Automatic and Robust Object Detection in X-Ray Baggage Inspection Using Deep Convolutional Neural Networks

Bangzhong Gu, Rongjun Ge, Yang Chen, *Senior Member, IEEE*, Limin Luo, *Senior Member, IEEE*, and Gouenou Coatrieux, *Senior Member, IEEE*

***Abstract*—For the purpose of ensuring public security, automatic inspection of X-ray scanners has been deployed at the entry points of many public places to detect dangerous objects. However, current surveillance systems cannot function without human supervision and intervention. In this article, we propose an effective method using deep convolutional neural networks to detect objects during X-ray baggage inspection. As a first step, a large amount of training data is generated by a specific data augmentation technique. Second, a feature enhancement module is used to improve feature extraction capabilities. Then, in order to address the foreground–background imbalance in the region proposal network, focal loss is adopted. Third, the multiscale fused region of interest is utilized to obtain more robust proposals. Finally, soft nonmaximum suppression is adopted to alleviate overlaps in baggage detection. As compared with existing algorithms, the proposed method proves that it is more accurate and robust when dealing with densely cluttered backgrounds during X-ray baggage inspection.**

***Index Terms*—Baggage detection, baggage inspection, convolutional neural networks (CNNs), X-ray images for security applications.**

Manuscript received December 20, 2019; revised April 10, 2020 and July 18, 2020; accepted August 28, 2020. Date of publication September 29, 2020; date of current version June 28, 2021. This work was supported in part by the National Key R&D Program of China under Grant 2017YFA0104302, Grant 2018YFA0704102, and Grant 2017YFC0109202, in part by the National High Technology Research and Development Program of China (863 Program, 2015AA043203), in part by the National Natural Science Foundation under Grant 81827805, Grant 61801003, Grant 61871117, and Grant 81530060 and in part by the R&D Projects in Key Technology Areas of Guangdong Province under Grant 2018B030333001. *(Corresponding author: Yang Chen.)*

Bangzhong Gu, Rongjun Ge, Yang Chen, and Limin Luo are with the Laboratory of Image Science and Technology, School of Computer Science and Technology, Southeast University, Nanjing 210096, China, with the School of Cyberspace Security, Southeast University, Nanjing 210096, China, with the Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing 210096, China, and also with the Centre de Recherche en Information Biomedicale Sino Francais, 35000 Rennes, France (e-mail: gu_bangzhong@163.com; 230169430@seu.edu.cn; chenyang.list@seu.edu.cn; luo.list@seu.edu.cn).

Gouenou Coatrieux is with the Institut Mines-Telecom, Telecom Bretagne, INSERM U1101 LaTIM, 29238 Brest, France (e-mail: gouenou.coatrieux@telecom-bretagne.eu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2020.3026285

I. Introduction

BAGGAGE inspection with security screening is a powerful tool to reduce the risk of potential terrorism and crime [1]. Security screening using X-ray scanners is widely used in public places [2]. These scans are visually inspected by a specifically trained human inspector to ensure there are no dangers. It is extremely tedious to manually perform this task since the baggage might actually be dangerous [3]. During rush hour, it only takes a few seconds to determine whether a piece of baggage contains any dangers or not [4]. Since each employee has to check a large amount of baggage, the possibility of human error over a long time is considerable, even with specialized training [5]. Automated X-ray analysis remains a crucial issue in baggage inspection.

X-ray imaging is quite different from natural optical imaging in several aspects. The main difference is that the X-ray image is formed by irradiating the object with X-rays, whereas the natural optical image is formed by the light reflection, which gives information about the surface of the objects [6], [7]. Thus, an X-ray image consists of shadows from overlapping transparent layers. The transparency of the image is determined by the material density along the X-ray path. The visibility of objects on X-ray images depends on the object's density: high-density objects (e.g., thick metal) behave substantially opaque and occlude all the other overlapping objects, whereas very low-density objects (e.g., clothes) are barely visible [8]. Fig. 1 shows some examples of baggage using X-ray scanners. If part of the image is too dark to be visible, the human inspector needs to open the baggage and check it manually.

Objects in X-ray baggage usually undergo in-plane and out-of-plane rotation, which makes X-ray baggage inspection to be a quite a difficult task. Despite these slight differences, object recognition using both imaging techniques suffers many similar issues, such as perspective projection, geometric distortion, pose problems, self-occasions, and large intraclass variability [9]. Observing the common problems, algorithms based on computer vision technology for optical imaging can also be used for X-ray baggage inspection.

In recent years, convolutional neural networks (CNNs) have been widely used in image analysis and interpretation. Methods based on deep learning have achieved state-of-the-art detection performance in many computer vision tasks [10]–[12], such as

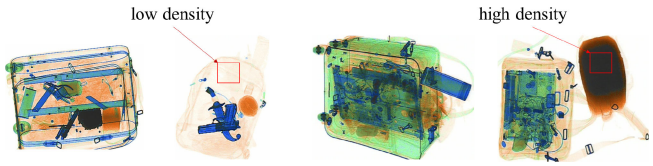


Fig. 1. Samples of X-ray baggage in inspection.

face recognition and automatic driving. However, few efforts have been dedicated to investigate object detection in X-ray baggage inspection due to many limitations. For the lack of training data, most of the existing methods fine-tune the networks [13] to achieve good performance. But this is not feasible in X-ray baggage inspection. Direct adoption of a pretrained network has little flexibility to adjust the structure. There might be bias in the learning process. A good solution to tackle these critical issues is to train the models from scratch. However, due to numerous parameters and inefficient training strategy with the limited training data, previous approaches are difficult to converge [14].

To address these issues, in this article, we propose an effective approach for object detection in X-ray baggage inspection. Compared with other detection methods for object surface, such as Faster Region-based CNNs (Faster RCNN) [15] and feature pyramid network (FPN) [16], our method has great advantages regarding object detection in X-ray baggage inspection for object interior character. The main contributions of the proposed method are as follows. First, a specific data augmentation pipeline is designed to accommodate the varied data. Second, an effective feature enhancement module is added to improve feature extraction capabilities; focal loss is adopted to address the foreground–background imbalance. Third, multiscale fused region of interest (RoI) is adopted to obtain more accurate region proposals. Finally, soft nonmaximum suppression (NMS) is used to reduce errors when detecting adjacent objects. Two new datasets are built for X-ray baggage object detection. To evaluate the method, a list of representative CNN-based methods is investigated on the task of object detection during X-ray baggage inspection. The results are reported as a useful performance baseline, and the proposed method outperforms the existing ones.

The rest of this article is organized as follows. Related work is explored in Section II. The proposed method in detail is presented in Section III. All methods evaluated in this work are reported in Section IV. Finally, Section V concludes this article.

II. Related Work

In this section, we briefly introduce traditional object detection methods in X-ray baggage inspection and CNN-based models for object detection.

A. Traditional Object Detection in X-Ray Baggage Inspection

Some approaches attempted to perform object detection in X-ray baggage images from a single view of a single energy. The adapted implicit shape model based on visual codebooks was proposed in [17]. This method used visual vocabulary and appearance structures that were generated from a training dataset, which includes representative X-ray images of the target object. Mery *et al.* [18] used adaptive sparse representations[19], [20] to automatically detect objects, when less restrictive conditions apply, including some contrast, pose, intraclass variability, and focal distance. The task presented in [21] considered a bag of visual words (BoVW) model with several hand-crafted feature representations. It achieved an average precision (AP) of 57%. Mery *et al.* [22] studied the applicability and efficiency of sparse local features on X-ray baggage object detection. This work investigated how material information given in multiview X-ray imaging affects detection performance.

As clearly seen, these methods are mostly based on hand-crafted features. However, the advances in automated baggage inspection are minimal and very limited compared to what is required for X-ray inspection systems, which rely less on human inspectors.

B. CNNs for Object Detection

Deep CNNs have made huge steps in object detection in recent years. State-of-the-art deep CNN based object detection methods can be divided into two groups: two-stage methods and single-stage methods. Two-stage methods, such as R-CNN [23], Fast R-CNN [24], Faster R-CNN [15], R-FCN [25], and FPN [16], achieve the detection through two steps. The first step generates a set of candidate region proposals and the second step classifies them into the target object category. To date, two-stage methods have achieved the highest accuracy among object detection methods. Single-stage methods, such as YOLO [26]–[28] and SSD [29], use a single feed-forward convolutional network to directly predict classes and bounding boxes. Although these methods have been tuned for speed, their accuracy is lower than two-stage methods.

Object detection during X-ray baggage inspection is a more challenging task than in natural optical images. To the best of our knowledge, most of the previous work used networks that were pretrained on the ImageNet classification dataset. The study presented in [30] compared a BoVW approach with a CNN approach, exploring the use of transfer learning by fine-tuning the weights of different layers. The layers were transferred from another network trained on a different task. Experiments show that CNN-based methods outperform BoVW methods. Akcay *et al.* [31] explored some framework on X-ray baggage image classification and detection. Their results showed that the CNN-based method outperforms hand-crafted methods.

III. Proposed Method

Our method is inspired by the design principles of the two-stage methods. Thus, it inherits the accuracy advantages of region proposal based methods. Fig. 2 illustrates the architecture of the proposed method, which can mainly be divided into two parts: X-ray proposal network (XPN) and X-ray discriminative network (XDN). XPN takes an image as input and outputs predicted region boxes. XDN is added after XPN. XDN takes the

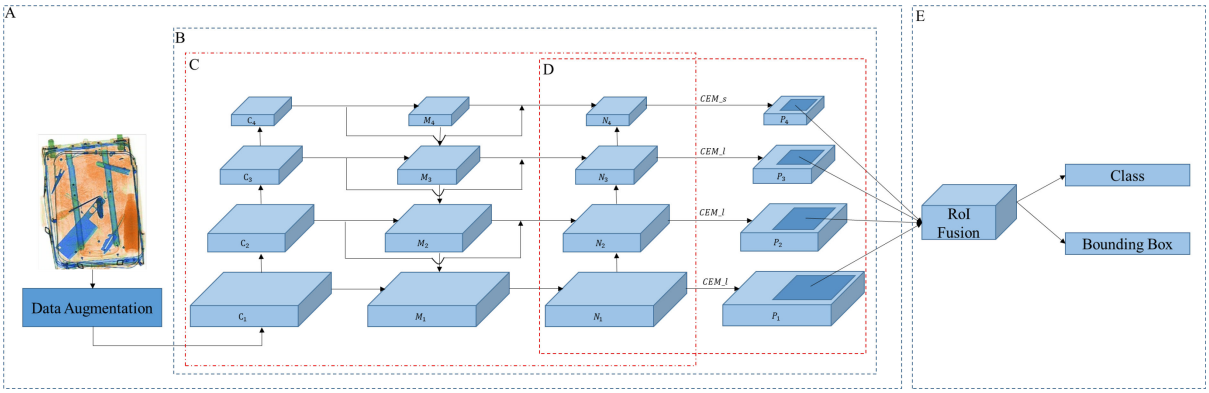


Fig. 2. Architecture of the proposed method. Part A represents the XPN. Part E represents the XDN.

coarse region boxes as input and outputs the refined category and position simultaneously. In XPN, data augmentation is used to accommodate the diversity of the input image. The followed feature enhancement module is utilized to make information easier to propagate. In XDN, the fused RoI layer allows each proposal to access information from all levels. Then, the bounding box regression and class prediction are processed. After XDN, soft NMS is used to alleviate object overlaps in the model.

A. Data Augmentation

Data augmentation plays an important role in increasing the network robustness against normal changes that might appear in X-ray images, such as density changes or changes in object orientation. Additionally, it can be used to achieve better generalization and to simulate different X-ray object conditions thus overcoming one of the main weaknesses of CNN: its heavy reliance on previous training data. X-ray images are quite different from natural images since they undergo severe geometric transformations and are densely cluttered, this makes it difficult to cover most situations in the training set. In this article, we use an online augmentation approach providing a virtually infinite dataset that does not require extra storage space on the disk. Many applications use basic geometric transformations for data augmentation, such as mirroring and flipping. In order to change the position of the objects, affine transformations are performed. Besides the basic data augmentation, we design an effective pipeline for X-ray images inspection to handle the problem of densely cluttered objects in X-ray images. Fig. 3(a) shows the details of the specific technique. We select two random images A and B from the database, image A belongs to the data that contain target objects and images B belongs to the data that contain no target objects. We cut the part containing the target object from image A, namely patch A. Then, we apply basic data augmentation to patch A. Finally, we combine image B with patch A to build the augmented data used for training. We define the combined operation as

$$C = \text{inv}(\lambda \times \text{inv}(\text{op}(A)) + (1 - \lambda) \times \text{inv}(B)) \quad (1)$$

where C represents the composited image, λ represents the combination ratio, which is sampled from the uniform

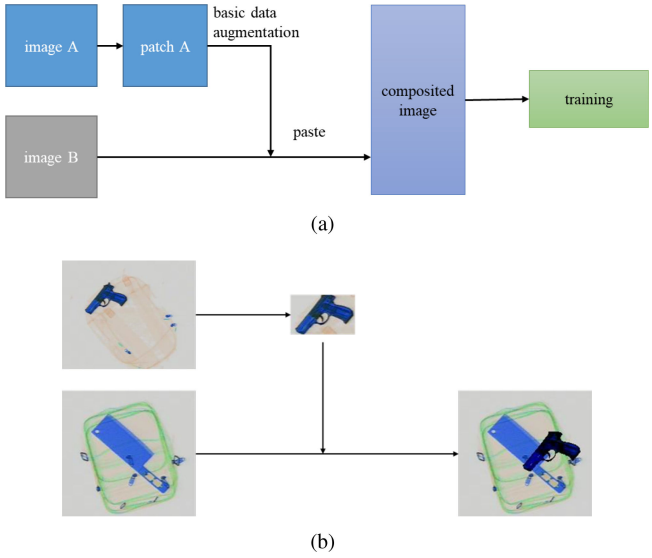


Fig. 3. Description of the proposed data augmentation technology. (a) the details of data augmentation pipeline. (b) an example of data augmentation.

distribution (0,1), inv$(\cdot)$ represents the complement of the image, and op$(\cdot)$ represents the basic augmentation operation. In this method, we use the following basic data augmentation: affine transformations, mirroring and flipping, cropping, and perspective transformations. Fig. 3(b) shows an example of our data augmentation. A gun patch is cut from image A. Then, we apply some basic augmentation to the gun patch. Finally, we paste the gun patch on another image B for training. This technology can make full use of the data that have no target object.

B. Feature Enhancement Module

Feature enhancement module is utilized to enhance the information flow, which is not trivial for traditional CNN-based detectors, especially for densely cluttered background and small objects. It has been found that in [32], low layers contain less semantic information compared with high layers, but they have higher localization accuracy. Since object detection requires

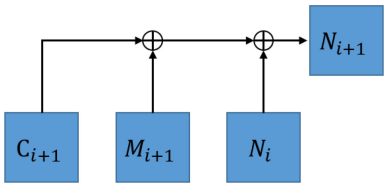


Fig. 4. Detail of the last bottom-up path that bring more combinations for the feature maps.

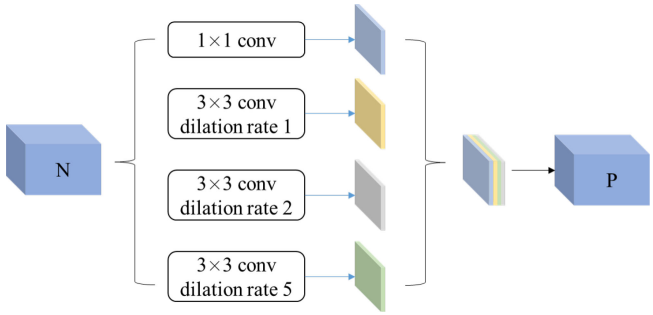


Fig. 5. Structure of CEM.

both accurate positions and precise categories, multiple layers fusion is required. Several recent models for object detection utilized different layers in a network. Some of them fused the high layers with low layers. Intuitively, low layers can provide more detailed information about X-ray objects with cluttered backgrounds. In contrast, high layers can capture more global context information. Although these methods have achieved excellent performance on some datasets, X-ray object detection is still a challenge for them with its characteristics, including the existence of disruptors and densely cluttered background.

Part B in Fig. 2 is the proposed feature enhancement module. The part B consists of parts C and D. The part C includes the first bottom-up path (indicated as $\{C_1, C_2, C_3, C_4\}$), the second top-down path (indicated as $\{M_4, M_3, M_2, M_1\}$), and the last bottom-up path (indicated as $\{N_1, N_2, N_3, N_4\}$). The bottom-up path $\{C_1, C_2, C_3, C_4\}$ and the top-down path $\{M_4, M_3, M_2, M_1\}$ are an FPN with the backbone of Resnet-50. The last bottom-up path $\{N_1, N_2, N_3, N_4\}$ is a combination of the first bottom-up path $\{C_1, C_2, C_3, C_4\}$ and the second top-down path $\{M_4, M_3, M_2, M_1\}$. The layers in the last bottom-up path $\{N_1, N_2, N_3, N_4\}$ are obtained by lateral connections, enhancing the spread of semantic information. The last bottom-up path starts from the low-level feature map N_1 and gradually reaches N_4. The spatial down-sampling factor is 2. The channel number is 256. They are consistent with the first bottom-up path $\{C_1, C_2, C_3, C_4\}$. The N_1 layer comes from the M_1 layer with 1×1 convolution. The N_{i+1} layer is defined as

$$N_{i+1} = N_i \oplus (C_{i+1} \oplus M_{i+1}) \tag{2}$$

where $\oplus$ is a concatenation operation. A 1×1 convolution operation is applied to the previous layers (C_{i+1} and M_{i+1}). Together with down-sampled layer of N_i, three components are added to produce fused layer N_{i+1}. With this design, the current layers can take full advantage of prior information to extract more discriminative representations. Fig. 4 shows the structure of the last bottom-up path $\{N_1, N_2, N_3, N_4\}$.

Dilated convolution layer has achieved progressive improvement in semantic segmentation, which can provide context information [33]–[35]. In this work, dilated convolution layers are utilized to enhance the feature maps for region proposal, thus making them more discriminable and robust. Fig. 5 shows the details of our context enhancement module (CEM). It takes the feature maps N as input, and outputs the feature maps P. The module contains one convolution layer and several dilated convolution layers with different dilation rates. We then concatenate the output feature maps of different convolution layers.

More details are shown in the part D of Fig. 2. We use two CEMs in this work, named as CEM_s and CEM_l. CEM_s have one 1×1 convolution layer and two 3×3 dilated convolution layers with dilation rate of 1 and 2. CEM_l has one more dilated convolution layer with a dilation rate of 5 than CEM_s. CEM_s is used for the feature map with small spatial resolution, whereas CEM_l is used for the feature map with large spatial resolution. The layers, indicated as $\{P_1, P_2, P_3, P_4\}$, have the same spatial resolution and channel size with the last bottom-up path $\{N_1, N_2, N_3, N_4\}$. This module effectively captures multiscale information especially for objects that are small or contained in a densely cluttered background.

C. Focal Loss in Region Proposal Network

Focal loss [36] is utilized to avoid class imbalance problem by down-weighting the losses of vast number of easy samples during the training of XPN. The basic region proposal network generates a large number of regions, including more negative regions than the positive ones. To compensate for this imbalance, sampling strategies, including random sampling and hard negative mining, are adopted in most models. Only a fixed number of anchors with a fixed ratio are sampled. However, the resulting sampled positives cannot fully represent the objects. In this article, we use focal loss to take all regions into account for training.

With the traditional definitions, the training regions of the nth proposal layer are defined as $S^n = \sum_i (p_i^*, b_i^*)$, where p_i^* and b_i^* are the corresponding label and ground truth coordinates, respectively. Similar to most CNN-based detectors, the loss of the ith sample in the nth detection layer is a combination of classification and bounding box regression, which is defined as follows:

$$L^n\left(p_i, b_i | W\right) = L_{\text{cls}}\left(p_i, p_i^*\right) + \lambda L_{\text{reg}}\left(b_i, b_i^*\right) \tag{3}$$

where W represents the parameters of the region proposal network, p_i is the probability distribution over the background and foreground object that is calculated by a softmax layer, λ is the balancing parameter, and b_i stands for the regressed bounding box. For regions that are positively labeled, the bounding box b_i^* is regressed from the corresponding region box b_i. The regression loss denotes a smooth L_1 loss defined as

$$L_{\text{reg}}\left(b_i, b_i^*\right) = \frac{1}{4} \sum_{j \in \{x,y,w,h\}} f_{L_1}\left(b_i^j - b_i^{*j}\right) \tag{4}$$

where

$$f_{L_1} = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}. \tag{5}$$

Most CNN-based detectors adopt cross-entropy (CE) loss as the softmax loss function, this loss is only effective when training minibatch samples, where the sampled positive and negative anchors have a ratio of up to 1:1 or 1:2. In order to naturally handle the foreground–background imbalance, we adopt the focal loss function that allows an efficient training on all anchors without any sampling strategies. The focal loss is defined as follows:

$$\begin{aligned} &\sum_{i \in S^m} L_{cls}\left(p_i, p_i^*\right) \\ &= \beta \sum_{i \in S_+^m} -(1-p_i)^2 \log p_i \\ &\quad + (1-\beta) \sum_{i \in S_-^m} -{p_i}^2 \log\left(1-p_i\right) \end{aligned} \tag{6}$$

where p_i is the probability confidence of the object and $1 - p_i$ is the probability confidence of the background, S_+^m and S_-^m represent the positive and negative anchors, respectively, and β is the balancing parameter to avoid domination of training loss by the negative anchors. Compared with CE loss, focal loss has two advantages. First, the loss is similar to CE loss for misclassified samples. For example, when a positive sample is misclassified and p_i is small, the modulating factor $(1 - p_i)^2$ is near 1, whereas in the case of misclassifying a negative sample and having a large p_i, the modulating factor ${p_i}^2$ is near 1. This results in unaffected loss. Second, the loss is smoothly downweighted for well-classified samples. For example, when a positive sample is well classified with a large p_i, the modulating factor $(1 - p_i)^2$ goes to 0, whereas a well-classified negative sample and a small p_i lead to a modulating factor ${p_i}^2$ that is near 0. Thus, it can prevent the large number of easy negatives from dominating the loss during training.

D. Multiscale RoI Fusion

Multiscale RoI fusion aims at combining different levels of RoIs for each proposal, which can make the proposals stronger. Since each feature map contains some specific information, the classification and regression of the bounding boxes are individually operated on different feature levels. However, a small region proposal can obtain more semantic information from the higher layers, which is helpful for classification, and a large region proposal can get better details from the lower layers to facilitate its localization ability. Thus, RoIs in our work are proposed from all levels in the layers $\{P_1, P_2, P_3, P_4\}$ for each proposal, they are shown as dark blue regions in part D of Fig. 2. After they are sent to RoI align pooling layer, we get all the feature grids with the same size. Then, an elementwise max fusion operation is utilized to fuse feature grids from all levels. The fused feature grid is used for the following prediction. Fig. 6 shows the details of multiscale RoI fusion module. Following mask R-CNN [37], RoI align pooling is utilized to pool feature grids from each level.

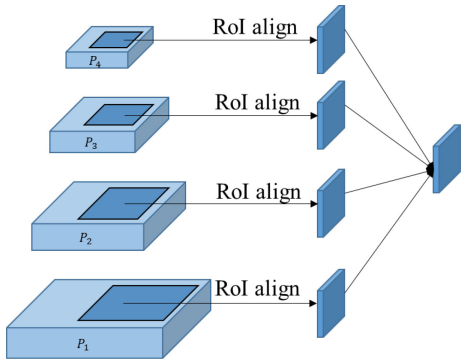


Fig. 6. Details of multiscale RoI fusion.

The original RoI pooling uses quantization to transfer the coordinates of the object in an image to its coordinates on the feature map. RoI align pooling uses bilinear interpolation to compute the exact values of the input features, which makes the results more accurate. By using bilinear interpolation operation, the RoI align pooling layer can take full use of the information learned from previous layers. However, other interpolation operations may cause a detail or semantic information loss. Nearest neighbor interpolation computes the sampling points by using the nearest neighbor point. This operation may lead to a detail information loss for the pooled feature grids. Bicubic interpolation computes the sampling points by using the nearest 16 points. Some of the points come from the points out of the RoI. This operation may cause a semantic information loss for the pooled feature grids.

E. Soft NMS

NMS has been a crucial part of many detection algorithms, in which it is used to obtain the final set of detections by significantly reducing the number of false positives. In nondense scenes, greedy NMS is applied to object scores, which can resolve overlapping detections. In densely cluttered images, however, multiple overlapping bounding boxes often reflect multiple, tightly packed objects among which many receive high object scores. NMS does not adequately discriminate between overlapping detections and suppress partial detections.

To address these problems, we adopt soft NMS [38] to alleviate the mistakes in detecting adjacent objects. Let $B = \{b_1, b_2, ..., b_N\}$ be a list of candidate bounding boxes and $S = \{s_1, s_2, ..., s_N\}$ be a list of corresponding detection scores. The detection box with the maximum score is denoted as b_m and the threshold is T. $\text{iou}(b_m, b_i)$ denotes the intersection over the union between b_m and b_i. The choice criterion in NMS can be written as follows:

$$s_i = \begin{cases} s_i, \text{iou}\left(b_m, b_i\right) < T \\ 0, \text{iou}\left(b_m, b_i\right) \geq T \end{cases}. \tag{7}$$

Hence, NMS sets a hard threshold when deciding what should be kept or removed. Instead of just removing the neighboring detections with an overlap greater than the threshold T, soft NMS assigns them great penalty with the purpose to impose a higher penalty on the bounding box with a higher overlap. Thus,

the choice criterion instead of (7) is proposed as follows:

$$s_i = s_i e^{-\frac{\text{iou}(b_m, b_i)^2}{\delta}}, i = 1, 2, ..., N \tag{8}$$

where δ is the variance.

IV. Experimental Results and Analysis

In this section, we evaluate our proposed method for object detection in X-ray baggage inspection. Experiments were implemented based on the deep learning framework PyTorch [39].

A. Dataset Description and Experimental Setup

1) Dataset Description: To perform this task, we use two types of datasets described below.

a) Xdb1: This dataset is collected from the simulated situation. To build a dataset for multiclass detection, baggage with random target objects is packed and then sent into a fixed X-ray machine. The X-ray tube voltage is 55 kV. The X-ray tube current is 5 mA. Finally, we get the simulated X-ray baggage image. In addition to baggage with target objects, we also build a set of images with no target objects. Following these approaches, this dataset consists of 4127 X-ray baggage images that contain target objects and 2352 X-ray baggage images without target objects. The target objects include scissors, bottle, mental cup, kitchen knife, knife, battery, and umbrella.

b) Xdb2: This dataset is collected from the subway security check. To build a dataset for multiclass detection, we manually select baggage images that contain target objects. In addition, we randomly select images from a large number of baggage images with no target objects. Following these approaches, this dataset consists of 21 538 X-ray sample images with target objects and 35 254 X-ray sample images with no target objects. The target objects include bottles, mental cup, knives, scissors, gun, battery, laptop, umbrella, lighter, and pressure cans.

2) Evaluation Metrics: The performance of our method is measured using the quality evaluation PASCAL criteria [40]: AP and mean AP (mAP). To calculate mAP, we perform the following: we first sort detections based on their confidence scores. Next, we calculate IoU for each detection. Assuming each detection as unique, and denoting the IoU area as a_i, we then threshold it by 0.5 giving a logical l_i, where

$$l_i = \begin{cases} 1, a_i > 0.5 \\ 0, \text{otherwise} \end{cases}. \tag{9}$$

This is followed by a prefix-sum giving both true positives t and false positives f, where

$$\begin{aligned} t_i &= t_{i-1} + l_i \\ f_i &= t_{i-1} + 1 - l_i. \end{aligned} \tag{10}$$

The precision p and recall r curves are calculated as

$$\begin{aligned} p_i &= \frac{t_i}{t_i + f_i} \\ r_i &= \frac{t_i}{n_p} \end{aligned} \tag{11}$$

where n_p is the number of positive samples. We then calculated AP based on the area under precision recall curve

$$\text{AP} = \sum_{i}^{n_p} p_i \Delta r. \tag{12}$$

As shown in (12), we finally find mAP by averaging AP values that calculate for N classes

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i. \tag{13}$$

3) Implementation Details and Parameter Optimization: Due to the large size of the X-ray images, training the proposed detector requires large amounts of memory resulting in a long training time. In our implementation of fast training, we randomly crop small patches around the target object from the input images used for training. This significantly reduces the used memory and enables training with a larger minibatch size, which can speed up the training.

In our method, the anchor size is empirically set according to the size distribution of the target objects. We use five scales with box areas of 16^2, 32^2, 64^2, 128^2, and 256^2 pixels and five aspect ratios of 4:1, 2:1, 1:1, 1:2, and 1:4. Since some target objects are long, such as knives, we added 4:1 and 1:4 aspect ratios to the anchors with aspect ratios of 2:1, 1:1, and 1:2. The branch weight was empirically set based on the effect of the corresponding detection branch on the gradient. Other parameters are set to default values, including the learning rate (0.02 with a linear warm-up) for the first 500k minibatches and (0.002) for the next 400k, the weight decay (0.0001) and momentum (0.9), and the tradeoff coefficient. In our training process, the X-ray region proposal network and XDN are jointly trained and all models are trained using synchronized stochastic gradient descent (SGD). The standard Faster R-CNN architecture typically adopts a fixed scale for all the training images. By randomly resizing the images to one of many scales, the detector is able to learn features across a wide range of sizes, thus improving its performance toward scale invariance. In this work, we randomly assign one of three scales (800, 600, and 400) to each image before it is fed into the network. In the XPN stage, the anchors whose overlaps with ground truth are greater than 0.6 are treated as positive samples, whereas the anchors whose overlaps with ground truth are less than 0.3 are treated as negative samples. The parameter of the focal loss is set to 0.25 and the soft NMS algorithm parameters are as follows: the threshold is 0.5 and δ is 1.

To compared with other methods, we split the datasets into training (80%), validation (10%), and test (10%) sets, such that the splits have similar class distributions but the unseen test set contains somewhat challenging samples never used for training.

B. Comparison Results on Dataset Xdb1

First, we train and evaluate our model using the Xdb1 training set and testing set, respectively. As comparisons in Table I, our model achieves promising results, i.e., 0.954 in terms of mAP, which outperforms most of the previous methods tested on this dataset by a large margin. As we can see, Faster R-CNN shows a similar performance with R-FCN. Region proposals based on

TABLE I
AP AND MAP ON THE TESTING SET OF THE XDB1 DATASET

model	mAP	scissors	bottle	mental cup	kitchen knife	knife	battery	scissors
Faster R-CNN (resnet50)	0.837	0.808	0.879	0.891	0.844	0.730	0.835	0.872
R-FCN(resnet50)	0.847	0.827	0.861	0.895	0.853	0.752	0.848	0.896
FPN(resnet50)	0.922	0.895	0.951	0.962	0.936	0.823	0.938	0.952
SSD(vgg)	0.823	0.788	0.873	0.866	0.801	0.714	0.827	0.893
YOLOv3(darknet53)	0.870	0.842	0.897	0.906	0.872	0.785	0.856	0.930
Ours	**0.954**	**0.938**	**0.981**	**0.989**	**0.963**	**0.880**	**0.951**	**0.974**

TABLE II
AP AND MAP ON THE TESTING SET OF THE XDB2 DATASET

model	mAP	bottles	mental cup	knives	scissors	gun	battery	laptop	umbrella	lighter	pressure cans
Faster R-CNN(resnet50)	0.706	0.808	0.809	0.528	0.614	0.710	0.683	0.818	0.781	0.557	0.749
R-FCN(resnet50)	0.706	0.817	0.801	0.525	0.623	0.722	0.688	0.815	0.762	0.551	0.752
FPN(resnet50)	0.797	0.860	0.858	0.677	0.690	0.834	0.766	0.927	0.853	0.669	0.831
SSD(vgg)	0.694	0.803	0.810	0.521	0.604	0.704	0.679	0.792	0.754	0.548	0.727
YOLOv3(darknet53)	0.713	0.826	0.822	0.532	0.612	0.731	0.696	0.811	0.795	0.552	0.755
Ours	**0.835**	**0.898**	**0.886**	**0.737**	**0.743**	**0.864**	**0.816**	**0.938**	**0.885**	**0.706**	**0.872**

them are collected from a single-scale feature map, leading to a loss in semantic and position information. FPN adopts the pyramid structure to deal with multiscale detection and fuse features from different levels. Therefore, it gets a higher mAP than that based on a single-scale feature map. For single-stage methods, SSD and YOLOv3 also generate RoIs from multiscale feature maps. However, large feature maps in SSD lead to a lack of semantic information, and small feature maps may lead to a lack of position information. Hence, SSD gets a lower performance than others. YOLOv3 takes advantage of the methods in FPN and gets a higher performance in terms of mAP than SSD. Compared to FPN, YOLOv3 falls in some categories because it does not have the proposal step. Single-stage methods still get a worse performance than two-stage methods. Table I shows that our method achieves impressive improvements in scissors and knife detection, which proves that our model can be more stable when dealing with densely cluttered backgrounds and small objects in X-ray detection. Fig. 7 shows some examples of the object detection results using the proposed method on the Xdb1 dataset.

To assess the robustness of our approach, we reduced the size of the training samples from 80 to 70% and 60%. Accordingly, the size of validation sets is all 10%, and the size of test sets is from 10 to 20% and 30%. Fig. 8 shows the mAP with different proportion on dataset Xdb1. As we can see, even with less training data and more test data, the proposed method also performs better than other representative methods.

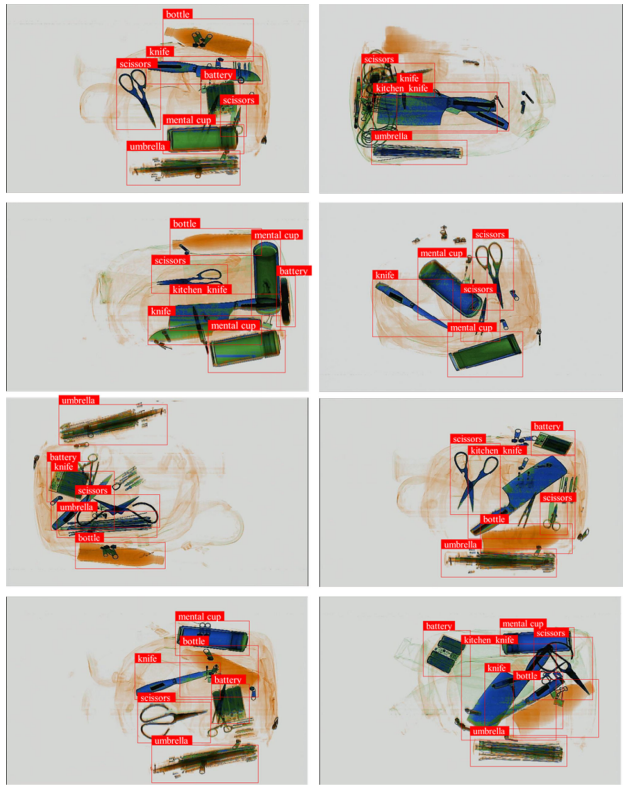


Fig. 7. Some qualitative detection results of our proposed method on the Xdb1 test dataset. Only detections with scores higher than 0.5 are shown.

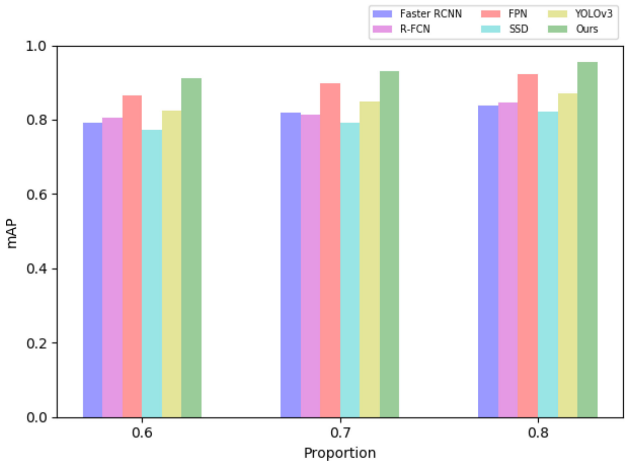


Fig. 8. Performance with different proportion on the dataset Xdb1.

C. Comparison Results on Dataset Xdb2

We also use the dataset Xdb2 training and testing sets to train and evaluate our method, respectively. As comparisons in Table II, our method achieves promising results, i.e., 0.835 in terms of mAP, which outperforms most of the tested methods on this dataset by a large margin. For two-stage methods, R-FCN shows a similar performance to Faster R-CNN. They use a single-scale feature map to generate region proposals. This may lead to a lack of semantic and localization information. FPN takes full use of the pyramid structure to deal with multiscale detection. The feature maps are fused from different levels. This structure contains more detail and semantic information. Therefore, FPN gets better performance than Faster R-CNN or R-FCN especially for knives, scissors, and lighter detections. For single-stage methods, SSD gets a worse performance than YOLOv3. Compared to Xdb1, Xdb2 is more complicated with densely cluttered backgrounds. SSD does not take enough advantage of different feature maps, which may lead to a lack of detail and semantic information. YOLOv3 also generate RoIs from multiscale feature maps. However, it gets a slight improvement when dealing with such a complicated background. Compared

Fig. 9. Some qualitative detection results of our proposed method on the Xdb2 test dataset. Only detections with scores higher than 0.5 are shown.

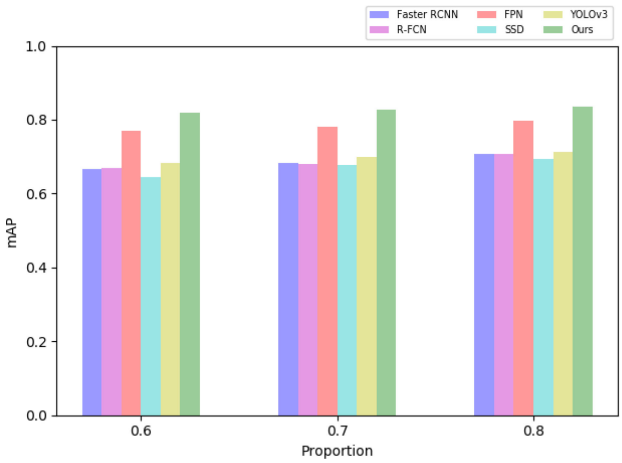


Fig. 10. Performance with different proportion on the dataset Xdb2.

to FPN, YOLOv3 still gets a worse performance, which may be due to the lack of region proposal step. Table II shows that the proposed method achieves impressive improvement especially in small size objects, such as knives and lighter. It proves that our method can be more stable when dealing with densely cluttered backgrounds and small objects during X-ray baggage inspection. Fig. 9 shows some qualitative detection results on the Xdb2 dataset.

To assess the robustness of our approach, we reduced the size of the training samples from 80 to 70% and 60%. Accordingly, the size of validation sets is all 10%, and the size of test sets is from 10 to 20% and 30%. Fig. 10 shows the mAP with different proportion on dataset Xdb2. As we can see, even with less training data and more test data, the proposed method also performs better than other methods.

TABLE III
AP AND MAP ON THE TESTING SET OF THE XDB1 DATASET

model	mAP	scissors	bottle	mental cup	kitchen knife	knife	battery	scissors
FPN	0.922	0.895	0.951	0.962	0.936	0.823	0.935	0.952
V1	0.936	0.919	0.964	0.973	0.950	0.845	0.941	0.963
V2	0.944	0.928	0.968	0.979	0.956	0.864	0.943	0.967
V3	0.945	0.929	0.970	0.980	0.956	0.866	0.944	0.967
V4	0.951	0.936	0.979	0.984	0.964	0.876	0.948	0.973
Ours	**0.954**	**0.938**	**0.981**	**0.989**	**0.963**	**0.880**	**0.951**	**0.974**

TABLE IV
AP AND MAP ON THE TESTING SET OF THE XDB2 DATASET

model	mAP	bottles	mental cup	knives	scissors	gun	battery	laptop	umbrella	lighter	pressure cans
FPN	0.797	0.860	0.858	0.677	0.690	0.834	0.766	0.927	0.853	0.669	0.831
V1	0.816	0.879	0.873	0.718	0.719	0.846	0.785	0.931	0.869	0.684	0.858
V2	0.827	0.887	0.879	0.727	0.735	0.857	0.805	0.935	0.877	0.699	0.864
V3	0.829	0.890	0.881	0.730	0.737	0.859	0.809	0.935	0.880	0.701	0.867
V4	0.833	0.895	0.885	0.735	0.743	0.862	0.813	0.938	0.883	0.705	0.870
Ours	**0.835**	**0.898**	**0.886**	**0.737**	**0.743**	**0.864**	**0.816**	**0.938**	**0.885**	**0.706**	**0.872**

TABLE V
ADDITIONAL EXPERIMENTS WITH DIFFERENT MODULES OF OUR PROPOSED METHOD

	Pyramid Structure	Data Augmentation	Feature Enhancement	Focal Loss	RoI Fusion	Soft NMS
FPN	✓	×	×	×	×	×
V1	✓	✓	×	×	×	×
V2	✓	✓	✓	×	×	×
V3	✓	✓	✓	✓	×	×
V4	✓	✓	✓	✓	✓	×
Ours	✓	✓	✓	✓	✓	✓

D. Analysis of the Proposed Modules

To examine the effectiveness and contributions of different modules that are used in the proposed method, we conduct an additional ablation experiment for studies listed in Tables IV and Tables V. We mainly analyze the data augmentation module, feature enhancement module, focal loss module and soft NMS module. They are all listed in Tables III where FPN represents our modified FPN, V1 represents the basic FPN with our data augmentation module, V2 represents V1 with our feature enhancement module, V3 represents V2 with our focal loss module, V4 represents V3 with our RoI fusion module, and ours represents V4 with soft NMS module. For Xdb1 dataset, the data augmentation module's improvement is 1.5% more than our modified FPN, feature enhancement module's improvement is 0.8% more than V1, focal loss module's improvement is 0.1% more than V2, RoI module's improvement is 0.6% more than V3, soft NMS module's improvement is 0.3% more than V4, and our proposed method's improvement is 3.4% more than our modified FPN. For Xdb2 dataset, data augmentation module's improvement is 2.3% more than our modified FPN, feature enhancement module gets an improvement of 1.3% more than V1, focal loss module's improvement is 0.2% more than V2, RoI module's improvement is 0.6% more than V3, soft NMS module gets an improvement of 0.3% more than V4, and our proposed method improves by 4.7% more than our modified FPN. The results of this comparison clearly reveal the advantages of our

Fig. 11. Miss detection situation on Xdb1 dataset. (Left) miss alarm and (Right) false alarm.

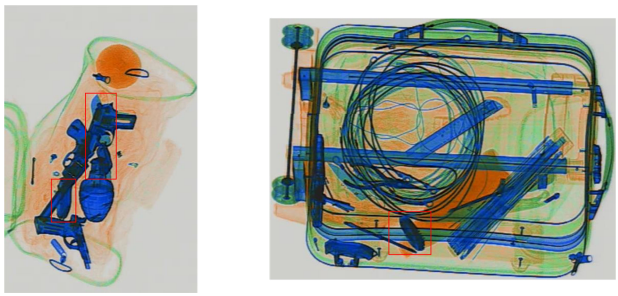


Fig. 12. Miss detection situation on Xdb2 dataset. (Left) miss alarm and (Right) false alarm.

method. The data augmentation module improves the most in mAP. It can generate diverse training data, which can make the detector more robust when it is fed with new data. Feature enhancement module is also effective, thanks to combining enhanced multiscale feature maps that can be helpful for small objects, such as knives. The focal loss module and soft NMS module slightly improve, which is still useful for detection. RoI module is also effective by fusing all feature maps to generate more robust results.

E. False Alarm and Miss Alarm

Although the proposed algorithm outperforms the relevant methods on object detection during X-ray baggage inspection, there are still some targets that are missed or misreported. This section will briefly analyze these situations. Test results show that most errors occur in situations like the ones shown in Figs. 11 and 12. Due to the impact of objects' cluttering, some objects are misreported during testing. The left parts of Figs. 11 and 12 show some samples that are missed marked in red rectangles. The misreported object may heavily suffer from other objects or similar objects. Due to the impact of different views, some objects have a similar shape and appearance of other objects. These reasons make it difficult for the model to correctly distinguish the targets. The right parts of Figs. 11 and 12 show some samples that are missed marked in red rectangles. In the right part of Fig. 11, the kitchen knife is recognized as a dagger. Due to different views, this sample has a similar shape of a dagger. The right part of Fig. 12 shows that the knives were recognized as a lighter. Due to the small size, they may have the same shape in this view. This situation may be alleviated by multiview detection.

In order to further show the effect of the proposed method, we also validate the number of baggage without target objects detected as containing any of target objects. For the dataset of Xdb1, 0.4% of baggage without target objects is detected as containing any of target objects. For the dataset of Xdb2, 0.2% of baggage without target objects is detected as containing any of target objects. The results show that the proposed method yields a very low percentage as the baggage without target objects is detected as containing any of target object. This proves that our method can be used in practice.

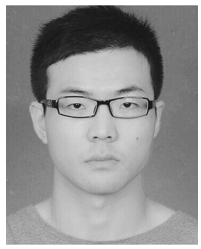
V. Conclusion

In this article, an effective approach was proposed to build a deep object detector and train it from scratch for X-ray image inspection. The novelties that distinguish the proposed work from previous works lied in two major aspects. First, instead of fine-tuning using ImageNet pretrained models, our method trained the deep detector from scratch, this provided the freedom to adjust or redesign the structures. Second, in order to improve the detection performance for clustered objects, we adopted focal loss to address the foreground–background imbalance and predicted multiscale object proposals from several enhanced intermediate layers to improve the accuracy. The proposed regions were scaled using RoI align, followed by element-level fusion and soft NMS postprocessing. The quantitative comparison results on the Xdb1 and Xdb2 datasets showed that the proposed method achieved better performance than comparative methods and it was more effective than existing algorithms for detecting small and densely cluttered X-ray objects. However, as stated earlier, our method still produced some false alarms and omissions in some severe situations. Hence, in our future studies, we will focus on discriminating the false alarms and learning the structure of the network adaptively. In addition, we will improve the transferability of our model using domain adaptation methods.

References

[1] G. Zentai, "X-ray imaging for homeland security," in *Proc. IEEE Int. Workshop Imag. Syst. Techn.*, 2008, pp. 1–6.
[2] E. Parliament, "Aviation security with a special focus on security scanners," in *Proc. Eur. Parliament Resolut. (INI)*, Oct. 2012, pp. 1–10.
[3] A. Schwaninger, A. Bolfing, T. Halbherr, S. Helman, A. Belyavin, and L. Hay, "The impact of image based factors and training on threat detection performance in X-ray screening," in *Proc. Int. Conf. Res. Air Transp.*, 2008, pp. 317–324.
[4] G. Blalock, V. Kadiyali, and D. H. Simon, "The impact of post-9/11 airport security measures on the demand for air travel," *J. Law Econ.*, vol. 50, no. 4, pp. 731–755, 2007.
[5] S. Michel, S. M. Koller, J. C. de Ruiter, R. Moerland, M. Hogervorst, and A. Schwaninger, "Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners," in *Proc. 41st Annu. IEEE Int. Carnahan Conf. Secur. Technol.*, 2007, pp. 201–206.
[6] Z. Chen, Y. Zheng, B. R. Abidi, D. L. Page, and M. A. Abidi, "A combinational approach to the fusion, de-noising and enhancement of dual-energy X-ray luggage images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Sep. 2005, p. 2.
[7] D. Mery, *Computer Vision for X-Ray Testing*. Cham, Switzerland: Springer, 2015.
[8] V. Rebuffel and J.-M. Dinten, "Dual-energy X-ray imaging: Benefits and limits," *Insight-Non-Destruct. Testing Condition Monit.*, vol. 49, no. 10, pp. 589–594, 2007.
[9] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern computer vision techniques for X-ray testing in baggage inspection," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2016.

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.
[11] Y. Zhu and S. Newsam, "DenseNet for dense flow," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 790–794.
[12] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron.*, vol. 147, pp. 70–90, 2018.
[13] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[14] M. Baştan, "Multi-view object detection in dual-energy X-ray images," *Mach. Vis. Appl.*, vol. 26, no. 7/8, pp. 1045–1060, 2015.
[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[17] V. Riffo and D. Mery, "Automated detection of threat objects using adapted implicit shape model," *IEEE Trans. Syst., Man, Cybern.: Syst.*, vol. 46, no. 4, pp. 472–482, Apr. 2015.
[18] D. Mery, E. Svec, and M. Arias, "Object recognition in baggage inspection using adaptive sparse representations of X-ray images," in *Image and Video Technology*. Berlin, Germany: Springer, 2015, pp. 709–720.
[19] J. Liu *et al.*, "3D feature constrained reconstruction for low-dose CT imaging," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1232–1247, May 2018.
[20] J. Liu *et al.*, "Discriminative feature representation to improve projection data inconsistency for low dose CT imaging," *IEEE Trans. Med. Imag.*, vol. 36, no. 12, pp. 2499–2509, Dec. 2017.
[21] M. Baştan, M. R. Yousefi, and T. M. Breuel, "Visual words on baggage X-ray images," in *Proc. Int. Conf. Comput. Anal. Images Patterns*, 2011, pp. 360–368.
[22] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer, "Automated X-ray object recognition using an efficient search algorithm in multiple views," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2013, pp. 368–374.
[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.
[24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
[25] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Visi. Pattern Recognit.*, 2016, pp. 779–788.
[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[29] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[30] S. Akçay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *Proc. IEEE Int. Conf. Image Process.*, 2016, pp. 1057–1061.
[31] S. Akcay, M. E. Kundegorski, C. G. Willcocks, and T. P. Breckon, "Using deep convolutional neural network architectures for object classification and detection within X-ray baggage security imagery," *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 9, pp. 2203–2215, Sep. 2018.
[32] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1451–1460.
[33] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 472–480.
[34] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[35] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[36] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[38] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5561–5569.
[39] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.
[40] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.



Bangzhong Gu received the M.E. degree in computer science and technology from Southeast University, Nanjing, China, in 2015. He is currently working toward the Ph.D. degree in computer science and technology, Southeast University, Nanjing, China.

His current research interests include image processing and machine learning.

Rongjun Ge received the M.E. degree in computer science and technology from Lanzhou University, Lanzhou, China, in 2016. He is currently working toward the Ph.D. degree in computer science and technology, Southeast University, Nanjing, China.

His current research interests include image encryption, chaos, medical image processing, and machine learning.

Yang Chen (Senior Member, IEEE) received the M.E. and Ph.D. degrees in biomedical engineering from the First Military Medical University, Guangzhou, China, in 2004 and 2007, respectively.

Since 2018, he has been a Professor with the Laboratory of Image Science and Technology, School of Computer Science and Engineering, Southeast University, Nanjing, China. His research interests include medical image reconstruction, image analysis, pattern recognition, and computerized-aid diagnosis.

Limin Luo (Senior Member, IEEE) received the Ph.D. degree in information processing from the University of Rennes, Rennes, France, in 1986.

He is currently a Professor with the Laboratory of Image Science and Technology, School of Computer Science and Engineering, Southeast University, Nanjing, China. His current research interests include medical imaging, image analysis, computer-assisted systems for diagnosis and therapy in medicine, and computer vision.

Gouenou Coatrieux (Senior Member, IEEE) received the Ph.D. degree in signal processing and telecommunication from the University of Rennes1, Rennes, France, in collaboration with Ecole Nationale Supérieure des Télécommunications, Paris, France, in 2002.

He is currently a Professor with the Information and Image Processing Department, Institut Mines-Télécom, Telecom Bretagne, Brest. His research is conducted in the LaTIM Laboratory, INSERM U1101, Brest, France. His research interests include data security, encryption, watermarking, secure processing of outsourced data, digital forensics in medical imaging, and electronic patient records.