Applied Intelligence (2022) 52:4807–4823
https://doi.org/10.1007/s10489-021-02582-1

Research on the application of high-efficiency detectors into the detection of prohibited item in X-ray images

Yuanxi Wei[1] · **Xiaoping Liu**[1] · **Yinan Liu**[2]

Accepted: 30 May 2021 / Published online: 28 July 2021

Abstract

X-ray imaging can be used to inspect the internal structure of the objects without destruction, so visual inspection based on X-ray images is widely used in the security check such as customs, airports, railway stations, and postal express. Especially in the postal express industry, fast and accurate inspection of express parcels can effectively improve logistics efficiency. This article studies the application of computer vision technology to detect prohibited items in X-ray images. Due to the multi-pose objects in the packages under multi-views, it is difficult to find out the prohibited item from the packages under a single view. This article explores how to solve this problem with the loss function of classification and the attention mechanism of convolutional neural network, and apply them to high-efficiency detectors. On the one hand, we proposed a new loss function named truncated loss for X-ray image classification task. In the proposed loss, we truncated input vector of loss layer to reduce the difference within the intra-classes and increase the difference between the inter-classes. On the other hand, we proposed two new architectures for the high-efficiency detectors for the purpose of obtaining the visual features of prohibited item more effectively. One of the new architectures named channel context block (CC block), and it is based on global context (GC block). It contains global context information on each channel through operations of global average pooling, which is different from global context (GC) block. The other one of the architectures named GCC block, it is formed by merging channel context block (CC block) and global context (GC) block, and it is used to further improve the detection accuracy of prohibited item. The results of experiments on the currently widely used high-efficiency detectors in GDXray dataset show that our proposed truncated loss can improve the detection accuracy of prohibited item to a certain extent, and the new architectures can improve detection accuracy to a greater extent. The algorithms proposed in this article are also state-of-the-art on GDXray dataset.

Keywords X-ray images · Prohibited item detection · Truncated loss · Self-attention mechanism · GCNet

1 Introduction

In recent years, with the rapid development of e-commerce and express delivery industry, mailing express parcels has become a part of our daily life. Express parcels are relatively closed, and some prohibited item can be easily hidden within them, which also brings some security risks that cannot be ignored. Manual inspection with X-ray machine is currently the most popular method of security inspection of prohibited item.

However, manual inspection is relatively inefficient. In operation, most of the parcels are normal (excluding prohibited item), and only a very small part of the parcels contain prohibited item, so inspectors need to inspect the normal parcels with circumspection for a long time. The efficiency of inspecting prohibited item from thousands of parcels with manual inspection is very low, and at the same time it also brings psychological and physical burden to inspectors.

In recent years, artificial intelligence technology based on neural networks has been widely used in the field of computer vision technologies, such as face recognition [1–3], object detection [4–7], medical imaging [8]. This technology has also been widely used in the field of automated control [9–12]. The inspection method based on computer vision can work for a long time without rest, so it can solve the efficiency problem caused by manual inspection.

However, it is relatively difficult and complicated for computer to identify prohibited item in X-ray images of

✉ Xiaoping Liu
liuxp@bupt.edu.cn

Extended author information available on the last page of the article.

airtight parcels. In addition to the lack of rich color (it's the characteristic of X-ray image), the following reasons are also very important [13, 14]. The object in the X-ray image presents different scales, angles of view and morphologies. The same object in a parcel looks different from different angles of view. This might cause the shapes of the same object to vary greatly, and also might cause the shapes of different objects to have looked very similar. As shown in Fig. 1, both the razor blade in (c) and the shuriken in (b) have an elongated shape, which is visually similar. But the razor blade and shuriken in (a) and (d) are quite different in shape. At the same time, due to the different visual angles, the shuriken in (b) also has a big difference in shape. The problem of multi-pose based on multi-view is an important issue that restricts the accuracy of recognition.

The 3D image recognition method based on multiple viewing angles can overcome the multi-pose problem of the same prohibited item in the X-ray image [15, 16]. However, in practice, compared with 2D image recognition technology, such method currently requires more additional computing resources (for example, the reconstruction process from 2D to 3D) and equipment (used to generate 3D data). For example, [17] uses a method of reconstructing 3D objects from its 2D silhouettes to improve the accuracy of threat object recognition; [18] uses cluttered dual-energy Computed Tomography (CT) data to achieve 3D segmentation of objects in the baggage, so as to achieve the purpose of identifying prohibited item.

At the same time, 2D images can be used in a deep neural network under single view conditions to realize end-to-end identification or detection of prohibited item without other auxiliary work. Therefore, in recent years, applied research in this direction has also achieved rapid development. [13] proposes a method to identify prohibited item in X-ray images based on deep learning. The method uses convolutional neural networks such as AlexNet [19] and GoogleNet [20] to extract the deep features of X-ray images. In addition, this method uses simple nearest neighbor (KNN) classifier to replace the fully connected layer, so as to avoid overfitting. [21] trains the SSD [4] object detection model to detect prohibited item in X-ray images with the method of transfer learning. Aiming at the complex scene of overlapping objects in X-ray images, [14] proposes a class-balanced hierarchical refinement (CHR) based on deep neural network. This method cleverly uses class balanced loss [8] technology to effectively reduce the training data of negative class, and improve the accuracy of the experiment through the class-balanced hierarchical refinement (CHR).

On the basis of the above, this article applies the currently widely used high-efficiency detectors such as Mask R-CNN [22], PANet [23], YOLOv3 [7] and Guassian YOLOv3 [24] into the detection of prohibited item based on 2D X-ray images. The experiments on GDXray dataset [25] have proved that our proposed algorithms achieve the detection accuracy of state-of-the-art on these detectors. The AP of Mask R-CNN is increased by 1.6%, the AP of PANet is increased by 2.1%, the AP of YOLOv3-416 is increased by 1.9%, and the AP of Guassian YOLOv3-416 is increased by 2.2%. The contributions of this article mainly include the following two aspects:

– This article proposes a new loss function, truncated loss. Taking softmax loss as an example, this article proves that truncated loss can shorten the cosine distance of the object relative to the label vector by a certain multiple. Experiments show that when the truncated loss is applied into the detectors, the accuracy of the detectors in detecting prohibited item in the X-ray image is effectively improved.

Fig. 1 Various poses of objects in X-ray images

– This article proposes two new architectures of self-attention mechanism: CC block and GCC block. Especially for GCC block, compared to GC block [26], the ablation experiments has proved that it can obtain higher detection accuracy at the cost of less consumption of computational resource.

Remark 1 This article explores an idea to express the problem of multiple postures of objects in 2D X-ray images as intra-class difference and inter-class difference of object features. Therefore, the problem of multi-pose is transformed into the problem of reducing the intra-class difference and increasing the inter-class difference of prohibited item in the X-ray image. For this reason, a new truncated loss is proposed.

Remark 2 On the design of the loss function for the detector, the improvement of the loss functions for classification is mainly to solve the problem of sample imbalance in object detection [27–30]. This article studies from a new aspect: based on the characteristics of the multi-pose objects in 2D X-ray images, the truncated loss is designed for the detector to improve the recognition accuracy.

2 Related work

2.1 Prohibited item detection

2.1.1 Architecture of the detector

In practice, we not only need to identify the prohibited item but also need to mark the location coordinates of the prohibited item in the images. This can be achieved by using object detection technology. At present, the object detection technology based on deep convolutional neural network (DCNN) has achieved rapid development. In summary, its network structure is divided into four parts. For the convenience of explanation, in this article we only list three parts: input(X-ray image), backbone and prediction. Among them, backbone is used to calculate the visual features of X-ray images, and prediction is combination of the locations prediction and the classes prediction.

As shown in Fig. 2, the branch of object locations prediction and the branch of object classes prediction share the backbone network to achieve multi-task prediction of deep neural networks. In the branch of object classes prediction, F_{cls} is expressed as calculation function of classes probability, which is used to predict the probability that different anchors of the input feature map belong to different classes. This function is generally implemented through the using of softmax [4–6], and sometimes the sigmoid is also used to achieve multi-label detection or

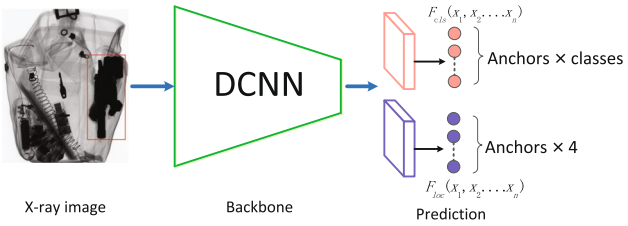

Fig. 2 Basic architecture of object detector based on deep convolutional neural network

calculation stability [7, 30]. In the branch of object locations prediction, F_{loc} is expressed as a position prediction function, which is used to predict the center coordinates (x, y) of the object and the width and height (w, h) of the rectangular box. Some detectors [6, 7] combine these two branches to improve the efficiency of object detection.

2.1.2 High-efficiency detectors

Faster R-CNN [5], a classic two-stage detector, realizes end-to-end object detection based on deep convolutional neural network (DCNN) through Region Proposal Networks (RPNs), which improves detection efficiency. Mask R-CNN [22] is based on Faster R-CNN and replaces the RoIPool layer with RoIAlign, which largely solves the misalignment problem caused by direct sampling only through pooling, and improves the accuracy of detection. The PANet [23] based on the Mask R-CNN further optimizes the detector, and introduces technologies such as bottom-up path augmentation (BPA) and adaptive feature pooling (AFP) to make the extracted features richer and more efficient. In the experiment, the Mask Branch in Mask R-CNN and PANet are not used, and they are only used as detectors to detect prohibited item in X-ray images.

Compared with the two-stage detectors, the detection speed of the one-stage YOLO [6] detectors has been greatly improved, but their positioning accuracy is not high, especially for some small objects. As the classic detector in the YOLO series, YOLOv3 [7] has made great improvements to this series. For example, the backbone architecture uses a more effective residual network and uses multi-scale feature fusion techniques, which greatly improves the detection efficiency. Guassian YOLOv3 [24] uses the characteristics of gaussian distribution to improve YOLOv3, so that the network can output the uncertainty of each detection box, thereby improving the detection efficiency.

2.2 Loss function

On the premise of not increasing the depth of the neural network model, the accuracy of recognition is improved by

optimizing the loss function, thereby achieving the purpose of improving the efficiency of the deep network model. At present, there are two main strategies for optimizing deep network models, namely deep metric learning and softmax loss.

Among which, the basic idea of metric learning is to learn the mapping from the original features to the low-dimensional dense embedding space, so that the input objects of same class are mapped in the embedding space with a short distance (small differences within the class), and the input objects of different classes are mapped in the embedding space with a long distance (large difference between the classes) [31, 32]. Many recent deep metric learning approaches are built on pairs of samples. Contrastive loss [33] is to constrain the similarity between two pairs of samples. This constraint only considers the similarity of the input samples relative to itself. Triplet loss [34, 35] improves Contrastive loss [33]. The input of Triplet loss consists of a triplet and each triplet has an anchor, a positive and a negative example, where the anchor and the positive have the same class labels and the negative has the different class label. Triplet loss not only considers the distance between the positive sample and the anchor point, but also considers the distance between the negative sample and the anchor point. The loss function of metric learning plays an important role in various computer vision applications.

The other type of optimization methods are based on the softmax, and these strategies are used to adjust the softmax loss so as to achieve the purpose of making inter-class difference increase and making intra-class difference decrease [1–3, 36]. These kind of strategies enlarge the angle distance of output feature vectors between different classes by adjusting the $\angle\theta$ between the input vector $\mathbf{x}$ and the weight $\mathbf{w}$.

2.3 Self-attention mechanism

In order to enable backbone to obtain the visual features of prohibited items in X-ray images more effectively, we have added a self-attention mechanism [37] to the detector. This mechanism can reduce dependence on external information, while capturing the internal correlation of features.

NLNet [38] cleverly uses the self-attention mechanism to calculate the dependence of long-range visual field. Researchers have applied this method to datasets such as Kinetics dataset [39] and COCO dataset [40] for video classification, object detection and semantic segmentation tasks. Experiments show that the method performs well in the above tasks. CCNet [41] improves NLNet [38] with two criss-cross blocks and applies it into semantic segmentation.

Compared with SENet [42], NLNet and CCNet learn query-independent attention maps for each position on the feature map, which consumes relatively little computing resources. However, for complex visual tasks, it is difficult for SENet to obtain the global context of the channel only through rescaling, and it is difficult to obtain an attention feature map with high efficiency. On these foundations, GCNet [26] combines the respective advantages of SENet and NLNet to propose the global context (GC) block, as shown in Fig. 3.

GCNet [26] uses 1x1 conv to reduce the number of channels of the branch K to 1, so the size of feature maps is reduced from [C × H × W] to [C × 1 × 1], formulated as

$$z_i = x_i + \delta\left(\sum_{j=1}^{N_P} a_j \cdot x_j\right), \tag{1}$$

where $a_j = \frac{e^{W_k x_j}}{\sum_m e^{W_k x_m}}$ is the weight for global attention pooling, and $\delta(.) = W_{v2} ReLU(LN(W_{v1}(.)))$ denotes the bottleneck transform.

3 Proposed algorithms

3.1 Truncated loss

In order to improve the accuracy of the model for identifying prohibited item, this article adopts the method of adjusting the classification loss to improve the robustness of the

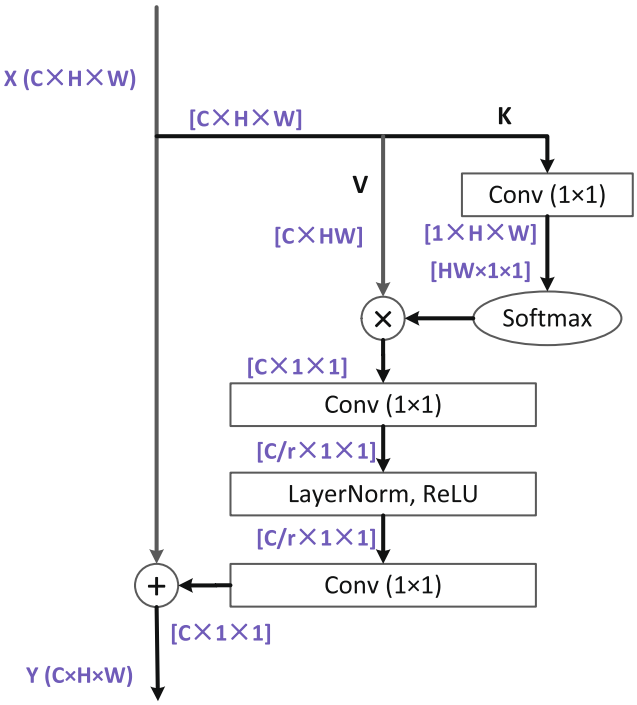

Fig. 3 Framework of global context (GC) block in GCNet [26]

recognition model and make it effectively converge during the training process.

3.1.1 Definition

We truncate the input vector of the classification loss function with a certain coefficient β. Assuming the classification loss function is $l_{cls}(\mathbf{x}, \mathbf{y})$, its truncated loss function can be defined as

$$l_t = l_{cls}(\mathbf{x} - (\beta|\mathbf{x}| \odot \mathbf{y}), \mathbf{y}). \tag{2}$$

In the above formula, $\mathbf{x}$ and $\mathbf{y}$ are the input vector and one-hot label vector, and $\odot$ is the corresponding element of the vector multiplied. For instance, truncated softmax can be defined as

$$l_{ts} = -\sum_{i=1}^{n} y_i \cdot log \frac{e^{x_i - \beta|x_i| \cdot y_i}}{\sum_{k=1}^{n} e^{x_k - \beta|x_k| \cdot y_k}}, \tag{3}$$

where x_i and y_i are the i-th element values of the $\mathbf{x}$ and $\mathbf{y}$, and $\beta|x_i|$ is the truncated value.

It can be found from the l_{ts} that in the case of positive samples, the value of l_{ts} increases relative to softmax loss. This shows that using truncated loss is equivalent to adding artificial noise disturbance to the loss layer and changing the original noise distribution.

Remark 3 In the training process, β is a hyper-parameter of the model, which needs to be set in advance. At the same time, $\beta|x_i|$ does not participate in the gradient update during each iteration, and only affects the loss value as a fixed noise. If the β is too large, the model will be difficult to converge due to excessive noise, and this will cause the training unstable. In order to achieve the desired effect, the Section 3.1.3 is used to qualitatively analyze the influence of β on the gradient during the training process, and in the Section 5.1, the experiment is used to make further verification and analysis, so as to provide method support for finding a suitable β setting.

3.1.2 Margin of truncated loss

Taking softmax loss as an example, cosine distance is used to analyze the margin value between inter-class and intra-class. Suppose $\mathbf{x}_1$ is the input vector of softmax of classes C_1, and $\mathbf{x}_{s1}$, $\mathbf{x}_{ts1}$ are the output vectors of softmax and truncated softmax of classes C_1. $\mathbf{y}_1$ and $\mathbf{y}_2$ are one-hot label vectors of classes C_1 and C_2. $\alpha_1 = \langle \mathbf{x}_{ts1}, \mathbf{y}_1 \rangle$, $\alpha_2 = \langle \mathbf{x}_{s1}, \mathbf{y}_1 \rangle$, $\alpha = \langle \mathbf{x}_{s1}, \mathbf{y}_2 \rangle$ are setup. Truncated softmax loss and softmax loss can make the inequalities of $cos(\alpha_1) \geq cos(\alpha)$ and $cos(\alpha_2) \geq cos(\alpha)$ true. We set $\mathbf{o}$ as a vector whose elements are all 1, $|\cdot|$ is the absolute value of the vector element and $\|\cdot\|$ is the module of the vector. For $\mathbf{x}_1$, the output of softmax is $\mathbf{x_{s1}} = \frac{e^{\mathbf{x}_1}}{e^{\mathbf{x}_1} \cdot \mathbf{o}^\top}$, and the output of truncated softmax is

$$\begin{aligned} \mathbf{x}_{ts1} = s(\mathbf{x}_1 - \beta|\mathbf{x}_1| \odot \mathbf{y}_1) &= \frac{e^{\mathbf{x}_1 - \beta|\mathbf{x}_1| \odot \mathbf{y}_1}}{(e^{\mathbf{x}_1 - \beta|\mathbf{x}_1| \odot \mathbf{y}_1}) \cdot \mathbf{o}^\top} \\ &= \frac{e^{-\beta|\mathbf{x}_1| \odot \mathbf{y}_1} \odot e^{\mathbf{x}_1}}{\theta(e^{\mathbf{x}_1} \cdot \mathbf{o}^\top)} \\ &= \frac{e^{-\beta|\mathbf{x}_1| \odot \mathbf{y}_1}}{\theta} \odot \mathbf{x_{s1}}, \end{aligned} \tag{4}$$

where θ satisfies the relationship of $0 < \theta \leq 1$. So there are $cos(\alpha_2) = \frac{\mathbf{x_{s1}} \cdot \mathbf{y}_1}{\|\mathbf{x_{s1}}\|}$ and $cos(\alpha_1) = \frac{e^{-\beta|x_1|}(\vec{x_{s1}} \cdot \vec{y_1})}{\theta\|\vec{x_{s1}}\|}$. $|x_1|$ is set to the absolute value of the $\mathbf{x}_1$ element, which is in accordance with the element of 1 in $\mathbf{y}_1$. From this, the relationship of

$$cos(\alpha_1) = \frac{e^{-\beta|x_1|}}{\eta} cos(\alpha_2) \tag{5}$$

can be calculated, where $e\beta|x| \leq \eta \leq 1$, where $e^{-\beta|x_1|} \leq \eta \leq 1$. When $\beta \geq 0$, $\frac{e^{-\beta|x_1|}}{\eta} \leq 1$, it can be seen that $cos(\alpha_1)$ has a margin relative to $cos(\alpha)$, as shown in Fig. 4.

At this time, since the objects of the same category are aggregated to the vectors of one-hot label, this margin can reduce the difference within the intra-classes and increase the difference between the inter-classes.

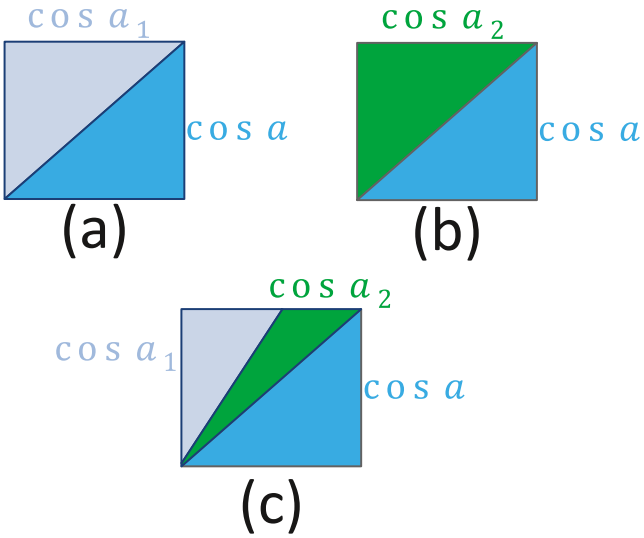

Fig. 4 (a) is represented as $cos(\alpha_1) \geq cos(\alpha)$ and (b) is represented as $cos(\alpha_2) \geq cos(\alpha)$. (c) indicates that the truncated softmax generates a margin of $(1 - \frac{e^{-\beta|x_1|}}{\eta})cos(\alpha_2)$ relative to $cos(\alpha)$, which is closer to the one-hot label

3.1.3 Gradient analysis

In order to study the effectiveness of this method, Softmax loss is taken as an example for calculation. The back-propagated gradient value of the softmax loss function is calculated to analyze the training process of the model. In the back-propagation process, the gradient of $\beta|x_i|$ is not calculated, and it is treated as a truncated variable.

Let $t_i = \frac{e^{x_i-\beta|x_i|\cdot y_i}}{\sum_{k=1}^{n} e^{x_k-\beta|x_k|\cdot y_k}}$, then $l_{ts} = -\sum_{i=1}^{n} y_i \cdot log\, t_i$. Next, the partial derivative of x_j is calculated in two cases, $i = j$ and $i \neq j$, respectively. if $i = j$, $\frac{\partial t_i}{\partial x_j} = t_i - t_i^2$, else $i \neq j$, $\frac{\partial t_i}{\partial x_j} = -t_i \cdot t_j$.

Finally, these are used to calculate the gradient of l_t, as shown in

$$\begin{aligned}\frac{\partial l_t}{\partial x_j} &= -\sum_{i=1}^{n} y_i \cdot \frac{1}{t_i} \cdot \frac{\partial t_i}{\partial x_j} \\ &= \begin{cases} \frac{e^{x_i-\beta|x_i|}}{e^{x_i-\beta|x_i|}+\sum_{k\neq i}^{n} e^{x_k}} - 1 & y_i = 1 \\ \frac{e^{x_i}}{e^{x_m-\beta|x_m|}+\sum_{k\neq m}^{n} e^{x_k}} & y_i = 0,\ y_i = 0 \end{cases}. \end{aligned} \tag{6}$$

It can be seen from the calculation result of $\frac{\partial l_t}{\partial x_j}$ that when $y_i = 1$, the gradient of l_t is smaller than the gradient of softmax loss $(\frac{e^{x_i}}{\sum_{j=k}^{n} e^{x_k}} - 1)$[43], and when $y_i = 0$, the gradient of l_{ts} is larger than the gradient of softmax loss $(\frac{e^{x_i}}{\sum_{j=k}^{n} e^{x_k}})$[43].

In the direction of $y_i = 1$ (positive class), truncated loss with softmax loss is relatively smoother than softmax loss. In the direction of $y_i = 0$ (negative classes), the gradient of truncated loss with softmax is larger than that of softmax loss. Therefore, in the training process, the increased gradient can make the model easier to converge.

However, when β is too large, it will have a negative impact on model training. Firstly, in the positive class direction, an excessively small gradient makes it difficult for the model to jump out of the local optimal solution; secondly, in the negative classes direction, an excessively large gradient increases the sensitivity of the learning rate, easily enhanced oscillations, and is not conducive to model convergence.

3.2 Design of self-attention mechanism

CC block This article further design the self-attention strategy based on GCNet [26], as shown in Fig. 5a. For the convenience of representation, this new architecture channel is named channel context block (CC block).

We first perform global average pooling (GAP) on the branch V, and change the feature map of size from C × H × W to C × 1. This feature map is matrix multiplied with the output of branch K. Finally, the feature map is input to a bottleneck transform, and its output is broadcast elementwise added to the X. The formula is

$$z_i = x_i + \epsilon\left(\frac{\sum_i (ga_i)}{N_p}\right), \tag{7}$$

where N_p is the number of positions in the feature map, and $g = \frac{\sum_i x_i}{N_p}$ is the output for global average pooling, and $a_i = \frac{e^{W_k x_i}}{\sum_m e^{W_k x_m}}$ is the output for global attention pooling, and $\epsilon(.) = Wv2ReLU(LN(Wv1g(.)))$ denotes the bottleneck transform in the channel context block.

It can be seen from z_i that compared with GC block [26], the CC block removes the accumulation operation, makes the size of the output of matrix multiplied to be C × H × W, so that CC block can obtain richer global attention information for each channel. In order to reduce the computational complexity of the bottleneck transform, a global average pooling is performed on the output of matrix multiplication. Like GC block [26], in the bottleneck of the CC block, LayerNorm is used to normalize the feature map.

GCC block In order to obtain richer visual features of prohibited item in X-ray image, we integrated GC block [26] and CC block. As shown in the Fig. 5b, the output feature map is added on channel context block to X, and enter the feature map accumulated by pixel into GC block, thus achieving the purpose of feature fusion. For the convenience of representation, this integrated architecture is named GCC block.

Set

$$y_j = x_j + ga_j \tag{8}$$

to integrate the output feature map of CC block at the input side of GC block. The output feature map of GCC block is shown in

$$z_i = x_i + \delta(\sum_{j=1}^{N_P} b_j y_j), \tag{9}$$

where, $a_j = \frac{e^{W_k x_j}}{\sum_m e^{W_k x_m}}$ is the output for global attention pooling of x_j and $b_j = \frac{e^{W_k y_j}}{\sum_m e^{W_k y_m}}$ is the output for global attention pooling of y_j.

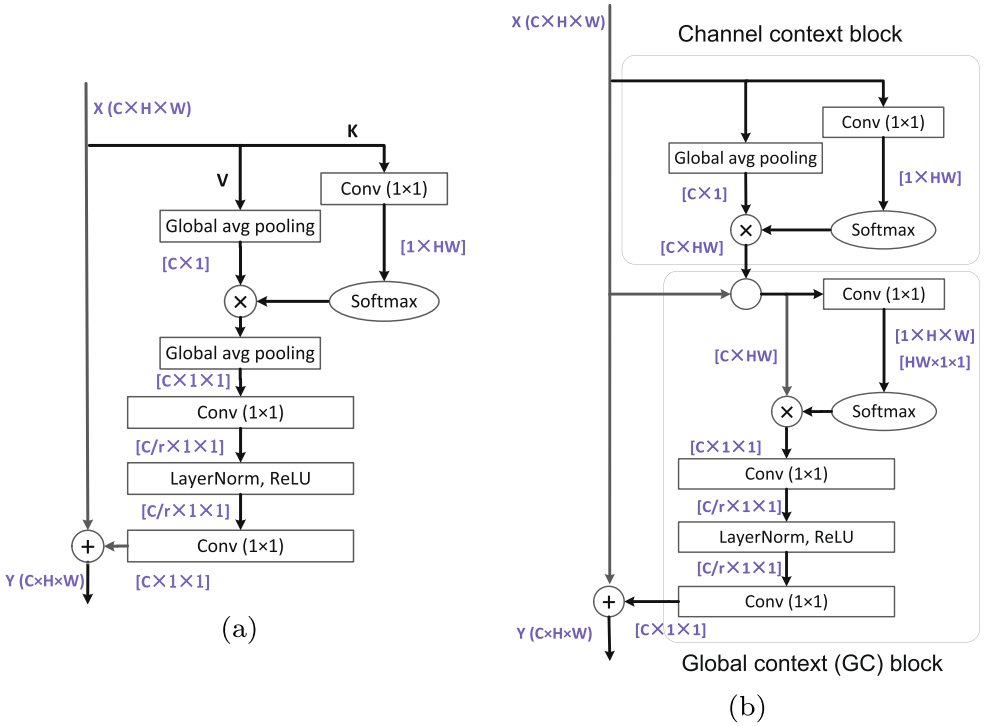

Fig. 5 Architecture of the context blocks. (a) is the architecture of channel context block (CC block) with bottleneck transform. (b) is the architecture that integrates channel context block into GC block (GCC block). ⊗ denotes matrix multiplication, ⊕ denotes broadcast elementwise addition, and ◯ denotes the addition of corresponding pixels. The feature maps are shown as feature dimensions, e.g. [CxHxW] denotes a feature map with channel number C, height H and width W

3.3 Pipeline of proposed algorithms

Truncated loss in detector In the training process, the loss function of the detector is used, denoted by $loss(\mathbf{x}, \mathbf{y}) = L_{cls}(\mathbf{x}, \mathbf{y}) + L_{loc}(\mathbf{x}, \mathbf{y})$. The classification loss value is truncated, as shown in $TL_{cls} = L_{cls}((\mathbf{x} - \beta(|\mathbf{x}| \cdot \mathbf{y})), \mathbf{y})$. For training stability, we perform logarithmic operations on TL_{cls}, as shown in $LTL = log(TL_{cls})$, the final loss is $loss = LTL_{cls} + L_{loc}$. In the above formulas, $\mathbf{x}$ and $\mathbf{y}$ represent the input feature vector and the label vector in one-hot encoding, respectively, and β is the truncated coefficient.

Multi-task loss In the practice of multi-task-based prohibited item detection, the loss function includes tasks of object classification and object location. Truncated loss is used for object classification in the experiment, which increased the loss value of object classification and destroyed the original multi-task loss value distribution.

Therefore, we introduce the multi-task homoscedastic uncertainty [44] in the loss function to balance the multi-loss. As shown in

$$MLTL = \frac{1}{\delta_1^2} LTL_{cls} + \frac{1}{\delta_2^2} L_{loc} + 2\log(\delta_1 \cdot \delta_2), \tag{10}$$

we use the learned δ_1 and δ_2 as balance parameters in the training.

ResNet-50 [45] with GCC block From GCNet [26], we know that adding a self-attention mechanism to the C3, C4, and C5 stages of ResNet-50 can achieve better detection accuracy with higher cost-effective. In this article, we learn from the practice of GCNet: insert the GCC block on C3, C4 and C5 of ResNet-50, as shown in Fig. 6.

4 Datasets

The CIFAR-10 [46] dataset is one of the most commonly used datasets for image recognition experiment analysis. The dataset has a total of 60,000 color images. These images are 32 × 32, divided into 10 categories, each with 6000 images. Among them, 50,000 images are used for training, and 10,000 images are used for testing.

The GDXray [25] dataset is an open comprehensive dataset based on X-ray images, among which, Group Baggage contains more than 8000 X-ray images. During the experiment, in order to facilitate the experiment of object detection of prohibited item, we re-marked the Group Baggage of GDXray according to the format of PASCAL VOC [47]. We made statistics on the objects distribution of the dataset in [21]. There are 1297 images with background and 2941 images with single object and no background.

Metric of recognition accuracy CIFAR-10 dataset is used to discuss the effect of changes in the truncated coefficient β. In the test, we calculate the accuracy of the model on CIFAR-10 dataset, as shown in

$$accuracy = \frac{\sum_{i=1}^{k} tp_i}{n} \cdot 100\%. \tag{11}$$

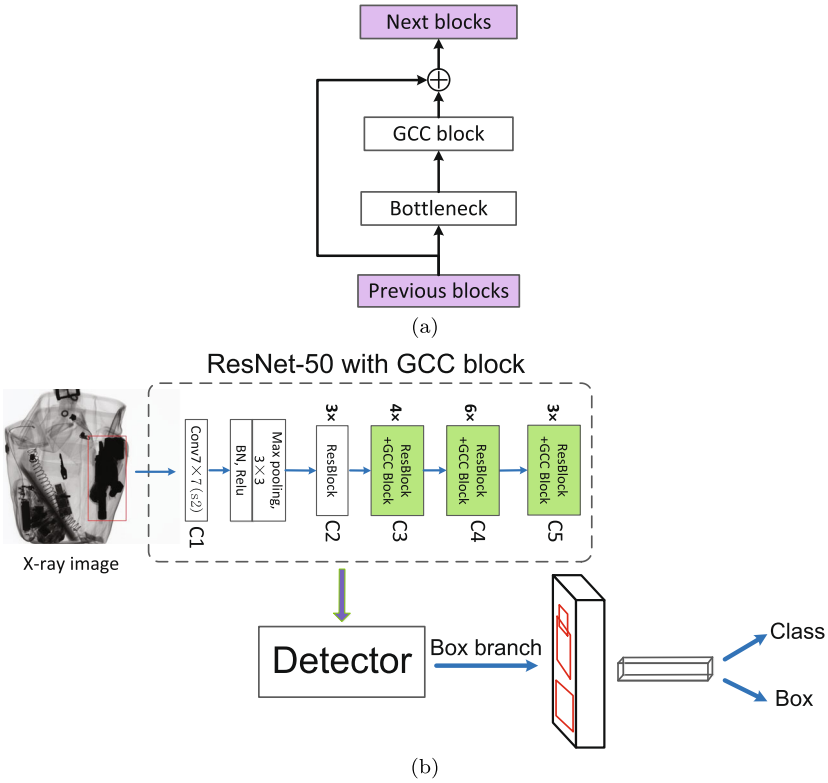

Fig. 6 **a** shows the inserting position of GCC block in residual block, which explains the 'ResBlock+GCC Block' in (**b**). GCC block is integrated after the last 1x1 conv inside the residual block (before Add). **b** is architecture of detector with GCC block

tp_i is the number of true positive of the i-th class, and n is the total number of images in the test set of CIFAR-10 dataset.

Detection metrics In the experiments of this article, we used cross-5-validation. GDXray dataset is randomly divided into five equal parts, 80% of them is used for training and the remaining 20% is used for testing.

In the experiment of detecting prohibited item on the GDXray, this article uses quantify predictive performance of the detection model to verify the effectiveness of our proposed method on the currently popular high-efficiency detector. Therefore, evaluate multiple metrics: $AP^{IoU=.50:.05:.95}$ (AP), $AP^{IoU=.50}$ (AP_{50}), and $AP^{IoU=.75}$ (AP_{75}), which are currently widely used on COCO dataset [40], are used for evaluation and comparison.

5 Experiment and discussion

5.1 Effect of truncated coefficient β

Settings MobileNetV1 [48] is used for training on CIFAR-10 [46], with the width multiplier(α) of the model being 0.25, and the resolution of the input images of the model being 128×128. In the training process, The SGD optimizer with momentum of 0.9 is used. At the same time, the exponential decay strategy with decay factor of 0.94 and initial learning rate of 0.005 are used to perform 160 epochs iterations.

5.1.1 Effect of β on recognition accuracy

Firstly, experiments are carried out on the CIFAR-10 and the results are recorded by every 6.4 ecpochs, as shown in Fig. 7.

The results prove that the truncated loss with softmax can improve the recognition accuracy to some extent. The change of recognition accuracy when changing the β from 0.1 to 0.7 is observed, when the β changes from 0.1 to 0.3, the recognition accuracy also increases and when the β changes from 0.4 to 0.7, the recognition accuracy has a downward trend. In extreme cases, when set β = 0.7, the recognition accuracy is smaller, lower than the case when softmax is not truncated.

This is because the introduction of the truncated values is to introduce artificial noise to the input of the softmax layer during the training process. If β is too large, it will have a negative impact on model training as mentioned in Section 3.1.

Secondly, the changes in Top-1 Accuracy and Top-2 Accuracy are observed when the value of β is changed. Experiments show that the effect of β on Top-1 is more obvious. The specific results are shown in Table 1.

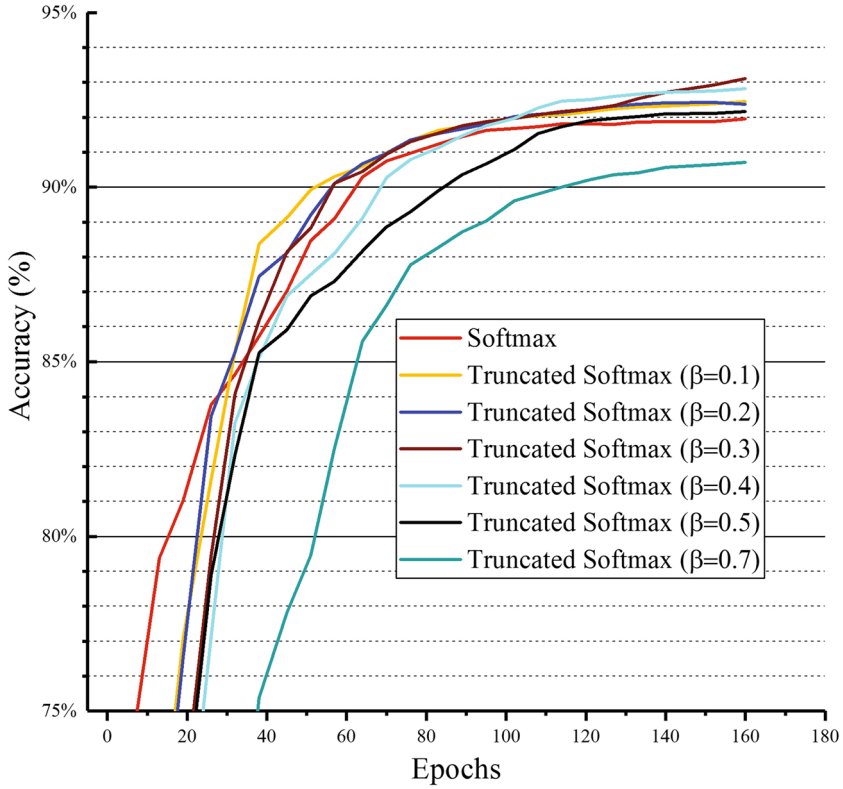

Fig. 7 Effect of the truncated coefficient β on recognition accuracy on CIFAR-10

The Top-1 Accuracy with a truncated coefficient (β) of 0.3 is 93.09%. This value is 1.12% higher than the Top-1 Accuracy of the softmax without truncated loss.

However, the test results show that the use of truncated loss with softmax has little effect on the accuracy of Top-2 Accuracy. In fact, the truncated loss is to add an artificial noise disturbance, which will change the predicted probability distribution of the softmax output. If Top-1 does not predict the target class, due to the presence of noise disturbance, this will make the probability value of other negative classes easily exceed the output probability value of the target class, and eventually cause Top-2 to make an incorrect prediction. Therefore, using truncated loss with softmax to improve the accuracy of Top-2 Accuracy recognition is not obvious.

Table 1 Experimental results of CIFAR-10 based on softmax and truncated loss with softmax

Model	α	β	Top-1 accuracy(%)	Top-2 accuracy(%)
MobileNetV1-128	0.25	–	91.97	97.71
	0.25	0.1	92.39	97.74
	0.25	0.2	92.39	97.73
	0.25	0.3	**93.09**	**97.85**
	0.25	0.4	92.82	97.50
	0.25	0.5	92.26	97.41
	0.25	0.7	90.75	97.43

5.1.2 Effect of $\boldsymbol{\beta}$ on prediction probability

In this section, the positive class probability of the softmax output layer during the training process is observed. As a comparison, we choose the softmax loss and the truncated loss with the β= 0.3. It is shown in Fig. 8.

During the training process of the model, the positive classes output probabilities of softmax for all images in a batch (batch size (n) is 32) are accumulated, and the average of these probabilities is calculated. The specific operation is shown in

$$prob = \frac{1}{n}\sum_{i=1}^{n} softmax(\mathbf{X}_i) \cdot \mathbf{Y_i}. \tag{12}$$

Among them, $\mathbf{X}_i$ is the softmax layer input vector of the ith image of a certain batch, and $\mathbf{Y_i}$ is the one-hot encoding vector of the corresponding label.

As can be seen from Fig. 8, the truncated loss with softmax can make the model's softmax output have a higher probability of positive class. This shows that the training of the truncated loss is more effective than softmax loss.

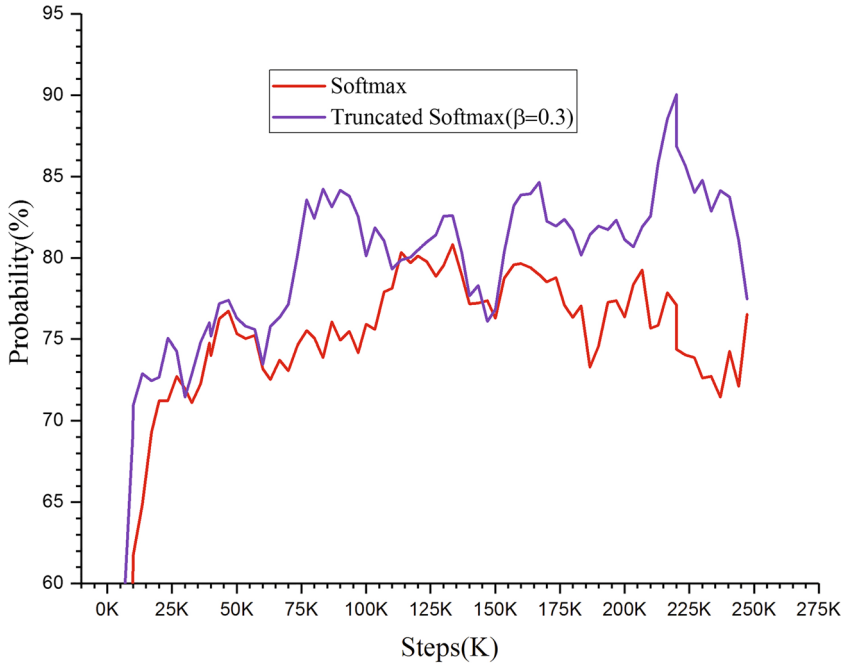

Fig. 8 The curves of positive class probability output by softmax during training

5.1.3 Summary analysis

The results of prediction probability also indirectly verify that using truncated loss with softmax can effectively improve Top-1 Accuracy, but the effect on Top-2 Accuracy is not obvious. In the training process, the input vector of softmax is truncated. This operation is to add artificial noise disturbance to the loss layer. This artificial noise disturbance improves the output probability of softmax of the positive classes, but due to the presence of noise, it also changes the probability distribution of softmax output of the negative classes. The changes of the negative classes' output probability distribution will affect the accuracy of Top-2 Accuracy.

Assume that the output of softmax is $\sum_i prob_i = 1$, the output of truncated loss with softmax is $\sum_i prob'_i = 1$, and t is the target class. If $prob_s$ and $prob_t$ are the two maximum probability values output by softmax, then Top-2 accurately predicts the target class.

In the prediction process, there will be such a situation: $prob_t$ is not the maximum value of the output probability and is relatively smaller than the maximum value. Although the truncated loss with softmax makes $prob'_t$ greater than $prob_t$, due to the presence of noise disturbances, other negative class probability values easily exceed $prob'_t$ value, so that $prob'_t$ is not in the two maximum values of the output probability, leading to the Top-2 prediction error. Therefore, the Top-2 prediction without using truncated loss is correct, and the prediction using truncated loss with softmax is incorrect.

5.2 Visualization of self-attention with Grad-CAM [49]

In order to further qualitatively analyze the effectiveness of the proposed self-attention architectures of CC block and GCC block, Grad-CAM [49] is applied to ResNet-50 [45] for visual analysis. Grad-CAM is a recently proposed qualitative visual analysis method, which analyzes the importance of spatial region in vision by calculating the gradient in the convolutional layer. Therefore, Grad-CAM is suitable for qualitatively analyzing the importance of the attented area in the network, based on which it can visually analyze the effectiveness of the attention mechanism.

5.2.1 Settings

Four images are randomly selected from the GDXray [25] dataset, where the images of shuriken and knife without background, and the images of razor blade and handgun contain background. These four images cover the four classes of objects in the dataset, so they are representative.

ResNet-50 [45] To illustrate the situation, ResNet-50 is used as the baseline for comparison. The pre-trained ResNet-50 on ImageNet [50] dataset is used to fine-tune on GDXray dataset. The experiment is divided into the following two situations:

- The ResNet-50 trained on GDXray is used as the baseline for qualitative comparison.
- With learning from the inserting position of thekattention structure of NLNet [38] and GCNet [26], CC block,

GC block and GCC block are inserted respectively before the adding operation of the ResBlock of C3, C4 and C5 of ResNet-50, as shown in Fig. 6a. However, in order to more effectively compare the effect of adding the attention module, the attention module is not embed in the structure of the last ResBlock on C5, so that it is convenient to qualitatively analyse other ResBlocks inserted into the attention module in ResNet-50.

Grad-CAM The gradient of the last convolutional layer on the above C5 is calculated respectively to obtain the Grad-CAM based heat map, as shown in Fig. 9. At the same time, we also calculate the predicted probability of the target category, denoted by P in Fig. 9.

5.2.2 Results and analysis

It can be seen from Fig. 9 that the region of red highlight in the heat maps after adding the self-attention mechanism can cover the target region of the X-ray image more effectively than the baseline. This shows that the self-attention mechanism is effective on X-ray images. From the perspective of the prediction probability P, compared with the object without background, the self-attention mechanism can effectively improve the prediction probability of the object with background.

From a vertical perspective of Fig. 9, the GCC block proposed in this article is most effective on ResNet-50. The region of red highlight in the heat map has better coverage, and at the same time, it has a higher prediction probability on the object than other models. Judging from the effect of the heat map on covering the object, the CC block is not as effective as the GC block. The probability of CC block predicting the target object is also lower than that of GC block. But after integrating CC block and GC block, a higher level of self-attention on GCC block is obtained.

6 Detection of prohibited item in X-ray images

According to the experimental analysis in Section 5.1, it can be found that when the truncated coefficient β is set to 0.3-0.4, the model can obtain better robustness. In order to facilitate observation and comparison, the truncated coefficient β in the experiments of this section is set to 0.3.

6.1 Ablation study

6.1.1 Training settings

In the training process, the standard configuration of Faster R-CNN [5] based on VGG16 [51] is used. The VGG16 used in the training process is pretrained from ImageNet dataset [50], and is used to initialize the weight of the Faster R-CNN. The input images of the detector are rescaled with a max-size of 1000 and a min-size of 600. The SGD with weight decay of 0.0001 and momentum of 0.9 is used for 7 epochs of iterations, where the initial learning rate is 0.001, and the learning rate decays to 0.0001 in the 5th epoch.

6.1.2 Ablation study on truncated loss

In order to study the effectiveness of the methods, the ablation experiments are carried out on Faster R-CNN based the VGG16, as shown in Table 2.

From the experimental results in Table 2, it can be found that in the training process, when MTL or LTL are separately applied to Faster R-CNN, the effect of improving the detection accuracy is not obvious. For example, compared with the Baseline, the use of MTL leads to the increase of the AP_{50} by 0.6%, and the use of LTL only improves the AP_{50} by 0.8%.

However, if MTL and LTL combined (it is expressed as MLTL), the AP_{50} of detection on prohibited item can be increased from 91.3% to 93.0%, and the AP has increased by 1%. Therefore, it proves that our truncated loss on Faster R-CNN is effective.

In order to prove the effectiveness of the truncated loss proposed in this article, the truncated loss with GIoU loss [52] and CIoU loss [53] are compared, which are used in the detector to optimize the regression loss of the bounding box. In the experiment, the truncated loss is used to calculate the object class loss, and GIoU loss, CIoU loss and Smooth L_1 loss [5] are used to calculate the regression loss of the bounding box, which are represented by MLTL (ours) + GIoU loss, MLTL (ours) + CIoU loss and MLTL (ours) + Smooth L_1 loss in Table 3.

The experimental results show that the article uses a combination of truncated loss and state-of-the-art regression loss to effectively improve the accuracy of prohibited item detection. Among them, the AP_{50} of GIoU loss increased by 1.8%, and the detection accuracy of MLTL (ours) + CIoU loss is the highest, with AP of 77.5%, AP_{50} of 94.4%, and AP_{75} of 82.3%.

6.1.3 Ablation study on GCC block

Settings The backbone of Faster R-CNN based on VGG16 is modified, and the self-attention mechanism is added to the output layer of the backbone. For comparison, a certain number of layers are added to GC block [26], CC block and GCC block. At the same time, the ratio of bottleneck transform is uniformly set to 8.

In order to measure the increment in model complexity after increasing the corresponding self-attention mechanism

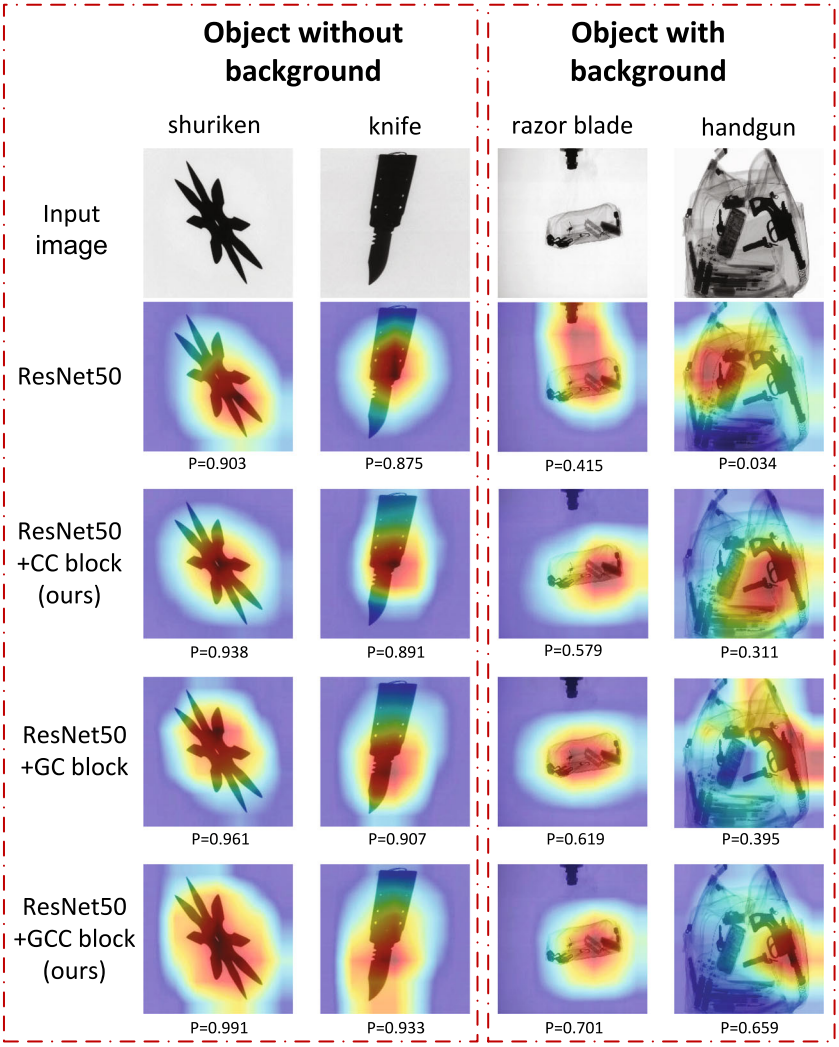

Fig. 9 The visualization results with Grad-CAM [49]

structure, the floating point operations (FLOPs) and parameters (# param) of the self-attention mechanism structure are calculated. Therefore, what is shown in Table 4 is only to calculate the increase of FLOPs and parameters of Faster R-CNN. In addition, the results of inserting the self-attention mechanism structure with different numbers of layers are compared, which are represented by layers in Table 4.

Table 2 Experiments of truncated loss based on Faster R-CNN

	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
Baseline	75.5	91.3	80.0
MTL [44]	75.7	91.9	80.7
LTL (ours)	75.7	92.1	80.9
MLTL (ours)	**76.5**	**93.0**	**81.3**

MTL denotes the balancing method [44] of loss function based on multitasking with the homoscedastic uncertainty, LTL denotes truncated loss without MTL. MLTL denotes combining MTL and LTL, as shown in (10)

Results From Table 4, we can see that the number of parameters that need to be learned in the structure of GC block [26] and CC block is the same. However, due to the existence of GAP, the FLOPs of CC block are higher than the FLOPs of GC block. This shows that the computational resources consumed by CC block are higher.

Since GCC block includes the architecture of CC block, the computational complexity of GCC block is higher than that of GC block. However, judging from the experimental

Table 3 Comparison of loss functions based on Faster R-CNN

	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
Smooth L_1 loss [5]	75.5	91.3	80.0
GIoU loss [52]	75.9	92.0	80.9
CIoU loss [53]	76.7	93.1	81.6
MLTL (ours) + Smooth L_1 loss	76.5	93.0	81.3
MLTL (ours) + GIoU loss	76.9	93.8	81.8
MLTL (ours) + CIoU loss	**77.5**	**94.4**	**82.3**

Table 4 Ablation experiments based on Faster R-CNN

	layers	# param(M)	FLOPs(M)	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
Self-attention mechanism in Faster R-CNN						
Baseline	-	-	-	75.5	91.3	80.0
GC block [26]	2	0.13	2.26	75.9	92.3	80.5
CC block (ours)	2	0.13	6.50	75.8	92.1	80.5
GC block [26]	4	0.27	4.53	76.5	**92.7**	81.1
CC block (ours)	4	0.27	13.00	76.4	92.5	81.3
GCC block (ours)	2	0.14	6.51	**76.8**	**92.7**	**81.6**
Combined truncated loss and self-attention mechanism in Faster R-CNN						
Baseline	-	-	-	75.5	91.3	80.0
GC block + MLTL	2	0.13	2.26	77.2	93.8	82.2
CC block + MLTL (ours)	2	0.13	6.50	77.1	93.8	82.3
GCC block + MLTL (ours)	2	0.14	6.51	**77.8**	**94.5**	**82.6**

results, the AP we obtained by adding two layers of GCC block is higher than that of adding four layers of GC block or GCC block. Compared with GC block and CC block, the highest detection accuracy is obtained at the cost of less computing resources. This shows that GCC block is more effective in feature extraction.

Finally, experiments with the result of combining the self-attention mechanism and truncated loss are also carried out, as shown in Table 4. GC block + MLTL, CC block + MLTL, and GCC block + MLTL are used to identify them, and found that the detection accuracy of the Faster R-CNN still has a further level of improvement at this time.

It is worth mentioning that the algorithm of GCC block + MLTL proposed in this article has respectively achieved 2.3%, 3.2% and 2.6% growth on AP, AP_{50} and AP_{75}.

6.2 Experiments on high-efficiency detectors

In order to verify the effectiveness of our algorithm, qualitative comparisons are conducted on high-efficiency detectors, such as Mask R-CNN [22], PANet [23], YOLOv3 [7], and Guassian YOLOv3 [24].

Mask R-CNN and PANet : During the training process, the mask branch of Mask R-CNN and PANet is removed, and only the box (object detection) branch is used to train these two detectors. ResNet-50 [45] is used as the backbone of the detectors, as shown in Section 3.3. GCC block is inserted in C3, C4 and C5 of ResNet-50, and the inserting position is shown in Fig. 6a.

YOLOv3 and Guassian YOLOv3 These two high-efficiency detectors are widely used in various fields based on visual detection. The GCC block is added to the Darknet-53 [7] based backbone of these two detectors. Darkenet-53 has a total of five residual units, which are named D1, D2, D3, D4, D5 in this article. With learning from the practice of GCNet [26], the GCC block is added to D3, D4, and D5 of Darknet-53 in YOLOv3 and Guassian YOLOv3.

The effectiveness of the truncated loss in these high-efficiency detectors are also verified. MLTL is used to represent the algorithm proposed in this article, as shown in (10), and the specific experimental results are shown in Table 5.

6.2.1 Training setting

A transfer learning strategy [21] is used to train these detectors separately: initialize these detectors with the weights pretrained from the COCO dataset [40]. With the specific hyper-parameter, The standard configuration of these four detectors is adopted respectively.

For these four detectors, SGD with weight decay of 0.0001 and momentum of 0.9 is used to perform 12 epochs of iteration on GDXray [25]. The initial learning rate of training is 0.01, and the 9th and 11th epochs decayed with the decay factor of 0.1. For Mask R-CNN and PANet, the input images of the detectors are rescaled with a maxsize of 1000 and a minsize of 600. For YOLOv3 [7] and Guassian YOLOv3 [24], the input images are uniformly resized to 416×416.

6.2.2 Discussion on high-efficiency detectors

From the experimental results in Table 5, the architecture of self-attention (GCC block) and truncated loss proposed in this article can effectively improve the accuracy of these detectors to detect prohibited item on the X-ray image.

Among them, the PANet detector has a high level of accuracy in detecting prohibited item. When we use GCC block and MLTL at the same time, PANet got the highest AP, AP_{50} and AP_{75} with the values of **92.3**%, **99.4**% and

Table 5 Experimental results of GDXray based on high-efficiency detector

Model	Backbone	GCC block	MLTL	$AP(\%)$	$AP_{50}(\%)$	$AP_{75}(\%)$
Mask R-CNN [22]	ResNet-50			88.1	95.6	92.6
		✓		89.2	97.1	93.3
			✓	88.9	96.3	93.4
		✓	✓	89.7	98.0	94.5
PANet [23]	ResNet-50			90.2	97.9	94.3
		✓		91.5	98.8	95.0
			✓	91.1	98.5	95.1
		✓	✓	**92.3**	**99.4**	**96.3**
YOLOv3-416 [7]	Darknet-53			86.4	94.1	90.3
		✓		87.5	95.8	91.1
			✓	87.6	95.3	91.0
		✓	✓	88.3	96.4	91.8
Guassian YOLOv3-416 [24]	Darknet-53			87.3	95.3	91.7
		✓		88.3	96.0	92.5
			✓	89.0	96.7	92.8
		✓	✓	89.5	97.2	93.5

96.3%. Compared with the standard PANet detector, GCC block increases AP by 1.3%, and MLTL increases AP by 0.9%. For the other three detectors, our proposed algorithms can also significantly improve the detection accuracy.

For these detectors, the one-stage detectors (YOLOv3, Guassian YOLOv3) and the two-stage detectors (Mask R-CNN, PANet) are compared, which shows that the algorithms proposed in this article have a wide range of applicability on these state-of-the-art detectors.

7 Conclusions

This article studies the detection method of prohibited item based on X-ray image. We propose a truncated loss function and two new architectures of self-attention mechanism of CC block and GCC block to detect prohibited item in X-ray images.

It can be seen from the mathematical analysis in Section 3.1 that the use of truncated loss with softmax can make the model convergence more effective, but the excessive truncated coefficient β makes it difficult for the model to converge to the optimal level. In other words, proper setting of the truncated coefficient is very important for training model. Therefore, in this article, the MobileNetV1 was used to discuss and analyze the truncated coefficient β in CIFAR-10 [46]. The results show that when the β is set to the values in the range of 0.3-0.4, the recognition accuracy of MobileNetV1 can be better improved. In the experiments of prohibited item detection, the β is set to 0.3 according to the experimental results in Section 5.1.

Experiments proves that the methods we designed is effective on the currently widely used high-efficiency detectors such as Mask R-CNN, PANet, YOLOv3 and Guassian YOLOv3.

In the experiment, YOLOv3 and Guassian YOLOv3 use sigmoid function based loss as the L_{cls} of the box branch, while Mask R-CNN and PANet use softmax loss as the L_{cls} of the box branch. Experiments prove that the truncated loss is effective for these two types of loss functions. However, in the analysis of the truncated loss, only softmax loss is used as an example to analyze its margin and gradient. In the future, we will further analyze the effectiveness and applicability of truncated loss.

This article designs CC block based on GC block [26]. Although the number of parameters to be learned is the same, the computational complexity of CC block exceeds that of GC block. In the future, we will continue to find ways to optimize the computational complexity of CC block without reducing the performance of the model.

Compliance with Ethical Standards

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

References

1. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition.

In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 212–220
2. Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, Li Z, Liu W (2018) Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 5265–5274
3. Wang F, Cheng J, Liu W, Liu H (2018) Additive margin softmax for face verification. IEEE Signal Process Lett 25(7):926–930
4. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp 21–37
5. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
6. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
7. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. CoRR arXiv:1804.02767
8. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2097–2106
9. Liu L, Liu Y-J, Chen A, Tong S, Chen CLPhilip (2020) Integral barrier lyapunov function-based adaptive control for switched nonlinear systems. Sci China Inf Sci 63(3):1–14
10. Liu L, Li X, Liu Y-J, Tong S (2021) Neural network based adaptive event trigger control for a class of electromagnetic suspension systems. Control Eng Pract 106:104675
11. Li D, Chen CLP, Liu Y, Tong S (2019) Neural network controller design for a class of nonlinear delayed systems with time-varying full-state constraints. IEEE Trans Neural Netw Learn Syst 30(9):2625–2636. https://doi.org/10.1109/TNNLS.2018.2886023
12. Liu Y-J, Zeng Q, Tong S, Chen CLP, Liu L (2019) Adaptive neural network control for active suspension systems with time-varying vertical displacement and speed constraints. IEEE Trans Ind Electron 66(12):9458–9466
13. Mery D, Svec E, Arias M, Riffo V, Saavedra JM, Banerjee S (2017) Modern computer vision techniques for x-ray testing in baggage inspection. IEEE Trans Syst Man Cybern Syst 47(4):682–692
14. Miao C, Xie L, Wan F, Su C, Liu H, Jiao J, Ye Q (2019) Sixray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2119–2128
15. Mery D, Saavedra D, Prasad M (2020) X-ray baggage inspection with computer vision: A survey. IEEE Access 8:145620–145633
16. Mery D, Pieringer C (2021) Applications in x-ray testing. In: Computer Vision for X-Ray Testing: Imaging, Systems, Image Databases, and Algorithms. Springer International Publishing, Cham, pp 375–436. https://doi.org/10.1007/978-3-030-56769-9_9
17. Riffo V, Godoy I, Mery D (2019) Handgun detection in single-spectrum multiple x-ray views based on 3d object recognition. J Nondestruct Eval 38(3):66
18. Mouton A, Breckon TP (2015) Materials-based 3d segmentation of unknown objects from dual-energy computed tomography imagery in baggage security screening. Pattern Recogn 48(6):1961–1978. https://doi.org/https://doi.org/10.1016/j.patcog.2015.01.010, https://www.sciencedirect.com/science/article/pii/S0031320315000291
19. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
20. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
21. Wei Y, Liu X (2020) Dangerous goods detection based on transfer learning in x-ray images. Neural Comput Appl 32(12):8711–8724
22. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
23. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
24. Choi J, Chun D, Kim H, Lee H-J (2019) Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. IEEE, pp 502–511. https://doi.org/10.1109/ICCV.2019.00059
25. Mery D, Riffo V, Zscherpel U, Mondragón G, Lillo I, Zuccar I, Lobel H, Carrasco M (2015) Gdxray: The database of x-ray images for nondestructive testing. J Nondestruct Eval 34(4):42
26. Cao Y, Xu J, Lin S, Wei F, Hu H (2019) Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp 0–0
27. Oksuz K, Cam BC, Kalkan S, Akbas E (2020) Imbalance problems in object detection: A review
28. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S (2019) Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 9268–9277
29. Li B, Liu Y, Wang X (2019) Gradient harmonized single-stage detector. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 33, pp 8577–8584
30. Lin T-Y, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988
31. Xing EP, Jordan MI, Russell SJ, Ng AY (2003) Distance metric learning with application to clustering with side-information. In: Advances in neural information processing systems, pp 521–528
32. Mika S, Ratsch G, Weston J, Scholkopf B, Mullers K-R (1999) Fisher discriminant analysis with kernels. In: Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468). IEEE, pp 41–48
33. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol 2. IEEE, pp 1735–1742
34. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
35. Weinberger KQ, Blitzer J, Saul LK (2006) Distance metric learning for large margin nearest neighbor classification. In: Advances in neural information processing systems, pp 1473–1480
36. Liu W, Wen Y, Yu Z, Yang M (2016) Large-margin softmax loss for convolutional neural networks. In: ICML, vol 2, p 7
37. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
38. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
39. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman

M, Zisserman A (2017) The kinetics human action video dataset. CoRR arXiv:1705.06950
40. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp 740–755
41. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp 603–612
42. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132–7141
43. Kanai S, Fujiwara Y, Yamanaka Y, Adachi S (2018) Sigsoftmax: Reanalysis of the softmax bottleneck. In: Advances in Neural Information Processing Systems, pp 286–296
44. Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7482–7491
45. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
46. Krizhevsky A, Hinton G, et al. (2009) Learning multiple layers of features from tiny images, Technical Report. U. Toronto
47. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. Int J Comput Vis 88(2):303–338
48. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR arXiv:1704.04861
49. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
50. Deng J, Dong W, Socher R, Li L, Kai Li, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255
51. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, ICLR 2015. Conference Track Proceedings, San Diego. arXiv:1409.1556
52. Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S (2019) Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 658–666
53. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-iou loss: Faster and better learning for bounding box regression. In: AAAI, pp 12993–13000

Yuanxi Wei received the Master's degree in Computer Science and Technology from Beijing University of Posts and Telecommunications, China, in 2014. He is now working for his Ph.D. at Beijing University of Posts and Telecommunication. His current research interests include machine learning, object recognition and detection of prohibited items based on computer vision in X-ray images.

Xiaoping Liu is a Professor at the School of Modern Post (School of Automation), Beijing University of Posts and Telecommunications (China). He received a Ph.D.in Mechanical Engineering in the Tianjin University (China). He was visiting scholar at the University of Southern California (US) and at the University of Washington (US). His current research interests include robot, logistics technology and equipment, detection technology. Yinan

Yinan Liu is a Senior Researcher at Tsingmicro (China). He received Master's degree at Beijing University of Posts and Telecommunication in 2014. His current research interests include object recognition, object detection and face recognition.

Affiliations

Yuanxi Wei[1] · Xiaoping Liu[1] · Yinan Liu[2]

Yuanxi Wei
weiyuanxi@bupt.edu.cn

Yinan Liu
liuyinan@tsingmicro.com

[1] Automation School, Beijing University of Posts and Telecommunications, No. 10, Xitucheng Road, Haidian District, Beijing 100876, China

[2] Tsingmicro, No. 1, Baosheng South Road, Haidian District, Beijing 100876, China