https://doi.org/10.26599/AIR.2025.9150053 CAAI Artificial Intelligence Research

Light2Ray: Lightweight dual-view prohibited item detection model in X-ray images based on Transformer

Haigang Zhang[1], Jingchang Gao[1,2], Xinxin Wang[1], Minglei Guan[1], Zhitao Wu[2], and Zhiwei Sun[1] ✉

ABSTRACT

Using computer vision technology to detect prohibited items in X-ray images is an effective method for realizing intelligent security checking. Dual-view security checking can capture the images from both vertical and horizontal perspectives of the same package at the same time, addressing issues such as unfavorable imaging angles and object occlusion at the image acquisition end. In this paper, we proposed a novel prohibited item detection model based on Transformer architecture in dual-view X-ray images. Two feature fusion module, named as feature selection module and corss-attention fusion module, are introduced to make interaction and enhancement. To improve the model inference efficiency, we use MobileViT as the backbone network to reduce the model size. Simulation results based on Dualray dataset has demonstrated the performance of the proposed model.

KEYWORDS

X-ray image detection; object detection; Transformer; feature fusion

Security checking serves as a vital defense line for safeguarding the lives and property of individuals[1]. The current security checking process is heavily dependent on manual operations. However, the high-intensity nature of manual security checking inevitably leads to instances of missed detections and false alarms. Using computer vision technology to address these issues, known as intelligent security checking, can alleviate the pressure on human operators and reduce the occurrence of missed detections[2].

Generally speaking, the intelligent analysis of X-ray security checking images through computer vision technology is considered the object detection task. The image analysis model not only needs to identify the categories of prohibited items in the X-ray images, but also determine their locations, so as to facilitate subsequent threat elimination operations[3]. X-ray security inspection images have some unique characteristics that increase the difficulty of detecting prohibited items[4]. (1) Cluttered background: The objects inside the package are placed in a disordered and chaotic manner, resulting in a complex and varied background; (2) Multi-scale and diversity of targets: The types of prohibited items are diverse and vary in size, which increases the difficulty of object detection; (3) Unfavorable imaging angles: In extreme cases, when the cross-section of the prohibited item is parallel to the X-ray beam, it becomes impossible to obtain a clear image of the prohibited item; (4) Target occlusion: Due to the X-ray transmission effect, occlusion between different targets is a common phenomenon, which can easily lead to missed detections.

The issues of cluttered backgrounds and multi-scale, diverse targets can be addressed by designing targeted network architectures and feature enhancement modules. Unfavorable imaging angles and target occlusion have a detrimental effect on the input image quality from the imaging side[5]. Figure 1 shows the imaging effect of security inspection images under unfavorable imaging angles and target occlusion conditions.

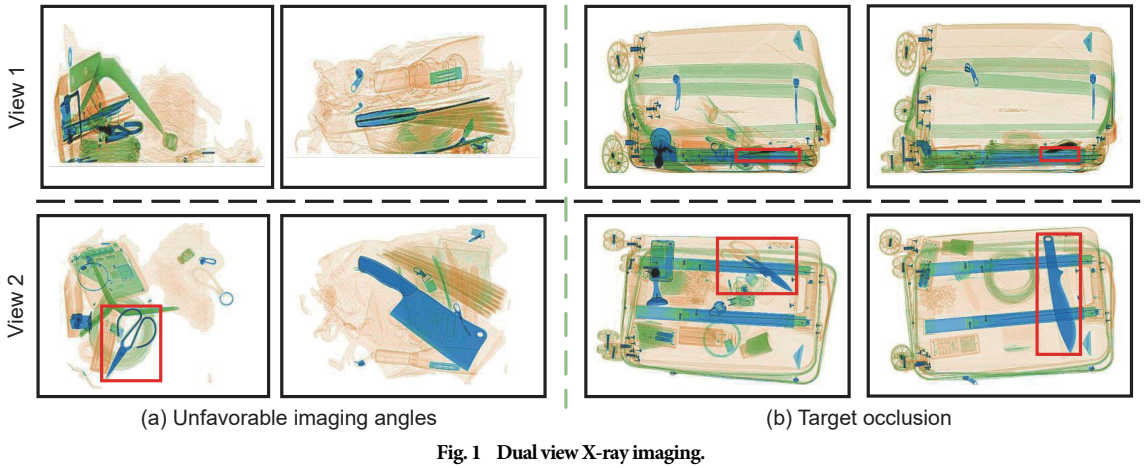

Fig. 1 Dual view X-ray imaging.

1 Shenzhen Polytechnic University, Shenzhen 518000, China
2 School of Electronic and Information Engineering, University of Science and Technology Liaoning, Anshan 114051, China
Address correspondence to Zhiwei Sun, smeker@szpu.edu.cn

First, because of the X-ray transmission effect, the imaging effect under extreme viewing angles cannot clearly depict the contour of prohibited items. Taking the knife as an example, when the surface of the knife is parallel to the X-ray, it is imaged as a blue line under the X-ray, which is not conducive to the processing of the visual inspection algorithm at all. As shown in Fig. 1(b), if occlusion occurs between different objects inside the package in one imaging view, it tends to be alleviated in the other view.

In our previous work, we released the first dual-view X-ray security inspection image dataset, named Dualray[5], which consists of 4371 pairs of images, including six categories of prohibited items of different sizes. At the same time, we established the first benchmark for feature fusion and object detection model using Dualray dataset. There have been previous works focused on prohibited item detection using multi-view X-ray security inspection images. Reference [6] applied multi-view collection to integrate image features from multiple views and adopts fast R-CNN[7] for detection. Their results show that pooling image features from multiple viewpoints leads to more accurate detection results than relying only on a single view. In addition, Ref. [8] implemented several approaches using single, two, and multiple X-ray views to make a reliable and practical model with varying levels of success in prohibited item detection in X-ray images. And the authors found that the approaches took in dual inputs to make more accuracy detection.

The key of dual-view X-ray security checking image analysis model is to fuse object features from different views. Due to the cluttered background of security checking images, the chaotic placement of objects, and mutual occlusion, it is necessary to achieve effective feature fusion module. Currently, visual analysis models based on the Transformer[9] framework have been brilliant in various vision tasks. The DETR model adopts the Transformer architecture and treats the object detection task as a sequence-to-sequence transformation problem. The main idea is to transform the object detection task into a set-element ranking problem. The model works by converting the vision of the input image into a set of feature vectors and modeling these features using a converter encoder. Then, the transform decoder predicts both the bounding box and the category of the object. In our previous work[10], DETR was applied to single-view X-ray security inspection image analysis, achieving satisfactory results in contraband detection accuracy.

We believe that the Transformer architecture, with its attention mechanism as a feature extractor, is well-suited for feature fusion of dual-view X-ray images. This paper attempts to build the first prohibited item detection model in dual-view X-ray images with the Transformer attention based on the DETR architecture. However, the dual-stream model will inevitably increase the weights of the network, and impacts the inference efficiency. In order to improve the lightweight performance of the object detection algorithm, this paper will improve MobileViT[11] and design a dual-view lightweight object detection network for X-ray image analysis, named as Light2Ray model. In order to achieve dual-view feature enhancement, the global attention upsample (GAU) module and the cross-attention module are proposed in the Light2Ray model to realize feature fusion and interaction. The high-level information of the GAU module can guide the low-level information and dynamically adjust the channel weight through the channel attention mechanism, so that the more important features get more attention and improve the accuracy of object detection[12]. In the simulation part, we conducted several comparative experiments based on the Dualray dataset to verify the state-of-the-art performance of the proposed Light2Ray model. The main contributions of this paper are as follows:

- We propose the Light2Ray model, which is the first work appling DETR model in the prohibited item detection task in dual-view X-ray images.
- The GAU module and the cross attention module are proposed to realize the information interaction between the horizontal perspective and the vertical perspective of security checking images, which improves performance to some extent.
- The proposed Light2Ray model is verified on the Dualray dataset, which shows that Light2Ray model has achieved satisfactory results in both detection accuracy and inference efficiency.

1 Related work

1.1 Object detection

In recent years, with the development of deep CNN, deep neural networks have achieved great success in various computer vision tasks, including image classification, semantic segmentation, object detection, and so on. Object detection is one of the most complex and challenging tasks in computer vision, which is not only necessary to identify the categories of objects in a given image, but also needs to locate the objects. Usually, object detector methods can be divided into one-stage and two-stage operations. The one-stage detector directly performs object classification and bounding-box based regression location on the obtained feature map. The speed of training and inference of the one-stage detector is fast. The famous YOLO[13] and its variants, RetinaNet and RefineDet[14], etc., belong to the one-stage detectors. Most one-stage object detectors are anchor-based or anchor-free. Although prior anchor boxes provide the bounding boxes with detected objects, the number of anchor boxes is much larger than that of the target, which brings extra computation. However, anchor-free detectors are the methods to predict target key points and locations directly, and with fewer calculations. On the contrary, the two-stage detectors generate a series of proposal boxes for samples by RPN. Then, the candidate boxes are classified and regressed by the CNN model, such as R-CNN[15] and its variants, R-FCN[16] and Cascade-RCNN[17], etc. The classification and regression operations are performed separately to produce more accurate results. However, the calculation process is complex, time-consuming, and slow. Currently, these object detectors have good detection accuracy because their feature extraction backbones have very deep CNN architectures, such as ResNet101. While these deep network models perform well, they are computationally expensive and not suitable for deployment on edge mobile devices. For the application of the detection model of prohibited items in X-ray images, we should design a more lightweight detection model.

1.2 Model lightweight

With the increasing prevalence of embedded and mobile devices, there is a growing demand for object detection algorithms that can operate efficiently within limited computational resources. As a result, the development of lightweight object detection algorithms has become imperative to meet these constraints.

Some commonly used model frameworks for lightweight object detection algorithms include the MobileNet series, ShuffleNet[18] series, and EfficientNet[19] series. Among these, the MobileNet series is a lightweight convolutional neural network model specifically designed to reduce parameters and overall

computation by employing depthwise separable convolutions. Applications of MobileNet in object detection include versions such as MobileNetv1[20], MobileNetv2[21], and MobileNetv3[22].

On the other hand, MobileViT is a lightweight variant based on the vision Transformer (ViT) model. ViT utilizes Transformer architecture for image classification by dividing images into patches and utilizing self-attention mechanism for feature extraction. MobileViT improves on ViT and incorporates various techniques to decrease model size and computational complexity, allowing it to be utilized in resource-constrained environments. It replaces the original ViT's multi-head self-attention mechanism with 1D convolutions and employs depthwise separable convolution layers to reduce parameters and computation. MobileViT can be conceptualized as a fusion of MobileNet and ViT, aiming to create a lightweight Transformer model suitable for target detection tasks. Its design principles combine aspects from MobileNet and the self-attention mechanism of ViT, resulting in commendable performance in lightweight target detection and image classification tasks.

1.3 X-ray image analysis

The mainstream research on X-ray security checking image analysis belongs to the field of object detection. Although the model obtains the category of prohibited item in the X-ray image, it must also provide location information of the prohibited item to facilitate the subsequent risk elimination operation[23]. The object detection algorithm based on deep learning has been maturely applied in the detection of prohibited items in X-ray security checking images, which is driven by the publicly available database of X-ray security checking images.

There are few studies on intelligent analysis of dual-view X-ray security checking images. Steitz et al.[24] proposed a multi-view prohibited item detection method based on deep learning. Faster R-CNN[7] is extended to multi-view X-ray security checking images using a late fusion method, and a new multi-view pooling layer is introduced to fuse features from a single view using the geometry of the imaging setup. This method requires only little additional knowledge to achieve more accurate detection accuracy. Hassan et al.[25] proposed a method to extract contour-based translation information from different directions and use the extracted information for target detection in a simple feed-forward convolutional neural network. The detection accuracy of this method is as high as 99%. Tuli et al.[26] designed a dual-channel input convolutional neural network architecture based on InceptionV3, VGG19[27], ResNet50[28], and other networks. Compared with the single-view network architecture, the performance of detection results using dual views is improved (5%–15%).

Due to the limitation of the lack of high-quality dual-view X-ray security checking image datasets, existing research results cannot be validated against public benchmarks. The release of our Dualray dataset addresses this gap. In addition, the feature fusion of dual-view images increases the network complexity, and in turn affects the model's inference efficiency, which is detrimental to security checking processes that emphasize timeliness. Currently, there is no research on lightweighting of object detection models of dual-view security inspection images.

2 Light2Ray model

Figure 2 illustrates the Light2Ray model proposed in this paper for detecting the prohibited items in dual-view security checking images. Light2Ray model uses MobileViT as the backbone architecture, aiming to minimize the model's weights while ensuring the accuracy of object detection. The Light2Ray model employs a dual-stream architecture, with dual-view security checking images as the input. In the Light2Ray model, the main channel is responsible for object detection, while the secondary channel provides valuable information to assist the main channel. Considering that the images of vertical view provide clearer features of prohibited items, the main channel of Light2Ray model takes vertical view images as input, while the secondary channel takes horizontal view images as input.

As shown in Fig. 2, MV2 block is the feature extraction module in MobileViT. After feature extraction, the dual-view features are passed into the feature selection module (FSM), which allows the main channel to filter out valuable features from the secondary channel. The retained features are then sent to the main channel for further feature extraction. After three rounds of feature selection, the dual-channel features are sent to the cross-attention fusion module (CFM), where high-level feature fusion from the dual views is performed. Finally, the fused features are passed into the YOLO head for object detection.

2.1 Feature selection module

Generally speaking, due to the effect of gravity, objects tend to scatter at the bottom of the container, which results in clearer features of prohibited items in the vertical view images, while horizontal view images often contain noise interference. We need to leverage the valuable visual features from the secondary channel while avoiding the introduction of noise that could interfere with the model. The proposed FSM in Light2Ray model is used to filter the secondary channel features based on the main channel features.

Figure 3 shows the proposed FSM, while we can see the main function of the FSM is to use the features from the main channel as a feature selector and apply the attention mechanism to the secondary channel for feature selection. Let x represent the features from the main channel and y represent the features from the secondary channel. First, a global pooling operation is performed on the main channel features x. After the 1×1 convolution, the main channel features are applied as attention

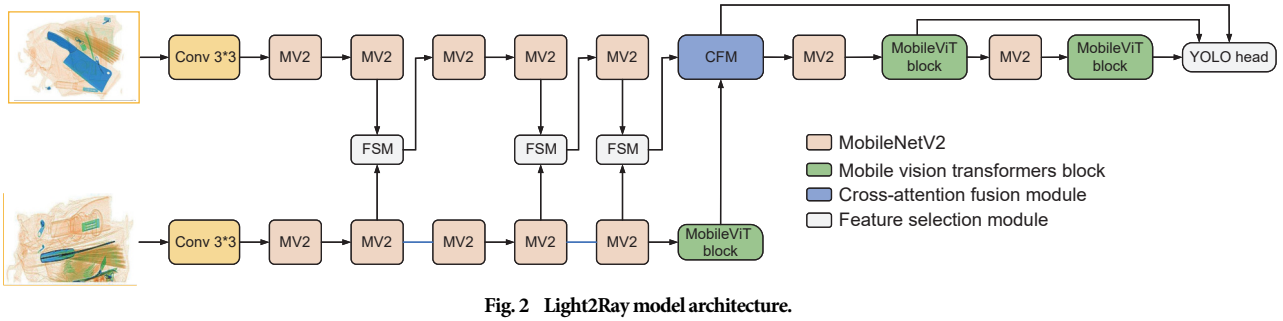

Fig. 2 Light2Ray model architecture.

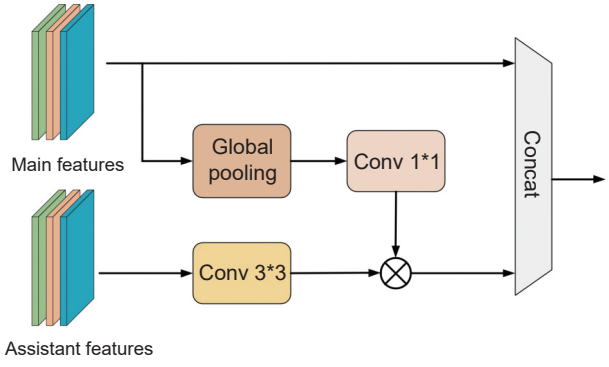

Fig. 3 Feature selection module.

weights to the secondary channel features. Finally, a concatenation operation is performed between the features of main channel and the filtered features of secondary channel. The entire computation process is as Eq.(1):

$$o = x \odot \{\mathrm{conv}_{1\times1}[\mathrm{GP}(x)] \otimes \mathrm{conv}_{3\times3}[y]\} \tag{1}$$

where $\mathrm{conv}()$ and $\mathrm{GP}()$ represent the convolution and global polling operations respectively.

2.2 Cross-attention fusion module

The CFM is built on Transformer technology, aiming to explore the interaction of dual-channel feature information from the perspective of the attention mechanism. Transformers were initially designed to handle sequential data, particularly for tasks in natural language processing (NLP). Unlike previous sequence-to-sequence models that relied heavily on recurrent neural networks (RNNs) or long short-term memory (LSTM) networks, the attention mechanism allows the model to focus on different parts of the input sequence when predicting each part of the output sequence. This mechanism significantly improves the model's ability to learn from and remember long sequences, which is crucial for understanding context in language.

Currently, the Transformer architecture is also widely applied in vision. In the Light2Ray model, we aim to leverage Transformer attention to facilitate information exchange between the dual-channel visual features. How to achieve attention interaction and response between the dual-channel features is a key issue. Four feature fusion strategies are presented in Fig. 4 and introduced as follows:

(a) Concat attention, which is a relatively simple and straightforward feature fusion strategy. We can aggregate the image patch tokens of the two channels together to form an integrated feature tokens. The subsequent Transformer attention operation can realize information interaction between all feature tokens.

(b) Pairwise attention, which is a position-sensitive feature fusion strategy. Given that each patch token occupies a unique spatial position within an image, a straightforward heuristic approach to fusion involves merging them according to their spatial locations.

(c) Cls-token attention. The CLS token acts as an abstract representation of the global features of a branch, given its role as the final embedding for predictions. Therefore, a basic strategy is to aggregate the CLS tokens from two separate branches.

(d) All-round attention, which achieves the best performance among four fusion strategies. All-round attention fusion comprehensively considers Cls-token and image patch tokens. Specifically, all-round attention fusion combines two Cls-tokens to form a high-performance Cls-token feature.

We use the x^c to represent the Cls-token, while x^p the image patch tokens. Therefore, the feature tokens of the two channels can be expressed as

$$x_i = [x_i^c || x_i^p], \quad i = 1, 2 \tag{2}$$

In Fig. 5, the $f(\cdot)$, $g(\cdot)$, and $p(\cdot)$ represents the single layer neural network, which is used to implement dimensional modification of input features. We combine the Cls-tokens of the two visual channels, and form new visual feature tokens of secondary channel, that is

$$\begin{cases} {x'_1}^c = f({x_1}^c) \\ {x'_{12}}^c = g([{x'_1}^c || {x_2}^c]) \\ {x'}_2 = [{x'_{12}}^c || {x_2}^p] \end{cases} \tag{3}$$

Then we employ the ${x'_1}^c$ as the *query* token, while ${x'}_2$ as the *key* and value token. The Transformer-based attention mechanism utilizes corresponding feature maps to construct the query, key, and value matrices. Such that

$$q = {x'_1}^c W_q, \quad k = x'_2 W_k, \quad v = x'_2 W_v \tag{4}$$

where, C represents the number of channels, and h denotes the number of attention heads. Dividing by this constant aims to stabilize the gradients during backpropagation. Incorporating residual connections also optimizes gradient issues. The final outputs y_i^c and x_i^p are aggregated through concatenated operation to facilitate the interaction of information between the primary visual channel and the auxiliary visual channel.

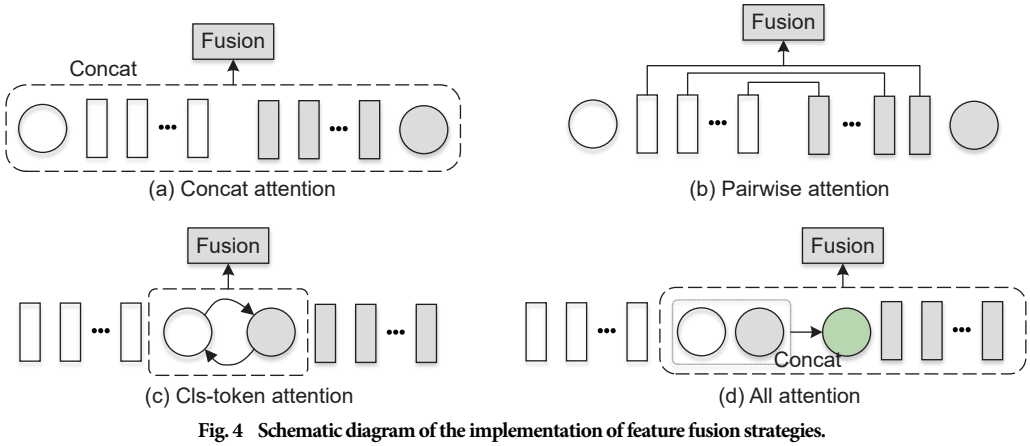

Fig. 4 Schematic diagram of the implementation of feature fusion strategies.

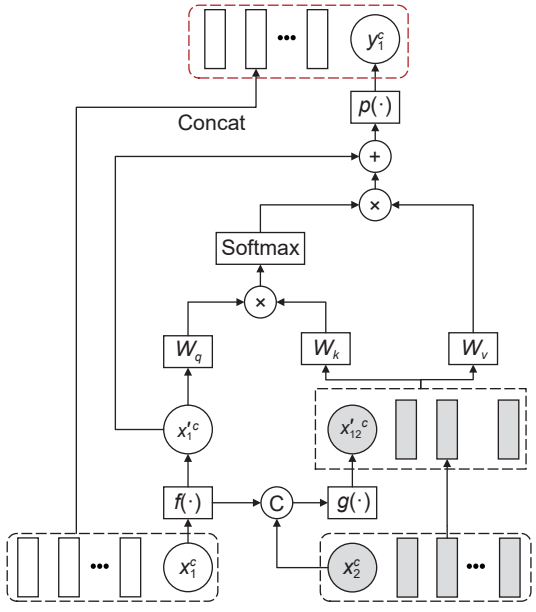

Fig. 5 All-round attention.

3 Simulation results

3.1 Experiment setting

The input image size is set to 384 × 384. The learning rate is set to 0.0001. And the batchsize is 8. In order to ensure that the original information of the two images is not modified, we do not use data enhancement methods to make the features align when the information of the two images is merged. In order to ensure the same experimental environment, we do not use pre-training weights in model training except DETR (the accuracy of 200 rounds is 0 when DETR does not use pre-training weights). For fair comparison, all models are trained on the same training set and evaluated on the same test set. The COCO AP is applied as the evaluation metrics.

3.2 Dualray dataset

The Dualray dataset is applied in the simulaiton experiments. The Dualray dataset consists of 4371 pairs of images, including six categories of prohibited items of different sizes. The items includes "Knives", "Liers", "Wrenches", "Scissors", "Lighters", and "Powerbanks". Figure 6 shows some examples for Dualray dataset.

3.3 Comparison results

Table 1 shows the accuracy comparison of different model network. After 200 epochs, the proposed Light2Ray model obtains 86.18% mAP, which is higher compared to the MobileNet series for prohibited item detection task in X-ray images. The performance of Light2Ray model increases by 37.9%, 43.1% and 6.65% mAP compared with MobileNetv1, MobileNetv2 and MobileNetv3, respectively. The mAP of our model is 2.08 percent higher than that of YOLOv5 and 24.34 percent higher than that of MobileViT. Compared with Dualray benchmark, the mAP of the propsed Light2Ray model increases by 4.74%. The classification accuracy of "Wrenches", "Powerbanks", "Scissors", "Knives", "Liers", "Lighters" by Light2Ray model is 99.97%, 98.4%, 92.25%, 91.64%, 84.83%, 49.97%, respectively. The detection accuracy is higher than any MobileNet algorithm and MobileViT algorithm, and some accuracy is higher than YOLOv5 and SSD algorithm. In some cases, small targets can also be detected.

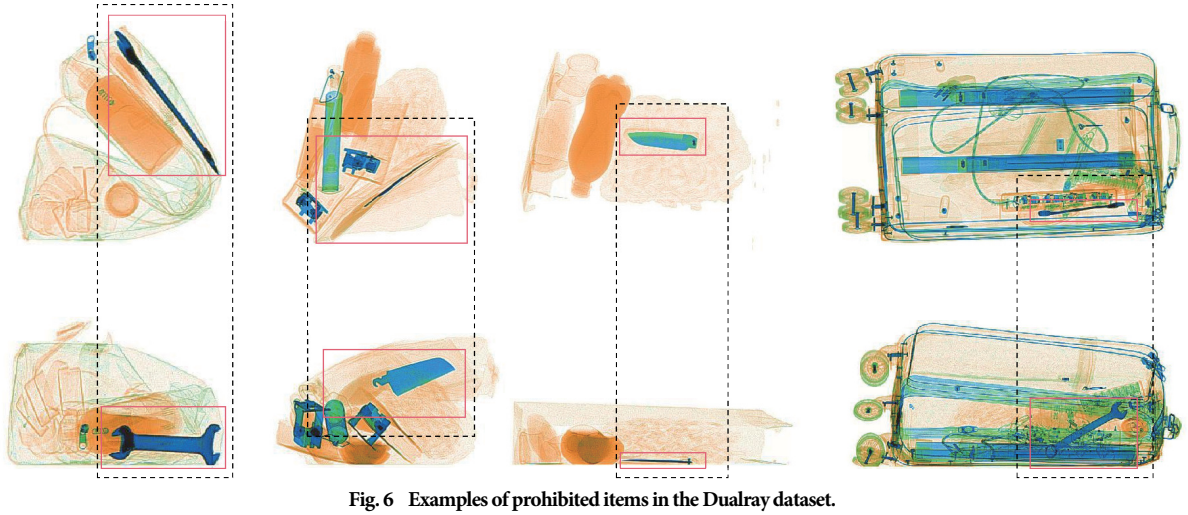

Fig. 6 Examples of prohibited items in the Dualray dataset.

Table 1 Comparison of classification accuracy of different model detection

Method	mAP	Knife	Lighter	Pliers	Powerbank	Scissors	Wrench
YOLOv5	84.10%	92.90%	30.30%	92.10%	96.60%	95.10%	97.40%
MobileViT	61.84%	65.05%	22.18%	56.40%	83.16%	62.78%	81.47%
MobileNetv1	48.24%	54.21%	24.59%	28.38%	70.20%	45.89%	66.15%
MobileNetv2	43.08%	43.52%	10.53%	30.80%	79.27%	33.29%	61.10%
MobileNetv3	79.53%	69.29%	43.61%	83.43%	96.94%	85.20%	98.73%
Dualray	81.44%	82.13%	47.33%	81.40%	95.95%	83.60%	98.22%
Ours	86.18%	84.83%	49.97%	92.25%	98.40%	91.64%	99.97%

Table 2 shows more detailed simulation results, and it can be seen that the detection accuracy of Light2Ray model for small objects is 20%, while other algorithms cannot detect small objects. Our model accuracy remains higher than that of the MobileNet series algorithms at different sizes and IoU thresholds. Compared to the Dualray algorithm, our algorithm improves the detection accuracy of small objects by 13.3%. The mAP for the detection of medium-sized objects is 42.5%, 6.4% higher than Dualray method. And the mAP for the detection of large objects improved by 5.7%. The recall rate is 20%, which is the same as before. The result for medium-sized objects is 52%, a 5.4-point improvement, and for large objects, it is 51.4%, a 2-point improvement. Compared to the single-channel MobileViT, both precision and recall rates have significantly improved. When using the DETR algorithm, we found that all the metrics are 0 when DETR is not using pre-trained weights. Therefore, compared to the DETR object detection algorithm, our network may not have high accuracy, but it converges more easily.

Table 3 illustrates the lightweight comparison of models. The model parameters for MobileNetv1 are 4.13 M, for MobileNetv2 are 3.50 M, and for MobileNetv3 are 5.48 M. MobileNetv3 increases the parameter quantity by 1.25 M compared to MobileNet1 and by 1.98 M compared to MobileNetv2. The model parameters for MobileViT are 5.64 M, while the parameter quantity for the Light2Ray model is 6.10 M. Compared to the increase of millions of parameters in the MobileNet series networks, the Light2Ray model only adds 0.46 M parameters compared to the single-channel MobileViT network. DETR has 36.74 M parameters, and the Light2Ray model reduces the parameter quantity by 30.64 M compared to the DETR model. The Frame Per Second (FPS) metrics is applied to verify the reasoning efficiency of different models. The proposed Light2Ray model achieves 46.59 FPS, which increases by 8 compared to the single-channel MobileViT algorithm, and by 3 to YOLOv5 algorithm. Our model is more lightweight on the basis of accuracy improvement. The experiments demonstrate that the proposed Light2Ray model for prohibited item detection in X-ray image has fewer parameters and higher accuracy.

3.4 Ablation experiment

In order to explore the influence of FSM and CFM on experimental results, the ablation experiments are carried out. The experimental results are shown in Table 4. After removing the FSM, the mAP metrica of the model is 84.80%, which decreases by 1.38 percent. After removing the CFM, the model obtains the mAP of 82.42%, a drop of 3.76%. Therefore, both the FSM and the CFM are helpful for improving the performance of the proposed Light2Ray model.

4 Conclusions

In this paper, we proposed Light2Ray model, which aims to make prohibited item detection in dual-view X-ray images. In the proposed Light2Ray model, the FSM is applied to filter the top-down view features based on the main visual features, in order to prevent the introduction of noise from the top-down view features into the network structure. The CFM, based on the Transformer

Table 2 Comparison of accuracy of different model detection

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
DETR	69.4%	96.9%	81.9%	−1	60.7%	78.1%	−1	72.3%	85.6%
MobileViT	25.9%	61.2%	13.4%	0	21.4%	29.1%	0	37%	45.9%
MobileNetv1	18.0%	47.8%	8.9%	0	16.2%	19.7%	0	29.6%	38.6%
MobileNetv2	19.3%	49.6%	10.8%	0	18.4%	20.6%	0	34.8%	40.9%
MobileNetv3	47.1%	79.5%	49.5%	−1	42.0%	53.4%	−1	48.1%	59.5%
Dualray	40.0%	80.2%	33.7%	6.7%	36.1%	39.3%	20%	46.6%	49.4%
Ours	45.4%	85.5%	41.6%	20%	42.5%	45%	20%	52%	51.4%

Table 3 Comparison of lightweight effects in different network structures

Method	*mAP*	Parameters	FPS
YOLOv5	84.1%	7.28 M	43.59
DETR	96%	36.74 M	–
MobileViT	61.84%	5.64 M	38.56
MobileNetv1	48.24%	4.23 M	74.88
MobileNetv2	43.08%	3.50 M	60.30
MobileNetv3	79.53%	5.48 M	53.21
Ours	86.18%	6.10 M	46.59

Table 4 Ablation experiment results

Method	mAP	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
-FSM+CFM	84.80%	45.3%	84.2%	42.2%	0	40.9%	44.3%	0	47.5%	54.6%
+FSM-CFM	82.42%	42.3%	81.9%	37.9%	10%	38.6%	43.6%	10%	47%	54.4%
+FSM+CFM	86.18%	45.4%	85.5%	41.6%	20%	42.5%	45%	20%	52%	51.4%

attention mechanism, is applied to enable interaction and enhancement between the visual image features in two visual channels. In the simulation experiments, the Dualray dataset is applied to verified the performance of the Light2Ray model, which shows the state-of-the-art performance.

Acknowledgements

This work was supported by the Research Projects of Department of Education of Guangdong Province (2023ZDZX1081, 2023KCXTD077, 2025ZDZX3068), and University-Enterprise Joint Research and Development Center (602531009PQ), and the Nature Science Foundation of Shenzhen Province (RCBS20221008093252090), and the Guangdong Basic and Applied Basic Research Foundation Project (2025A1515011370), , the Phase III Program (Grant No. 6025310002Q) of the Institute for Economic and Social Development at Shenzhen Polytechnic University.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Article History

Received: 12 May 2025; Revised: 23 June 2025; Accepted: 13 August 2025

References

[1] Akcay, S.; Breckon, T., Towards automatic threat detection: A survey of advances of deep learning within X-ray security imaging, *Pattern Recognition*, vol. 122, pp. Article No. 108245, 2022.

[2] Aydin, I.; Karakose, M.; Akin, E. A new approach for baggage inspection by using deep convolutional neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Data Processing, 1–6, 2019.

[3] Gaus, Y. F. A.; Bhowmik, N.; Akcay, S.; Guillen-Garcia, P. M.; Barker, J. W.; Breckon, T. P. Evaluation of a dual convolutional neural network architecture for object-wise anomaly detection in cluttered X-ray security imagery. In: Proceedings of the International Joint Conference on Neural Networks, 2019.

[4] Zhang, H.; Zhao, Z.; Yang, J., Attention-based prohibited item detection in X-ray images during security checking, *IET Image Processing*, vol. 18, no. 5, pp. 1119–1131, 2024.

[5] Wu, M.; Yi, F.; Zhang, H.; Ouyang, X.; Yang, J. Dualray: Dual-view X-ray security inspection benchmark and fusion detection framework. In: *Pattern Recognition and Computer Vision*, Springer Nature Switzerland, 721–734, 2022.

[6] Steitz, J. -M. O.; Saeedan, F.; Roth, S. Multi-view X-Ray R-CNN. In: Proceedings of the 40th German Conference on Pattern Recognition, 2019.

[7] Ren, S.; He, K.; Girshick, R.; Sun, J., Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[8] Tuli, A.; Bohra, R.; Moghe, T.; Chaturvedi, N.; Mery, D.; Dhiraj. Automatic threat detection in single, stereo (two) and multi view X-ray images. In: Proceedings of the IEEE 17th India Council International Conference, 1–7, 2020.

[9] Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for end-to-end object detection. *arXiv preprint* arXiv: 2010.04159, 2021.

[10] Wang, X.; Zhang, H.; Cao, W.; Xue, Y.; Xu, Y. DETR based prohibited item detection in X-ray security checking images. In: Proceedings of the IEEE 19th International Conference on Automation Science and Engineering, 1–7, 2023.

[11] Mehta, S.; Rastegari, M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision Transformer. *arXiv preprint* arXiv: 2110.02178, 2021.

[12] Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint* arXiv: 1805.10180, 2018.

[13] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 779–788, 2016.

[14] Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S. Z. Single-shot refinement neural network for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4203–4212, 2018.

[15] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587, 2014.

[16] Dai, J.; Li, Y.; He, K.; Sun, J. R-FCN: Object detection via region-based fully convolutional networks. *arXiv preprint* arXiv: 1605.06409, 2016.

[17] Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6154–6162, 2018.

[18] Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6848–6856, 2018.

[19] Tan, M.; Le, Q. V. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv preprint* arXiv: 1905.11946, 2019.

[20] Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint* arXiv: 1704.04861, 2017.

[21] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4510–4520, 2018.

[22] Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L. C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 1314–1324, 2019.

[23] Mery, D.; Svec, E.; Arias, M.; Riffo, V.; Saavedra, J. M.; Banerjee, S., Modern computer vision techniques for X-ray testing in baggage inspection, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 4, pp. 682–692, 2017.

[24] Steitz, J. -M. O.; Saeedan, F.; Roth, S. Multi-view X-ray R-CNN. *arXiv preprint* arXiv: 1810.02344, 2018

[25] Hassan, T.; Khan, S. H.; Akcay, S.; Bennamoun, M.; Werghi, N. Deep CMST framework for the autonomous recognition of heavily occluded and cluttered baggage items from multivendor security radiographs. *arXiv preprint* arXiv: 1912.04251, 2019.

[26] Tuli, A.; Bohra, R.; Moghe, T.; Chaturvedi, N.; Mery, D.; Dhiraj. Automatic threat detection in single, stereo (two) and multi view X-ray images. In: Proceedings of the IEEE 17th India Council International Conference, 1-7, 2020.

[27] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv: 1409.1556, 2014.

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