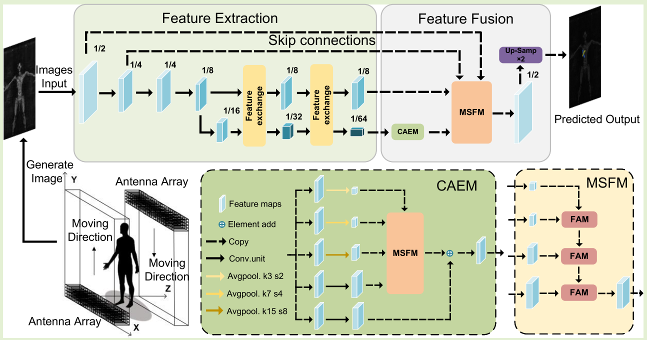
Object Recognition for Millimeter Wave SAR Images Based on Dual-Branch Multiscale Fusion Network

Junhua Ding, Bofeng Su, and Minghui Yuan



***Abstract*—There are several major challenges in the detection and identification of contraband in millimeter wave synthetic aperture radar (SAR) security images: the complexities of small target sizes, partially occluded targets, and overlap between multiple targets are not conducive to the accurate identification of contraband. To solve these problems, a contraband detection method based on the dual branch multiscale fusion network (DBMFnet) is proposed. In the feature extraction stage, the feature extraction network is designed with a dual-branch parallel output structure. One branch preserves high resolution, while the other branch extracts rich semantic information through multiple downsampling operations. Bilateral connections are established between high-resolution and low-resolution branches to facilitate repeated feature exchange, which enhances the ability to detect small and fuzzy targets. In the feature fusion stage, a multiscale fusion module (MSFM) and a context aggregation extraction module (CAEM) are devised; MSFM is utilized to enhance target edge perception, and CAEM is employed to extract contextual information from low-resolution feature maps for enhancing model segmentation performance while reducing computational complexity. The experimental results show that the proposed semantic segmentation method outperforms the existing semantic segmentation methods in mean intersection over union (mIoU).**

***Index Terms*— Contraband detection, deep learning, dual-branch multiscale fusion network (DBMFnet), millimeter wave synthetic aperture radar (SAR) security image.**

I. Introduction

WITH the advancement of millimeter wave technology, millimeter wave security inspection systems have reached a higher level of maturity [1], [2]. Traditional security inspection technologies exhibit varying degrees of limitations. For instance, metal detectors are effective only for metallic objects, such as knives and firearms, rendering them ineffective against nonmetallic dangerous goods. X-ray possesses strong transmission capabilities. However, its ionizing radiation can damage cellular structures and pose risks to human health, making it unsuitable for human security inspection. Infrared has limited penetration capabilities and fails to effectively detect deeply concealed contraband. Compared with traditional security inspection technologies, millimeter wave security imaging not only enables the detection of metallic objects hidden under fabrics but also identifies dangerous items, such as plastic firearms, knives, explosives, and so on. Significantly, it is crucial to note that millimeter waves are nonionizing and do not cause harm to the human body [3], [4], [5], [6], [7]. Therefore, millimeter wave imaging equipment is widely utilized for human security inspection.

The emergence of active millimeter wave human screening devices based on MIMO-SAR architecture has significantly improved the image accuracy and speed of active millimeter wave screening devices through the development of multiple-input-multiple-output (MIMO) technology and synthetic aperture radar (SAR) technology [8]. After obtaining a more advanced millimeter wave imaging system, the system generates millimeter wave security images that can be identified by computers to enable automatic detection and positioning of hidden contraband. Accurately identifying hidden contraband in millimeter wave images has become a focus of current research.

The early detection techniques for millimeter wave images are typically based on traditional computer vision

Manuscript received 15 January 2024; accepted 5 February 2024. Date of publication 16 February 2024; date of current version 16 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61988102. The associate editor coordinating the review of this article and approving it for publication was Dr. Fabrizio Santi. *(Corresponding author: Minghui Yuan.)*

The authors are with the Terahertz Technology Innovation Research Institute, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China (e-mail: yuanminghui@usst.edu.cn).

Digital Object Identifier 10.1109/JSEN.2024.3363690

technologies. Zheng et al. [9] employed scale-invariant feature transform (SIFT) and histogram of oriented gradient (HOG) as feature extractors for the support vector machine (SVM)-based classification of hazardous materials. Dai et al. [10] employed K-nearest neighbor (KNN) to identify hazardous objects based on 3-D reconstructed images. Du et al. [11] performed whole image segmentation by fast wavelet transform and then detected the hidden target using the gray value difference between hidden objects and the human body in the image. These methods are less reliable and result in lower target detection rates.

In recent years, deep convolutional neural networks have achieved significant progress in the fields of image classification [12], [13], target detection [14], [15], [16], [17], [18], and image segmentation [19], [20], [21], [22], [23], [24], [25], [26], with the development of artificial intelligence. Consequently, numerous efficient deep learning algorithms have been employed for concealed object detection in millimeter wave images. Nonetheless, these algorithms primarily focus on detecting red, green, blue (RGB) images and high-resolution images, which differ significantly from millimeter wave images. The millimeter wave images typically exhibit low resolution and a grayscale appearance. In addition, the performance limitations of millimeter wave detectors and imaging algorithms result in the signal-to-noise ratio of these images being significantly lower compared to optical images, reducing image contrast, and incomplete object representation.

Previous research has shown that these problems significantly hinder the identification and localization of hidden targets and reduce the effectiveness of millimeter wave image detection [27]. Consequently, several deep learning approaches have been proposed to detect millimeter wave images with SAR imaging properties. Liu et al. [28] attempted to increase the detection rate of contraband by combining millimeter wave images with additional spatial depth maps and designing a new loss function, but only one category was identified. Sun et al. [29] designed a multisource aggregation transformer with two different attention mechanisms to improve hidden target detection performance. The above study could only detect the location of the contraband on the surface of the human body and could not identify the type of contraband, which resulted in for the security personnel having to conduct a secondary inspection to determine the type of contraband. Pang et al. [30] utilized the YOLO v3 algorithm to instantly detect hidden metal weapons on humans using passive millimeter wave images. They primarily identified handguns and humans, both of which have distinct shape differences that can be easily distinguished by the deep learning network. Although some types of contraband can be detected, the detection range is limited and there is still scope for improving the detection accuracy and reducing the false alarm rate.

The methods used in the aforementioned studies are anchor-based target detection. The anchor box contains background noise that can affect target detection. Wang et al. [31] improved the receptive field by using dilated convolutions to improve detection performance and achieved good results in the experiment. Liang et al. [32] achieved the segmentation of contraband in the human body by combining a generative adversarial network with a selected connected U-net network. At present, the research focus in SAR image detection primarily revolves around the issues of target imaging and background interference. However, there still exist certain issues within SAR image detection, such as the mutual interference between multiple targets and the detection of small targets. Hence, achieving accurate recognition and localization of concealed targets in SAR images continues to pose significant challenges.

In this article, we focus on the problem of target localization and detection in low signal-to-noise ratio and low-resolution human security SAR images. To overcome the problems of small targets detection and object interaction, we propose a multitarget semantic segmentation method for millimeter wave SAR images based on a dual-branch multiscale fusion network (DBMFnet). The network detects contraband using semantic segmentation. We used the MIMO-SAR imaging system independently developed by our laboratory for data acquisition and created two millimeter wave SAR image datasets named MW-SAR1 and MW-SAR2. Experimental results on two datasets show that the proposed method achieves better segmentation results compared to the existing semantic segmentation methods.

The main contributions are as follows.

1) A contraband detection model with a DBMFnet is proposed for millimeter wave SAR security images.
2) A context aggregation extraction module (CAEM) is proposed to extract rich contextual information, and it adds little to the inference time when the richest low-resolution feature maps are used as an input.
3) A multiscale fusion module (MSFM) is proposed to integrate multiple low-resolution feature maps into one high-resolution feature map, which can reduce the loss of semantic information associated with traditional sampling methods, decrease the computational complexity of the model, improve edge segmentation effectiveness, and enhance target positioning accuracy.

The structure of this article is as follows: in Section II, the DBMFnet for detecting contraband in MIMO-SAR security screening images is described in detail. In Section III, the experimental results and corresponding image analysis under different conditions are described. In Section IV, we discuss the problems faced by detection algorithms for security screening systems and the direction of subsequent improvements. Finally, relevant conclusions are given in Section V.

II. Methods

A. Dual Branch Multiscale Fusion Network (DBMFnet)

This article proposes a DBMFnet semantic segmentation model for contraband detection in millimeter wave SAR images. In the feature extraction stage, we employ a double branch parallel feature extraction network (DBPFEN). In the process of feature extraction, one branch keeps the high resolution, the other branch extracts rich semantic information through multiple downsampling operations, and bilateral connections are established between the two branches for repeated

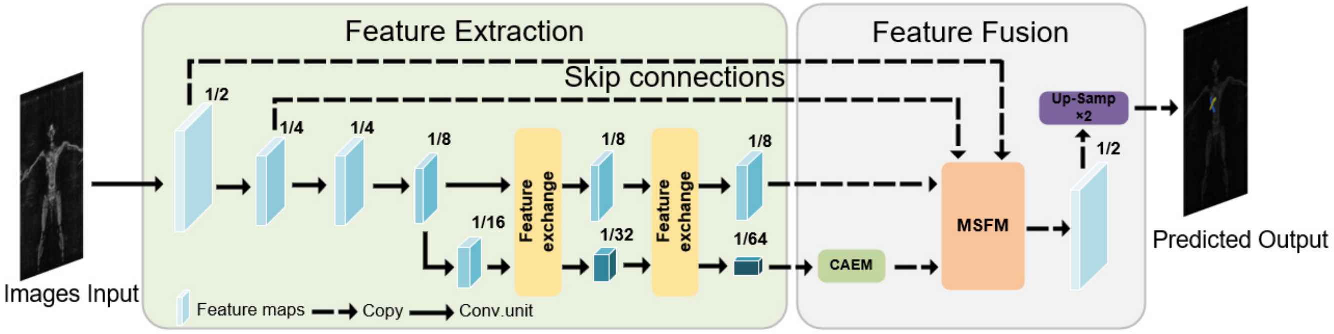


Fig. 1. Network structure diagram.

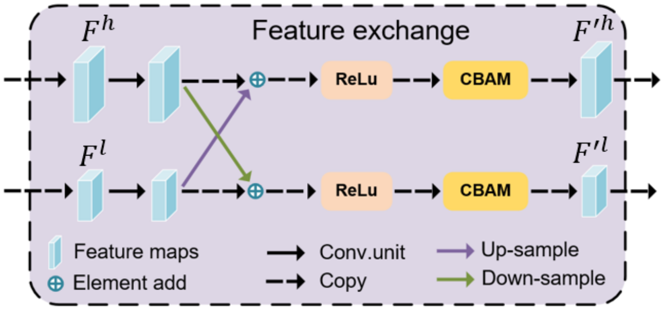


Fig. 2. Feature exchange process.

feature exchange. In the feature fusion stage, the final feature map containing the most semantic information in the low-resolution branch is incorporated into the CAEM to extract contextual information and enhance detection accuracy. To enhance the detection of target edges, the MSFM incorporates low-resolution branching feature maps, high-resolution branching feature maps, and higher resolution feature maps obtained from the skip connections layer and achieves mutual fusion between multiple feature maps of different resolutions. Finally, the prediction map is output by upsampling the image twice to restore its original size. The network framework is shown in Fig. 1.

1) Double Branch Parallel Feature Extraction Network (DBPFEN): Generally, shallow features contain greater spatial information and preserve more detailed features of small targets; deep features contain more semantic information, which helps improve the classification accuracy of pixel points of different targets in the image, but are less perceptive of details. During the feature extraction process, the downsampling operation of the network can reduce the size of the feature map. Although the feature map becomes richer in the semantic information it contains, the more detailed information about the target is lost, which can significantly affect the fine segmentation of an image, especially when detecting small objects.

To solve the above problems, the feature extraction process of DBMFnet one branch preserves high resolution while the other branch extracts rich semantic information through multiple downsampling operations. Bilateral connections are established between high-resolution and low-resolution branches to facilitate repeated feature exchange, ensuring that high-resolution branch feature maps integrate into the low-rate branch feature maps across different scales, which facilitates the combination of rich semantic information and fine-grained details to improve the detection of small and interfering targets in images.

The specific structure of the DPFEN is shown in the feature extraction stage of Fig. 1. The input image goes through two convolutional layers with a convolutional kernel size of 3 × 3 and a stride size of 2 and becomes 1/2 and 1/4 of the input resolution. The 1/4 resolution feature map is passed through four stacked basic block residual modules; the first two residual modules do not change the image size, and the last two residual modules double the number of image channels. The resolution will be reduced to 1/8 of the input image to get the initial feature map of the high-resolution branch. The initial feature map of the high-resolution branch is also passed through two stacked basic block residual modules; the number of channels is doubled, and the resolution will be reduced to 1/16 of the input resolution to obtain the initial feature map of the low-resolution branch. The two branch feature maps perform feature exchange, and the specific exchange process is shown in Fig. 2. The feature map F^h of the high-resolution branch is downsampled and the number of channels changed by convolution with a 3 × 3 kernel size and stride 2 and then combined with the feature map from the low-resolution branch, followed by passing through both the rectified linear unit (ReLu) layer and the convolutional block attention module (CBAM) [33] for the final output F'^h. Similarly, the low-resolution branch F^l undergoes the aforementioned process, resulting in the output F'^l. The process is repeated subsequently, with the resolution of the high-resolution branch kept constant and the low-resolution branch continuously downsampling, both fused. The feature map resolution of the low-resolution branch corresponds to 1/16, 1/32, and 1/64 of the input resolution, respectively, with corresponding convolution channel numbers 256, 512, and 1024. Meanwhile, the resolution of the feature map of the high-resolution branch is maintained at 1/8 of the input resolution, with 128 channel numbers.

2) Multiscale Fusion Module (MSFM): The restoration of the low-resolution feature map to the resolution of the original image is gradually necessary during the feature fusion stage of semantic segmentation. A common operation is to fuse the feature maps from the downsampling process by skip connections during the recovery process. The feature splice module (FSM) [21], [23] is applied in both the U-net and the DeeplabV3+, as shown in Fig. 3, and the feature addition module (FDM) [20] is applied in the fully convolutional networks (FCNs), as shown in Fig. 4. However, these approaches ignore the misalignment between feature maps with different resolutions, resulting in the potential loss of significant semantic information and consequently leading to subpar segmentation performance at object boundaries. The high-resolution branch and the low-resolution branch output feature maps at 1/8 of the input resolution and 1/64 of the input resolution, respectively, resulting in a significant size discrepancy between the two. If the traditional linear interpolation method is employed for upsampling and fusion, significant semantic information may be lost. Similarly, the resolution of the high-resolution branch feature map is too different from the resolution of the original feature map, and the direct upsampling will also lose a lot of semantic information.

MSFM is proposed to reduce the loss of semantic information in feature fusion for finer segmentation boundaries, as shown in Fig. 5. The module consists of the feature alignment module (FAM) [34], which allows multiple low-resolution feature maps to merge into high-resolution maps. The FAM is inspired by the optical flow for motion alignment between adjacent video frames [35], where the feature maps F^h and F^l of different resolutions are used as an input, and changed to the same number of channels by a 1 × 1 convolutional layer, respectively, Subsequently, the high-resolution feature map F^h is concatenated with the low-resolution feature map F^l by a bilinear interpolation upsampling layer. Finally, the concatenated feature maps undergo a 3 × 3 convolution layer to produce an offset field $\Delta \in R^{2\times H^h \times W^h}$ with the same size as F^h. Mathematically, the aforementioned steps can be written as

$$\Delta = \text{conv}(\text{cat}(\text{upsample}(F^l), F^h)) \tag{1}$$

where upsample(.) is the bilinear interpolation upsampling, cat(.) represents the concatenate operation, and conv(.) denotes a 3 × 3 convolutional layer.

After obtaining the offset field, each position $F^l(X_l, Y_l)$ of low-resolution feature map is then mapped to a point offset field $\Delta(X_h, Y_h)$ by a simple addition operation. The expression is as follows:

$$\begin{cases} X_l = \dfrac{(1+\Delta x)}{N} X_h \\ Y_l = \dfrac{(1+\Delta y)}{N} Y_h \end{cases} \tag{2}$$

where Δx and Δy indicate the learned 2-D transformation offsets for position (X_h, Y_h) and N denotes the multiple of the difference between the high and low resolutions. Then, the value of the position of the warped high-resolution feature map $U(F^h(X_h, Y_h), \Delta)$ is obtained using a four-neighborhood interpolation of (X_l, Y_l) by the differentiable bilinear sampling mechanism. The mathematical expressions are as follows:

$$\begin{aligned} U(F^l(X_l, Y_l), \Delta) = &\sum_{Y_l=1}^{H^l} \sum_{X_l=1}^{W^l} \mathbf{F}^{\mathrm{l}}(X_l, Y_l) \\ &\times \max(0, 1 - |X_l - X_h|) \\ &\times \max(0, 1 - |Y_l - Y_h|) \end{aligned} \tag{3}$$

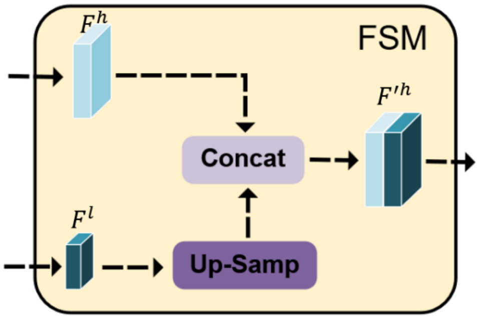


Fig. 3. FSM.

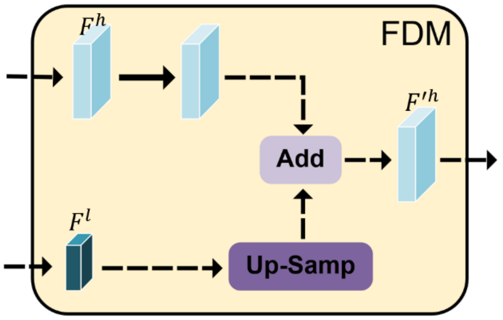


Fig. 4. FDM.

where H^l and W^l denote the size of the low-resolution feature maps. The mathematical expression of the entire FAM process is as follows:

$$F'^h = \text{conv}(F^h) + U(\text{conv}(F^l), \Delta). \tag{4}$$

In MSFM, 1/64 input resolution feature maps of the low-resolution branch, 1/8 input resolution feature maps of the high-resolution branch, 1/4 input resolution feature maps, and 1/2 input resolution feature maps obtained by skip connections are introduced as F_3^l, F_2^l, F_1^l, and F_1^h, respectively, and the lowest resolution feature map F_3^l is sequentially fused upward to obtain a 1/2 input resolution high-resolution feature map $F_1'^h$. The proposed MSFM refines edge segmentation.

3) Context Aggregation Extraction Module (CAEM): Another key to semantic segmentation is how to capture richer contextual information. Contextual information can provide rich semantic guidance for overall scene images, thus minimizing error occurrences. Atrous spatial pyramid pooling (ASPP) [20] is composed of parallel atrous convolutional layers with different rates, which can capture multiscale contextual information. Pyramid pooling module (PPM) [22] in PSPNet attends to multiscale contextual information by implementing pyramid pooling ahead of convolutional layers. However, these context extraction modules are computationally intensive and particularly time-consuming.

CAEM is proposed to reduce computation and time, as shown in Fig. 6. Taking feature maps of 1/64 input resolution has the richest semantic information as an input, large pooling kernels with exponential strides are performed to generate feature maps of 1/128, 1/256, and 1/512 input resolution. These feature maps are inputted into the MSFM mentioned above, fused, and then added to the shortcut of 1 × 1 convolution. Although CEAM has many internal operations, it hardly

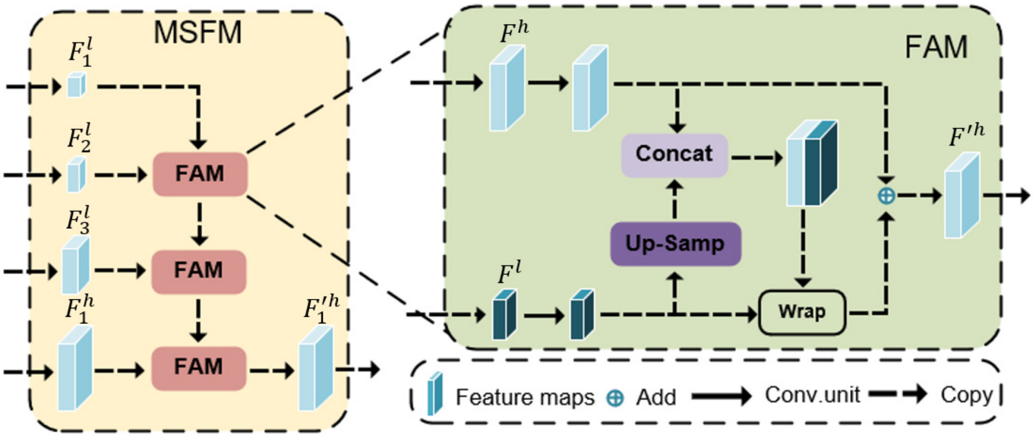


Fig. 5. MSFM.

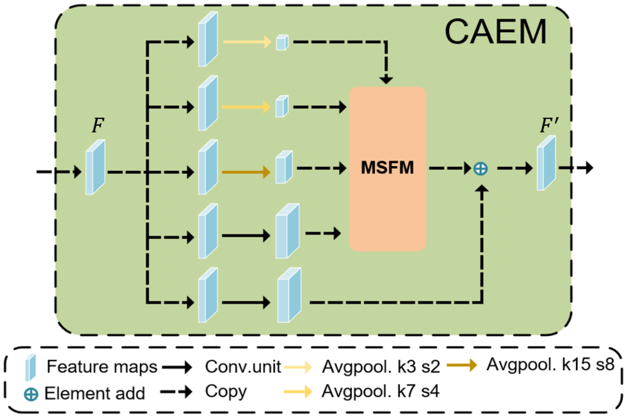


Fig. 6. CAEM.

increases the inference time because the input resolution is only 1/64 of the input resolution. Considering an input F, each scale y_i can be written as

$$y_i = \begin{cases} \mathrm{conv}_{1\times1}(F), & i = 1 \\ K(P_{2^i-1,2^{i-1}}(F), y_{i-1}), & 1 < i \le n. \end{cases} \tag{5}$$

The mathematical expression of the entire CAEM process is as follows:
where $\mathrm{conv}_{1\times1}$ is 1×1 convolution, K denotes the MSFM operation, and $P_{j,k}$ denotes the pool layer of which kernel size is j and stride is k.

B. Loss Function

The loss function is used during the training phase of the model. The predicted value is generated by forward propagation after each batch of training data is input into the model, and the loss function determines the difference between the predicted value and the actual value. To achieve the goal of learning, the model updates each parameter by backpropagation after obtaining the loss value in order to minimize the difference between the true value and the predicted value. This allows the predicted value produced by the model to be as close to the true value as possible.

The most commonly used loss function in this article is the cross-entropy (CE) loss function. The expression of the loss function is

$$\mathrm{Loss}_{\mathrm{CE}} = -\frac{1}{N}\sum_{l\in L}\sum_{i=1} y_l^i \log(\hat{y}_l^i) \tag{6}$$

where L is the number of classes, N is the number of pixels, and y_l^i and $\hat{y}_l^i$ represent the label value and predictive value of pixel i in class l, respectively.

III. Experiment and Analysis

A. Dataset

The laboratory has developed two datasets, MW-SAR1 and MW-SAR2, based on the active millimeter wave human security imaging system with MIMO-SAR architecture. The system structure diagram is shown in Fig. 7. The system is a flat-scan system with an operating frequency of 35 GHz, signal bandwidth 5.04 GHz, sampling rate 27.6 GHz, the antenna array is placed on the X-axis, and the system scans up and down along the Y-axis, which can simultaneously scan the front and back of the human body. The imaging results are shown in Fig. 8(a) and (b). The scanned images were saved in jpg format, and each image was fixed at 200 × 400 pixels. The imaging system operates on the near-field imaging radar principle, where the contraband of metal and ceramic materials reflects higher millimeter wave intensities than those of human bodies, resulting in their appearance as brighter areas in the image. When the plain scanning system detects contraband near the surface of the body, there will be a certain angular distortion in the contraband image, and the distance between the contraband and the scanning antenna is not fixed. Due to a fixed imaging focal length, some areas of the contraband close to the body surface may appear blurred. As shown in Fig. 8(b), the contraband in the image is all affected by imaging noise, resulting in significant blurriness and an indistinct boundary with the human body. In addition, contraband is a small target relative to the whole image, and the difficulty of detecting small-target contraband is increased due to the low resolution of the image.

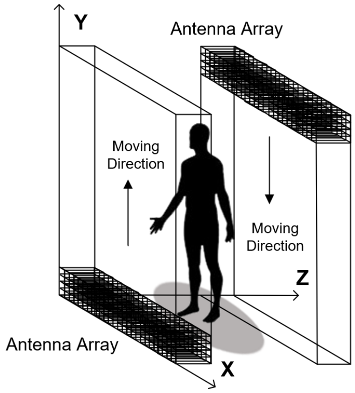


Fig. 7. MIMO-SAR security inspection system structure diagram.

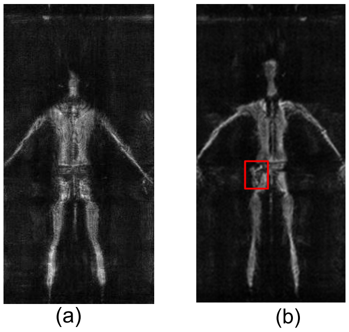


Fig. 8. MIMO-SAR security images. (a) Back scanning image of the human body. (b) Frontal scanning image of human body.

The images in the dataset are labeled using Labelme, which assigns labels with different colors to different objects, and all the remaining unlabeled ones are classified as background classes. During the data collection process, the targets were randomly hidden on the surface of the human body along the edges of the body. We used four types of contraband as recognition targets, which are wrenches, hammers, pistols, and knives, as shown in Fig. 9(a)–(d). Targets with a resolution of fewer than 32 × 32 pixels are defined as small targets. All four contrabands have less than 32 × 32 pixels in the image.

The MW-SAR1 dataset has 1400 images, 90% of which are for training and 10% for testing. The postures of human bodies in this dataset are relatively consistent, with hands down and apart, as shown in Fig. 10(a). The MW-SAR2 dataset contains 700 images, all for testing. This dataset comprises images with incomplete contraband imaging and varying human postures, as shown in Fig. 10(b) and (c), its usage in testing model robustness under more complex conditions.

B. Evaluation Metrics

For quantitative evaluation, the performance metrics used to evaluate the model include mean pixel accuracy (MPA), intersection over union (IoU), and mean IoU (mIoU). MPA denotes the result of averaging the class pixel accuray (CPA) over all classes. IoU denotes the intersection and union ratio of the true and predicted masks for each class. mIoU denotes the average value of IoU over all classes. The higher values of mIoU indicate better overlap between the predicted and true values of the model, indicating better segmentation performance of the model

$$\mathrm{MPA} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{7}$$

$$\mathrm{IoU} = \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{8}$$

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{9}$$

where k denotes the total number of pixel classes, p_{ii} denotes the number of pixels whose original class i is predicted to be class i, p_{ij} denotes the number of pixels whose original class i is predicted to be class j, and p_{ji} denotes the number of pixels whose original class j is predicted to be class i. In addition, parameters, floating-point operations (FLOPs), and reasoning speed [frames per second (FPS)] are used to evaluate the computational cost of the model. Parameters denote the number of parameters to be learned during training. FLOPs denote the number of FLOPs performed during reasoning, which is usually in units of GFLOPs (billions of FLOPs). FPS denotes the number of frames transmitted per second and also the number of images that the model infers per second.

To compare the performance of the segmentation model with the anchor box target detection model, we use the IoU value to calculate the number of accurately segmented target instances by referring to the method of judging anchor boxes in target detection [36]. When the prediction mask of the target is the same as the semantic class of the ground truth mask, the target with an IoU value exceeding the predefined threshold is recorded as a true positive (TP); otherwise, it is recorded as a false positive (FP). When the prediction mask of the target is inconsistent with the semantic class of the ground truth mask, it is directly recorded as an FP. A false negative (FN) indicates that the target is not recognized. The predefined threshold in our study is set to 0.5. For the object-level analysis, we focus more on the overall model performance, as well as the model's miss and false rate. The evaluation metrics used in our study are precision, recall, and $F1$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{10}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

$$F1 = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{12}$$

where Precision represents the proportion of the actual positive samples in all the predicted positive samples, and the sum of the false rate and the precision is 1. Recall represents the proportion of the actual positive samples that are predicted to be positive, and the sum of the miss rate and the Recall is 1. Precision and Recall constrain each other, and $F1$ is the harmonic mean of Precision and Recall.

C. Implementation Details

The proposed model in this article is implemented in the Pytorch 1.12.0 framework and Cuda version 11.6, using a

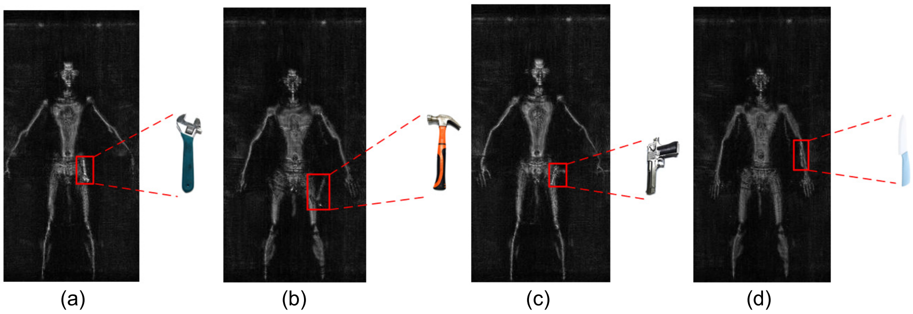


Fig. 9. SAR image of contraband. SAR image of (a) wrench, (b) hammer, (c) pistol, and (d) knife.

TABLE I
COMPARISON OF SEGMENTATION PERFORMANCE OF EACH MODEL IN THE MW-SAR1 AND MW-SAR2 DATASET

Dataset	Network model	MPA(%)	mIoU(%)	Params(M)	FLOPs(G)	Speed(FPS)
MW-SAR1	U-net	80.29	70.35	24.89	452.31	32
	Pspnet	82.98	72.32	46.70	118.43	33.5
	Deeplabv3+	81.05	70.58	54.71	166.87	21
	HRnet-v2	82.33	72.90	29.55	80.18	11.5
	DDRnet	83.21	73.26	20.15	**31.66**	**40**
	HR-FRAFnet(Su)	85.10	75.02	**9.66**	40.17	7
	DBMFnet(our)	**85.26**	**76.23**	19.54	47.36	26
MW-SAR2	U-net	75.66	59.91	-	-	-
	Pspnet	76.82	62.11	-	-	-
	Deeplabv3+	73.23	59.22	-	-	-
	HRnet	78.83	63.89	-	-	-
	DDRnet	75.46	61.78	-	-	-
	HR-FRAFnet(Su)	79.21	62.71	-	-	-
	DBMFnet(our)	**82.13**	**67.00**	-	-	-

TABLE II
COMPARISON OF TARGET SEGMENTATION PERFORMANCE OF EACH MODEL IN THE MW-SAR1 DATASET

Class	U-net	Pspnet	Deeplabv3+	HRnet-v2	DDRnet	HR-FRAFnet(Su)	DBMFnet(our)
	IoU	IoU	IoU	IoU	IoU	IoU	IoU
Hammer	61.98	63.70	63.99	67.35	65.25	66.14	**67.67**
Pistol	66.78	71.84	66.57	66.15	66.74	69.67	**72.17**
Wrench	63.77	64.21	62.65	69.47	71.75	70.96	**74.14**
Knife	59.40	62.01	59.84	61.68	74.46	67.92	**75.81**

dataset in VOC2007 format and end-to-end training of the network using Adam optimizer with an initial learning rate of 5×10^{-4} and a minimum learning rate of 5×10^{-6}. All experiments were performed on a desktop computer with a single NVIDIA GeForce RTX 3090 GPU, with a training number of 300 iterations, a batch size of 8, and an input picture size of 512×512.

D. Experimental Results

In the experiments, we compare the proposed model with the previous state-of-the-art semantic segmentation models and analyze the model results from the perspective of pixel and instance, respectively, which demonstrates the superiority of our proposed model, and the effectiveness of MSFM and CAEM is illustrated by ablation experiments.

1) Comparison With Semantic Segmentation Model: U-net, Pspnet, Deeplabv3+, HRnet-v2, and DDRnet are selected as reference models to evaluate the performance of DBMFnet, all of which have great influence in the field of semantic segmentation. In addition, our laboratory's previous research works the high-resolution feature recursive alignment fusion network (FR-FRAFnet) [26] will also as reference models to evaluate the performance of DBMFnet. We used 90% of the images in the dataset MW-SAR1 as the training set to train the model and 10% of the images as the test set to evaluate the performance of the model.

The upper part of Table I shows the results of all the models trained and tested in the dataset MW-SAR1. The segmentation performance of DBMFnet is better than other models, with the values of MPA and mIoU at 85.26% and 76.23%, respectively. The higher the mIoU value of the model, the closer the predicted mask of the target is to the ground truth; the higher the MPA value of the model, the more accurate the localization of the target.

Table II shows the segmentation results of each model for different objects in the dataset MW-SAR1. Our model outperforms other models in IoU. Compared with other models, the IoU of the hammer improved by 0.32%–5.69%, respectively. The IoU of the wrench improved by 0.33%–6.02%, respectively. The IoU of the pistol increased by 2.39%–11.49%, respectively, and the IoU of the knife increased by 1.35%–16.41%, respectively, relative to the other objects, with

TABLE III
COMPARISON OF TARGET SEGMENTATION PERFORMANCE OF EACH MODEL IN THE MW-SAR2 DATASET

Class	U-net	Pspnet	Deeplabv3+	HRnet-v2	DDRnet	HR-FRAFnet(Su)	DBMFnet(our)
	IoU	IoU	IoU	IoU	IoU	IoU	IoU
Hammer	48.66	47.97	44.66	51.81	49.65	47.68	**52.16**
Pistol	58.67	57.2	55.8	58.6	56.26	56.4	**59.81**
Wrench	54.51	58.49	52.63	60.05	57.85	55.79	**62.45**
Knife	37.85	47.06	43.15	49.09	45.26	53.78	**60.72**

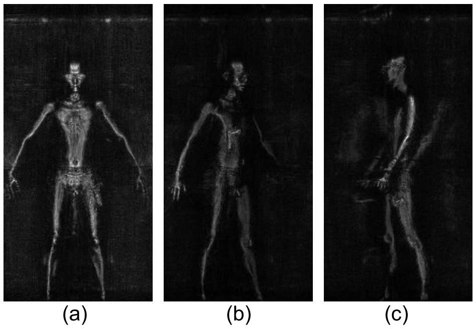


Fig. 10. Some images for two datasets. (a) Images from the MW-SAR1 dataset. (b) and (c) Images from the MW-SAR2 dataset.

the IoU of the knife improving by 16.41% at the most since the knife is the smallest in shape and closest to the human body's contour relative to the other objects, thus improving the difficulty in detecting it. The incorporation of the MSFM enhances our model's capability to detect small targets and object contours, resulting in improving the detection performance.

The segmentation results are tested using five representative images from the MW-SAR1 dataset. Fig. 11 shows the test results of each model. In the first row, the head part of the hammerhead is almost integrated with the human body, and the segmentation results of each model are different from each other; only the segmentation result of DBMFnet is the closest to ground truth. In the fourth row, the pistol's grip is so thin that only the U-net, FR-FRAFnet, and DBMFnet can segment the pistol's grip correctly, and the DBMFnet can segment the wrench's opening. In the fifth row, there are overlapping regions between the targets in the test image, which will increase the difficulty of segmentation. From the segmentation results, all other models have incorrect segmentation or incomplete segmentation, and only the segmentation results of DBMFnet are closest to the ground truth. The above experimental results show that our model has a more accurate pixel classification ability and the ability to identify the outline of small objects.

In practice, the screening system will face more complex situations, such as the irregular posture of the detected person and the random location of the contraband. To further verify the effectiveness and robustness of our method, we test the model on the dataset MW-SAR2, and the test results are shown in the lower part of Table I. Since the dataset MW-SAR2 contains more complex situations and the human pose in the image is different, the segmentation performance of each model has different degrees of decline. DBMFnet still gives the best segmentation results among all tested models, only its MPA is greater than 80%. This shows that DBMFnet has excellent pixel classification ability even in complex situations, and its mIoU is also the highest among all the models, at 67.00%.

Table III shows the segmentation performance of each model for various types of objects. DBMFnet has the highest IoU for various types of objects.

The segmentation results are tested using five representative images from the MW-SAR2 dataset. Fig. 12 shows the test results of each model. In the first row, the pistol is perpendicular to the scanning plane of the security system, and only the grip of the pistol can be imaged, with some models showing segmentation errors. In the third row, the knife is completely integrated into the human body contour, and some models have the same segmentation error. In the fourth row, the hammer is barely visible, and most of the models have segmentation errors or are not detected, only DDRnet, FR-FRAFnet, and DBMFnet are segmented and correctly categorized, while DBMFnet also accurately segments the head of the hammer. In the fifth row, all models detect the wrench, but the segmentation region is missed, and the segmentation region of DBMFnet is closest to the ground truth. The above results show that our proposed model can perform accurate segmentation even in complex scenarios, while all other models suffer from segmentation errors.

In terms of model complexity, the number of parameters in our proposed model has increased compared to our previous work FR-FRAFnet, but it remains minimal compared to other models, and in terms of FLOPs, our proposed model is also smaller than most of the models. Therefore, our model requires less hardware performance and can be easily used in a variety of security systems. In terms of inference speed, our proposed model has improved 19 frames/s compared to our previous work FR-FRAFnet. Although it did not have the best results, under the premise of guaranteeing detection accuracy, the detection speed of our model can already meet the actual security needs.

2) Comparison With Object Detection Model: Faster-RCNN, SSD, and YOLO v4 are selected as the comparison models, all of which have great influence in the field of object detection. In addition, our laboratory's previous research works the FR-FRAFnet will also as reference models to evaluate the performance of DBMFnet. The false and miss rates of models are mainly compared, which are of great importance in practical engineering. False detection indicates that the target is incorrectly detected by the device, and miss detection indicates that the target is not detected by the device. We also

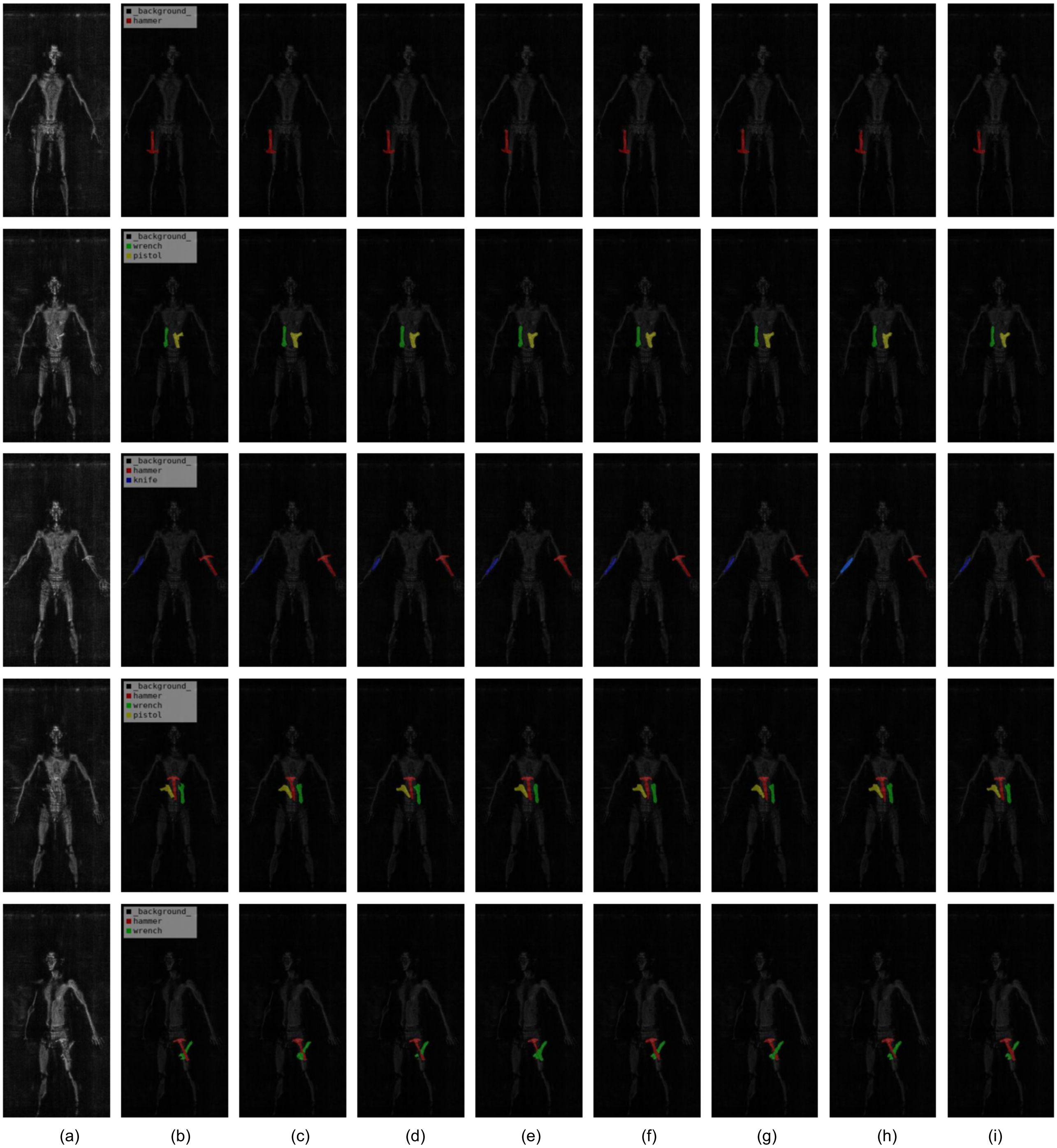


Fig. 11. Visualization test results for five examples in MW-SAR1. Black denotes the background, green denotes the wrench, yellow denotes the pistol, red denotes the hammer, and blue denotes the knife. (a) Test image. (b) Ground truth. Results of (c) Pspnet, (d) U-net, (e) Deeplabv3+, (f) HRnet, (g) DDRnet, (h) FR-FRAFnet, and (i) Results of DBMFnet.

conduct experiments on two datasets, and all models are trained on 90% of the images in the MW-SAR1 dataset and tested on the MW-SAR1 dataset and the MW-SAR2 dataset, respectively.

TABLE IV
COMPARISON OF OBJECT DETECTION PERFORMANCE OF EACH MODEL IN THE MW-SAR1 DATASET

Network model	F1(%)	False rate(%)	Miss rate(%)
Faster-RCNN	84.24	23.72	5.63
SSD	72.35	6.05	39.71
YOLO v4	73.01	12.56	35.92
HR-FRAFnet(Su)	93.17	6.45	7.2
DBMFnet	**95.65**	**5.47**	**3.2**

From Tables IV and V, our proposed DBMFnet achieved the best performance in both MW-SAR1 and MW-SAR2 datasets, with $F1$ score of 95.65% and 92.54%, respectively, and with the lowest values of misdetection rate and omission rate among all models. In addition, when testing the MW-SAR2 dataset with more complex scenes, each model showed a certain degree of decrease in $F1$ score compared to the MW-SAR1 dataset, with an increase in false rate and missed rate. However, there was no significant difference in

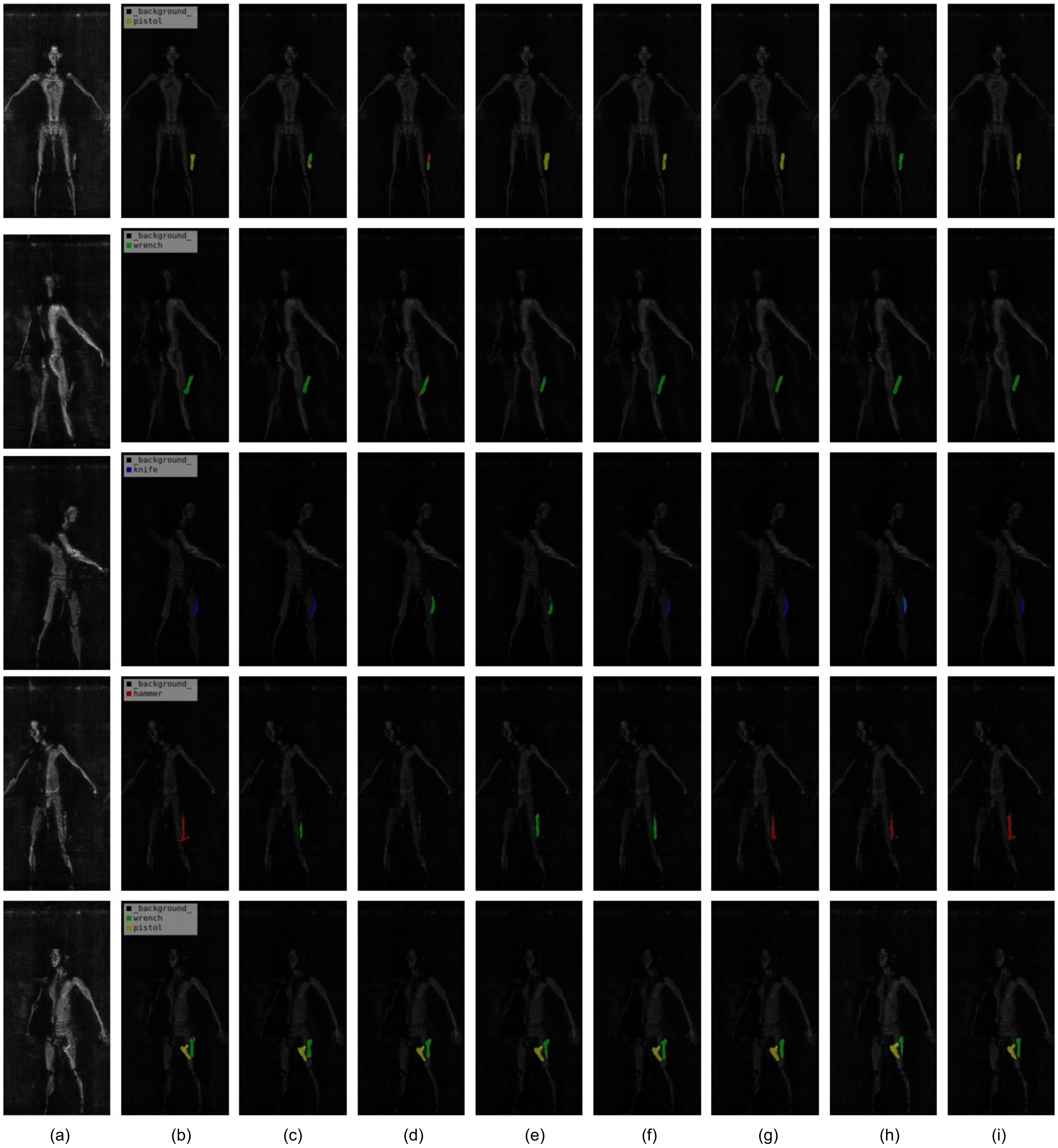


Fig. 12. Visualization test results for five examples in MW-SAR2. Black denotes the background, green denotes the wrench, yellow denotes the pistol, red denotes the hammer, and blue denotes the knife. (a) Test image. (b) Ground truth. Results of (c) Pspnet, (d) U-net, (e) Deeplabv3+, (f) HRnet, (g) DDRnet, (h) FR-FRAFnet, and (i) DBMFnet.

TABLE V
COMPARISON OF OBJECT DETECTION PERFORMANCE OF EACH MODEL IN THE MW-SAR2 DATASET

Network model	F1(%)	False rate(%)	Miss rate(%)
Faster-RCNN	65.26	41.57	12.80
SSD	48.86	28.92	62.30
YOLO v4	55.09	29.03	59.87
HR-FRAFnet(Su)	90.56	10.2	8.67
DBMFnet(our)	**92.54**	**7.37**	**7.55**

TABLE VI
MODEL PERFORMANCE COMPARISON USING DIFFERENT FEATURE FUSION MODULES

Network model	mIoU(%)	Params(M)	FLOPs(G)
Baseline	72.61	23.15	38.78
Baseline+FSM	74.1	22.44	100.80
Baseline+FDM	73.16	**21.65**	**45.27**
Baseline+MSFM	**75.11**	23.06	47.86
Baseline+FSM+CAEM	74.60	21.55	94.31
Baseline+FDM+CAEM	74.64	**19.32**	**44.92**
Baseline+MSFM+CAEM	**76.21**	19.54	47.36

the testing performance of DBMFnet between MW-SAR1 and MW-SAR2, and it still maintained high $F1$ score and low false rate and miss rate.

3) Ablation Experiment: In this article, we have used the MSMF and CAEM to improve the performance of DBMFnet.

TABLE VII
COMPARISON OF CAEM WITH OTHER CONTEXT EXTRACTION MODULES

PPM	ASPP	RES2	CAEM	mIoU(%)	Speed(FPS)	FLOPs(G)
				75.11	23	47.86
√				73.95	**28.5**	47.93
	√			74.38	28	47.48
		√		75.55	27.5	48.28
			√	**76.21**	26	**47.36**

To verify the effectiveness of the MSFM and CAEM, we conduct ablation experiments on them separately. All experiments are performed on the MW-SAR1 dataset.

We first analyze the effectiveness of the MSFM, using the DBMFnet without a feature fusion module as a baseline for comparison, directly splicing the output feature maps of the high-resolution branch and the low-resolution branch, and then upsampling to get the prediction results. Then, the feature fusion module is added, and the features are fused in the manner of FSM as shown in Fig. 3, FDM as shown in Fig. 4, and MSFM as shown in Fig. 5, respectively, and the fused feature maps are upsampled to obtain the prediction results.

Subsequently, we analyze the effectiveness of the CAEM and add CAEM to the feature fusion module in FSM, FDM, and MSFM modes, respectively, where only the lowest resolution feature map is allowed to access the CAEM and the other resolutions are left unchanged, and then, we perform feature fusion in the same way as the original one and upsample the fused feature maps to get the prediction results.

The results are shown in Table VI, where the mIoU of each model is improved relative to the baseline model after the addition of the feature fusion module, with MSFM improving the most by 2.50%. The CAEM module is added based on the existing feature fusion model. The results showed a slight increase in mIoU for each model after its addition, as well as a slight decrease in FLOPs, and the combination of MSFM and CAEM obtained the highest mIoU value by 76.21%.

We compare the CAEM with the PPM-based methods, ASPP-based methods, and the RES2Net [37] module. The results in Table VII show that the proposed module improves the performance of the model from 75.11% mIoU to 76.21% mIoU, and there is also a slight increase in the inference speed. Compared to the ASPP, it also achieves a 0.66% mIoU gain, while PPM and RES2 showed a slight decrease in model performance.

IV. Discussion

At present, the security system needs to detect fewer and fewer hidden objects, and most detection methods are based on the anchor boxes target detection algorithm. These algorithms achieve target detection and positioning by generating a bounding box around the object. However, these algorithms have limitations in detecting small targets or multiple overlapping targets in complex situations, increasing the likelihood of missed or false detections. We employ a semantic segmentation algorithm that annotates targets in pixels, reducing the significance of contour information in target identification and enhancing the capability of separating minuscule, multiple, and overlapping targets.

Millimeter wave security system usually uses the operating frequency range of 30–100 GHz. The higher the frequency of millimeter wave, the weaker the penetration. On the other hand, the higher the frequency of the millimeter wave, the higher the imaging resolution can be achieved and the higher the frequency of the millimeter wave generator and receiver, the higher the manufacturing cost and maintenance cost.

The resolution of 35-GHz systems is relatively low, especially compared to millimeter waves in higher frequency bands. Therefore, in some high-precision security tasks, the 35-GHz system may not meet the requirements. Although the resolution can be improved by increasing the number of antennas, the cost and complexity will increase. For the human security screening scenario, personal privacy should be considered. The low resolution of 35-GHz system can image the contraband outline and location without exposing the sensitive body information of the individual. Therefore, we choose 35 GHz as the operating frequency of the millimeter wave security inspection system, which is mainly a tradeoff between performance, cost, and complexity.

There are additional types of contraband in the security system, and we must create more data sets for training new contraband models. The current detection accuracy does not fully meet the demands of practical detection; hence, further improvement in the accuracy of detection and reduction of false alarms is necessary.

V. Conclusion

In this article, we propose a semantic segmentation network DBMFnet for the detection of contraband in millimeter wave SAR images. DBMFnet includes a DBPFEN, an MSFM, and a CAEM. The parallel output structure of DBPFEN is able to continuously fuse and exchange the information of high-resolution features and low-resolution features in the process of feature extraction, which reduces the feature loss in the process of downsampling and facilitates the recognition of small and overlapping targets, while the dual-branching structure greatly reduces the model complexity and lowers the hardware cost of deployment. MSFM enhances the model's ability to detect target edges, and CAEM can extract rich contextual information while reducing model computation and inference time. Our proposed model improves IoU by 2.97% compared to the existing best-performing semantic segmentation model when tested using the regular MW-SAR1 dataset and by 3.11% when tested using the complex MW-SAR2 dataset.

Overall, our proposed model has the better performance and robustness. Ablation experiments show that both the proposed MSFM and CAEM can effectively improve the mIoU value. Our method can be extended to other remote sensing scenarios, such as detecting ships, land classification, and water body detection. In the future work, we expect to apply instance segmentation to the detection of SAR images and hope to use as much information as possible in SAR images to promote the research of target recognition in SAR images.

References

[1] M. S. Saadat, S. Sur, S. Nelakuditi, and P. Ramanathan, "MilliCam: Hand-held millimeter-wave imaging," in *Proc. 29th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Aug. 2020, pp. 1–9.

[2] H. Jing, S. Li, X. Cui, G. Zhao, and H. Sun, "Near-field single-frequency millimeter-wave 3-D imaging via multifocus image fusion," *IEEE Antennas Wireless Propag. Lett.*, vol. 20, pp. 298–302, 2021.

[3] T. Nozokido, M. Noto, and T. Murai, "Passive millimeter-wave microscopy," *IEEE Microw. Wireless Compon. Lett.*, vol. 19, no. 10, pp. 638–640, Sep. 2009.

[4] R. Appleby and R. N. Anderton, "Millimeter-wave and submillimeter-wave imaging for security and surveillance," *Proc. IEEE*, vol. 95, no. 8, pp. 1683–1690, Aug. 2007.

[5] H. Işiker, I. Ünal, M. Tekbaş, and C. Özdemir, "An auto-classification procedure for concealed weapon detection in millimeter-wave radiometric imaging systems," *Microw. Opt. Technol. Lett.*, vol. 60, no. 3, pp. 583–594, Feb. 2018.

[6] X. Zang et al., "Metasurfaces for manipulating terahertz waves," *Light, Adv. Manuf.*, vol. 2, no. 2, p. 148, 2021.

[7] Y. Peng et al., "Three-step one-way model in terahertz biomedical detection," *PhotoniX*, vol. 2, no. 1, pp. 1–18, Jul. 2021.

[8] Z. Wang, T. Chang, and H.-L. Cui, "Review of active millimeter wave imaging techniques for personnel security screening," *IEEE Access*, vol. 7, pp. 148336–148350, 2019.

[9] L. Zheng, J. Yingkang, S. Zongjun, G. Jianping, W. Ziye, and Z. Ziran, "A synthetic targets detection method for human millimeter-wave holographic imaging system," in *Proc. 7th Int. Conf. Cloud Comput. Big Data (CCBD)*, Nov. 2016, pp. 284–288.

[10] P. Dai, Y. Yang, M. Wang, and R. Yan, "Combination of DNN and improved KNN for indoor location fingerprinting," *Wireless Commun. Mobile Comput.*, vol. 2019, pp. 1–9, Mar. 2019.

[11] K. Du, L. Zhang, W. Chen, G. Wan, and R. Fu, "Concealed objects detection based on FWT in active millimeter-wave images," *Proc. SPIE*, vol. 10322, pp. 386–391, Jan. 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Aug. 2015, pp. 770–778.

[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[15] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 21–37.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[17] Y. Zhu, X. Zang, H. Chi, Y. Zhou, Y. Zhu, and S. Zhuang, "Metasurfaces designed by a bidirectional deep neural network and iterative algorithm for generating quantitative field distributions," *Light, Adv. Manuf.*, vol. 4, no. 2, pp. 1–11, 2023.

[18] M. Yuan, Q. Zhang, Y. Li, Y. Yan, and Y. Zhu, "A suspicious multi-object detection and recognition method for millimeter wave SAR security inspection images based on multi-path extraction network," *Remote Sens.*, vol. 13, no. 24, p. 4978, Dec. 2021.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, May 2015, pp. 234–241.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[23] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2018, pp. 801–818.

[24] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[25] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, Mar. 2023.

[26] B. Su and M. Yuan, "Object recognition for millimeter wave MIMO-SAR images based on high-resolution feature recursive alignment fusion network," *IEEE Sensors J.*, vol. 23, no. 14, pp. 16413–16427, Jul. 2023.

[27] X. Wang et al., "Self-paced feature attention fusion network for concealed object detection in millimeter-wave image," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 224–239, Jan. 2022.

[28] C. Liu, M.-H. Yang, and X.-W. Sun, "Towards robust human millimeter wave imaging inspection system in real time with deep learning," *Prog. Electromagn. Res.*, vol. 161, pp. 87–100, 2018.

[29] P. Sun, T. Liu, X. Chen, S. Zhang, Y. Zhao, and S. Wei, "Multi-source aggregation transformer for concealed object detection in millimeter-wave images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6148–6159, Sep. 2022.

[30] L. Pang, H. Liu, Y. Chen, and J. Miao, "Real-time concealed object detection from passive millimeter wave images based on the YOLOv3 algorithm," *Sensors*, vol. 20, no. 6, p. 1678, Mar. 2020.

[31] C. Wang, K. Yang, and X. Sun, "Precise localization of concealed objects in millimeter-wave images via semantic segmentation," *IEEE Access*, vol. 8, pp. 121246–121256, 2020.

[32] D. Liang, J. Pan, Y. Yu, and H. Zhou, "Concealed object segmentation in Terahertz imaging via adversarial learning," *Optik*, vol. 185, pp. 1104–1114, May 2019.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jul. 2018, pp. 3–19.

[34] X. Li et al., "Semantic flow for fast and accurate scene parsing," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 775–793.

[35] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4141–4150.

[36] D. Yu, S. Ji, X. Li, Z. Yuan, and C. Shen, "Earthquake crack detection from aerial images using a deformable convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022.

[37] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2Net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2019.

Junhua Ding received the B.S. degree from Shanghai DianJi University, Shanghai, China, in 2022. He is currently pursuing the M.S. degree with the Shanghai University of Science and Technology, Shanghai.

His research interests include deep learning, object detection, and semantic segmentation.

Bofeng Su received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2021. He is currently pursuing the M.S. degree with the Shanghai University of Science and Technology, Shanghai.

His research interests include deep learning, object detection, and semantic segmentation.

Minghui Yuan received the Ph.D. degree from Southeast University, Nanjing, China, in 2006.

At present, he is an Associate Professor with the University of Shanghai for Science and Technology. His research interests include terahertz technology.