The Visual Computer (2023) 39:2865–2877
https://doi.org/10.1007/s00371-022-02498-y

ORIGINAL ARTICLE

Material-aware Cross-channel Interaction Attention (MCIA) for occluded prohibited item detection

Man Wang[1] · Huiqian Du[2] · Wenbo Mei[1] · Shaui Wang[1] · Dasen Yuan[3]

Accepted: 10 April 2022 / Published online: 9 May 2022

Abstract

For security inspection, detecting prohibited items in X-ray images is challenging since they are usually occluded by non-prohibited items. In X-ray images, different materials present different colors and textures. On this basis, we exploit the material characteristics to detect occluded prohibited items. Moreover, the occlusion mainly exists between prohibited items and non-prohibited ones, belonging to inter-class occlusion. We propose a Material-aware Cross-channel Interaction Attention (MCIA) module which can use the material information of X-ray images to deal with the inter-class occlusion. Specifically, MCIA is composed of Material Perception (MP) and Cross-channel Interaction (CI). MP captures distinctive material information of X-ray images and CI gets the local cross-channel interaction to convert material information into channel-wise weights. By combining MP and CI, MCIA effectively helps the network to highlight the core features of prohibited items while suppressing non-prohibited items. Meanwhile, we design the MCIA-Net and MCIA-FPN by placing our MCIA module behind each stage in ResNet. Our MCIA-Net and MCIA-FPN can be used as backbones to detect occluded prohibited items. Note that MCIA-FPN also takes into account the prohibited items of various sizes. Our MCIA-Net and MCIA-FPN have been comprehensively validated on the SIXray dataset and OPIXray dataset. The experimental results prove the superiority of our method. Furthermore, our proposed MCIA module outperforms several widely used attention mechanisms and effectively improves the performance of Faster R-CNN and Cascade R-CNN in detecting occluded prohibited items.

Keywords Object detection · Prohibited items · X-ray images · Occlusion

1 Introduction

Security inspection is extremely important to maintaining airport and traffic safety. Currently, X-ray scanners are usually used to detect prohibited items in baggage. Even though the scanners can provide detailed insight into the baggage content, existing X-ray security inspection still relies on cumbersome manual detection. Manual detection is easily affected by eye fatigue, which leads to missed detection and consumes a lot of manpower and time. It is a trend to develop a reliable and accurate method for automatically detecting prohibited items.

Machine learning approaches have been proposed to automatically detect prohibited items in X-ray images. These methods are mainly divided into two categories, non-deep learning methods and deep learning methods [1].

Non-deep learning methods exploit hand-crafted features such as SURF [4], FAST-SURF [19], SIFT [30], based on which the items are classified into the prohibited or safe ones by using support vector machines (SVM) [4,19], Random Forest [18], and K-Nearest Neighbor (K-NN) [37]. However, non-deep learning methods heavily rely on features extracted manually to classify prohibited items.

Recently, deep learning (DL) has made great achievements in image classification and object detection. A deep convolutional neural network can automatically extract low-level and high-level features, which yields a significant improvement in object detection. The mainstream object detection can be divided into one-stage and two-stage. One-stage detectors do not rely on region proposals, the most

✉ Huiqian Du
duhuiqian@bit.edu.cn

Man Wang
3120190807@bit.edu.cn

1 School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China

2 School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China

3 The Inner Mongolia Autonomous Region Public Security Bureau, Hohhot 010051, China

representative models are YOLO [5,33–35], SSD [27], and RetinaNet [25]. Two-stage detectors rely on region proposals, pioneered by RCNN architecture, including Fast R-CNN [12], Faster R-CNN [36], and follow-up R-FCN [10]. More recently, FPN [24] has made full use of multi-scale feature maps to solve the difficulty of detecting small size objects.

Researchers exploited the aforementioned detectors to detect prohibited items and presented promising results. For instance, Liu et al. [26] employed a two-stage approach Faster R-CNN [36] to detect subway X-ray images and achieved the mAP of 0.77. Akcay et al. [2] compared the gun detection accuracy of Faster R-CNN [36] and R-FCN [10] based on DBF2/6 datasets [3]. A follow-up work [40] introduced a multi-view pooling layer based on Faster R-CNN [36] to achieve better performance.

As shown in Fig. 1, we can see that prohibited items in X-ray images have three main characteristics. First, items of different materials in X-ray images appear with different pseudo colors and textures. This characteristic of material information facilitates the detection of prohibited items. Second, the size of prohibited items in X-ray images varies. It is difficult to design a network that can detect large and small objects at the same time. Third, prohibited items in the baggage may be seriously occluded by messy non-prohibited items. The occlusion mainly belongs to inter-class occlusion, which is easy to cause missed detection. Although the emergence of deep learning has greatly promoted the development of prohibited item detection, the latter two characteristics are still the main difficulties confronted by prohibited item detection in X-ray images.

To solve the problem of detecting items of various sizes, Liang et al. [22,23] used SSD [27] to generate multi-scale feature maps. Liu et al. [29] utilized YOLOv2 [34] to train with multi-scale, and Cui et al. [9] adopted RetinaNet [25] with FPN [24] as the backbone to detect gun. A recent work [46] added a semantic enrichment module (SEM) and a residual module (Res) to FSSD [21] for detecting prohibited items with small size.

To cope with the occlusion problem, Hassan et al. [13,14] applied structure tensors to extract contours of prohibited items, but they need elaborate parametric tuning. Miao et al. [31] proposed a dataset named Security Inspection X-ray (SIXray) and employed a class-balanced hierarchical framework (CHR) to detect occluded prohibited items in SIXray. Wei et al. [43] proposed to combine edge detection and attention model named De-occlusion Attention Module (DOAM), they tested DOAM on their proposed Occluded Prohibited Items X-ray (OPIXray) dataset. Nevertheless, the methods in [31,43] have only been validated on one dataset, and their performance under occlusion needs to be improved.

In this paper, we deal with the problem of inter-class occlusion by introducing the attention mechanism. In the scenario of heavy occlusion, the shape and appearance of overlapped prohibited items are incomplete, while the material characteristics are still preserved. Thus, we leverage the material information to enhance local features of occluded prohibited items. We propose a channel attention module named Material-aware Cross-channel Interaction Attention (MCIA). By determining channel-wise weights according to the material information of X-ray images, our proposed MCIA recalibrates the features to emphasize important local features of occluded prohibited items and suppress unnecessary background information accordingly. In addition, we design the MCIA-Net and MCIA-FPN by placing our MCIA module behind each stage of ResNet. Note that our MCIA-FPN also takes into account the various sizes of prohibited items. Comprehensive experiments on the X-ray datasets prove that our MCIA-Net and MCIA-FPN can bring a promising improvement to detectors in detecting occluded prohibited items.

Overall, our main contributions are summarized as follows:

(1) We propose a novel channel attention module named Material-aware Cross-channel Interaction Attention (MCIA). Our MCIA consists of two sub-modules, Material Perception (MP) for capturing the material information of X-ray images from each channel, and Cross-channel Interaction (CI) for capturing local cross-channel interaction to convert the material information into channel-wise weights. By combining MP and CI, with slightly computational overhead, MCIA can lay particular emphasis on local features of occluded prohibited items and suppress non-prohibited items accordingly.
(2) In contrast to the prior works, we propose to place our MCIA module behind each stage in ResNet, instead of inserting it into the residual blocks. In our designed MCIA-Net and MCIA-FPN, just four MCIA modules are needed to obtain performance improvement, increasing negligible model complexity. In addition to solving the problem of inter-class occlusion, our MCIA-FPN also takes into account the prohibited items of different sizes.
(3) The experimental results on the SIXray dataset [31] and OPIXray dataset [43] demonstrate that by using MCIA-Net or MCIA-FPN as the backbone, the performance of Faster R-CNN and Cascade R-CNN can be effectively improved. Besides, our MCIA is superior to several widely used attention mechanisms in detecting occluded prohibited items.

2 Related work

2.1 Style information extracting

Extracting style information from convolution feature maps has been extensively studied in the field of style transfer. If

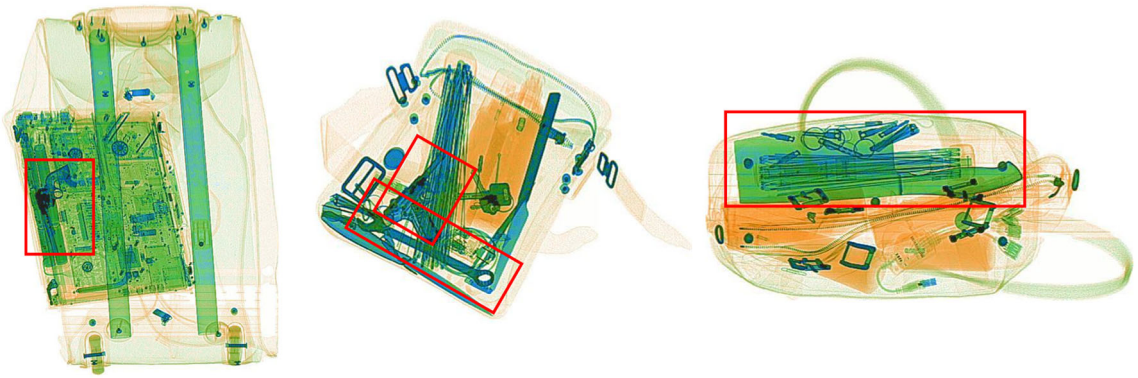


Fig. 1 Examples of X-ray images from SIXray dataset [31], prohibited items are in the area marked with a red box

we can extract the style information successfully, the style of one image can be transferred into another. As illustrated in [32], the style information of an image can be encoded by the feature statistics of a convolutional neural network (CNN). The pioneering work [11] introduced a new algorithm and exploited second-order feature statistics as style information to perform style transfer. A follow-up work [17] presented a novel adaptive instance normalization (AdaIN) layer and showed that style information including colors and textures of an image can be arbitrarily transferred by changing channel mean and standard deviation. On this basis, the recent work [45] takes the seasonal features of a remote sensing image as its style features and extracts them by channel-wise mean and standard deviation as in style transfer.

Similar to style information, the material information contained in X-ray images is also related to colors and textures. Thus, we adopt the channel-wise mean and standard deviation of each feature map as material information. By leveraging the material characteristics, our proposed MCIA can help the network to emphasize or suppress information according to its importance to prohibited item detection.

2.2 Attention mechanism

It has been proven that the attention mechanism has the potential to improve the performance of several tasks [28,38,39], including object detection. Attention mechanisms allow the network to concentrate more on useful information and suppress useless ones. One of the heuristic approaches is squeeze-and-excitation (SE) [16], shown in Fig. 2d. As a channel attention module, SE helped the network to improve its performance. Some works attempted to combine the SE block with other blocks. The convolutional block attention module (CBAM) [44] provided considerable performance gains over the SE block by emphasizing channel attention and spatial attention simultaneously. Global context network (GCNet) [8] simplified the non-local (NL) neural network [42] and integrated the NL with the SE block, which can model the global context.

Some works made efforts on modifying the structures of the SE block and effectively achieving competitive performance. As illustrated by Fig. 2b, the style-based recalibration module (SRM) [20] adopted global average pooling and global standard deviation pooling. SRM [20] verified that the combination of global average pooling and global standard deviation pooling outperforms global average pooling. Efficient channel attention (ECA) [41] aimed to learn channel attention effectively with low model complexity. ECA [41] proved that channels and weights need to correspond directly, while SE [16] used 2D convolution resulting in indirect correspondence between the channels and weights. As illustrated in Fig. 2c, it replaced the 2D convolution in SE [16] with a 1D convolution to learn channel attention efficiently. Both SRM [20] and ECA [41] effectively improved the performance of SE [16].

As shown in Fig. 2a, to better extract material information, we combine global average pooling and global standard deviation pooling. However, unlike SRM [20], we do not simply integrate information by a fully connected layer. Since the 1D convolution has an advantage in capturing channel dependencies, we design two 1D convolutions to subsequently process material information for better converting the material information into channel-wise weights. Our two 1D convolutions can achieve more adequate cross-channel interaction to properly assign the weights to different channels.

Furthermore, we place our MCIA modules behind each stage in ResNet, instead of inserting them into residual blocks as in SRM [20] and ECA [41]. By placing the proposed MCIA behind the stages, the network can address the problem of inter-class occlusion with a negligible computational burden, which is further verified in experiments.

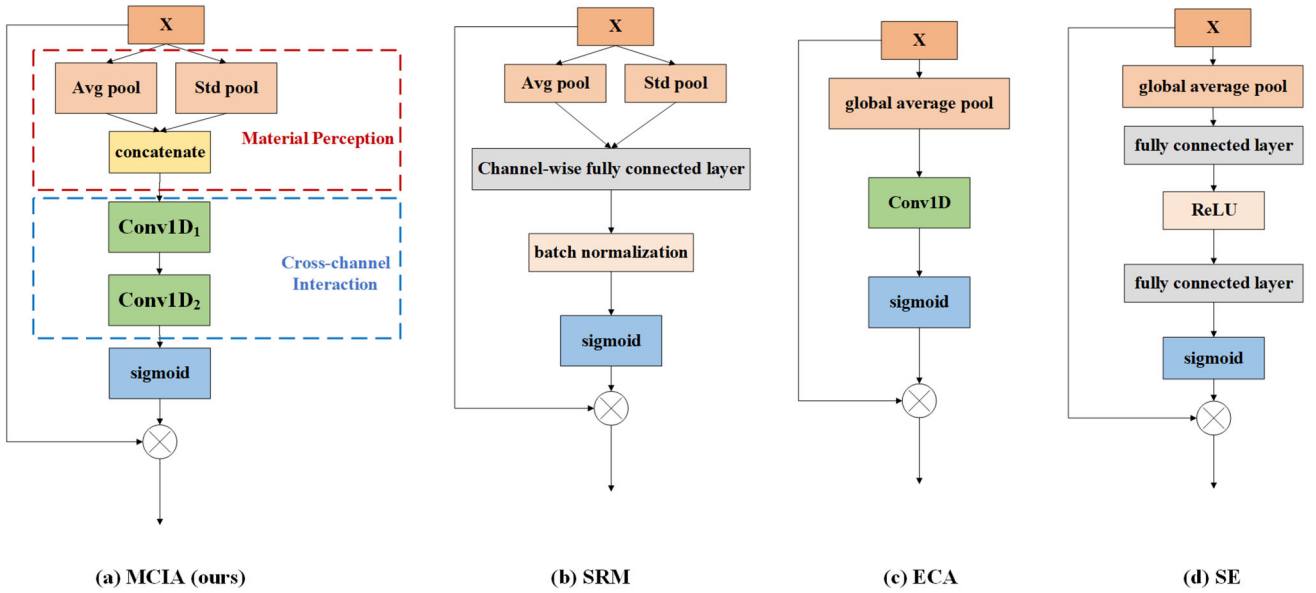


Fig. 2 Comparison of different attention modules, including ECA [41], SRM [20], SE [16], ⊗ represents element-wise multiplication

3 Proposed method

In this section, we first describe the proposed Material-aware Cross-channel Interaction Attention (MCIA) module in detail and analyze its two submodules, Material Perception (MP) and Cross-channel interaction (CI). Then we further elucidate the framework of MCIA-Net and MCIA-FPN. Finally, we analyze the module complexity in terms of floating-point operations (FLOPs) and parameters.

3.1 Material-aware Cross-channel Interaction Attention (MCIA)

The detailed structure of our proposed MCIA is shown in Fig. 3, MCIA consists of Material Perception (MP), Cross-channel interaction (CI), and the followed Sigmoid function. To tackle the problem of inter-class occlusion in X-ray images, we attempt to enhance the local features of occluded prohibited items and suppress unnecessary channel features by modeling the channel dependencies. Considering the characteristic of X-ray images, we facilitate the utilization of material information in designing our attention module. Specifically, we design a Material Perception (MP) unit to capture the material information of X-ray images and a cross-channel interaction (CI) submodule to capture the local cross-channel interaction for generating channel weights. The weights relating to material information are supposed to model the importance of feature channels so as to emphasize or suppress them accordingly. Hence, by combining MP and CI, the proposed MCIA can highlight local features regarding their relevance to occluded prohibited items and remove the influence of irrelevant information.

Specifically, given an input tensor $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$, where C, H and W, respectively, indicate the number of channels, the height and width. MCIA assigns different channel-wise weights to different channels, the channel-wise weights $\mathrm{W}_{\mathrm{MCIA}} \in \mathbb{R}^{C\times 1\times 1}$ can be computed as

$$W_{MCIA} = \sigma(CI(MP(X))) \tag{1}$$

Where σ denotes sigmoid function, $MP(\cdot)$ is Material Perception (MP) operation to extract the material information $\mathrm{S} \in \mathbb{R}^{C\times 2}$ from the input tensor $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$. $CI(\cdot)$ represents Cross-channel Interaction (CI) operation which accepts $\mathrm{S} \in \mathbb{R}^{C\times 2}$ as input and converts it into channel-wise weights $\mathrm{W}_{\mathrm{CI}} \in \mathbb{R}^{C\times 1\times 1}$ based on the local cross-channel interaction.

The final output of MCIA is the weighted feature map $\hat{\mathrm{X}} \in \mathbb{R}^{C\times H\times W}$ which can be computed as:

$$\hat{X} = X \otimes W_{MCIA} \tag{2}$$

Where ⊗ represents element-wise multiplication.

3.1.1 Material Perception(MP)

Inspired by SRM [20], we adopt the global average pooling and global standard deviation pooling to better extract the material information of X-ray images. After obtaining the channel-wise mean and standard deviation, we concatenate them together to represent material information. Specifically, for each input feature map $\mathrm{X} \in \mathbb{R}^{C\times H\times W}$, the channel-wise

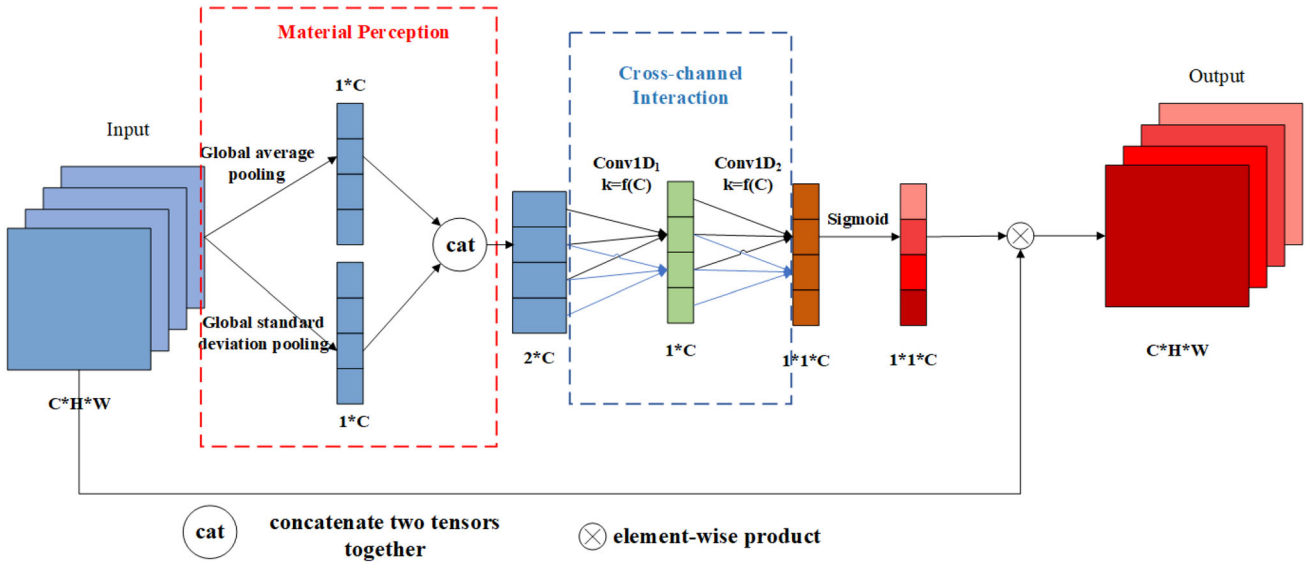


Fig. 3 Material-aware Cross-channel Interaction Attention (MCIA). $Conv1D_1$ and $Conv1D_2$ represent the two 1 D convolutions of $k = f(C)$

mean and standard deviation can be expressed as follows.

$$\mu_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} x_{chw} \tag{3}$$

$$\sigma_c = \sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(x_{chw} - \mu_c)^2} \tag{4}$$

The output of Material Perception (MP) is defined as:

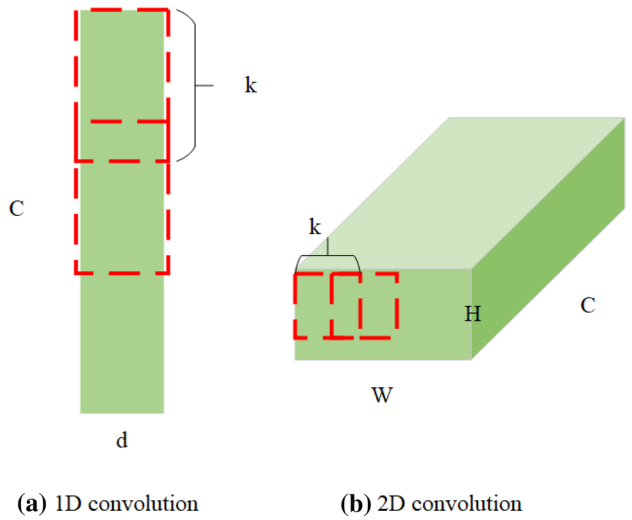
$$S = cat(\mu_c, \sigma_c) \tag{5}$$

Where $cat(\cdot)$ means concatenating two tensors together. As shown in Fig. 3, after getting channel-wise mean and standard deviation, Material Perception (MP) concatenates the channel-wise mean $\mu_c \in \mathbb{R}^{C\times 1}$ and standard deviation $\sigma_c \in \mathbb{R}^{C\times 1}$ to obtain S $\in \mathbb{R}^{C\times 2}$.

3.1.2 Cross-channel Interaction (CI)

Cross-channel Interaction (CI) takes the material information S $\in \mathbb{R}^{C\times 2}$ as input. Based on the fact that 1D convolution can capture local cross-channel interaction to make the channels and their weights directly related [16]. In Cross-channel Interaction (CI), we design two 1D convolutions to make different channels sufficiently interact with nearby channels.

As shown in Fig. 4b, given an input tensor X $\in \mathbb{R}^{C\times H\times W}$, the filter (red dotted box) slides along the H*W, thus 2D convolution does not involve interaction between channels. On the contrary, as shown in Fig. 4a, the filter slides along the dimension C in 1D convolution, thus one sliding involves interaction between k channels.



(a) 1D convolution **(b)** 2D convolution

Fig. 4 Schematic diagram of one-dimensional convolution and two-dimensional convolution

The local cross-channel interaction effectively converts the material information into channel-wise weights, which enables the network to focus on important local features of prohibited items to remove the influence of invalid information. Specifically, we set the same kernel size for the two 1D convolutions as ECA [41]. The kernel size k is computed via a function related to the channels.

$$k = f(C) = int|\frac{\log_2(C) + 1}{2}| \tag{6}$$

According to the above function, channels of different dimensions can get different interaction distances, and the interaction distance of high-dimensional channels is longer

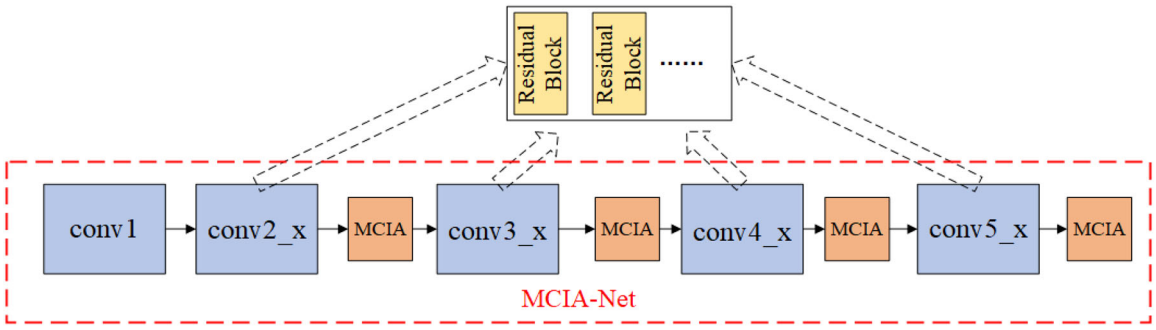


Fig. 5 The diagram of MCIA-Net. ResNet consists of a conv1 and four subsequent stages, including conv2_x, conv3_x, conv4_x and conv5_x, our MCIA module is placed behind each stage. The four stages of ResNet (conv2_x to conv5_x) are stacked by residual blocks

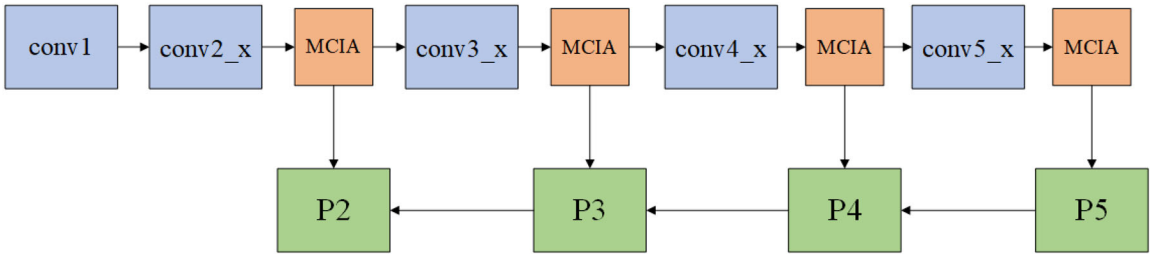


Fig. 6 The diagram of MCIA-FPN. {P2, P3, P4, P5} represents pyramid features of FPN [31]

than that of low-dimensional channels. In Table 5, we verify the practical benefits of the selected k for our two 1D convolutions compared to other kernel sizes.

Finally, the output of Cross-channel interaction (CI) is formulated as:

$$W_{CI} = Conv1D_2(Conv1D_1(S)) \tag{7}$$

Where $Conv1D_1(\cdot)$ and $Conv1D_2(\cdot)$ represent the two 1 D convolutions, respectively.

3.2 MCIA-Net and MCIA-FPN

3.2.1 MCIA-Net

The block diagram of the MCIA-Net is shown in Fig. 5. ResNet composes of a conv1 and four subsequent stages, named conv2_x, conv3_x, conv4_x and conv5_x [15]. Except that conv1 is a convolution with the kernel size of 7, con2_x to conv5_x are stacked by residual blocks. In general, attention modules are inserted into residual blocks of the four stages. The computational costs and parameters inevitably increase with the growing number of modules, hence rendering suboptimal performance.

Instead of inserting the proposed MCIA module into residual blocks, we place our MCIA module behind each of the four stages (conv2_x to conv5_x) in ResNet, so that only four MCIA modules are needed to achieve superior performance. Compared with the original ResNet-101, our MCIA-Net introduces negligible parameters and computational burden.

3.2.2 MCIA-FPN

Considering the different sizes of prohibited items in X-ray images, we design the MCIA-FPN based on MCIA-Net. Fig. 6 illustrates the detailed architecture of our MCIA-FPN. MCIA-FPN takes the output of MCIA to obtain {P2, P3, P4, P5} while the original FPN [24] utilizes the output of each stage to obtain {P2, P3, P4, P5}.

According to FPN [24], the semantic information of the low-level features is relatively insufficient, while the object location is accurate. In contrast, the semantic information of high-level features is richer, while the target location is relatively rough. FPN [24] makes full use of the complementary information from different layers to extract features of different dimensions effectively. Therefore, in our MCIA-FPN, {P2, P3, P4, P5} are scale attention feature maps containing both multi-scale information and attention information. Using MCIA-FPN as a backbone, the network emphasizes the core features of prohibited items and also takes into account various sizes.

3.3 Module complexity analysis

Take ResNet-101 as an example, the output of conv2_x is calculated by two methods to demonstrate the superiority of our method.

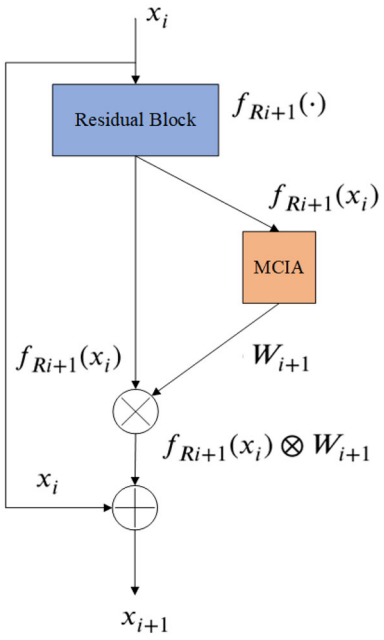


Fig. 7 The diagram of inserting our MCIA module into the residual block, namely ResNet+MCIA

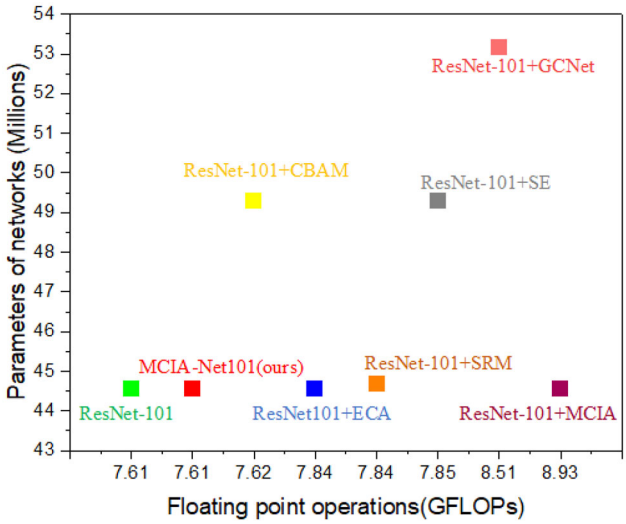


Fig. 8 Comparison of various attention modules in terms of network parameters and FLOPs, including ECA [41], SRM [20], SE[16], GCNet [8], CBAM [44], and ResNet-101+MCIA which represents inserting our MCIA module into residual blocks. Note that our MCIA-Net increases the lowest model complexity

Suppose an input x_0. Conv2_x is composed of three residual blocks. As shown in Fig. 7, if our MCIA is inserted into the residual block, the output of the first residual block can be computed as follows.

$$x_1 = x_0 \oplus (f_{R1}(x_0) \otimes W_1) \tag{8}$$

Where $f_{Ri}(\cdot)(i = 1, 2, 3)$ represents the residual block operation, $W_i(i = 1, 2, 3)$ represents the output weights of our proposed MCIA.

By analogy, the output of the second and third residual block can be, respectively, calculated as:

$$x_2 = x_1 \oplus (f_{R2}(x_1) \otimes W_2) \tag{9}$$

$$\begin{aligned} x_3 &= x_2 \oplus (f_{R3}(x_2) \otimes W_3) \\ &= (x_1 \oplus (f_{R2}(x_1) \otimes W_2)) \oplus \\ &\quad (f_{R3}(x_1 \oplus (f_{R2}(x_1) \otimes W_2)) \otimes W_3) \end{aligned} \tag{10}$$

To calculate the output of the next residual block, it is necessary to iterate the output of the previous residual block and iterate the weights continuously. With the accumulation of residual blocks, the computational cost also increases.

As shown in Fig. 5, if we place our MCIA behind the conv2_x, the output of the three residual blocks can be defined as:

$$x_1 = x_0 \oplus f_{R1}(x_0) \tag{11}$$

$$x_2 = x_1 \oplus f_{R2}(x_1) \tag{12}$$

$$x_3 = x_2 \oplus f_{R3}(x_2) \tag{13}$$

x_3 represents the original output of conv2_x, by placing our MCIA behind conv2_x, the output can be finally calculated as:

$$\hat{x}_3 = x_3 \otimes W_3 = (x_2 \oplus f_{R3}(x_2)) \otimes W_3 \tag{14}$$

We can find that our MCIA-Net involves fewer parameters.

As illustrated by Fig. 8, our MCIA-Net is lightweight in terms of floating-point operations (FLOPs) and parameters. We fairly compare MCIA-Net with several attention mechanisms based on ResNet-101 and analyze the model complexity. From the comparison between our MCIA-Net and the other five attention mechanisms, we can find that our MCIA-Net brings the least additional parameters.

In terms of computational complexity, MCIA-Net only introduces negligible extra computation to the original ResNet-101. For example, given a single forward pass of a 224×224 pixel image, MCIA-Net hardly increases the relative computational burden. This phenomenon may be attributed to the fact that our MCIA-Net only places our MCIA modules behind each stage, while the other five attention mechanisms and ResNet-101+MCIA insert attention modules into the residual blocks, resulting in more computational burden.

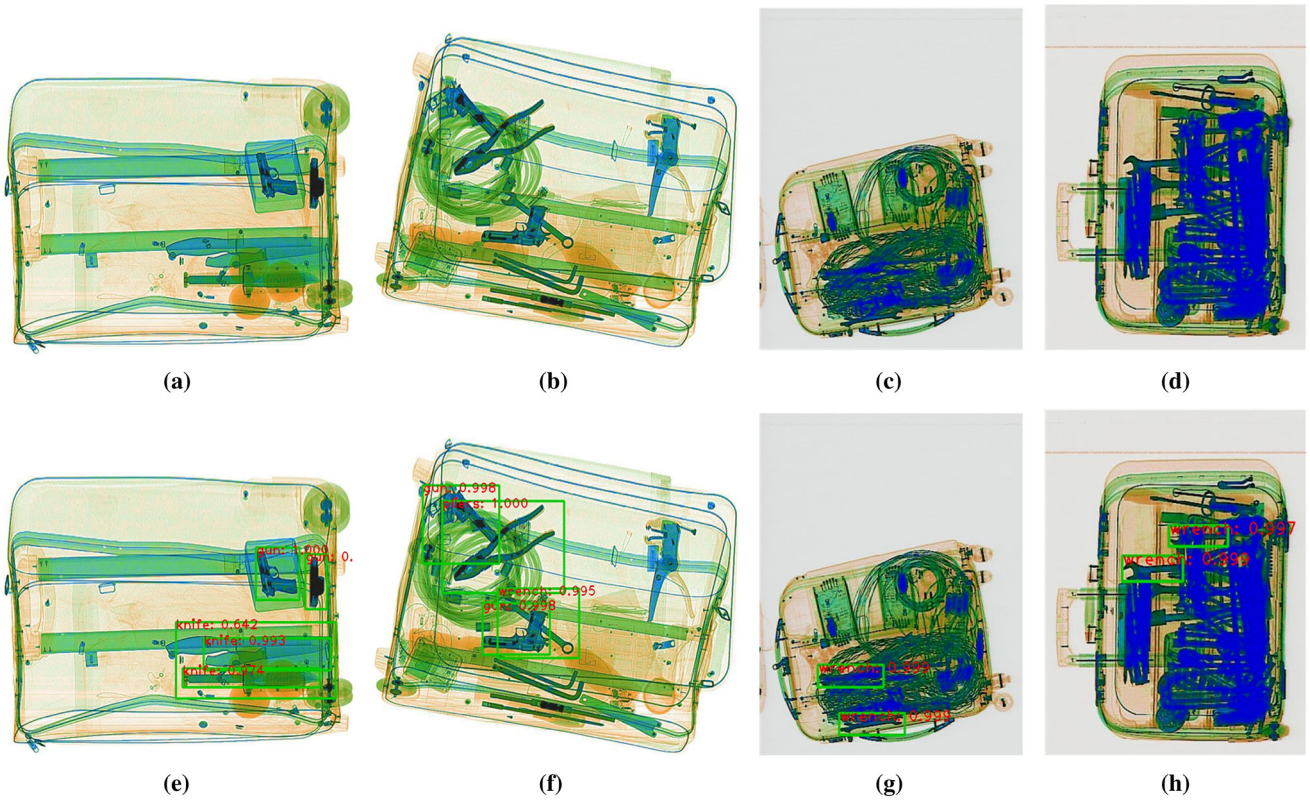


Fig. 9 Visual examples showcasing the performance of Cascade R-CNN+MCIA-FPN (ours) on SIXray dataset. (**a**) (**b**) (**c**) (**d**) represent the original images, and (**e**) (**f**) (**g**) (**h**) represent the images with detecting boxes

To sum up, our proposed MCIA-Net is more efficient than several attention mechanisms in terms of parameters and computational consumption.

4 Experiment

In this section, we systematically elucidate our experiments. First, we describe our implementation details, including datasets, evaluation metrics and training setting. Second, we verify the superiority of our MCIA-Net and MCIA-FPN on the SIXray dataset [31] and OPIXray dataset [43] comprehensively. Finally, in our ablation studies, we analyze three aspects, respectively, including the effectiveness of our MCIA module, the effects of the kernel size, and the influence of the placement of MCIA.

4.1 Implementation details

4.1.1 Datasets

To provide evidence for the effectiveness of our proposed MCIA module, we conduct comprehensive experiments on two publicly available X-ray image datasets, including SIXray dataset [31] and OPIXray dataset [43].

SIXray dataset [31] contains 1,059,231 complex X-ray images, of which 8929 contain six categories of prohibited items, namely, gun, knife, wrench, pliers, scissors, and hammer. Each of the 8929 X-ray images contains several prohibited items that are multi-scale and highly occluded. Of the 8929 X-ray images, about 80% are used for training and 20% for testing. Note that the hammer class with merely 60 samples is not used in our experiments.

OPIXray dataset [43] contains 8885 X-ray images, of which 7109 are used for training while the remaining 1776 are used as a test set. The test set is divided into three levels of occlusion, namely, OL1, OL2, OL3. Furthermore, the dataset contains five classes of prohibited items, including folding knives, straight knives, utility knives, multi-tool knives, and scissors. Different from the SIXray dataset [31], merely 35 images of the OPIXray dataset [43] contain more than one prohibited item, while the vast majority of images only contain one.

4.1.2 Evaluation metrics

Mean Average Precision (mAP) is the index for measuring recognition accuracy in object detection. We use the mAP to evaluate the performance of our method in detecting occluded prohibited items.

4.1.3 Training setting

We employ two detectors including Faster R-CNN [36] with IoU threshold of 0.5 and Cascade R-CNN [7] with IoU threshold of {0.5, 0.6, 0.7}, where ResNet-101 and ResNet-101 with FPN [24] are taken as backbones. Specifically, we implement all programs by PyTorch on a PC equipped with NVIDIA 1080Ti GPUs. During training, the training data is random shuffling and horizontal flipping. The optimizer is stochastic gradient descent (SGD) with a momentum of 0.9, weight decay of 0.0001. We set the initial learning rate 0.001 which is divided by 10 with a learning rate decay step of 5.

4.2 Evaluations on SIXray dataset

We use Faster R-CNN [36] and Cascade R-CNN [7] as baselines to evaluate our MCIA-Net and MCIA-FPN on the SIXray dataset [31]. We also compare our method with the CHR [31]. Table 1 shows the results, our method achieves better performance than CHR [31]. We can observe that Cascade R-CNN+MCIA-FPN achieves the mAP of 0.8370 leading the CHR [31] by 4.33%. Some visual results of our method are shown in Fig. 9. Note that when X-ray images contain some disruptive effects, our method can also detect prohibited items. Meanwhile, with our MCIA-Net or MCIA-FPN as the backbone, the performance of basic detectors is obviously improved. For example, MCIA-Net and MCIA-FPN can improve the performance of Faster R-CNN by 1.47% and 1.52%, respectively. We attribute the improvement to the capability of our MCIA module in dealing with inter-class occlusion.

In addition to inter-class occlusion, intra-class occlusion also exists in X-ray images. Soft-NMS [6] is helpful to solve the problem of intra-class occlusion, it increases the mAP of Cascade R-CNN+MCIA-FPN from 0.8370 to 0.8523.

4.3 Evaluations on OPIXray dataset

To further verify the effectiveness of MCIA-Net and MCIA-FPN, we use Faster R-CNN [36] and Cascade R-CNN [7] to conduct experiments on the OPIXray dataset [43]. In particular, we compare our methods with DOAM [43]. As can be observed from Table 2, Faster R-CNN+MCIA-Net outperforms FCOS+DOAM [43] by 3.48%. The visual results of our method are shown in Fig. 10. Furthermore, we can clearly find that MCIA-FPN brings impressive improvements to the baselines Faster R-CNN [36] and Cascade R-CNN [7]. Using the proposed MCIA-FPN as a backbone, the performance of the two detectors can be improved by 1.66% and 2.37%, respectively.

According to the work [43], the OPIXray dataset is divided into three levels of occlusion, OL1, OL2, OL3. Under the three levels of occlusion, we compare our methods with DOAM [43]. As shown in Table 3, using MCIA-Net and MCIA-FPN, both detectors can achieve considerable improvement, regardless of the levels of occlusion. The results prove that our MCIA-Net has great potential in detecting occluded prohibited items.

4.4 Ablation studies

In this part, we conduct three experiments. The first one examines the effectiveness of our proposed MCIA module itself. The second one explains the superiority of the kernel size of our two 1D convolutions. The last experiment proves why the MCIA module is placed behind each stage. In ablation studies, we employ Cascade R-CNN [7] as a baseline and train the detector on the SIXray dataset [31].

4.4.1 Effectiveness of MCIA module

To verify the effectiveness of the proposed MCIA module, in this subsection, we insert our MCIA module into the residual blocks, which is shown in Fig. 7. We compare our method with 5 attention mechanisms, including ECA [41], SRM [20], SE [16], GCNet [8], CBAM [42]. For a fair comparison, we implement all attention mechanisms according to the original paper. From Table 4, we can see that the proposed MCIA achieves the mAP of 0.8274 and acquires performance gains of 1.45% compared with the baseline ResNet-101. We attribute the superior performance to the ability of our MCIA module in dealing with inter-class occlusion. It can highlight the features of prohibited items by capturing the material information and converting it into channel-wise weights correspondingly.

Additionally, it is worth noting that the ResNet-101+MCIA (mAP: 0.8274) performs worse than the MCIA-FPN (mAP: 0.8370). It proves that for the proposed MCIA module, placing it behind each stage (conv2_x to conv5_x) is the best choice.

4.4.2 Number of kernel size

As illustrated in the previous section, our MCIA module involves two 1D convolutions with the same kernel size $f(C)$. To demonstrate why we choose $f(C)$, we set kernel size $\{3, 5, 7, 9, f(C)\}$, while other settings are exactly the same. The results are shown in Table 5, which shows the influence of kernel size on the MCIA module. The results indicate that our kernel size outperforms other kernel size values in detecting occluded prohibited items.

4.4.3 Influence of the placement of MCIA

We finally explore the effect of placing the MCIA module behind different stages. As shown in Table 6, first, we only

Table 1 Evaluations on SIXray Dataset [31].

Method	Backbone	Knife	Pliers	Gun	Wrench	Scissors	mAP
Faster R-CNN	ResNet-101	0.7365	0.8085	0.8122	0.7841	0.8172	0.7917
Faster R-CNN+MCIA-Net(ours)		0.7588	0.8008	0.8076	0.7740	0.8909	**0.8064**
Faster R-CNN+FPN	ResNet-101	0.7127	0.8101	0.8175	0.7911	0.8395	0.7942
Faster R-CNN+MCIA-FPN(ours)		0.7182	0.7965	0.8543	0.7979	0.8675	**0.8069**
Cascade R-CNN+FPN	ResNet-101	0.7739	0.8621	0.8136	0.7867	0.8284	0.8129
Cascade R-CNN+MCIA-FPN(ours)		0.7839	0.8547	0.8926	0.7954	0.8587	**0.8370**
Cascade R-CNN+MCIA-FPN(ours)+Soft-NMS		0.8375	0.8679	0.8575	0.8150	0.8834	**0.8523**
ResNet101+CHR[31]	ResNet-101	0.8721	0.8828	0.8545	0.7123	0.6468	0.7937

Table 2 Evaluations on OPIXray Dataset [43].

Method	Backbone	Folding	Straight	Scissor	Utility	Multi-tool	mAP
Faster R-CNN	ResNet-101	0.8906	0.7241	0.9044	0.8619	0.8986	0.8559
Faster R-CNN+MCIA-Net(ours)		0.8908	0.7448	0.8999	0.8613	0.8975	**0.8589**
Faster R-CNN+FPN	ResNet-101	0.8512	0.6422	0.8924	0.7754	0.8784	0.8079
Faster R-CNN+MCIA-FPN(ours)		0.8841	0.6475	0.8993	0.8440	0.8474	**0.8245**
Cascade R-CNN+FPN	ResNet-101	0.8301	0.6406	0.9011	0.7558	0.7720	0.7799
Cascade R-CNN+MCIA-FPN(ours)		0.8478	0.6299	0.8952	0.7749	0.8704	**0.8036**
SSD+DOAM[43]		0.8137	0.4150	0.9512	0.6821	0.8383	0.7401
YOLOv3+DOAM[43]		0.9023	0.4173	0.9696	0.7212	0.9523	0.7925
FCOS+DOAM[43]		0.8671	0.6858	0.9023	0.7884	0.8767	0.8241

Table 3 Comparison between our methods and DOAM [43] in three occlusion levels of OPIXray dataset [43].

Method	OL1	OL2	OL3
Faster R-CNN+MCIA-Net(ours)	**0.8689**	**0.8424**	**0.8537**
Faster R-CNN+MCIA-FPN(ours)	0.8224	0.8171	0.7958
Cascade R-CNN+MCIA-FPN(ours)	0.8223	0.7903	0.7747
SSD+DOAM[43]	0.7787	0.7245	0.7078

attach our MCIA module behind the conv2_x and then add one stage at a time, totaling four experiments. From the fourth row of the table, we can observe that our MCIA-Net brings the best performance benefits. We conclude the reason is that placing our MCIA module behind each stage in ResNet can combine multi-level features, which is helpful for the network to deal with the inter-class occlusion in X-ray images.

5 Conclusion

In this paper, we focus on detecting highly occluded prohibited items in X-ray images, which is promising in security inspection but is understudied. To this end, we proposed the Material-aware Cross-channel Interaction Attention (MCIA) module to capture the material information and convert it into channel-wise weights based on local cross-channel interaction. MCIA is effective for dealing with inter-class occlusion in X-ray images. According to the material information, MCIA can lead the network to concentrate on the local features of occluded prohibited items and remove the influence of unnecessary non-prohibited items information. Meanwhile, by placing the MCIA behind each stage (conv2_x to conv5_x) in ResNet, we designed MCIA-Net and MCIA-FPN. Our MCIA-FPN takes into account both multi-scale information and attention information. To comprehensively exhibit the superiority of our proposed method, we evaluated our MCIA-Net and MCIA-FPN on the SIXray dataset [31] and OPIXray dataset [43]. Experimental results demonstrate that our MCIA-Net and MCIA-FPN bring obvious improvement in detecting occluded prohibited items. Moreover, our

Table 4 Comparison of various attention mechanisms on SIXray dataset [31]. ResNet-101+MCIA means inserting our MCIA module into residual blocks, as shown in Fig. 7.

Method	Knife	Pliers	Gun	Wrench	Scissors	mAP
ResNet-101+FPN	0.7739	**0.8621**	0.8136	0.7867	0.8284	0.8129
ResNet101+ECA+FPN	0.7664	0.8343	0.8778	0.7775	0.8431	0.8198
ResNet-101+SRM+FPN	0.7777	0.8133	0.8815	0.7987	0.8441	0.8231
ResNet-101+SE+FPN	**0.7895**	0.8363	0.8197	0.7982	0.8376	0.8162
ResNet-101+GCNet+FPN	0.7445	0.8184	0.8622	0.7916	0.8111	0.8056
ResNet-101+CBAM+FPN	0.7670	0.8328	0.8709	0.7846	0.8488	0.8208
ResNet-101+MCIA+FPN	0.7759	0.8424	0.8632	**0.8147**	0.8407	**0.8274**
MCIA-FPN(ours)	0.7839	0.8547	**0.8926**	0.7954	**0.8587**	**0.8370**

Table 5 Results of our MCIA module with different numbers of k.

Kernel Size	Knife	Pliers	Gun	Wrench	Scissors	mAP
k=3	0.7734	**0.8564**	**0.8934**	0.7783	**0.8643**	0.8331
k=5	0.7674	0.8293	0.8083	0.7828	0.8349	0.8046
k=7	0.7765	0.8558	0.8673	**0.8020**	0.8577	0.8319
k=9	0.7794	0.8526	0.8904	0.7818	0.8635	0.8335
k=f(C)(ours)	**0.7839**	0.8547	0.8926	0.7954	0.8587	**0.8370**

Table 6 Comparison of placing our MCIA module behind different stages.

Method	Knife	Pliers	Gun	Wrench	Scissors	**mAP**
conv2_x	0.7662	0.8332	0.8510	0.7795	0.8110	0.8082
conv2_x-conv3_x	0.7758	0.8373	0.8508	0.7989	0.8315	0.8189
conv2_x-conv4_x	0.7667	**0.8560**	0.8773	**0.8116**	**0.8647**	0.8353
conv2_x-conv5_x(MCIA-FPN)	**0.7839**	0.8547	**0.8926**	0.7954	0.8587	**0.8370**

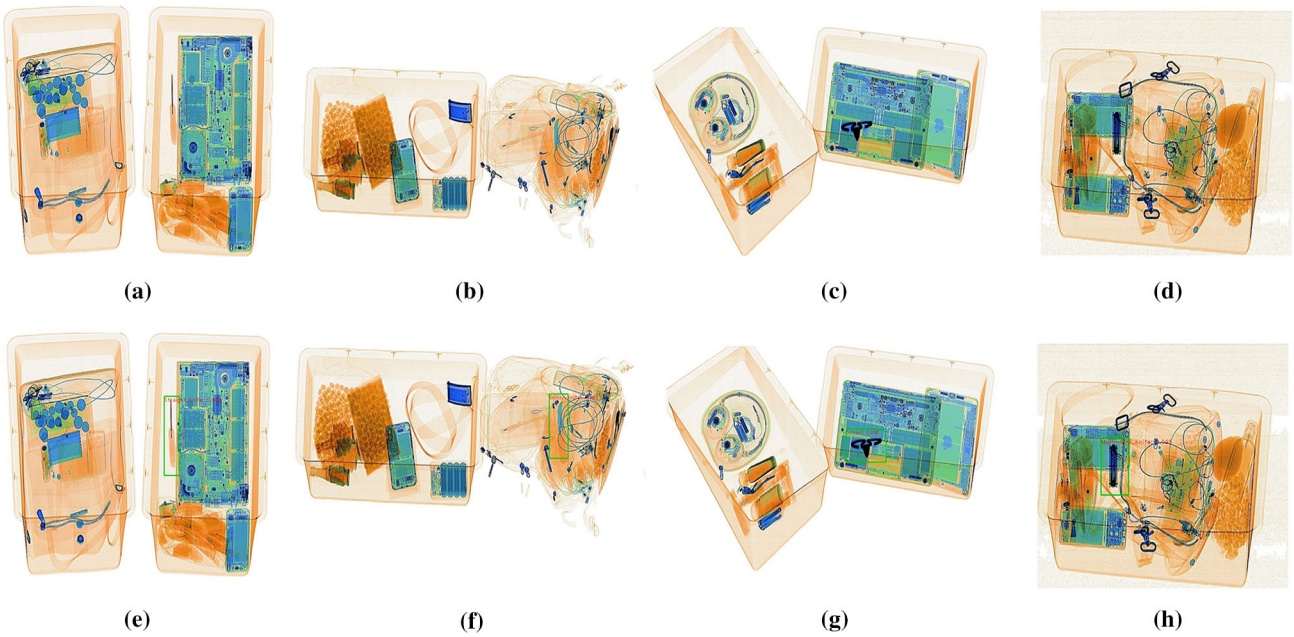


Fig. 10 Visual examples showcasing the performance of Faster R-CNN+MCIA-Net (ours) on OPIXray dataset. (**a**) (**b**) (**c**) (**d**) represent the original images, and (**e**) (**f**) (**g**) (**h**) represent the images with detecting boxes

MCIA-Net outperforms several attention mechanisms with lower model complexity.

Funding No funding was received to assist with the preparation of this manuscript.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Akcay, S., Breckon, T.: Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. arXiv preprint arXiv:2001.01293 (2020)
2. Akcay, S., Breckon, T.P.: An evaluation of region based object detection strategies within x-ray baggage security imagery. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1337–1341. IEEE (2017)
3. Akçay, S., Kundegorski, M.E., Devereux, M., Breckon, T.P.: Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1057–1061. IEEE (2016)
4. Baştan, M., Yousefi, M.R., Breuel, T.M.: Visual words on baggage x-ray images. In: International Conference on Computer Analysis of Images and Patterns, pp. 360–368. Springer (2011)
5. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
6. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms–improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017)
7. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
8. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0 (2019)
9. Cui, Y., Oztan, B.: Automated firearms detection in cargo x-ray images using retinanet. In: Anomaly Detection and Imaging with X-Rays (ADIX) IV, vol. 10999, p. 109990P. International Society for Optics and Photonics (2019)
10. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. arXiv preprint arXiv:1605.06409 (2016)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
12. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
13. Hassan, T., Bettayeb, M., Akçay, S., Khan, S., Bennamoun, M., Werghi, N.: Detecting prohibited items in x-ray images: A contour proposal learning approach. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 2016–2020. IEEE (2020)
14. 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)
15. 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, pp. 770–778 (2016)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
17. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
18. Jaccard, N., Rogers, T.W., Griffin, L.D.: Automated detection of cars in transmission x-ray images of freight containers. In: 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 387–392. IEEE (2014)
19. Kundegorski, M.E., Akçay, S., Devereux, M., Mouton, A., Breckon, T.P.: On using feature descriptors as visual words for object detection within x-ray baggage security screening (2016)
20. Lee, H., Kim, H.E., Nam, H.: Srm: A style-based recalibration module for convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1854–1862 (2019)
21. Li, Z., Zhou, F.: Fssd: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960 (2017)
22. Liang, K.J., Heilmann, G., Gregory, C., Diallo, S.O., Carlson, D., Spell, G.P., Sigman, J.B., Roe, K., Carin, L.: Automatic threat recognition of prohibited items at aviation checkpoint with x-ray imaging: a deep learning approach. In: Anomaly Detection and Imaging with X-Rays (ADIX) III, vol. 10632, p. 1063203. International Society for Optics and Photonics (2018)
23. Liang, K.J., Sigman, J.B., Spell, G.P., Strellis, D., Chang, W., Liu, F., Mehta, T., Carin, L.: Toward automatic threat recognition for airport x-ray baggage screening with deep convolutional object detection. arXiv preprint arXiv:1912.06329 (2019)
24. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
25. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
26. Liu, J., Leng, X., Liu, Y.: Deep convolutional neural network based object detector for x-ray baggage security imagery. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1757–1761. IEEE (2019)
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37. Springer (2016)
28. Liu, Z., Duan, Q., Shi, S., Zhao, P.: Multi-level progressive parallel attention guided salient object detection for rgb-d images. The Visual Computer pp. 1–12 (2020)
29. Liu, Z., Li, J., Shu, Y., Zhang, D.: Detection and recognition of security detection object based on yolo9000. In: 2018 5th International Conference on Systems and Informatics (ICSAI), pp. 278–282. IEEE (2018)
30. Mery, D., Svec, E., Arias, M.: Object recognition in baggage inspection using adaptive sparse representations of x-ray images. In: Image and Video Technology, pp. 709–720. Springer (2015)
31. Miao, C., Xie, L., Wan, F., Su, C., Liu, H., Jiao, J., Ye, Q.: Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2119–2128 (2019)
32. Nam, H., Kim, H.E.: Batch-instance normalization for adaptively style-invariant neural networks. Adv. Neural. Inf. Process. Syst. **31**, 2558–2567 (2018)

33. 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, pp. 779–788 (2016)
34. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
35. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
36. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2016)
37. Riffo, V., Mery, D.: Automated detection of threat objects using adapted implicit shape model. IEEE Trans. Syst. Man, Cyber: Syst. **46**(4), 472–482 (2015)
38. Shajini, M., Ramanan, A.: An improved landmark-driven and spatial–channel attentive convolutional neural network for fashion clothes classification. The Visual Computer pp. 1–10 (2020)
39. Shi, W., Du, H., Mei, W., Ma, Z.: (sarn) spatial-wise attention residual network for image super-resolution. The Visual Computer pp. 1–12 (2020)
40. Steitz, J.M.O., Saeedan, F., Roth, S.: Multi-view x-ray r-cnn. In: German Conference on Pattern Recognition, pp. 153–168. Springer (2018)
41. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542 (2020)
42. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
43. Wei, Y., Tao, R., Wu, Z., Ma, Y., Zhang, L., Liu, X.: Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module. arXiv preprint arXiv:2004.08656 (2020)
44. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
45. Zhang, T., Gao, F., Dong, J., Du, Q.: Remote sensing image translation via style-based recalibration module and improved style discriminator. IEEE Geosci. Remote Sens. Lett. **19**, 1–5 (2021)
46. Zhang, Yt., Zhang, Hg., Zhao, If., Yang, Jf.: Automatic detection of prohibited items with small size in x-ray images. Optoelectron. Lett. **16**(4), 313–317 (2020)

Man Wang received her B.S degree in Communication Engineering from Beijing Institute of Technology, Beijing, China in 2019. She is currently pursuing her M.S. degree in the School of Information and Electronics, Beijing institute of technology, Beijing, China. Her main research interests include image processing, object detection and deep learning.

Huiqian Du received her BS, Master and Dr.Eng. degrees in Electrical Engineering from Beijing Institute of Technology, Beijing, China, in 1995, 1998 and 2007 respectively. In April 1998, she joined Beijing Institute of Technology, where she is currently an associate professor. She was a visiting research scholar at University of Illinois at Urbana Champaign. Her research interests lie in the fields of image processing, image reconstruction, image fusion, object detection, machine learning.

Wenbo Mei (M'87) received the B.S. degree in Electronic Engineering and M.S. degree in Communication and Electronic System from Beijing Institute of Technology(BIT), China, in 1982 and 1993, respectively. He is a professor at School of Information and Electronics, BIT. He is also a Visiting Research Fellow at University of Central Lancashire, a Visiting Professor at University of Reading UK. His research interests include wavelet transform, time-frequency analysis, compressive sensing in signal and image processing for radar, communication system and channel, MRI, etc.

Shuai Wang received his B.S. degree from Harbin Institute of Technology and his M.S. degree from the Second Institute of the China Aerospace Science and Industry Group. He is currently pursuing his Ph.D. degree in the school of Information and Electronics, Beijing Institute of Technology, Beijing, China. His main research interests include object detection and deep learning.

Dasen Yuan received his B.S. degree in Electronic Information Engineering from Naval Aeronautical Engineering College, his M.S. degree from Changchun University of Science and Technology, and his Ph.D. degree from Beijing Institute of Technology. He currently works in The Inner Mongolia Autonomous Region Public Security Bureau. His main research interests include the application of big data in social management and control, the application of artificial intelligence and image processing in security inspection.