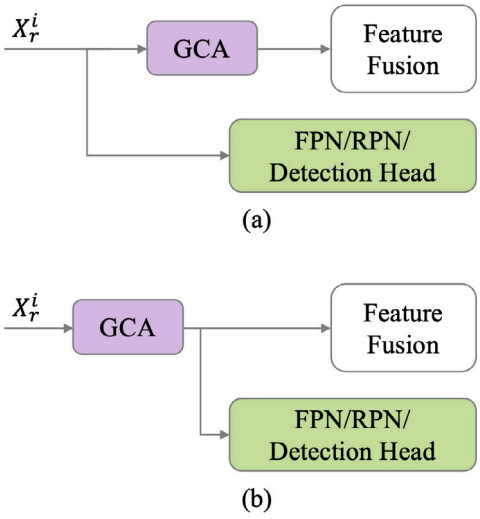


Fig. 7. Different positions of GCA.

4.6. Inference complexity

To demonstrate the superior detection speed of FDTNet, we have provided a comparison of computational complexity and model parameters for different object detection methods in Table 8. GFLOP and the number of parameters are commonly used as indicators of a model's computational complexity. It is evident from the table that our proposed FDTNet has lower computational complexity and model parameters compared to existing object detection methods.

5. Conclusion

In this paper, we propose a dual-stream frequency-aware network to combine the RGB feature and the frequency feature of the image to detection prohibited items. We design a frequency-aware module (FAM) to focus on the frequency information of the prohibited items, and use SRM filter to extract the high-frequency in the X-ray image. Meanwhile, in order to better combine the feature maps from RGB branch, the global and channel attention module (GCA) is used to enhance the representation of the feature map. Experiments on the PIDray and OPIXray dataset demonstrate the superiority of the method. Compared with other algorithms, the proposed FDTNet performs well in detection and segmentation evaluation.

Although there has been some progress in this study, the persistent challenge of dealing with hard samples, especially those with random stacking and placement in X-ray images, remains. The existence of these challenging examples continues to hinder the overall performance of algorithms and limit their full potential. Moving forward, we intend to conduct further research on hard and hidden test sets to improve the network's ability to adapt to various environments. Additionally, we will gradually enhance the theoretical and practical framework of this algorithm.

CRediT authorship contribution statement

Ziming Zhu: Conceptualization, Methodology, Software, Writing – original draft. **Yu Zhu:** Project administration, Supervision, Writing – review & editing. **Haoran Wang:** Data curation, Formal analysis, Validation. **Nan Wang:** Investigation, Visualization. **Jiongyao Ye:** Project administration, Supervision. **Xiaofeng Ling:** Formal analysis, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to thank the (anonymous) reviewers for their constructive comments. This work was supported by the Shanghai Automotive Industry Science and Technology Development Foundation, China (2304).

References

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., 2019. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Chen, P., Xu, C., Ma, Z., Jin, Y., 2023. A mixed samples-driven methodology based on denoising diffusion probabilistic model for identifying damage in carbon fiber composite structures. IEEE Trans. Instrum. Meas. 72, 1–11. http://dx.doi.org/10.1109/TIM.2023.3267522.

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., Ji, R., 2021. Local relation learning for face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 2. pp. 1081–1088.

Ding, J., Ye, C., Wang, H., Huyan, J., Yang, M., Li, W., 2023. Foreign bodies detector based on DETR for high-resolution X-Ray images of textiles. IEEE Trans. Instrum. Meas. 72, 1–10. http://dx.doi.org/10.1109/TIM.2023.3246510.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W., 2021. Tood: Task-aligned one-stage object detection. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV, IEEE Computer Society, pp. 3490–3499.

Fridrich, J., Kodovsky, J., 2012. Rich models for steganalysis of digital images. IEEE Trans. inf. Forensics Secur. 7 (3), 868–882.

Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R., 2021. Exploiting fine-grained face forgery clues via progressive enhancement learning. arXiv preprint arXiv: 2112.13977.

Hassan, T., Akcay, S., Bennamoun, M., Khan, S., Werghi, N., 2022. Tensor pooling-driven instance segmentation framework for baggage threat recognition. Neural Comput. Appl. 34 (2), 1239–1250.

Hassan, T., Werghi, N., 2020. Trainable structure tensors for autonomous baggage threat detection under extreme occlusion. In: Proceedings of the Asian Conference on Computer Vision.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Lee, Y., Kim, J., Willette, J., Hwang, S.J., 2022. MPViT: Multi-path vision transformer for dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7287–7296.

Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H., 2021. Cbnetv2: A composite backbone network architecture for object detection. arXiv preprint arXiv:2107.00420.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, J., Liu, X., Qu, F., Zhang, H., Zhang, L., 2022. A defect recognition method for low-quality weld image based on consistent multiscale feature mapping. IEEE Trans. Instrum. Meas. 71, 1–11. http://dx.doi.org/10.1109/TIM.2022.3171609.

Liu, Y., Wang, Y., Wang, S., Liang, T., Zhao, Q., Tang, Z., Ling, H., 2020. Cbnet: A novel composite backbone network architecture for object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07. pp. 11653–11660.

Ma, C., Zhuo, L., Li, J., Zhang, Y., Zhang, J., 2022. EAOD-net: Effective anomaly object detection networks for X-ray images. IET Image Process..

Ma, C., Zhuo, L., Li, J., Zhang, Y., Zhang, J., 2023. Occluded prohibited object detection in X-ray images with global context-aware multi-scale feature aggregation. Neurocomputing 519, 1–16.

Mao, M., Zhang, R., Zheng, H., Ma, T., Peng, Y., Ding, E., Zhang, B., Han, S., et al., 2021. Dual-stream network for visual recognition. Adv. Neural Inf. Process. Syst. 34, 25346–25358.

Mery, D., Riffo, V., Zscherpel, U., Mondragón, G., Lillo, I., Zuccar, I., Lobel, H., Carrasco, M., 2015. GDXray: The database of X-ray images for nondestructive testing. J. Nondestruct. Eval. 34 (4), 1–12.

Mery, D., Svec, E., Arias, M., Riffo, V., Saavedra, J.M., Banerjee, S., 2016. Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans. Syst. Man Cybern. Syst. 47 (4), 682–692.

Miao, C., Xie, L., Wan, F., Su, C., Liu, H., Jiao, J., Ye, Q., 2019. Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2119–2128.

Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q., 2021. Conformer: Local features coupling global representations for visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 367–376.

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J., 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision. Springer, pp. 86–103.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al., 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14454–14463.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Tao, R., Wei, Y., Jiang, X., Li, H., Qin, H., Wang, J., Ma, Y., Zhang, L., Liu, X., 2021. Towards real-world X-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10923–10932.

Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9627–9636.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, M., Du, H., Mei, W., 2021a. Information-exchange enhanced feature pyramid network (IEFPN) for detecting prohibited items in X-ray security images. In: 2021 7th International Conference on Computer and Communications. ICCC, IEEE, pp. 731–735.

Wang, M., Du, H., Mei, W., Wang, S., Yuan, D., 2022. Material-aware cross-channel interaction attention (MCIA) for occluded prohibited item detection. Vis. Comput. 1–13.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2021b. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578.

Wang, B., Zhang, L., Wen, L., Liu, X., Wu, Y., 2021c. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5412–5421.

Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L., Liu, X., 2020. Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 138–146.

Wei, Y., Wang, Y., Song, H., 2021. CFPA-net: Cross-layer feature fusion and parallel attention network for detection and classification of prohibited items in X-ray baggage images. In: 2021 IEEE 7th International Conference on Cloud Computing and Intelligent Systems. CCIS, IEEE, pp. 203–207.

Wei, Y., Wang, H., et al., 2024. Cooperative distillation with X-ray images classifiers for prohibited items detection. Eng. Appl. Artif. Intell. 127, 107276.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 3–19.

Zhang, S., Chi, C., Yao, Y., Lei, Z., Li, S.Z., 2020. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9759–9768.

Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N., 2021. Varifocalnet: An iou-aware dense object detector. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8514–8523.

Zhao, C., Zhu, L., Dou, S., Deng, W., Wang, L., 2022. Detecting overlapped objects in X-Ray security imagery by a label-aware mechanism. IEEE Trans. Inf. Forensics Secur. 17, 998–1009.

Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., Yang, R., 2019. Iou loss for 2d/3d object detection. In: 2019 International Conference on 3D Vision. 3DV, IEEE, pp. 85–94.

Zhu, C., He, Y., Savvides, M., 2019. Feature selective anchor-free module for single-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 840–849.