ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 3 - 2026

Research and Application of High Resolution Infrared Image Rapid Detection Technology Based on Deep Learning in Human Body Binding and Smuggling Detection

Ying Chen[1,*], *Jun Li*[2], *Zeliang Luo*[1] *and Ming Hui*[3,*]

[1] Wuhan Guide Infrared Co.,Ltd, Wuhan 430070, Hubei, China
[2] Shenzhen Customs Information Center, Shenzhen 518045, Guangdong, China
[3] Shenzhen Customs District P.R.China, Shenzhen 518026, Guangdong, China

SUMMARY: *In response to the low recognition rate of smuggled dangerous goods in real security check scenarios and the shortcomings of traditional image classification and existing deep learning algorithms in the field of security check, effective recognition algorithms are proposed to improve security check efficiency and accuracy, and ensure passenger safety. This article uses deep convolutional neural networks to propose a deep learning network-based algorithm for identifying dangerous goods in human body smuggling and security checks. By combining infrared images with VI, using image segmentation and unsupervised registration techniques, a self-encoding deep learning network is used as the registration network skeleton for human contour registration learning. The network is optimized by measuring the similarity of the registered images and minimizing the similarity cost. The experimental results show that, compared with other algorithms on the public dataset SIXray, our algorithm achieves an average accuracy (mAP) of 93.21%, which is higher than other compared algorithms, especially 38.7 percentage points higher than Inception V3, with small increases in parameter quantity and model size and almost unchanged detection time. The algorithm in this article performs excellently in terms of parameter quantity, model size, detection time, and average accuracy, and has higher feasibility and practicality, providing efficient and accurate solutions for security inspection work.*

KEYWORDS: *Deep learning; High resolution infrared images; Human body bondage and smuggling; Deep Convolutional Neural Network; Image segmentation; Self-ensembling autoencoder; Average precision*

1 Introduction

The hidden dangerous goods in the human body pose a great threat to the safety of passengers. The application of infrared imaging technology in personal security check is relatively mature [1, 2]. However, the random placement and disorderly overlap of objects in the human body can lead to differences in the proportion and observation perspective of dangerous goods in the image, resulting in low recognition rate of existing safety inspection dangerous goods recognition algorithms in real scenes [3]. Most places such as stations and airports first use security cameras to capture infrared images before conducting manual inspections. This method consumes a lot of manpower and material resources, especially during peak passenger flow periods [4]. Long term high-intensity work can cause extreme fatigue among security personnel,

*hw_szc@163.com
https://doi.org/10.65102/is20261200
Published: 10 June, 2026

leading to false or missed detections of dangerous goods and causing public safety accidents.

Traditional security inspection image dangerous goods recognition, as a branch of image classification, is consistent with traditional image classification algorithms. It mainly involves feature extraction through manual processing such as image color segmentation, edge detection, image enhancement, and image denoising, and then uses decision trees, support vector machines (SVM), etc. to recognize dangerous goods in security inspection images. Reference [5] proposed using the Scale Invariant Feature Transform (SIFT) algorithm to extract keypoint features, which achieved good recognition results. The bag of words (BOW) algorithm proposed in reference [6] first uses SIFT to extract features from each image, then clusters the features using K-means algorithm to construct BOW visual word sequences, and finally uses SVM for recognition. The adaptive sparse representation algorithm proposed in reference [7] first extracts several random blocks from the image to construct an adaptive dictionary, then performs proxy optimization on the dictionary, and performs classification and recognition based on the sparse representation algorithm. Since the proposal of deep learning networks, various neural networks with different levels of structure have been widely applied in fields such as facial recognition, detection of human smuggling marks, and human pose recognition. However, the application of neural networks in the field of security inspection is relatively limited. Reference [8] applied deep convolutional neural networks to the recognition of infrared security inspection images through transfer learning. The feature extraction, feature representation, and recognition of this algorithm were all completed using convolutional neural networks, but the images used were relatively simple and did not match the real scene. Reference [9] proposed an image classification algorithm based on deep convolutional features, which obtains significant features through computer self-learning and has high recognition accuracy. Reference [10] compared the recognition accuracy of traditional visual bag of words, sparse transformation, classical pattern recognition, and deep learning algorithms, and the results showed that the deep learning algorithm had the highest recognition accuracy. Reference [11] proposed an algorithm based on Class Balanced Hierarchical Refinement (CHR), which combines deep convolutional neural networks and the principle of infiltration assumption to model security check images in layers. The algorithm achieved good results in dealing with a large amount of imbalanced data in real scenes, but the recognition accuracy was low.

In summary, traditional image classification algorithms rely on manual feature extraction, and the extracted features are limited. They are only suitable for security check images in simple scenes and cannot effectively identify dangerous goods in real scene images. However, deep learning algorithms use simple image data, have poor recognition efficiency and accuracy, and are prone to false positives and false negatives, making them difficult to apply to practical problems. Therefore, this article proposes a deep learning network-based algorithm for identifying dangerous goods in human body bundling, smuggling, and security checks using deep convolutional neural networks.

2 Related research

In terms of detecting hidden prohibited objects carried by the human body in infrared images, reference [12] extracted pixel centered image blocks from the image and used Haar operator to extract features. Then, the random forest method was used to detect hidden objects on the preprocessed infrared image and achieved certain results. Reference [13] uses deep neural network methods to identify prohibited objects in image blocks, while using image segmentation methods to detect and locate prohibited objects. Reference [14] used the YOLO v3 (You OnlyLookOncev3) algorithm to detect prohibited object targets in infrared images and

achieved real-time detection on its self-built small dataset. Due to the limited imaging quality of infrared images, using a single millimeter wave image can easily lead to false positives. Therefore, this article adopts a combination of infrared images and VI to efficiently detect hidden prohibited items carried by the human body, mainly using image segmentation and image registration techniques.

In terms of image semantic segmentation, methods based on Fully Convolutional Network (FCN) have achieved superior performance. Especially with the emergence of deep neural networks, researchers have begun to use encoder decoder structures to fuse low and high-level features to obtain more contextual information and achieve better segmentation results. Reference [15] combines the advantages of region based and full convolution-based methods, and uses a residual network with dilated convolution for feature extraction. Finally, the segmentation results are obtained through multi model fusion. Reference [16] proposes a statistical texture learning network for image segmentation to better utilize the low-level texture information of the network and achieve better performance.

Image registration is currently widely used in the field of image processing, and deep learning-based registration methods are mainly divided into supervised registration and unsupervised registration. Registration based on supervised learning generally utilizes existing algorithms to generate labels or simulates deformation to generate labels, which is relatively complex. Given this, most researchers nowadays tend to use registration methods based on unsupervised learning. Reference [17] proposes a Voxel Morph framework for fast learning of image registration, which uses convolutional neural networks to learn deformation fields and optimizes the network by minimizing the similarity cost between images. Reference [18] proposes a registration network with an encoder decoder structure, and calibrates the features based on their performance and the relationships between them. Then, a hierarchical cost function is designed for network training. This article draws on the experience of processing 3D images from reference [19] and combines it with lightweight deep neural networks to apply unsupervised registration learning to 2D infrared and VI images.

3 Problem description

3.1 Unsupervised learning strategy

Training deep neural networks usually requires a large amount of labeled data as support. In supervised registration for identifying dangerous goods in human body smuggling and security checks, the following two methods are generally used to obtain deformation field labels: (1) using traditional methods to register security check images to obtain deformation fields as the true values of the labels. For example, in reference [20], the symmetric image normalization method is first used to preliminarily register the images, and then the boundaries are further aligned to generate the final deformation field. (2) Artificially generating deformation fields, as described in reference [21], involves synthesizing deformation fields and applying them to the original images to generate deformation images for training. It can be seen that both methods are complex and inefficient, making them difficult to promote. At the same time, the accuracy of labels also has an impact on the effectiveness of network training. Unlike supervised learning that requires labeled information, unsupervised learning does not require data labels, saving time for manual annotation. Taking into account the above considerations, this article adopts unsupervised learning methods to register and learn human body contours in security check images.

Unsupervised learning can be divided into various methods, such as clustering and dimensionality reduction. In this paper, we adopt unsupervised learning based on similarity

measurement, which compares the similarity between the model output and the registered image, and minimizes the similarity cost function to optimize and adjust the network parameters.

3.2 Image registration strategy

On the one hand, due to the different imaging mechanisms of infrared images and VI imaging for smuggling and security checks, there are significant differences in imaging effects. Specifically, the imaging field of infrared images for smuggling and security checks with personal restraints is smaller than VI, and the imaging quality is poor with low resolution. But passive millimeter waves can detect hidden objects and have penetrability that visible light does not have [22]. And the VI for smuggling and security checks with personal restraints has better clarity and resolution than infrared images. The imaging differences between the two images make it possible to complement each other's advantages. This article simultaneously utilizes the clarity of the personal restraint smuggling security VI and the penetrability of infrared images to make the detection of concealed prohibited items more efficient. On the other hand, there are black holes in the human body area in both the infrared images of human body bundling and smuggling security checks and the human body contour segmentation results of VI. The holes at the same position are generated by the gap between the arm and the body, while the holes that exist in the infrared images of human body bundling and smuggling security checks but are not in the VI are the locations of hidden objects. Therefore, only by registering the segmentation results of the two can false alarm targets be effectively removed and the location of hidden prohibited objects be obtained. In addition, due to privacy concerns caused by directly displaying the infrared images of human body bundling and smuggling security checks, it is necessary to mark the location of hidden objects in the VI, which also requires the registration of the infrared images/VI of human body bundling and smuggling security checks to be achieved.

3.3 Unsupervised learning method for human body contour registration

Taking inspiration from the image registration method in reference [23], this paper designs an unsupervised learning registration method suitable for infrared and VI images of human body bundling and smuggling security checks, as shown in **Figure 1**. This article uses a Self-ensembling autoencoder deep learning network as the registration network skeleton, with the only difference compared to segmentation networks being the removal of the final classification layer and sigmoid layer. Define the segmentation image p and VI segmentation image v of the infrared image of smuggling and security check hidden in the body as a 2D image space $W \subset R^2$. The Self-ensembling autoencoder deep learning network registration network is denoted as $f_\theta(\cdot)$, and θ is the neural network parameter. Therefore, the registration displacement calculation process based on the Self-ensembling autoencoder deep learning network can be represented as $f_\theta(p,v) = u$. Among them, u is a 3D displacement field. For any pixel m, $u(m)$ is a displacement that causes $v(m)$ and $[p \circ \varphi](m)$ to represent the same position, where mapping $\varphi = \mathrm{Id} + u$, Id represents the identity transformation, $v(m)$ represents the position of m in the middle, $p \circ \varphi$ represents the registered image of p after mapping transformation φ, and $[p \circ \varphi](m)$ represents the position of m in the human body smuggling security check image after registration.

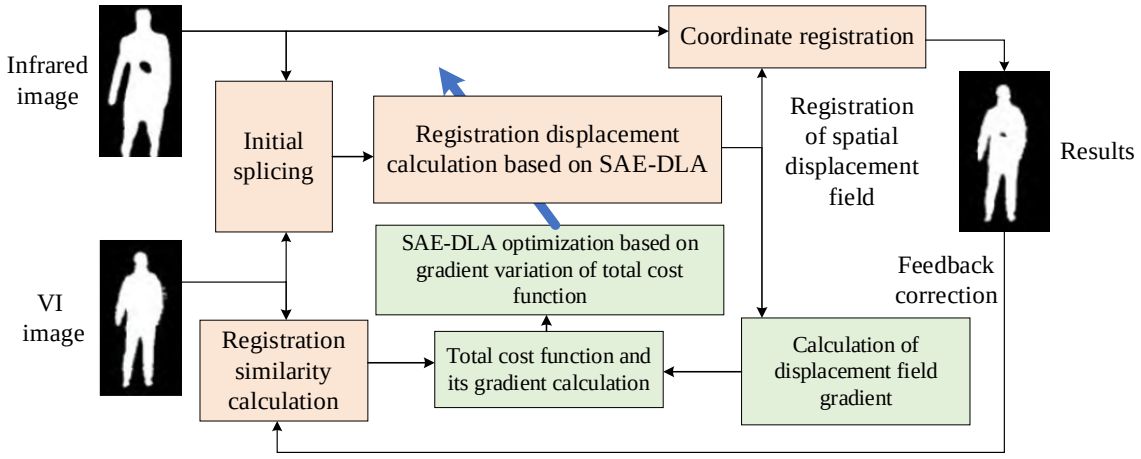

Figure 1: Unsupervised human contour registration based on similar Self-ensembling autoencoder deep learning

As shown in Figure 1, the infrared image and VI segmentation image of the human body bound smuggling security check to be registered are concatenated and input into a self-encoding deep learning network registration network. The network calculates the registration spatial displacement field *u*, and then coordinates the segmented image *p* of the human body bound smuggling security check infrared image to transform it into $p \circ \varphi$, obtaining the registered infrared image p_r. This article is based on the idea of spatial transformation networks and uses the method described in reference [24] to calculate $p \circ \varphi$. For each pixel *m* in the registered human body smuggling security check image, this paper calculates its corresponding pixel position in the original human body smuggling security check image *p*, and uses the pixel values of 8 adjacent points for linear interpolation to obtain $[p \circ \varphi](m)$.

3.4 Learning process optimization based on similarity measurement

This article measures the similarity between the registered infrared images of human body smuggling and security checks and VI, and optimizes the registration network by minimizing the cost of similarity. At the same time, a penalty term for registering displacement field gradients is added to the cost function to make the coordinate transformation smoother. The total cost function is given by equation (1) [25]:

$$C = \frac{1}{|W|}\sum_{m \in W}\left[v(m) - [p \circ \varphi](m)\right]^2 + \lambda \sum_{m \in W}\left\|\nabla u(m)\right\|^2 \tag{1}$$

where, λ is the regularization coefficient, and $\nabla u(m)$ represents the gradient calculation for $u(m)$.

The first half of equation (1) calculates the Mean Square Error (MSE) between the registered infrared image and VI, while the second half applies regularization to the gradient of the displacement field. As shown in Figure 1, by calculating the total cost function and its gradient changes, and then using the adaptive moment estimation (Adam) optimizer based on the gradient changes of the total cost function to optimize the parameters of the Self-ensembling autoencoder deep learning network, the network performance can be continuously improved.

In addition, during the training process, on the one hand, this article trains the input network with infrared images/VI segmented images of people being bound and smuggled for security checks; On the other hand, for the hollow areas in the human body generated during the segmentation process, they are filled to generate segmented images that are used as an

augmented set for training to improve the robustness of network performance.

4 Self-ensembling autoencoder deep learning algorithms (SAE-DLA)

Traditional models use parameterized methods for prediction, however, due to the randomness and nonlinearity of human trafficking and smuggling, parameterized methods cannot make accurate predictions. Therefore, non-parametric machine learning methods have become the preferred choice for predicting human trafficking and smuggling. However, with the advent of the big data era, traditional machine learning models have some limitations when dealing with high-dimensional datasets and complex architectures due to their relatively simple structure. Therefore, the superiority of deep learning technology in predicting short-term human trafficking volume is gradually emerging. Deep learning not only provides stable prediction performance independent of model hyperparameters, but also effectively mines the hidden features between human trafficking data, making it an important advantage in modeling complex human trafficking data. In addition, deep neural networks can effectively capture the spatiotemporal characteristics of human trafficking data, making them the mainstream method for short-term prediction of human trafficking.

4.1 Autoencoder

Self-ensembling autoencoder (SAE) is an unsupervised neural network that extracts features by learning compressed representations of input data. SAE defines Self-ensembling autoencoderor auto association configuration as a discriminative graphical model that reconstructs its input signal, where the output is constrained to be the same as the input. The SAE model, as shown in **Figure 2**, stacks the output of the Self-ensembling autoencoderin the lower layer as the current input. Faced with the situation where the size of SAE in the hidden layer is the same as or larger than that of the input layer, the model shows obvious drawbacks, such as the possibility of being copied as input and not extracting useful features. At present, there are two main methods to solve this situation: one is to use denoising autoencoders, and the other is to use sparse autoencoders. Due to the strong feature extraction ability of sparse autoencoders, their ability to handle high-dimensional data, and their ability to be combined with prediction models such as LSTM and CNN, this article mainly discusses sparse autoencoders.

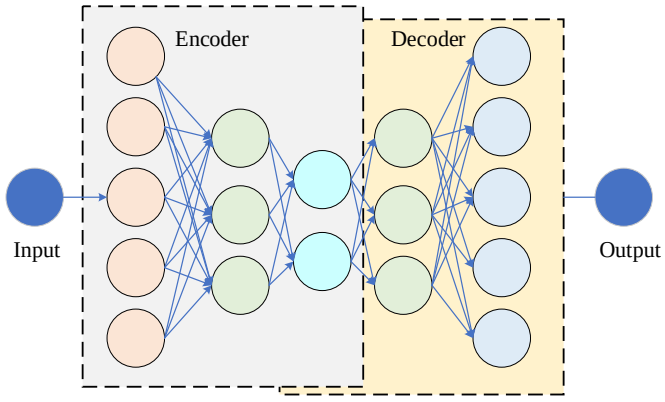

Figure 2: SAE Model Diagram

In order to evaluate the smuggling volume of human trafficking during different time periods, reference [26] further evaluated the performance of the proposed SAE model in

predicting human trafficking during the day and night through 250 experiments. By integrating the optimal hyperparameter combination of weekdays and non-workdays, it was shown that the deep learning based human trafficking prediction method can be a combination of multiple SAEs, with different parameters adapted to different time periods, and can capture the traffic characteristics of different time periods. In order to solve the problem of large fluctuations in the volume of human trafficking and insufficient traffic data leading to overfitting of the model, reference [27] proposed a new SAE-LE model, which improves the accuracy of human trafficking prediction by introducing a scheme into the SAE neural network architecture. This approach retrains SAEs by rearranging the training data to enhance the network's generalization ability, and improves prediction accuracy by constructing a set of SAEs. At present, in the field of predicting human body trafficking, a hybrid model is often formed using the advantages of deep neural networks and SAEs to capture the spatiotemporal relationships in human body trafficking data. For example, in reference [28], LSTM and SAE architectures were used for multi-step prediction of human trafficking volume, capturing spatiotemporal features and improving the accuracy of prediction. Reference [29] proposes a hybrid human body smuggling prediction model combining SAE and LSTM. This fusion model fully utilizes the features of SAE and LSTM to predict human body smuggling through a fully connected layer, and obtains the spatiotemporal characteristics of human body smuggling.

4.2 Recurrent neural networks

Although CNN has shown good performance in predicting human trafficking, it may face challenges in dealing with the spatiotemporal characteristics of human trafficking. Due to the complex spatiotemporal correlations of human trafficking data, RNNs and their variants have been widely used in predicting human trafficking, as shown in **Figure 3**. Due to the existence of gradient vanishing and gradient explosion problems, standard RNNs perform poorly in handling tasks that require long-term time dependencies.

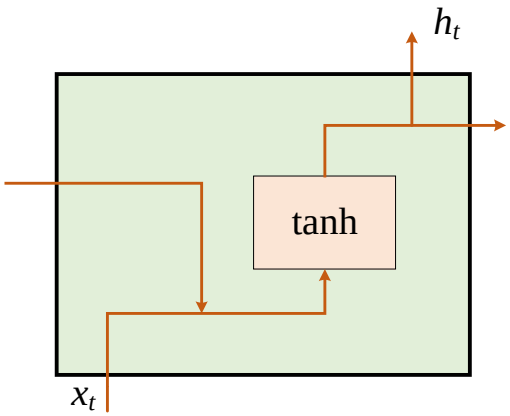

Figure 3: RNN Structure Diagram

In traditional standard RNNs, repeated neuron modules contain a very simple structure. This structure integrates the input value of the hidden layer with the output value of the previous time step, and then feeds the integrated result into an activation function (such as tanh() function). The final output obtained is the output of the current neuron. This design allows RNNs to process sequential data and has memory capabilities to capture long-term dependencies in sequential data. The formula is expressed as follows [30]:

$$h_t = \sigma_h\left(W_{xh}x_t + W_{hh}h_t - 1 + b_h\right) \tag{2}$$

where, W_{xh} is the weight matrix from the input layer to the hidden layer, W_{hh} is the weight matrix between neurons inside the hidden layer, b_h is the bias vector of the hidden layer, and σ_h is the activation function of the hidden layer.

Long Short Term Memory Recurrent Neural Networks (LSTM-RNN) are used to solve gradient vanishing and exploding problems in RNNs. Compared to traditional RNNs, LSTM has a gating mechanism that selectively stores and forgets information, enabling it to more effectively handle long-term time-dependent tasks. The LSTM structure is shown in **Figure 4**. LSTM uses memory units to replace the hidden units of RNN, which mainly include three gate units: forget gate EE, input gate FF, and output gate GG.

The forget gate determines whether to retain or discard information from previous cell states at the current moment, and its formula is expressed as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, X_t\right] + b_f\right) \tag{3}$$

The input gate determines the incorporation of new information into the cell state at the current time step, and its formula is expressed as follows:

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, X_t\right] + b_i\right) \tag{4}$$

$$\tilde{C}_t = \tanh\left(W_C \cdot \left[h_{t-1}, X_t\right] + b_c\right) \tag{5}$$

The output gate determines the value that the neuron structure of the current time step needs to output, and its formula is expressed as follows:

$$o_t = \sigma\left(W_o\left[h_{t-1}, X_t\right] + b_o\right) \tag{6}$$

$$h_t = o_t \times \tanh\left(C_t\right) \tag{7}$$

The prediction of human trafficking and smuggling based on recurrent neural networks has been widely applied. As the length of the time series increases, the LSTM model becomes difficult to solve the overly long-term dependency relationship in predicting human trafficking and smuggling. Unlike traditional prediction methods, LSTM networks consider the spatiotemporal correlation of human trafficking systems through a two-dimensional network composed of multiple memory units. By comparing with other typical prediction models, it has been verified that the proposed new model for predicting human body trafficking and smuggling can achieve better performance. In addition, the encoder decoder structure in RNN networks is used for traffic prediction, where the encoder encodes historical data into a vector representation and the decoder uses this vector to generate future traffic predictions. The encoder decoder structure can capture temporal dependencies, achieve accurate prediction, and provide support for decision-making.

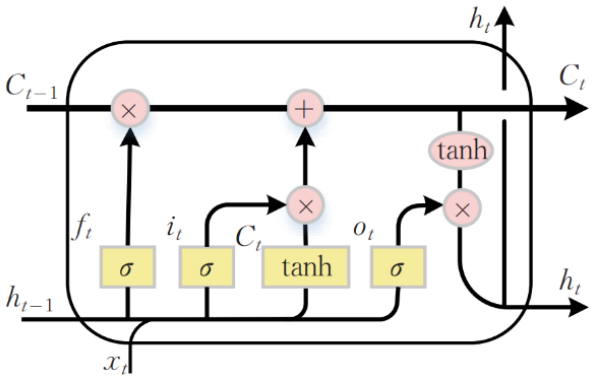

Figure 4: LSTM Structure Diagram

5 Experimental analysis

5.1 Dataset

The algorithm proposed in this article was tested on the public dataset SIXray, which collected 17641 annotated images of human trafficking and security checks. Compared with other datasets, SIXray has more categories and a relatively large amount of data. Randomly divide the dataset into three parts, with the validation set and test set each accounting for 39.5% (6964 images), and the remaining (10677 images) being the training set, with a ratio of approximately 1:1:4. In this study, we removed the detection of scissor like items in the experiment because the sample size was too small and the inter class data was imbalanced. The detailed distribution of each category in the dataset is shown in **Table 1**. In addition, many of the smuggled security images in the dataset contain multiple prohibited items, so the total number of items is much higher than the number of smuggled security images in the dataset.

Table 1: Distribution of each category in the SIXray dataset

SIXray	category					total
	pistol	knife	wrench	pliers	hammer	
training set	3014	1879	1896	3208	680	10677
validation set	982	600	594	1080	226	3482
test set	982	600	594	1080	226	3482
total	4978	3079	3084	5368	1132	17641

5.2 Evaluation indicators

The evaluation of the detection performance of the detector needs to consider both accuracy and recall. In object detection, the mean average precision (mAP), macro precision (MP), macro recall (MR), and macro F1 (MF1) at IoU=0.5 are used to evaluate the performance of the network model. The definition of accuracy is equation (8), and the definition of recall is equation (9). Among them, T_{TP}, T_{TN}, F_{FP}, and F_{FN} represent true positive, true negative, false positive, and false negative, respectively.

$$A_{Aceuracy} = \frac{F_{TP}}{T_{TP} + F_{FP}} \times 100\% \tag{8}$$

$$R_{\text{Recall}} = \frac{T_{\text{TP}}}{T_{\text{TN}} + F_{\text{FN}}} \times 100\% \tag{9}$$

Average precision (AP) is obtained by combining accuracy and recall, and is used to evaluate the accuracy of a model in detecting a single category. The mAP measurement model detects the accuracy of all categories by taking the average A_{AP} value of all categories, as defined in equation (10). The F1 score is the weighted average of accuracy and recall, defined as equation (11), where a higher value indicates better performance.

$$M_{\text{mAP}} = \left(\frac{1}{n}\right)\sum_{i=0}^{n} A_{\text{AP}}(i) \tag{10}$$

$$F1 = \frac{2 \cdot A_{\text{Aceuracy}} \cdot R_{\text{Recall}}}{A_{\text{Aceuracy}} + R_{\text{Recall}}} \times 100\% \tag{11}$$

Similar to the average precision mean, macro precision, macro recall, and macro F1 are obtained by taking the average of all category accuracy, recall, and F1 scores, respectively. In addition, confusion matrices can also be used to assist in result analysis.

5.3 Result analysis

The hardware configuration for the experiment is Core (TM) i7-10720X processor, GeForce RTX 1060 graphics card, and software configuration is torch 1.8.2. During the training process, the stochastic gradient descent (SGD) method is used to optimize the network parameters with a value of 200 epochs. The input image size is 640 × 640, and the batch size is 64. Conduct experiments on multiple models while keeping other parameters consistent.

This article presents the experimental results of four models: YOLOv5s algorithm R-CNN, Yolo v11, The algorithm presented in this article. **Table 2** shows the overall performance comparison of four algorithms on the SIXray dataset.

From the data in Table 2, it can be clearly seen that there are significant differences in various evaluation indicators among different algorithms. In terms of average precision (AP), Yolo v11 algorithm performs the best, reaching 98.42%, followed closely by our algorithm at 98.34%, R-CNN at 98.32%, and YOLOv5s relatively low at 97.53%. This indicates that Yolo v11 has a slight advantage in detection accuracy for a single category, but the gap between our algorithm and it is very small, indicating that our algorithm also has strong competitiveness in single category detection accuracy. In terms of the Mean Average Precision (mAP) metric, the algorithm proposed in this paper has a significant advantage, reaching 89.76%, which is different from YOLOv5s' 86.04%, R-CNN's 88.37%, and Yolo v11's 87.18%, all of which show varying degrees of improvement. MAP measures the accuracy of the model in detecting all categories, and the outstanding performance of our algorithm in this indicator indicates that it has higher accuracy in overall detection of various categories and can more comprehensively identify different prohibited items in the dataset. In terms of macro precision (MP), our algorithm achieves 91.20%, which is higher than YOLOv5s' 86.35%, R-CNN's 87.87%, and Yolo v11's 88.59%. The macro precision reflects the proportion of samples predicted as positive by the model that are actually positive. The algorithm in this paper leads in this indicator, which means that the proportion of correctly predicted prohibited items in its prediction results is higher, reducing the occurrence of misjudgments. In terms of macro recall (MR), our algorithm achieved 95.55%, which is also superior to the other three algorithms. The macro recall rate reflects the ability of the model to correctly detect samples that are actually positive examples. The high macro recall rate of the algorithm in this paper indicates that it can more effectively

identify prohibited items in the dataset, reducing the possibility of missed detections. Macro F1 (MF1), as a weighted average of accuracy and recall, comprehensively reflects the performance of the model. The MF1 of the algorithm in this article is 92.08%, which is higher than the 89.01% of YOLOv5s, 91.02% of R-CNN, and 90.06% of Yolo v11. This further proves the superiority of the algorithm in overall performance, as it can accurately identify prohibited items while also balancing false positives and false negatives. Overall, the deep learning network-based algorithm proposed in this article for identifying dangerous goods in human bound smuggling and security checks performs well in various evaluation indicators on the SIXray dataset. Compared with other classic algorithms, it has higher detection accuracy and better comprehensive performance, and can be more effectively applied to the identification task of dangerous goods in human bound smuggling and security checks.

Table 2: Performance Comparison of Detection Algorithms (%)

Method	AP					mAP	MP	MR	MF1
	pistol	knife	wrench	pliers	hammer				
YOLOv5s	97.53	86.04	86.35	93.84	89.32	90.63	93.21	85.28	89.01
R-CNN	98.32	88.37	87.87	94.51	92.27	92.29	94.34	87.09	91.02
Yolo v11	98.42	87.18	88.59	94.70	90.18	91.81	93.30	87.75	90.06
proposed algorithm	98.34	89.76	91.20	95.55	91.34	93.26	95.21	89.25	92.08

According to **Table 3**, the number of algorithm parameters in this article only increased by 0.28%, the model size increased by 2.82% (0.4MByte), and the detection time remained almost unchanged.

From the time comparison data of different algorithms presented in Table 3, we can deeply analyze the characteristics and advantages of our algorithm in key dimensions such as parameter quantity, model size, and detection time. In terms of parameter count, YOLOv5s and R-CNN both have a parameter count of 7.05×10^{6}, while Yolo v11 and our algorithm both have a parameter count of 7.07×10^{6}. Compared to YOLOv5s and R-CNN, the algorithm parameters in this paper have only increased by 0.28%. This slight increase means that the algorithm proposed in this paper does not impose a significant burden on the complexity of the model. In the field of deep learning, having too many parameters often leads to overfitting of the model, where it performs well on training data but performs poorly on new, unseen data. And with the addition of a small number of parameters, the algorithm in this article can effectively avoid this problem, ensuring the model's ability to learn data features while not losing generalization due to too many parameters. This is crucial for practical applications facing various complex and ever-changing security check scenarios, ensuring that the algorithm can stably identify smuggled items hidden by people in different environments and data distributions. In terms of model size, YOLOv5s is 13.73 Mbyte, R-CNN is 14.09 Mbyte, Yolo v11 is 13.86 Mbyte, and our algorithm is 14.21 Mbyte, an increase of 2.82% (0.4 Mbyte). A moderate increase in model size has little impact on actual deployment. At security checkpoints such as stations and airports, the storage resources of equipment are limited, and oversized models may occupy too much space, affecting the overall operational efficiency of the system. The limited growth of the algorithm model size in this article enables it to be deployed and applied smoothly in environments with tight storage resources. Moreover, with the continuous development of storage technology, a growth of 0.4Mbyte will not cause significant storage pressure on most modern devices. Detection time is an important indicator for measuring the real-time performance of algorithms. From Table 3, it can be seen that the detection time of YOLOv5s is 2.54ms, R-CNN is 2.52ms, Yolo v11 is 2.51ms, and our algorithm is 2.50ms. Our

algorithm is almost on par with other algorithms in terms of detection time, and even has a slight advantage. In practical security check scenarios, especially in stations and airports with high passenger flow, it is crucial to quickly and accurately detect dangerous goods. If the detection time is too long, it may increase the waiting time for passengers, reduce the efficiency of security checks, and even cause congestion and safety accidents. This algorithm can achieve fast detection while ensuring high accuracy, meeting the needs of real-time security checks and providing passengers with more efficient and convenient security services.

Table 3: Time Comparison of Different Algorithms

Method	parameter count/($\times 10^6$)	model size /(Mbyte)	Detection Time /ms
YOLOv5s	7.05	13.73	2.54
R-CNN	7.05	14.09	2.52
Yolo v11	7.07	13.86	2.51
proposed algorithm	7.07	14.21	2.50

Table 4 presents a comparison between our algorithm and five other algorithms, including a comparison of time and accuracy. From the results in Table 4, it can be seen that the algorithm proposed in this paper not only takes much less time than other algorithms, but also has higher average accuracy than other algorithms. Specifically, the accuracy is 38.7 percentage points higher than Inception V3.

From the data presented in Table 4, a comprehensive and in-depth comparative analysis can be conducted between our algorithm and the other five algorithms in key indicators such as parameter quantity, model size, detection time, and mean accuracy (mAP). In terms of parameter count, InceptionV3 has a parameter count of up to 24.65 × 10 ⁶, VGG19 has a parameter count of 45.09 × 10 ⁶, DenseNet has a parameter count of 57.01 × 10 ⁶, CNN-ABiGRU has a parameter count of 14.34 × 10 ⁶, OctConv ABiGRU has a parameter count of 121.47 × 10 ⁶, while our algorithm only has a parameter count of 7.07 × 10 ⁶. Compared with other algorithms, this paper significantly reduces the number of algorithm parameters. Excessive parameter count can increase the computational complexity of model training and inference, require higher hardware resources, and easily lead to overfitting. The lower parameter count of the algorithm in this article means that its model complexity is lower. While ensuring performance, it is easier to deploy on devices with limited resources, such as some small security devices, and has stronger practicality and adaptability. In terms of model size, InceptionV3 is 107Mbyte, VGG19 is 343Mbyte, DenseNet is 436Mbyte, CNN-ABiGRU is 107Mbyte, OctConv ABiGRU is as high as 1381Mbyte, while our algorithm is only 37Mbyte. The size of the model directly affects the storage and transmission costs of the algorithm in practical applications. Larger models require more storage space and consume more bandwidth and time during transmission. The smaller model size of the algorithm in this article makes it easy to deploy in storage resource constrained environments and more efficient in network transmission, reducing the cost and difficulty in practical applications. Detection time is a key indicator for measuring the real-time performance of algorithms. The detection time of InceptionV3 is 0.11ms, VGG19 is 41.54ms, DenseNet is 24.87ms, CNN-ABiGRU is 75.09ms, OctConv ABiGRU is 36.74ms, and the algorithm in this paper is only 0.0023ms. In practical security check scenarios, especially in places with high passenger flow such as stations and airports, fast detection is crucial. A longer detection time can cause passengers to queue up, reduce security check efficiency, and even lead to congestion and safety accidents. The algorithm in this article has a very short detection time, which can meet the needs of real-time security check, achieve fast and efficient dangerous goods detection, and provide passengers with a more convenient security check experience. Mean Precision (mAP) is an important

indicator for measuring the accuracy of algorithm detection. The mAP of InceptionV3 is 54.54%, VGG19 is 81.46%, DenseNet is 78.48%, CNN-ABiGRU is 87.69%, OctConv ABiGRU is 91.32%, and our algorithm achieves 93.21%. Compared with other algorithms, this algorithm has significant advantages in mAP. Especially when compared with InceptionV3, the accuracy is 38.7 percentage points higher. This indicates that the algorithm proposed in this article can more accurately identify targets when detecting smuggled dangerous goods hidden in people's bodies, reduce false positives and omissions, and greatly improve the accuracy and reliability of security checks. Overall, the algorithm in this article performs well in terms of parameter quantity, model size, detection time, and average accuracy. Compared to other algorithms, this algorithm achieves extremely short detection time and high average accuracy while ensuring low parameter count and small model size. This makes the algorithm in this article more feasible and practical in practical applications, providing more efficient and accurate solutions for security checks and effectively ensuring the safety of passengers.

Table 4: Comparison of Accuracy and Time between Our Algorithm and Other Algorithms

Method	parameter count/($\times 10^6$)	model size /(Mbyte)	Detection Time /ms	mAP
InceptionV3	24.65	107	0.11	54.54
VGG19	45.09	343	41.54	81.46
DenseNet	57.01	436	24.87	78.48
CNN-ABiGRU	14.34	107	75.09	87.69
OctConv-ABiGRU	121.47	1381	36.74	91.32
proposed algorithm	7.07	37	0.0023	93.21

6 Conclusion

This article focuses on the problem of low recognition rate of smuggled dangerous goods in real security check scenarios, as well as the shortcomings of traditional image classification and existing deep learning algorithms in the field of security check. A series of studies have been conducted with the aim of proposing an effective recognition algorithm to improve security check efficiency and accuracy, and ensure passenger safety. In the research process, this article fully utilizes the advantages of deep convolutional neural networks and proposes a deep learning network-based algorithm for identifying dangerous goods in human body binding, smuggling, and security checks. This algorithm innovatively combines infrared images with VI, uses image segmentation and unsupervised registration techniques, and uses a self-encoding deep learning network as the registration network skeleton for human contour registration learning. By measuring the similarity of registered images and minimizing the similarity cost, the network is continuously optimized to achieve efficient and accurate identification of smuggled dangerous goods hidden in human bodies. The experimental results show that on the public dataset SIXray, compared with other algorithms, our algorithm achieves an average accuracy (mAP) of 93.21%, which is higher than other comparison algorithms, especially 38.7 percentage points higher than Inception V3, with small increases in parameter quantity and model size and almost unchanged detection time. This result indicates that the algorithm proposed in this paper performs well in terms of parameter quantity, model size, detection time, and average accuracy, and has higher feasibility and practicality. It can provide efficient and accurate solutions for security inspection work.

Although the algorithm proposed in this article has achieved significant results in the detection of human trafficking, there are still some directions for further research and

improvement. Future research can further optimize algorithm structures, reduce model complexity and computational complexity, and improve algorithm efficiency; Explore more image fusion techniques, fully utilize the information of different modal images, and further improve the accuracy and robustness of recognition; Conduct application research in practical scenarios, apply the algorithm to more security check scenarios, and verify its performance and effectiveness in actual environments.

Funding

This work was supported by Research Project of the General Administration of Customs (2024HK199).

Author's Profile

Ying Chen was born in Wuhan, Hubei, China, in 1982. He obtained a bachelor's degree from Huazhong University of Science and Technology in China. He is currently working at the Wuhan Guide Infrared Co., Ltd. His main research areas are Sensor Automation Control Technology.

Jun Li was born in Yaan, Sichuan, China, in 1975. He obtained a bachelor's degree from Zhejiang University in China. He is currently working at the Shenzhen Customs Information Center. His main research areas are electronic information technology and intelligent customs inspection equipment.

Zeliang Luo was born in Wuhan, Hubei, China, in 1987. He obtained a bachelor's degree from Huanggang Normal University in China. He is currently working at the Wuhan Guide Infrared Co., Ltd. His main research areas are Sensor Automation Control Technology.

Ming Hui was born in Qinhuangdao, Hebei, China.He currently works at Shenzhen Customs. His core research directions are electronic information technology and intelligent customs inspection equipment, and he focuses on improving the efficiency of customs inspection and the accuracy of supervision through technological optimization.

References

[1] Park T J, Kim K, Moon S. Securing Infrared Communication in Nuclear Power Plants: Advanced Encryption for Infrared Sensor Networks[J]. Sensors, 2024, 24(7): 2054.

[2] Khor W L, Chen Y K, Roberts M, et al. Automated detection and classification of concealed objects using infrared thermography and convolutional neural networks[J]. Scientific reports, 2024, 14(1): 8353.

[3] Kim S, Lim J, Kim S, et al. Near-Infrared Luminescent Imaging-Based 3D QR Cube Platform for Spatial Information Storage and Security[J]. Advanced Materials, 2025, 37(6): 2416121.

[4] Khor W L, Chen Y, Roberts M, et al. Enhanced visualisation of concealed target objects by infrared thermography and machine learning[J]. Infrared Physics & Technology, 2025: 106186.

[5] Amuta E O, Sobola G O, Eseabasi O, et al. Motion detection system using passive infrared

technology[C]//IOP Conference Series: Earth and Environmental Science. IOP Publishing, 2024, 1342(1): 012001.

[6] Blancaflor E B, Pulgar J R B, Chan J R S, et al. BMAT (Baggage Management Against Theft): An Automated Baggage Management System Using IoT Sensors with QR Code Verification Against Theft[C]//Proceedings of the 2024 6th Blockchain and Internet of Things Conference. 2024: 1-8.

[7] D'Accardi E, Dell'Avvocato G, Masciopinto G, et al. Evaluation of typical rail defects by induction thermography: experimental results and procedure for data analysis during high-speed laboratory testing[J]. Quantitative InfraRed Thermography Journal, 2025, 22(2): 173-194.

[8] Jayachitra J, Khan A A, Yuvaraj V. IoT-Driven Robotic System for Enhanced Nighttime Perimeter Security and Real-Time Surveillance in Smart Environments[C]//2025 International Conference on Visual Analytics and Data Visualization (ICVADV). IEEE, 2025: 440-445.

[9] Didier P, Zaminga S, Spitz O, et al. Data encryption with chaotic light in the long wavelength infrared atmospheric window[J]. Optica, 2024, 11(5): 626-633.

[10] Charsley J M, Farrell C, Rutkauskas M, et al. Mid-infrared optical coherence tomography with a stabilized OP-GaP optical parametric oscillator[J]. Optics Letters, 2024, 49(11): 2882-2885.

[11] Wredh S, Dai M, Hamada K, et al. Sb2Te3–Bi2Te3 direct photo–thermoelectric mid-infrared detection[J]. Advanced Optical Materials, 2024, 12(31): 2401450.

[12] Yadav S, Tomar M, Singhal T, et al. Near-infrared reflectance spectroscopy (NIRS): An innovative, rapid, economical, easy and non-destructive whole grain analysis method for nutritional profiling of pearl millet genotypes[J]. Journal of Food Composition and Analysis, 2025, 142: 107373.

[13] Kaur S, Singh N, Sharma P, et al. Optimizing protein content prediction in rice bean (Vigna umbellata L.) by integrating near-infrared reflectance spectroscopy, MPLS, deep learning, and key wavelengths selection algorithms[J]. Journal of Food Composition and Analysis, 2024, 135: 106655.

[14] Abdelhakim M M, Khalil A A I, Salah A, et al. Exploring the impact of high-power infrared lasers on electro-optical systems performance: A field study with different wavelengths[J]. Infrared Physics & Technology, 2024, 139: 105348.

[15] Feick M, Tang X, Garcia-Martin R, et al. Imprinto: Enhancing Infrared Inkjet Watermarking for Human and Machine Perception[C]//Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025: 1-18.

[16] Nguyen T H, Tran T T T. Synthesis of near-infrared absorbing materials based on copper phosphate compounds[J]. Bulletin of Materials Science, 2024, 47(4): 283.

[17] Gazzano O, Chambon M, Ferrec Y, et al. Long-Term Radiometric Stability of Uncooled and Shutterless Microbolometer-Based Infrared Cameras[J]. Sensors, 2024, 24(19): 6387.

[18] Pissard A, Gofflot S, Baeten V, et al. Chitin Assessment in Insect-Based Products from Reference Methods to Near-Infrared Models[J]. Insects, 2025, 16(9): 924.

[19] Abo-Zahhad M M, Abo-Zahhad M. Smart Parking System in Smart Cities Based On the Internet of Things and Machine Learning[C]//2024 International Telecommunications Conference (ITC-Egypt). IEEE, 2024: 190-195.

[20] Vejdanivahid S, Salehi F. Enhancing the Quality and Nutritional Properties of Gluten-Free Pancakes Using Sprouted Quinoa Flour Treated With Magnetic Fields, Ultrasound, and Infrared Drying[J]. Food Science & Nutrition, 2025, 13(7): e70502.

[21] Singh K, Aalam U, Mishra A, et al. Spectroscopic and imaging considerations of THz-TDS and ULF-Raman techniques towards practical security applications[J]. Optics Express, 2024, 32(2): 1314-1324.

[22] Reinegger R D, Bissessur P, Meereechpersad I, et al. Improving primate detection using thermal infrared imaging: availability and observer errors in drone surveys of mixed tropical forests[J]. International Journal of Remote Sensing, 2025, 46(11): 4345-4373.

[23] Shokrollahi A, Persson J A, Malekian R, et al. Passive infrared sensor-based occupancy monitoring in smart buildings: a review of methodologies and machine learning approaches[J]. Sensors, 2024, 24(5): 1533.

[24] Khan M, Aljuaydi F, Said L, et al. A secure chaotic cryptosystem for thermal Imaging: Logistic map-based encryption with substitution-diffusion and spatial decorrelation[J]. Journal of Radiation Research and Applied Sciences, 2025, 18(2): 101546.

[25] Elhussein M, Almuhaideb A, Alholyal F, et al. Efficient Entry: A Stateful Authentication Approach in Health-Aware Smart Gate Systems[J]. IEEE Access, 2024, 12: 70634-70645.

[26] Argirusis N, Achilleos A, Alizadeh N, et al. IR Sensors, Related Materials, and Applications[J]. Sensors, 2025, 25(3): 673.

[27] Zinna F, Botta C, Luzzati S, et al. Near-Infrared Circularly Polarized Electroluminescence with Switchable Handedness in Organic LEDs[J]. Advanced Functional Materials, 2025, 35(22): 2423077.

[28] Chaukhande P, Luthra S K, Patel R N, et al. Development and validation of near-infrared reflectance spectroscopy prediction modeling for the rapid estimation of biochemical traits in potato[J]. Foods, 2024, 13(11): 1655.

[29] Žiljak Gršić J, Bogović T, Plehati S, et al. New Integrations of Designers' Solutions for Images in the VIS/NIR Spectral Area[C]//Intelligent Systems Conference. Cham: Springer Nature Switzerland, 2024: 101-108.

[30] Mishra R K, Mondal A, Mathew J. Nystromformer based cross-modality transformer for visible-infrared person re-identification[J]. Scientific Reports, 2025, 15(1): 16224.