remote sensing

MDPI

Article

A Suspicious Multi-Object Detection and Recognition Method for Millimeter Wave SAR Security Inspection Images Based on Multi-Path Extraction Network

Minghui Yuan, Quansheng Zhang, Yinwei Li *, Yunhao Yan and Yiming Zhu

Terahertz Technology Innovation Research Institute, University of Shanghai for Science and Technology, Shanghai 200093, China; yuanminghui@usst.edu.cn (M.Y.); 193730618@st.usst.edu.cn (Q.Z.); 202330364@st.usst.edu.cn (Y.Y.); ymzhu@usst.edu.cn (Y.Z.)
* Correspondence: liyw@usst.edu.cn

Abstract: There are several major challenges in detecting and recognizing multiple hidden objects from millimeter wave SAR security inspection images: inconsistent clarity of objects, similar objects, and complex background interference. To address these problems, a suspicious multi-object detection and recognition method based on the Multi-Path Extraction Network (MPEN) is proposed. In MPEN, You Only Look Once (YOLO) v3 is used as the base network, and then the Multi-Path Feature Pyramid (MPFP) module and modified residual block distribution are proposed. MPFP is designed to output the deep network feature layers separately. Then, to distinguish similar objects more easily, the residual block distribution is modified to improve the ability of the shallow network to capture details. To verify the effectiveness of the proposed method, the millimeter wave SAR images from the laboratory's self-developed security inspection system are utilized in conducting research on multi-object detection and recognition. The detection rate (probability of detecting a target) and average false alarm (probability of error detection) rate of our method on the target are 94.6% and 14.6%, respectively. The mean Average Precision (mAP) of recognizing multi-object is 82.39%. Compared with YOLOv3, our method shows a better performance in detecting and recognizing similar targets.

Keywords: SAR image; deep learning; object recognition; multipath feature pyramid (MPFP); residual block distribution

check for updates

Citation: Yuan, M.; Zhang, Q.; Li, Y.; Yan, Y.; Zhu, Y. A Suspicious Multi-Object Detection and Recognition Method for Millimeter Wave SAR Security Inspection Images Based on Multi-Path Extraction Network. *Remote Sens.* **2021**, *13*, 4978. https://doi.org/10.3390/rs13244978

Academic Editor: Dusan Gleich

Received: 9 November 2021
Accepted: 6 December 2021
Published: 8 December 2021

1. Introduction

With the development of millimeter wave imaging technology, the millimeter wave security scheme is becoming more and more sophisticated [1–3]. The current milli-meter wave imaging systems can be divided into two categories: passive imaging and active imaging. Passive imaging systems use the object's own radiation for imaging. Because the object's radiation power is meager, the imaging effect is terrible. Active imaging systems use the reflection characteristics of objects for imaging. Since the power of the emission source is high, the reflected radiation can obtain a clear image of the human body. Upon the introduction of the concept of Multiple-Input Multiple-Output (MIMO), SAR imaging-based security inspection systems are also in development, which greatly improves the imaging accuracy and rate [4–9]. Different from X-ray, active millimeter-wave security equipment does not cause harm to the human body [10].

At present, there are two scanning methods for SAR imaging security inspection systems. One is circular scanning, in which the linear array antenna is placed on a vertical plane. Although the circular scanning system has a better imaging effect on the target, it has the disadvantages of slow imaging speed and complex mechanical and electrical structure. The flat scanning system places the antenna on a horizontal plane to scan vertically, contributing to a fast-scanning speed and a simple mechanical and electrical structure. However, the antenna distance from the human body is equal at any given moment, so the flat scanning system has different imaging clarity at different human body positions.

With the availability of more sophisticated millimeter wave imaging systems, the conditions for target detection in millimeter wave images are facilitated. Yeom et al. [11] extract geometric feature vectors from Principal Component Analysis (PCA) transform target shapes and use the geometric feature vectors to analyze the segmented binary images for target recognition. This method uses traditional image processing methods, and the detection target is limited.

In recent years, deep learning has been widely used in target detection from SAR images. Chen et al. [12,13] use multi-scale fusion to enhance the method of extracting features to achieve a high accuracy detection of bridges and aircrafts. Meng et al. [14] propose a human pose segmentation algorithm based on deep Convolutional Neural Network (CNN) detection, which enables human images to be divided into several parts for recognition. Lopez-Tapia et al. [15] enhance the passive millimeter wave image to improve the detection rate of the target. Guo et al. [16] use human contour segmentation combined with deep learning to achieve the detection of targets in passive millimeter wave images. Liu et al. [17] analyze millimeter wave imaging data, then extract a familiar millimeter wave image and a new spatial depth map for further detection. There is only one detect object in these studies.

Del Prete et al. [18] used the region-based convolutional neural network to detect ship wakes. Because some ship wakes are similar to water surface membranes, the mAP is only 67.63%. Ghaderpour et al. [19] studied the various frequency and time-frequency decomposition methods based on Fourier and least square analysis. Arivazhagan et al. [20] used wavelet transform and Fourier transform for target/change detection. The wavelet-based algorithms proved to be successful in detecting changes/targets within images.

Most of the existing multi-object recognition is the recognition of a human body and a hidden object. Pang et al. [21] use the YOLO v3 algorithm for the real-time detection of human concealed metal weapons from passive millimeter wave images. They mainly identify pistols and people. The shapes of the two targets are different, so the deep learning network can readily identify them. Zhang et al. [22] propose an improved Faster Region-Based Convolutional Neural Network (R-CNN) for target detection from the millimeter wave image of a circular scanning system. The mean Average Precision of the improved network reaches 69.7%. The human body in their image occupies more than 50% of the pixel area, and hidden objects are placed inside the human body. This recognition has a distinctive feature: the recognition rate of the human body is very high, which can reach 98.75%, but the recognition rate of hidden objects is very low, such as the recognition rate of mobile phones is only 47.18%. This mAP cannot objectively express the recognition rate of hidden objects. This article does not consider the human body as the detection target, but only recognizes targets with inconsistent clarity and similar targets, such as a pistol, hammer, wrench.

In this article, a suspicious multi-object detection and recognition method for millimeter wave SAR security inspection images based on multi-path extraction network is proposed. We overcome the problem of inconsistent definition targets and similar targets difficult to identify. We use a flat scanning imaging system to obtain a SAR image of the human body. The imaging focal length of the flat scanning system is a fixed value, but the body surface is not flat. Due to the change of the placement position of the target, the target area will be in different imaging distances, and the scanning imaging will have inconsistent clarity on the same target. Targets with inconsistent clarity are at least 1/4 of the targets blurred, and the blurred area is not limited to the boundary. At the same time, SAR imaging is often a grayscale image with a single color. The identification of dangerous articles mainly depends on the shape of the object. To the best of our knowledge, most of the current research focuses on detecting hidden objects in the human body, but few studies identify the names of hidden objects. This article combines the imaging characteristics to solve the interference problem of inconsistent object clarity and objects similar, and a multi-path extraction network for multi-target recognition of SAR imaging security inspection images is proposed.

The main research of this article is as follows:

(1) A suspicious multi-object detection and recognition method based on multi-path extraction network (MPEN) is proposed for millimeter wave SAR security inspection images. Combining the output characteristics of the deep and shallow networks, it realizes the recognition of inconsistent sharpness targets and similar targets.

(2) A Multi-Path Feature Pyramid (MPFP) model and an improved residual block distribution are proposed. We use MPFP to output the semantic information of the deep network independently. The feature map contains the semantics of a deep independent network, which enhances the feature extraction of different targets and better distinguishes different targets. At the same time, we adjust the number of repetitions of the residual block, and we focus on using the receptive field of the shallow network. The modified residual block distribution helps distinguish the differences in the contours of different targets and improve the recognition accuracy.

(3) This article provides a new idea for automatically identifying multi-objects from the SAR imaging of security inspection systems.

The structure of this article is as follows. In Section 2, the main problems of multi-target recognition are introduced. In Section 3, the MPEN of millimeter wave SAR security inspection images proposed in this paper is described in detail. In Section 4, the experimental results and corresponding image analysis under different conditions are described. In Section 5, the specific differences between our method and existing research are discussed. Finally, the research conclusions are given in Section 6.

2. SAR Imaging System of Security Inspection

At present, different security inspection systems with SAR imaging have different imaging effects, and there is no relevant dataset. Therefore, this study is based on the laboratory self-developed security inspection system with SAR imaging. The system structure is shown in Figure 1. The system places all sources and detectors in parallel on the X axis, and scans up and down along the Z axis. The operating frequency of the system is 35 GHz. Our security inspection system has a simple mechanical structure, and the scanning structure can complete scanning in one cycle of the Z-axis movement. The system can run continuously, scanning imaging is saved in jpg format, and the ratio of each image is fixed at 200×400 pixels.

In principle, the imaging system is a near-field imaging radar. The intensity of the millimeter waves reflected by metal contraband is higher than that of the human body. The contraband can be seen as a brighter area in the imaging. In the datasets, the target placement position is the center and the edge of the human body. The recognition target is set to common contraband, such as pistols, hammers, and wrenches.

When the flat scanning system detects the target close to the body surface, the target image will be deformed at a certain angle. At the same time, the distance between the target and the scanning antenna is not a fixed value. Because the imaging focal length is fixed, the target close to the human body surface will be blurred in some areas. Take the pistol imaging result as an example. The imaging is shown in Figure 2. The pistol is placed on the edge of the body, and the pistol image is clear, as shown in Figure 2a. When the pistol is placed on the surface of the human body, the outline of the pistol merges with the outline of the human body, and only the area of the pistol grip is clear, as shown in Figure 2b.

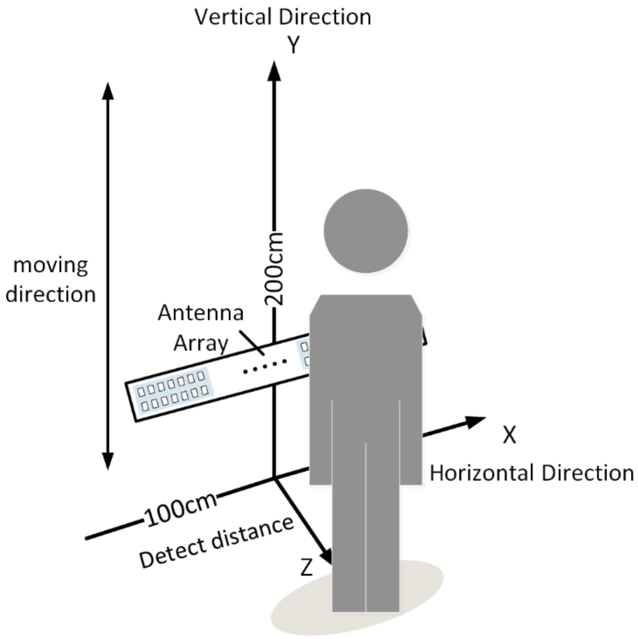


Figure 1. System structure diagram.

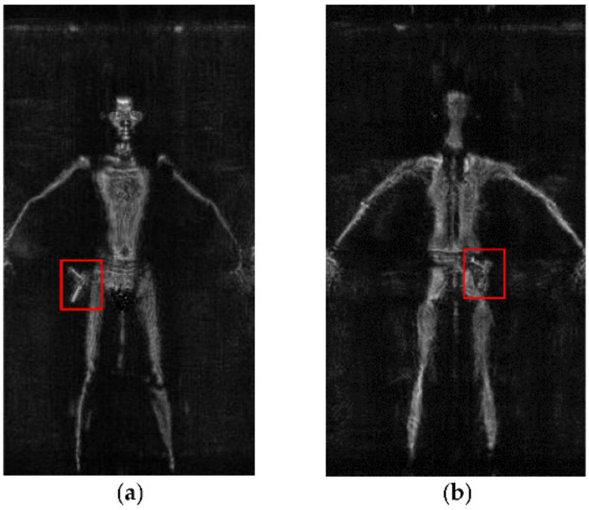


Figure 2. Body image. (**a**) Target position 1, (**b**) Target position 2.

In addition, this article mainly uses hammers and wrenches to identify similar targets. As shown in Figure 3, both the hammer and wrench have a long handle. The handles of the two targets are similar, and they are similar to the human limbs. The difference is that the hammer has a head, and the wrench has an opening. Therefore, this article is aimed at the characteristics of the flat scanning millimeter wave imaging system, and a suspicious multi-object detection and recognition method for millimeter wave SAR security inspection images based on multi-path extraction network is proposed. Our method improves the recognition rate of targets with inconsistent clarity and similar targets in SAR images.

The data used in this experiment is a laboratory-developed security inspection system with 1000 SAR images, 90% of which are used for training and 10% for testing. We consider using most of the dataset for training, and the test dataset contains all actual recognition situations. We have three recognition targets: a pistol, wrench, and hammer. We produced three types of datasets: inconsistent clarity of objects, similar objects, and mixed targets (including inconsistent clarity of objects and similar objects, at the same time). Each category is divided into two situations: the target is placed on the body's trunk, and the target is placed on the limbs. We use a method of including multiple recognition targets in one image to make up for the problem of a small dataset. Although the number of test images is slight, many targets need to be identified, thereby ensuring the reliability of the test results. The images in the dataset are labeled using LabelImg, with one to three targets in each image, as shown in the figure after LabelImg. The target is placed at a random position on the human body, and the training image shown in Figure 4 is when the target is placed at the thigh position. The targets in all test images are randomly placed, and the number of targets in the test images ranged from one to three. The object of this research is to identify the inconsistent clarity of objects and similar targets. It is required that the targets in the dataset do not block each other. Therefore, when placing the target on the human torso, the number of objects placed generally does not exceed three. When the target is placed on the limbs, the number of places is generally one. All test images are used as independent tests to verify the network performance, and the test images are all 200 × 400 pixels.

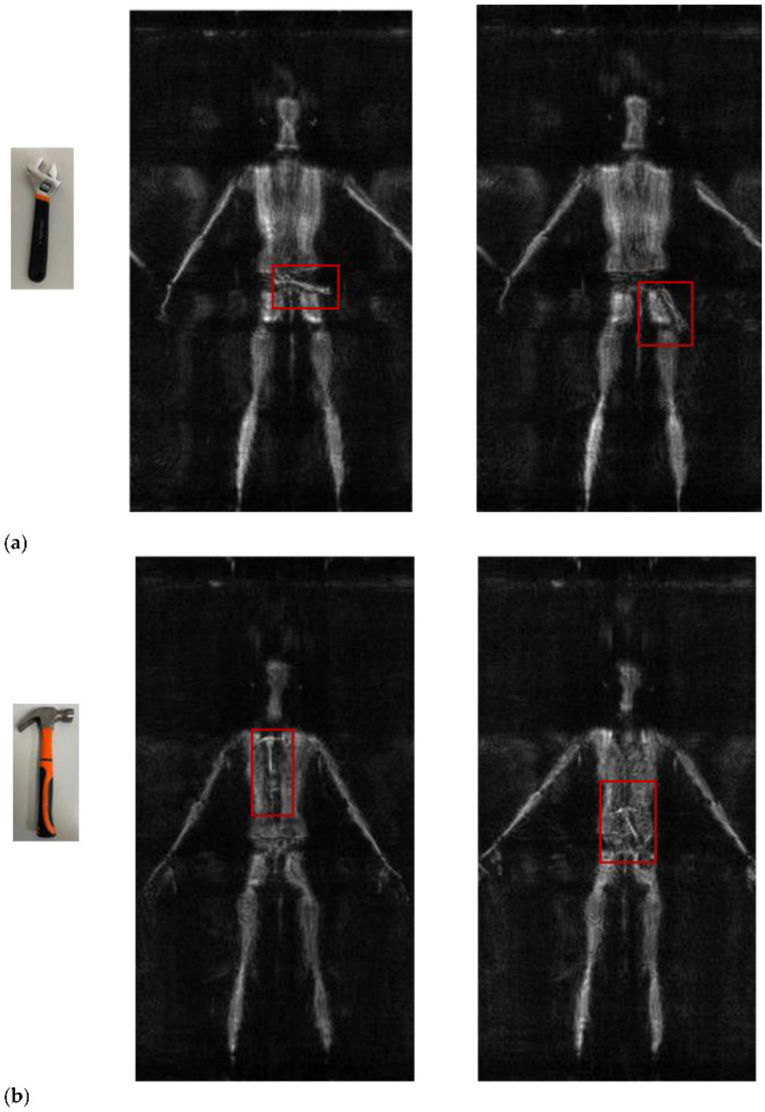


Figure 3. Image comparison of similar targets (**a**) Wrench imaging effect diagram. (**b**) Hammer imaging effect diagram. (The red box is the target).

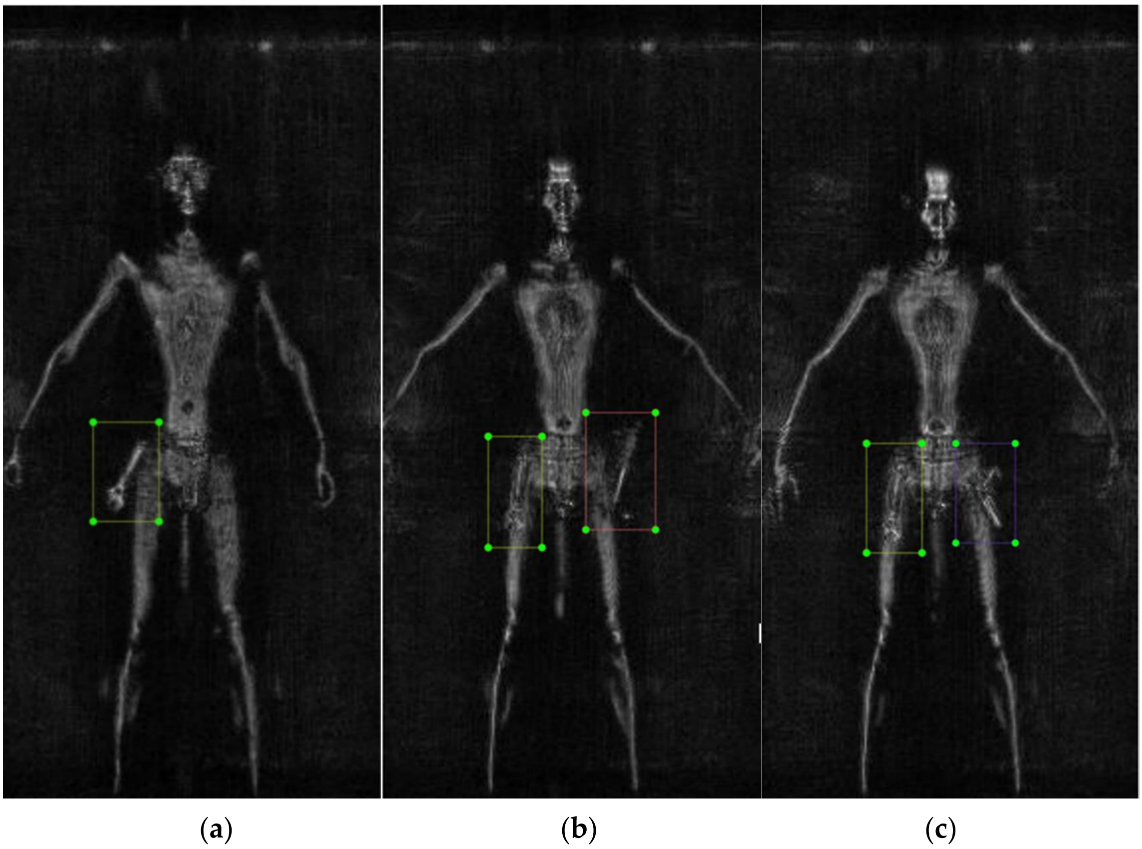


Figure 4. LabelImg images. (**a**) Wrench image marking diagram. (**b**) Wrench and hammer image marking diagram. (**c**) Wrench and pistol image marker diagram.

3. Methods

3.1. Overall Network Framework

If the security inspection system wants to identify multi-object, it is crucial to identify targets with similar appearances accurately. SAR images are mainly grayscale images, so the color characteristics of dangerous goods are not prominent. Therefore, the requirement for the recognition ability of the target contour is relatively high. Moreover, some areas of the same object in the image are clear, and some are blurred. It is necessary to improve the ability of contour extraction and the interference of speckle noise. This article proposes the method based on the YOLO v3 backbone network. First, propose a Multi-Path Feature Pyramid (MPFP) module and modify the number of repetitions of residual blocks. MPFP outputs the deep network separately, and the shallow network is combined with the deep network to output. The feature scale of the deep network is only affected by the output of the current residual block, and the output feature of the deep network is not affected. In the Darknet-53 network, adjusting the number of residual block repetitions in the shallow and deep networks allows the network to generalize to different angles, scales, or new objects. The network structure is shown in Figure 5.

3.2. YOLO v3 Backbone

In the YOLO v3 algorithm, an image is divided into S × S grids. If the target object to be detected exists in a particular grid or some grids, then these grids are responsible for detecting the target [23]. YOLO v3 uses three scales of feature maps for target detection. Each feature map uses anchor boxes as priori boxes, borrowing the idea of the anchor from Faster RCNN. It is a set of candidate boxes with a fixed aspect ratio, which is equivalent to a set of templates, with which the subsequent detection is performed [24]. Then, a predicted bounding box covering the entire feature map is generated on the feature map, and finally, the generated predicted bounding box is subjected to regression processing. The recognition process of YOLO v3 is shown in Figure 6.

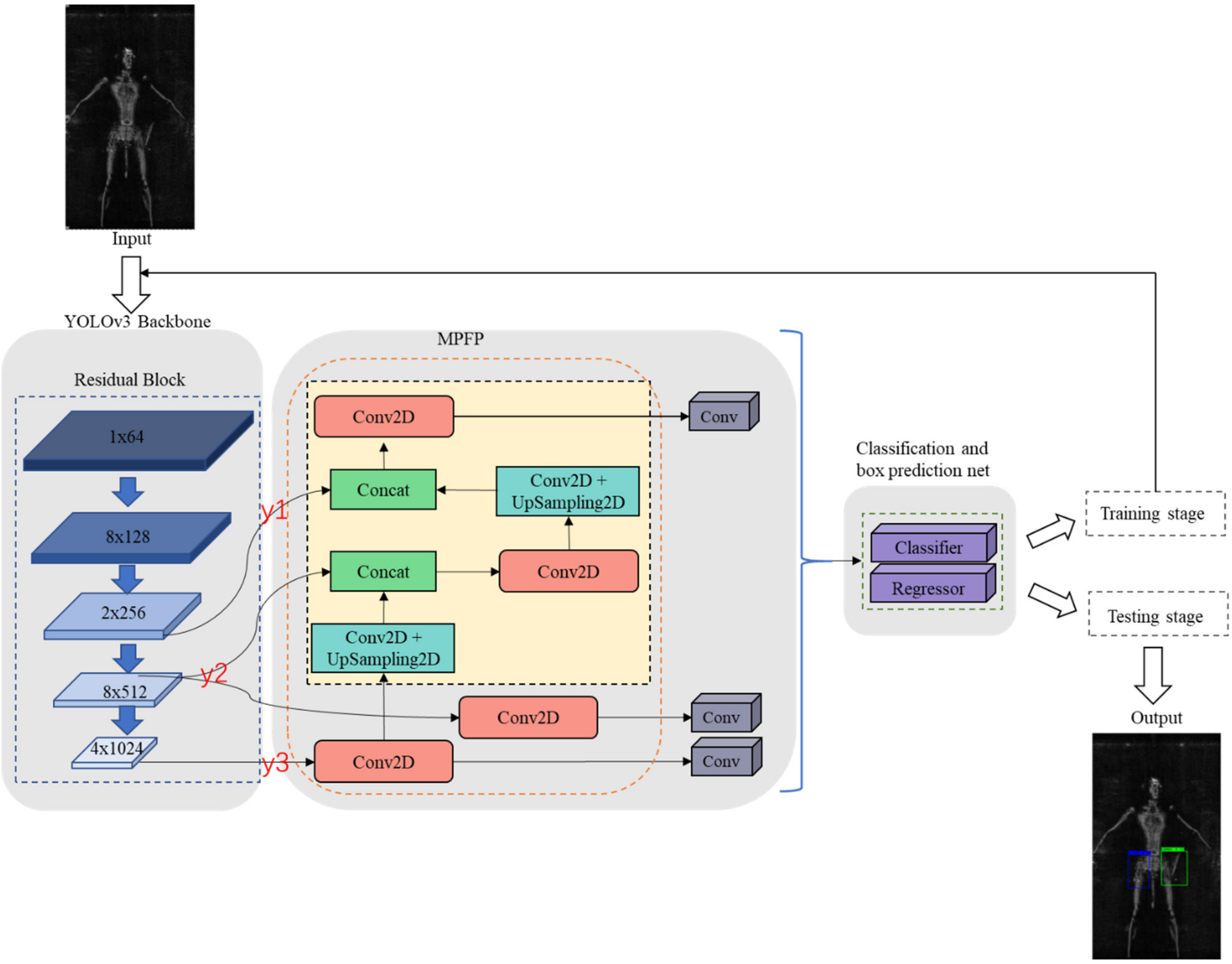


Figure 5. Network structure diagram.

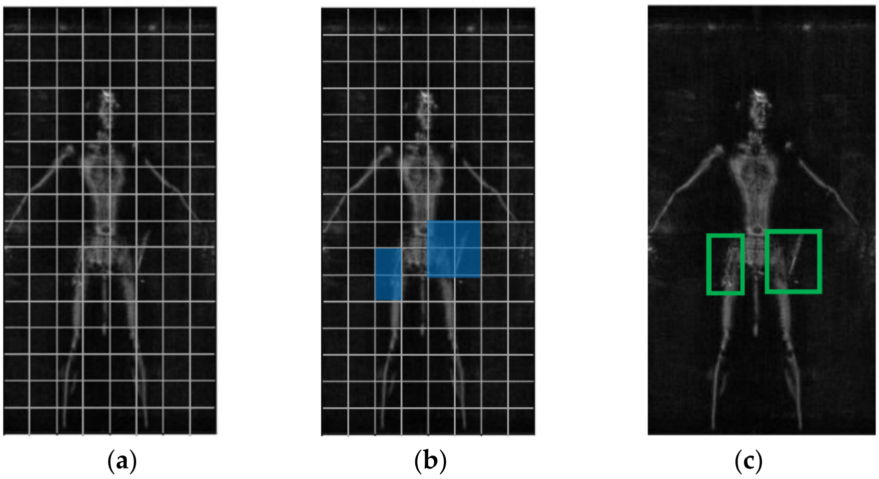


Figure 6. Recognition process of yolov3. (**a**) Grid image. (**b**) Classification and box prediction net. (**c**) Result image.

The first 52 layers in Darknet-53 are used for feature extraction, and the last layer is the output layer. The specific structure follows: first is a convolution kernel with 32 filters, then 5 sets of repeated residual units, and finally, the output layer outputs the results. There are 5 groups of residual units, and each unit is composed of a separate convolutional layer and residual block. The residual block is repeated 1, 2, 8, 8, and 4 times. In each residual block, a 1 × 1 convolution operation is performed first, and then a 3 × 3 convolution operation is performed. The number of filters is first halved and then restored to the original number.

3.3. Multi-Path Extraction Network (MPEN)

3.3.1. Multi-Path Feature Pyramid (MPFP) Module

In deep learning, the receptive field usually refers to the size of the corresponding area of the feature pixel in the image on the input image. As shown in Figure 7, convolutional neural networks are often composed of multiple convolutional layers, and the receptive field sizes of neurons in different convolutional layers are also different from each other. The deeper the convolutional layer is, the larger the receptive field is. Since shallow neurons pass through a small number of convolutional layers, the receptive field that is mapped to the original image is relatively small. Therefore, the extracted features contain more detailed information, such as the contour and color of the target in the image. Deep layer neurons have undergone more complicated calculations than shallow layer neurons, and the receptive field mapped on the original image is more extensive. Therefore, the extracted features are more abstract, and more detailed information is lost [25–30].

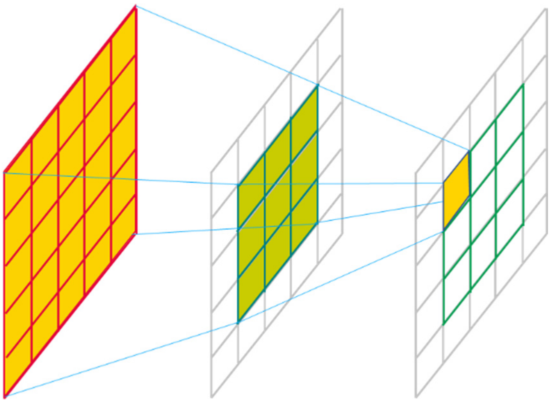


Figure 7. Receptive filed.

The SAR image is the gray image. Next in the process is to identify multi-object hidden on the surface of the human body, relying mainly on the shape of the target profile. Generally speaking, the feature map of the shallow network in the convolutional neural network has a solid ability to respond to the details of the target. Therefore, the feature map of the shallow network has a better advantage for detecting the detailed information of the target. The deep layer network responds to the abstract semantic information, which loses more detailed features but expresses the features better. The more extensive contour information can identify a recognition target when the deep network is combined with the shallow network. In contrast, similar targets can be better distinguished based on detailed information.

In the YOLOv3 network, three features of y1, y2, and y3 are output, as shown in Figure 5. The final output of y2 is the fusion of the y2 feature layer and the y3 feature layer. The advantage of this linking method is that it can identify the target better. However, due to the fusion and utilization of the two deep network feature maps, the deep network overuses the deep semantics, resulting in the poor discrimination of target types. A Multi-Path Feature Pyramid (MPFP) module is proposed, as shown in the MPFP section of Figure 5. Since the detected target types are more than two and the target similarity is high, the y2 and y3 scale features are used for direct output. The purpose is to use the rich expressive power of the deep network to obtain the characteristic differences between multi-object. The y1 scale is then output after fusing with y2 and y3 scales. The shallow network has a small receptive field and a solid ability to capture the original features of the image. Next, we let the network learn the characteristics of the shallow network and the deep network simultaneously, and the expression effect of the network is better. Ultimately, MPFP increases the connection path, only adds a small amount of calculation, and improves the extraction ability of the target feature.

3.3.2. Residual Block Modification

Our method is modified from the YOLO v3 feature extraction network Darknet-53, as shown in Figure 8. The residual block arrangement is modified to the 1, 8, 2, 8, 4 structure on the original 1, 2, 8, 8, 4 structure. After the numbers of the second and third residual blocks are exchanged, the number of repetitions when the tensor dimension is 52 × 52 × 256 is reduced, and the number of repetitions when the tensor dimension is 104 × 104 × 128 is increased. The effect is to increase the number of feature map channels of the shallow network and extract the semantic information of the shallow layer. Decreasing the feature map size of the deep network extracts abstract semantic information. Finally, with MPFP, the feature maps contain deep and shallow multiple semantic information to enrich the multi-scale feature maps of the prediction recognition frame.

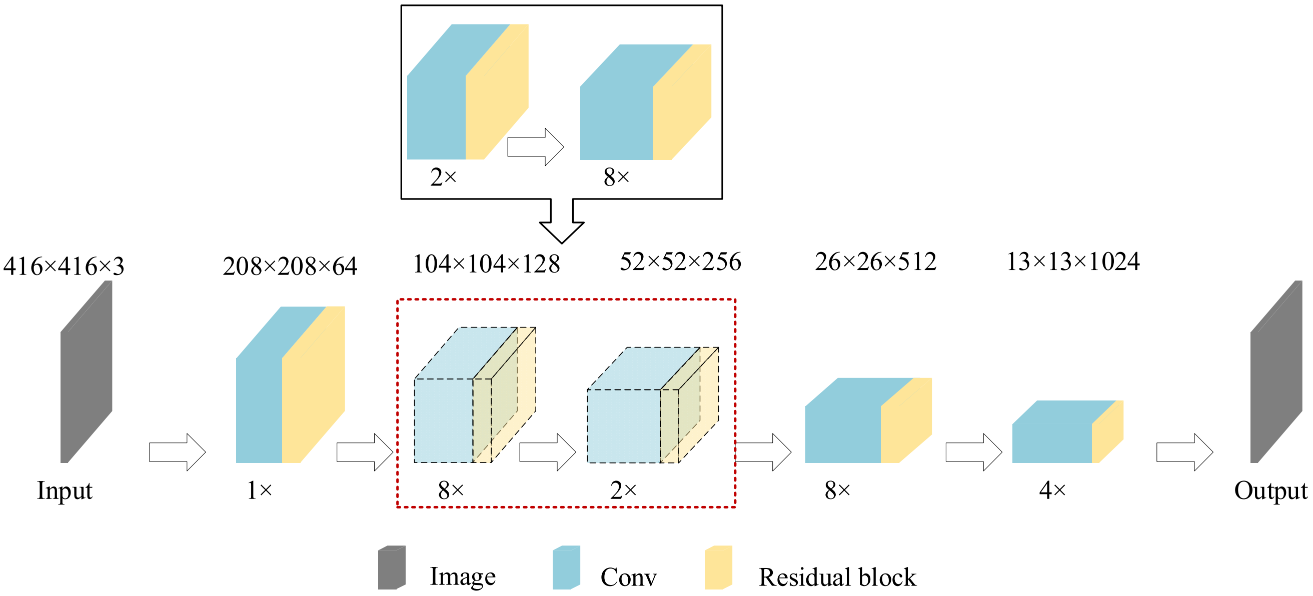


Figure 8. Residual block. (The black box is the original residual block, and the red box is the modified residual block).

3.4. Classification and Forecasting

After the target features are extracted, MPEN outputs a valid prediction feature map at three scales, where the grid is divided into 52 × 52, 26 × 26 and 13 × 13. For each grid center, multiple prior boxes are created. Further, the prediction result of the network determines whether these boxes contain the target and the type of target. However, this prediction result does not correspond to the final prediction box on the image, as shown in Figure 9. The final result is then obtained by score sorting with Non-Maximum Suppression (NMS) filtering.

3.5. Implementation Details

The network is trained on Google COLAB with the deep learning framework Pytorch 1.8.0, using the VOC2007 dataset, with 300 iterations. Freezing training can speed up the training and prevent the weights from being destroyed in the early stage of training, so the first 50 iterations of freezing training are used. The Learning Rate (LR) of the first

50 iterations is 1 × 10^{-3}, and the batch size 8, the learning rate of the 51st to 300th iterations is 1 × 10^{-4}, and the batch size 4.

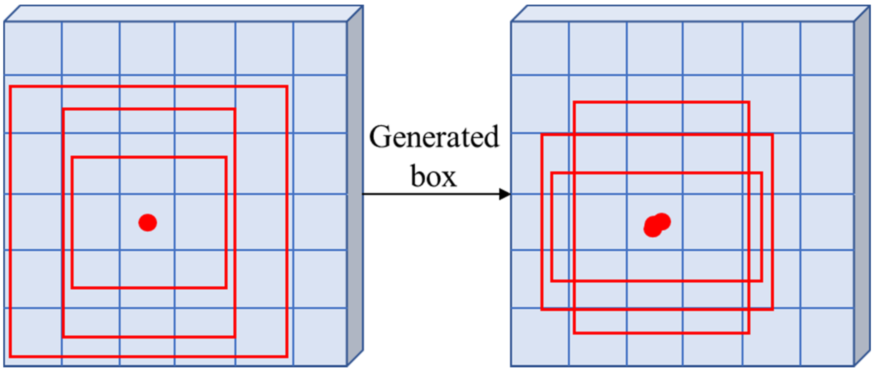


Figure 9. Priori box.

3.6. Evaluation Rules

At present, there are some differences in the performance of SAR imaging system, and there are some differences in the imaging effect. The size of the target in the human body is small, and at the same time, to be a higher recognition rate, the comparison network is chosen as YOLO, Single-Shot MultiBox Detector (SSD) such regression recognition based network. Therefore, we compare the recognition accuracy and recognition speed of YOLO v3, SSD, and MPEN under the same conditions. Test network performance parameters are mainly P*recision*, R*ecall*, *mAP*, *accuracy*, and *F*1.

$$\mathrm{P}recision = \frac{TP}{TP + NP} \tag{1}$$

$$\mathrm{R}ecall = \frac{TP}{TP + FN} \tag{2}$$

$$mAP = \frac{\sum AP}{NC} \tag{3}$$

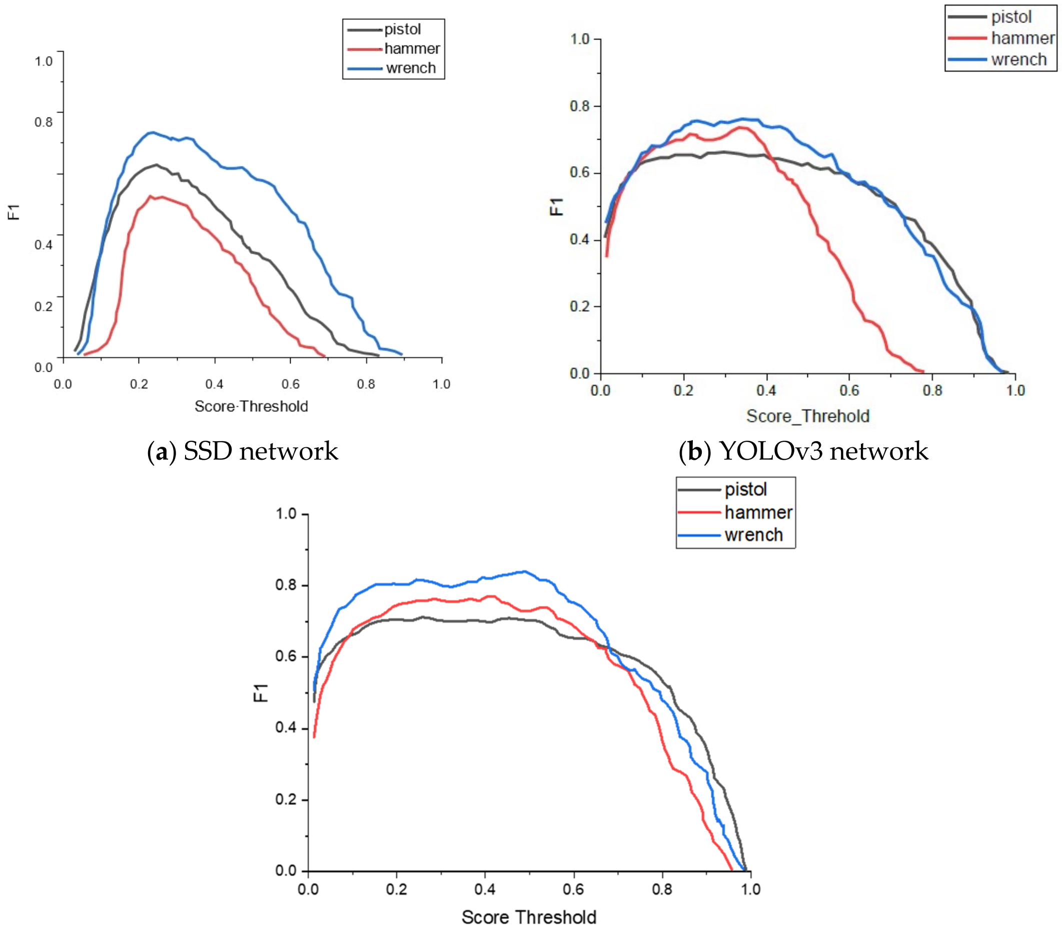
$$Accuracy = \frac{TP + TN}{S} \tag{4}$$

$$F1 = 2 \times \frac{\mathrm{P}recision \times \mathrm{R}ecall}{\mathrm{P}recision + \mathrm{R}ecall} \tag{5}$$

In the formula, P*recision* is the proportion of the correct retrieved target to all the actual retrieved targets. *TP* predicts the positive class to the positive class number. *FP* predicts the negative class to the positive class number. R*ecall* is the proportion of correctly retrieved objects in all that should be retrieved. *FN* predicts the positive class to be negative. The *mAP* is the average of *AP* values of all classes. *AP* is the average precision of each class. *NC* is the total number of classes. *Accuracy* is the proportion of correctly retrieved targets to all targets. *TN* predicts negative classification as negative classification. S is the total number of samples. Moreover, *F*1 is the harmonic average of P*recision* and R*ecall*.

4. Results

As shown in Figure 10, *F*1 combines the results of Precision and Recall. As shown in Figure 4, we first compare the relationship between the network's Score Threshold and *F*1. In MPEN, when the Score Threshold value is 0.5, all three recognition targets can maintain a high *F*1 value. The *F*1 value of the hammer begins to decline rapidly when the Score Threshold value is 0.4. When the Score Threshold value of the SSD network is 0.5, only the *F*1 value of the wrench exceeds 0.5. In MPEN, the three recognition targets can maintain a high *F*1 value when the Score Threshold value is 0.5. In other words, when the Score Threshold value is 0.5, both Precision and Recall can be maintained at a higher level. Therefore, this study will mainly discuss the performance changes of all aspects of the network when the Score Threshold value is 0.5.



(**a**) SSD network

(**b**) YOLOv3 network

(**c**) MPEN

Figure 10. Schematic diagram of each network *F*1.

The overall comparison of mAP and recognition time between YOLO v3 network and MPEN can be seen in Table 1. Compared with the YOLO v3 network, the mAP of MPEN is increased by 7.57% under the same time-consuming condition. Although the recognition time of the SSD network is short, its mAP is too low, and *F*1 performs poorly when the Score Threshold value is 0.5. Therefore, the following analysis does not compare to the SSD network. In Table 2, we compare the recognition accuracy of a pistol, wrench, and hammer. Even though the similarity of wrench and hammer is high, the accuracy of MPEN is still higher than that of the YOLO v3 network. The recognition accuracy of a hammer is lower than that of the YOLO v3 network. Among the three targets, the improvement is the highest, 11.73%, and the recognition accuracy of the pistol is the lowest, 4.98%. At the same time, compared with the AP of each target, the AP of the wrench is the highest among the three, which is 9.72%. The AP increase is the lowest among the three, at 5.32%.

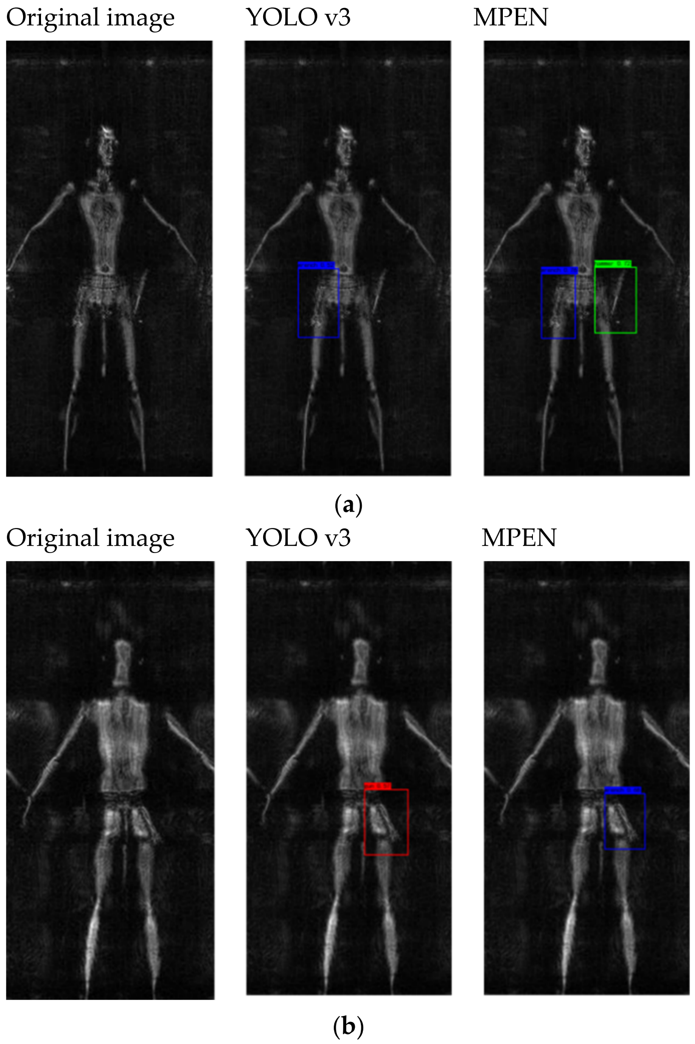
Next in the process is to choose 4 SAR images that are scanned by a typical flat scanning system with inconsistent targets and similar targets, and use YOLO v3 and MPEN to detect targets. Figure 11 shows the test results. Among them, in Figure 11a, MPEN can effectively detect the hammer and wrench with a confidence level greater than 0.7. While the hammer is missing in the YOLO v3 network, the confidence level of the wrench is only 0.52. In Figure 11b, MPEN can detect the target wrench with a confidence level greater than 0.6, while the YOLO v3 network misidentifies the wrench as a pistol. In Figure 11c, both the MPEN and YOLO v3 network can detect the wrench, but the confidence of MPEN is 0.96, while the confidence of the YOLO v3 network is only 0.81. In Figure 11d, MPEN can detect the handgun with a confidence level of 0.58, while the handgun is missing in the YOLO v3 network. It can be seen that the YOLO v3 network failed to detect or misclassified the target, and MPEN has further improved its ability to distinguish the blurred targets with low pixels and similar targets.

Table 1. Comparison of target recognition performance in millimeter wave images by each network.

Network Model	mAP	Recognition Time (ms)
SSD	65.30%	30.5
YOLO v3	74.82%	22.4
MPEN (ours)	82.39%	24.2

Table 2. Performance comparison of recognition targets.

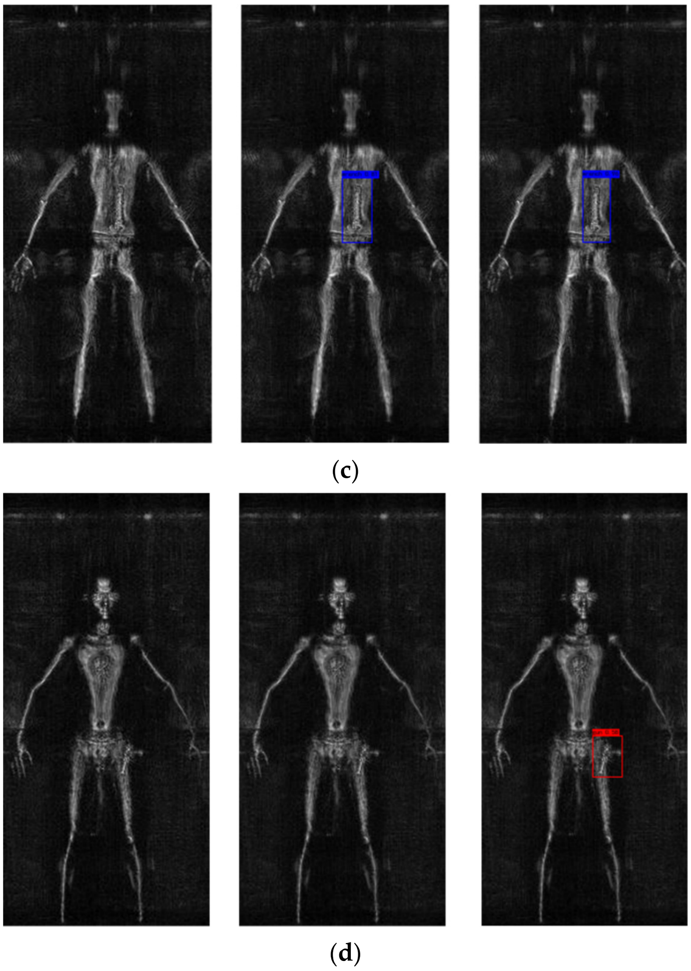
Target	Yolo v3 Accuracy	MPEN Accuracy	Yolo v3 AP	MPEN AP
Pistol	74.44%	79.42%	71.33%	76.65%
Hammer	74.12%	85.85%	74.71%	82.37%
Wrench	84.73%	91.00%	78.42%	88.14%



(**a**)

(**b**)

Figure 11. *Cont.*



(**c**)

(**d**)

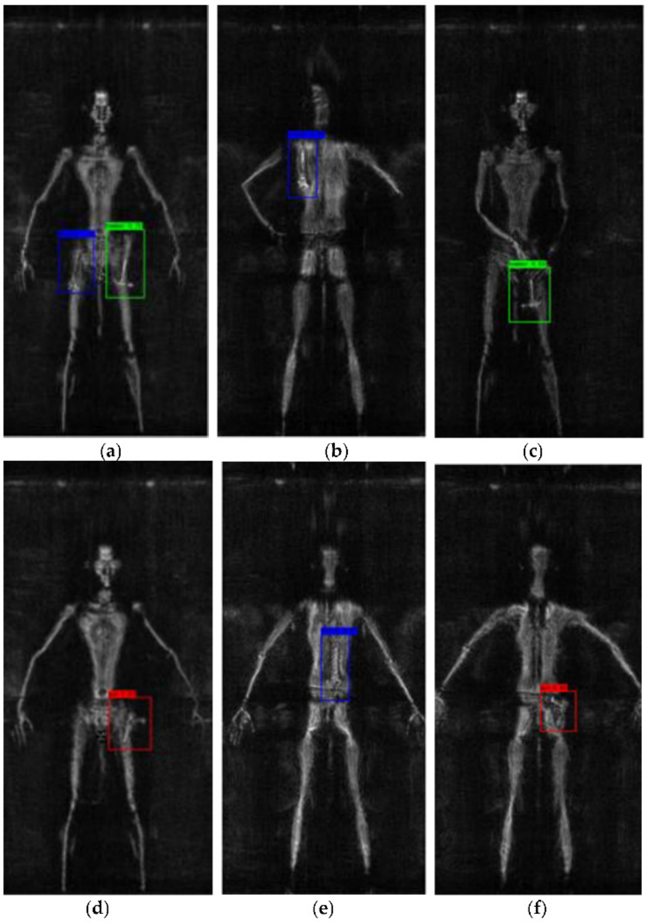
Figure 11. Schematic diagram of each network *F*1. (**a**) Test photo 1. (**b**) Test photo 2. (**c**) Test photo 3. (**d**) Test photo 4.

We list the false alarm rate and miss alarm rate of the network recognition target in Table 3. The false alarm rate is the ratio of the number of misclassified targets to the total number of samples. The main reason for a false recognition is that the human limbs are similar to the target handle, and the YOLO network has a low ability to distinguish similar targets. The miss alarm rate is the ratio of the number of missed targets to the total number of samples. Missing targets are mainly due to inconsistent target clarity, which leads to insufficient network acquisition capabilities. The false alarm rates of pistols, hammers, and wrenches for the network that we propose to MPEN is lower than the YOLO v3 network, and the gun miss rate is slightly higher than that of the YOLO v3 network. The false rate and miss rate can reflect the performance of the network from another angle. According to the data in Table 3, MPEN can reduce the error rate when detecting similar targets. In terms of missed detection rate, the missed detection rate of pistols has increased slightly, but the missed detection rate of hammers and wrenches has decreased. According to the increase in error rate and missed detection rate, it can be considered that MPEN has a particular improvement in the recognition of inconsistent clarity targets. At the same time, it has a particular improvement in the recognition ability of similar interference targets (such as human limbs).

In this study, contraband is placed in different human body parts, deliberately imaging some blurred contraband images. The purpose is to simulate some extreme usage scenarios of security inspection equipment in practical applications as much as possible. The deep learning network has no requirements for human posture. The recognition results are shown in Figure 12. When the contraband is on the human torso, the appearance of the contraband is more obvious. In Figure 12a,b,e,j, the blue box is selected as the wrench. When the wrench is on the outer thigh, back, waist, and leg of the human body, compared with the other three images, the wrench in Figure 12a has a different appearance, clarity, and edge. In Figure 12e, the left side of the middle wrench is more blurred than the right side. In Figure 12d,f,g, when the pistol is on the legs, buttocks, and chest of the human body, the muzzle in Figure 12f is more blurred than the handle of the firearm. MPEN can still identify targets. Where the target may be confused with the human body, the tester will give instructions by hand to provide a mark for verification. In Figure 12c, the hammer is on the inner thigh, and the hammerhead can confirm the position, but the handle position is indicated by hand. In a situation where the hammer is placed in the human body under the armpit, the shape of the hammer itself may be confused with the human body, and the handle position with hand instructions. Another example: putting a hammer on a person's hand, so that the handle of the hammer almost merges with the hand. At this time, the handle is indicated by hand. In order to show that the hand's instructions will not affect the recognition, place the wrench on the shank. At this time, the wrench on the shank can still be detected.

Table 3. Comparison of false alarm rate and miss alarm rate.

Target	Yolo v3 False Alarm Rate	MPEN False Alarm Rate	Yolo v3 Miss Alarm Rate	MPEN Miss Alarm Rate
Pistol	25.56%	20.58%	8.50%	9.50%
Hammer	25.88%	14.15%	6.30%	5.30%
Wrench	15.27%	9.00%	11.00%	5.50%


(a) (b) (c)

(d) (e) (f)

Figure 12. *Cont.*

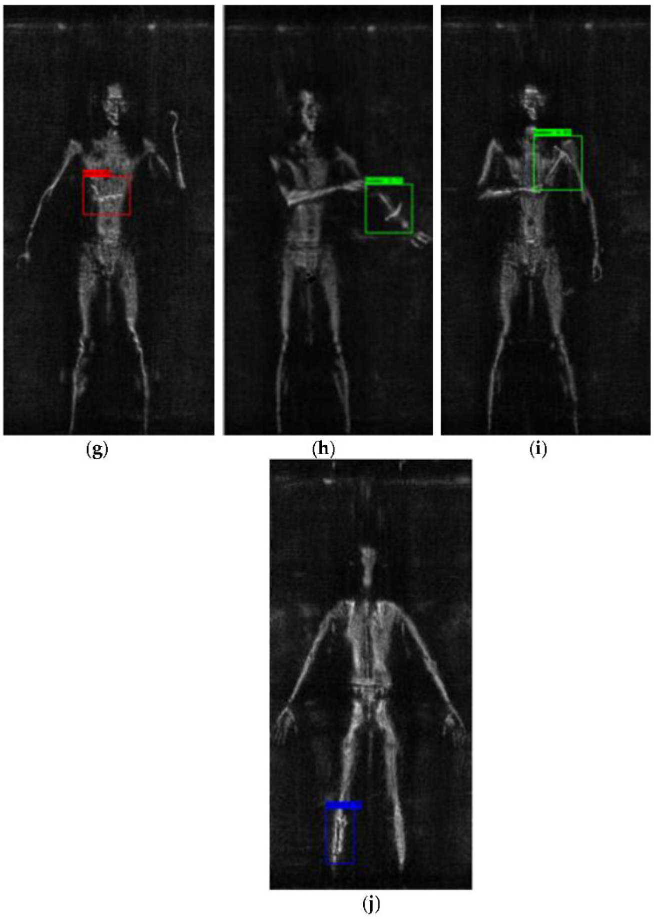


Figure 12. Recognition results of SAR images using MPEN. (**a**) The targets are placed on the outer thigh. (**b**) The target is placed on the back. (**c**) The target is placed on the inner thigh. (**d**) The target is placed on the outer thigh. (**e**) The target is placed on the waist. (**f**) The target is placed on the buttock. (**g**) The target is placed on the chest. (**h**) The target is placed on the arm. (**i**) The target is placed under the armpit. (**j**) The target is placed on the shank.

5. Discussion

There are many researches on target detection using deep learning. Chen et al. [12,13] inspected the aircraft. Their detection network has a solid ability to extract targets to detect aircraft in the image. The purpose of our method to improve the detection capability is mainly to distinguish similar targets. Del Prete et al. [18] detected the wake of the ship. This method is interfered with by other similar elements, resulting in low mAP. Our method can distinguish similar targets even when the target is similar to the body part. Ghaderpour et al. [19] and Arivazhagan et al. [20] used wavelet transform and Fourier transform for target/change detection. This detection method can detect multiple targets. Due to the limitation of the algorithm, it can only detect multiple target selections in sequence. Our network uses a deep learning solution to identify and label multiple targets at the same time.

At present, multi-object often needs to be identified in the security inspection system. Most of the current research focuses on detecting a single target, and the research on the recognition of multi-object is relatively immature. Since the SAR imaging systems used by each researcher vary, there may be differences in the data sets used. Due to the performance limitations of the flat scanning imaging system, there is inconsistency in the imaging of foreign objects on the surface of the human body.

The current related studies mainly focus on the detection of caches in the human body, and there are relatively few studies on the identification of cache types. Compared with the [14,16,17] network, we found that the three researchers mainly detected hidden objects, and did not detect the shape and size of the hidden objects, so the detection accuracy of this type of research is relatively high. MPEN can detect multi-object at the same time, and similar objects (hammer and wrench are listed in this article) can be well distinguished again.

Less research results on multiple target recognition on the human surface. In [17], they focus on the recognition of handguns and human bodies. The two targets differ significantly in shape, and the human body occupies the majority of the image area. When the pistol is placed on the human body surface, the pixel share of the two targets is approximately constant, so it is easier to recognize the two targets. In [18], the researcher's purpose of the study is similar to our goal. The researchers conduct tests on human bodies, knives, pistols, bottles, and mobile phones. Since the recognition rate of the human body is as high as 98.75%, the recognition rate of the mobile phone is only 47.18%. At this time, the detection performance of the network is affected by the recognition rate of the human body. Although the mAP of this target detection is 69.7%, it does not reflect the recognition rate of dangerous targets. In security inspection equipment, the recognition rate of dangerous targets should be the first consideration. We only consider dangerous targets, such as pistols, hammers, and wrenches. Our mAP can objectively reflect the identification of dangerous targets.

Since the imaging system we use operates at 35 GHz, if a system with higher imaging resolution is used to obtain SAR images, there will be a difference in performance. There are many prohibited items in the security inspection system, and now it is necessary to train more contraband models to establish datasets. Moreover, as far as the existing recognition accuracy is concerned, it still cannot meet the actual detection requirements. In terms of recognition accuracy and false alarm rate, there is still room for further improvement.

6. Conclusions

A suspicious multi-object detection and recognition method for millimeter wave SAR security inspection images based on multi-path extraction network has been proposed. The method includes proposing a multi-path feature pyramid (MPFP) module and modifying the distribution of residual blocks. Flat scanning system had fixed focal length, blurred imaging, and difficult target recognition. We modified the number of repetitions of the residual block in Darknet-53 and simultaneously increased the path of three scale features. The deep output scale was output separately, and the shallow network was combined with the output after sampling on the deep scale. Our method was the extraction of target boundaries and improve the ability to recognize clarity targets. Compared with the YOLO v3 network, mAP increased by 7.57%, and the recognition accuracy increased by 11.73% at the highest.

In general, our method performed better in SAR images for poor sharpness and similar target recognition. However, when our method recognized a target with specific unique characteristics (such as a pistol), the contour feature of the pistol was relatively regular. It was easy to coincide with the contour of the human body when placed on the edge of the human body. In this case, there still has room for improvement in recognition accuracy. The millimeter wave imaging accuracy was related to the operating frequency of the system. The multi-object recognition method proposed in this article can be applied to devices with lower operating frequencies. It still had high recognition accuracy under low accuracy, which can reduce the dependence of millimeter wave recognition system on working frequency and improve equipment economy. Our method had a better ability to distinguish similar targets. It was used in the field of security inspection and can refine the recognition targets, such as identifying the models of aircraft, ships, and vehicles. The core of research had shifted from identifying a certain type of target to identifying a certain target. And we hope to utilize as much information as possible in SAR images to promote SAR image target recognition research.

Author Contributions: Conceptualization, M.Y. and Y.L.; methodology, M.Y., Y.L. and Q.Z.; software, Q.Z.; validation, M.Y., Q.Z. and Y.Y.; formal analysis, Q.Z. and Y.L.; investigation, Q.Z.; resources, Y.Z.; data curation, Q.Z. and Y.Y.; writing—original draft preparation, M.Y. and Q.Z.; writing—review and editing, Y.L. and Q.Z.; visualization, M.Y. and Q.Z.; supervision, Y.Z.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the National Natural Science Foundation of China (Grant No. 61988102, 61731020), in part by the Natural Science Foundation of Shanghai (Grant No. 21ZR1444300), in part by the National key R&D Project of China (Grant No. 2018YFF01013003), in part by the 111 Project (Grant No. D18014), the Key project supported by Science and Technology Commission Shanghai Municipality (Grant No. YDZX20193100004960).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: We sincerely thank the anonymous reviewers for their critical comments and suggestions for improving the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following are the abbreviations used in this article:

MPEN	Multi-Path Extraction Network
YOLO	You Only Look Once
MPFP	Multi-Path Feature Pyramid
mAP	mean Average Precision
MIMO	Multiple-Input Multiple-Output
PCA	Principal Component Analysis
CNN	Convolutional Neural Network
RCNN	Region-Based Convolutional Neural Network
NMS	Non-Maximum Suppression
SSD	Single-Shot MultiBox Detector

References

1. Saadat, M.S.; Sur, S.; Nelakuditi, S.; Ramanathan, P. Millicam: Hand-held millimeter-wave imaging. In Proceedings of the 2020 29th International Conference on Computer Communications and Networks (ICCCN), Honolulu, HI, USA, 3–6 August 2020; pp. 1–9.
2. Jing, H.D.; Li, S.Y.; Cui, X.X.; Zhao, G.Q.; Sun, H.J. Near-field single-frequency millimeter-wave 3d imaging via multifocus image fusion. *IEEE Antennas Wirel. Propag. Lett.* **2021**, *20*, 298–302. [CrossRef]
3. Zhang, F.; Wu, C.S.; Wang, B.B.; Liu, K.J.R. Mmeye: Super-resolution millimeter wave imaging. *IEEE Internet Things J.* **2020**, *8*, 6995–7008. [CrossRef]
4. Wu, S.Y.; Gao, H.; Li, C.; Zhang, Q.Y.; Fang, G.Y. Research on MIMO THz azimuth imaging algorithm based on arc antenna array. *J. Electron. Inf. Techn.* **2018**, *40*, 860–866.
5. Rubani, Q.; Gupta, S.H.; Rajawat, A. A compact MIMO antenna for WBAN operating at terahertz frequency. *Optik* **2020**, *207*, 164447. [CrossRef]
6. Gao, H.; Li, C.; Wu, S.Y.; Geng, H.B.; Zheng, S.; Qu, X.D.; Fang, G.Y. Study of the extended phase shift migration for three-dimensional MIMO-SAR imaging in terahertz band. *IEEE Access* **2020**, *8*, 24773–24783. [CrossRef]
7. Sun, C.; Chang, Q.G.; Zhao, R.; Wang, Y.H.; Wang, J.B. Terahertz imaging based on sparse MIMO array. In Proceedings of the 2020 International Conference on Microwave and Millimeter Wave Technology (ICMMT), Shanghai, China, 20–23 September 2020; pp. 1–3.
8. Yang, G.; Li, C.; Gao, H.; Fang, G.Y. Phase shift migration with SIMO superposition for MIMO-sidelooking imaging at terahertz band. *IEEE Access* **2020**, *8*, 208418–208426. [CrossRef]
9. Zhang, Y.; Wang, H.; Zeng, Y.; Deng, B.; Qin, Y.; Yang, Q. Three-dimensional surface reconstruction of space targets using a terahertz MIMO linear array based on multi-layer wideband frequency interferometry techniques. *IEEE Trans. Terahertz. Sci. Technol.* **2021**, *11*, 353–366. [CrossRef]
10. Isiker, H.; Unal, I.; Tekbas, M.; Ozdemir, C. An auto-classification procedure for concealed weapon detection in millimeter-wave radiometric imaging systems. *Microw Opt. Technol. Lett.* **2018**, *60*, 583–594. [CrossRef]

11. Yeom, S.; Lee, D.S.; Jang, Y.; Lee, M.K.; Jung, S.W. Real-time concealed-object detection and recognition with passive millimeter wave imaging. *Opt. Express* **2012**, *20*, 9371–9381. [CrossRef] [PubMed]
12. Chen, L.F.; Weng, T.; Xing, J.; Li, Z.H.; Yuan, Z.H.; Pan, Z.H.; Tan, S.Y.; Luo, R. Employing deep learning for automatic river bridge detection from SAR images based on adaptively effective feature fusion. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102245. [CrossRef]
13. Luo, R.; Chen, L.F.; Xing, J.; Yuan, Z.H.; Tan, S.Y.; Cai, X.M.; Wang, J.L. A fast aircraft detection method for SAR images based on efficient bidirectional path aggregated attention network. *Remote Sens.* **2021**, *13*, 2940. [CrossRef]
14. Meng, Z.C.; Zhang, M.; Wang, H.X. CNN with pose segmentation for suspicious object detection in MMW security images. *Sensors* **2020**, *20*, 4974. [CrossRef]
15. Lopez-Tapia, S.; Molina, R.; de la Blanca, N.P. Using machine learning to detect and localize concealed objects in passive millimeter-wave images. *Eng. Appl. Artif. Intell.* **2018**, *67*, 81–90. [CrossRef]
16. Guo, L.; Qin, S.Y. High-performance detection of concealed forbidden objects on human body with deep neural networks based on passive millimeter wave and visible imagery. *J. Infrared. Millim. Terahertz. Waves* **2019**, *40*, 314–347. [CrossRef]
17. Liu, C.Y.; Yang, M.H.; Sun, X.W. Towards robust human millimeter wave imaging inspection system in real time with deep learning. *Prog. Electromagn. Res.* **2018**, *161*, 87–100. [CrossRef]
18. Del Prete, R.; Graziano, M.D.; Renga, A. First results on wake detection in SAR images by deep learning. *Remote Sens.* **2021**, *13*, 4573. [CrossRef]
19. Ghaderpour, E.; Pagiatakis, S.D.; Hassan, Q.K. A survey on change detection and time series analysis with applications. *Appl. Sci.* **2021**, *11*, 6141. [CrossRef]
20. Arivazhagan, S.; Ganesan, L. Automatic target detection using wavelet transform. *EURASIP J. Adv. Signal Process.* **2004**, *2004*, 853290. [CrossRef]
21. Pang, L.; Liu, H.; Chen, Y.; Miao, J.G. Real-time concealed object detection from passive millimeter wave images based on the YOLOv3 algorithm. *Sensors* **2020**, *20*, 1678. [CrossRef] [PubMed]
22. Zhang, J.S.; Xing, W.J.; Xing, M.D.; Sun, G.C. Terahertz image detection with the improved faster region-based convolutional neural network. *Sensors* **2018**, *18*, 2327. [CrossRef] [PubMed]
23. Rajchl, M.; Lee, M.C.; Oktay, O.; Kamnitsas, K.; Passerat-Palmbach, J.; Bai, W.; Damodaram, M.; Rutherford, M.A.; Hajnal, J.V.; Kainz, B. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. *IEEE Trans. Med. Imag.* **2016**, *36*, 674–683. [CrossRef] [PubMed]
24. Drolet, G. Contraband detection program. In Proceedings of the Chemistry and Biology-Based Technologiesfor Contraband Detection, Boston, MA, USA, 17 February 1997; pp. 162–172.
25. Kaiming, H.; Xiangyu, Z.; Shaoqing, R.; Jian, S. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Liu, Y.G.; Yu, J.Z.; Han, Y.H. Understanding the effective receptive field in semantic image segmentation. *Multimed. Tools Appl.* **2018**, *77*, 22159–22171. [CrossRef]
27. Guo, P.C.; Su, X.D.; Zhang, H.R.; Wang, M.; Bao, F.L. A multi-scaled receptive field learning approach for medical image segmentation. In Proceedings of the 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1414–1418.
28. Tang, Q.L.; Sang, N.; Liu, H.H. Learning nonclassical receptive field modulation for contour detection. *IEEE Trans Image Process.* **2020**, *29*, 1192–1203. [CrossRef]
29. Behboodi, B.; Fortin, M.; Belasso, C.J.; Brooks, R.; Rivaz, H. Receptive field size as a key design parameter for ultrasound image segmentation with U-Net. In Proceedings of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 2117–2120.
30. He, Z.W.; Cao, Y.P.; Du, L.; Xu, B.B.; Yang, J.X.; Cao, Y.L.; Tang, S.L.; Zhuang, Y.T. MRFN: Multi-receptive-field network for fast and accurate single image super-resolution. *IEEE Trans. Multimed.* **2020**, *22*, 1042–1054. [CrossRef]