ISSN 1003-501X CN 51-1346/O4

光电工程

Opto-Electronic Engineering

中文核心期刊 中国科技核心期刊
Scopus CSCD

基于双分支多尺度融合网络的毫米波SAR图像多目标语义分割方法

丁俊华，袁明辉

引用本文:

丁俊华，袁明辉. 基于双分支多尺度融合网络的毫米波SAR图像多目标语义分割方法[J]. 光电工程，2023， **50**(12): 230242.

Ding J H, Yuan M H. A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network[J]. *Opto-Electron Eng*, 2023, **50**(12): 230242.

https://doi.org/10.12086/oee.2023.230242

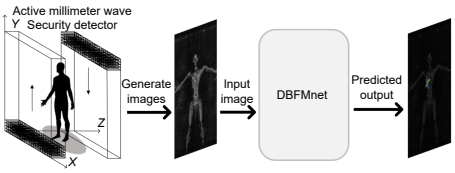
收稿日期: 2023-09-28; 修改日期: 2023-11-30; 录用日期: 2023-11-30

Opto-Electronic Engineering
光 电 工 程

Article
2023 年, 第 50 卷, 第 12 期

DOI: 10.12086/oee.2023.230242

基于双分支多尺度融合网络的毫米波 SAR 图像多目标语义分割方法



丁俊华[1,2]，袁明辉[1,2*]

[1]上海理工大学太赫兹技术创新研究院，上海 200093；
[2]上海理工大学光电信息与计算机工程学院，上海 200093

摘要： 在毫米波合成孔径雷达 (SAR) 安检成像违禁品的检测与识别中，存在着目标尺寸过小、目标被部分遮挡和多目标之间重叠等复杂情况，不利于违禁品的准确识别。针对这些问题，提出了一种基于双分支多尺度融合网络 (DBMFnet) 的违禁品检测方法。该网络使用 Encoder-Decoder 的结构，在 Encoder 阶段，提出一种双分支并行特征提取网络 (DBPFEN) 来增强特征提取；在 Decoder 阶段，提出一种多尺度融合模块 (MSFM) 来提高对目标的检测能力。实验结果表明，该方法的均交并比 (mIoU) 均优于现有的语义分割方法，降低了漏检与错检率。

关键词： 毫米波合成孔径雷达；违禁品检测；深度学习；语义分割；双分支多尺度融合网络

中图分类号： TP391　　**文献标志码：** A

丁俊华，袁明辉. 基于双分支多尺度融合网络的毫米波 SAR 图像多目标语义分割方法 [J]. 光电工程，2023，**50**(12): 230242
Ding J H, Yuan M H. A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network[J]. *Opto-Electron Eng*, 2023, **50**(12): 230242

A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network

Ding Junhua[1,2], Yuan Minghui[1,2*]

[1]Terahertz Technology Innovation Research Institute, University of Shanghai for Science and Technology, Shanghai 200093, China;
[2]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: There are several major challenges in the detection and identification of contraband in millimetre-wave synthetic aperture radar (SAR) security imaging: the complexities of small target sizes, partially occluded targets and overlap between multiple targets, which are not conducive to the accurate identification of contraband. To address these problems, a contraband detection method based on dual branch multiscale fusion network (DBMFnet) is proposed. The overall architecture of the DBMFnet follows the encoder-decoder framework. In the encoder stage, a dual-branch parallel feature extraction network (DBPFEN) is proposed to enhance the feature

收稿日期：2023-09-28；修回日期：2023-11-30；录用日期：2023-11-30
基金项目：国家自然科学基金资助项目 (61601291)；上海市科委专项资助 (14dz1206602)
*通信作者：袁明辉，yuanminghui@usst.edu.cn。

extraction. In the decoder stage, a multi-scale fusion module (MSFM) is proposed to enhance the detection ability of the targets. The experimental results show that the proposed method outperforms the existing semantic segmentation methods in the mean intersection over union (mIoU) and reduces the incidence of missed and error detection of targets.

Keywords: millimetre-wave synthetic aperture radar; contraband detection; deep learning; semantic segmentation; dual-branch multi-scale fusion network

1 引 言

随着毫米波技术的发展，毫米波安检方案越来越成熟[1-2]，与传统的安检技术相比，如 X 射线安检、红外线安检、金属探测仪安检等，毫米波安检成像不仅可以检测出织物下隐匿的金属物，还可以检测出塑料枪支、陶瓷刀具、炸药等危险品，最重要的是，毫米波具有非电离性，对人体不会造成伤害[3-5]。毫米波安检能获得清晰的图像信息，可大大降低虚警率，因而毫米波成像设备在人体安检场合得以广泛应用。

近几年，随着人工智能的发展，深度卷积神经网络在图像分类[6-7]、目标检测[8-9]、图像分割[10-15] 等领域都获得取得重大突破。因此，许多高效的深度学习算法被应用到毫米波图像的隐藏目标检测中。然而，这些算法一般检测的目标都是 RGB 图像与高分辨率图像，这与毫米波图像存在着较大的差异。毫米波图像普遍为灰度图像，且分辨率较低，同时由于毫米波探测器性能以及成像算法的影响，毫米波图像的信噪比远低于光学图像，图像对比度低，物体成像不完整。现有的研究已经证明，这些问题严重干扰了隐藏目标的识别与定位，降低了毫米波图像的检测性能[16]。因此，一些针对 SAR (millimetre-wave synthetic aperture radar) 成像特性的毫米波图像检测的深度学习方法被陆续提出。Liu 等人[17] 将毫米波图像与额外的空间深度图结合，并设计了一个新的损失函数，提高了对违禁品的检测率，但检测类别只有一个。Sun 等人[18] 设计了具有两种不同注意力机制的多源聚合 Transformer，能有效提升隐藏目标的检测性能。然而，上述研究只能检测出违禁品在人体表面的位置，无法对违禁品的种类进行识别，这导致安检人员需要进行二次检查，确定违禁品种类，以对违规人员进行处置。如果能直接对违禁物品进行识别，将会大大提升安检效率。Pang 等人[19] 使用了 YOLO v3 算法，对被动毫米波图像的人体隐蔽金属武器进行实时检测。他们主要针对手枪和人进行识别，两个目标形状差异明显，便于深度学习网络识别，但识别种类较少，识别精度和误报率仍有进一步改进的空间。

上述方法都是基于锚框的目标检测方法。锚框中不仅包含检测目标，还有背景噪声，这会影响目标的识别。同时，由于毫米波图像的可读性差，只用锚框框出检测目标，不利于安检人员的查看。语义分割方法则可以解决此类问题，它不同于目标检测和识别，语义分割可以实现图像像素级的分类。“语义”指具有人们可用语言探讨的意义，“分割”指图像分割。语义分割能够将整张图的每个部分分割开，使每个部分都有一定类别意义。与目标检测不同的是，目标检测只需要找到图片中目标，打上框然后分出类别。语义分割是将图像中的所有像素点分类，然后以描边的形式，将整张图不留缝隙地分割成各个区域。每个区域是一个类别，没有类别的默认为背景 (background)，这大大避免了背景噪声的干扰，可以将整个违禁品分割出来进行种类划分，同时上色，即便在可读性差的毫米波图像中，安检人员也可轻易地看出物体的形状，快速获取因违禁品的信息，以便后续的处理。Wang 等人[20] 通过扩展卷积来提高感受野，以用来提高检测性能，并且在实验用取得良好的结果。Liang 等人[21] 通过选定连接 U-net 网络结合生成对抗网络，实现在人体中分割出违禁物品。然而，目前 SAR 图像检测的研究重心主要是目标成像问题和背景干扰问题，但 SAR 图像检测还存在多目标间的相互干扰、小目标如何检测等问题。因此，要实现 SAR 图像中隐藏目标的精确识别与定位还存在着一些挑战。针对以上难点，本文主要贡献有：1) 提出双分支特征提取网络 (DBPFEN, dual-branch parallel feature extraction network)，该网络为双分支并行输出的结构，两个分支之间建立双边连接，进行重复的信息融合，能够大大减小模型复杂度，同时提高模型对成像不完整的目标、相互阻挡的目标和小目标的识别能力；2) 提出一个多尺度融合模块 (MSFM, multi-scale fusion module)。该模块能够将多个不同层的高分辨率

特征图融入低分辨率特征图中，从而去获得更多的位置细节信息以增加对目标的检测能力。

2 基本原理

2.1 双分支多尺度融合网络 (DBMFnet)

在本文中，我们提出了 DBMFnet 语义分割模型来识别 SAR 图像中的违禁品，该网络为 Encoder-Decoder 结构。在 Encoder 阶段，我们采用双分支并行特征提取网络 (DBFEN)，在特征提取的过程中一个分支保持高分辨率不变，另一个分支通过多次下采样操作提取丰富的语义信息，两个分支之间建立双边连接，进行重复的特征融合。在 Decoder 阶段，为了增加对目标的检测能力，低分辨率分支特征图、高分辨率分支特征图以及通过 Skip 连接层获得的更高分辨率的特征图同时引入多尺度融合模块 (MSFM)，实现多个不同分辨率特征图之间的相互融合。之后通过上采样二倍恢复原图大小后输出预测图。其网络结构如图 1 所示。

双分支并行特征提取网络具体实现流程如图 1 的 Encoder 阶段所示，输入图片大小为 512×512，经过两个卷积核大小为 3×3、步距为 2 的卷积层，变为输入分辨率的 1/2 与 1/4。1/4 分辨率的特征图经过 2 个堆叠的 basic block 残差模块，通道数翻倍，分辨率降为输入图像的 1/8，得到高分辨分支的初始特征图。将高分辨率分支的初始特征图同样经过 2 个堆叠的 basic block 残差模块，通道数翻倍，分辨率降为输入图像的 1/16，得到低分辨率分支的初始特征图。两个分支的特征图进行特征融合，其具体融合过程如图 2 所示，高分辨率分支的特征图F^{h}通过一个卷积核大小为 3×3 步距 2 的卷积层来降低分辨率与改变通道数，之后与低分率分支的特征图相加得到F'^{h}。同样地，低分辨率分支的特征图F^{l}需要通过双线性插值与一个卷积核大小为 1×1 步距为 1 的卷积层来提高分辨率与改变通道数，之后与高分辨分支的特征图相加得到F'^{l}，融合完成后输出F'^{h}与F'^{l}。后续重复此过程，高分辨率分支分辨率保持不变，低分辨率分支不断地进行下采样。低分辨分支的特征图分辨率分别为输入分辨率的 1/16、1/32 和 1/64，对应的通道数为 256、512 和 1024，高分辨率分支的特征图分辨率则保持为输入分辨率的 1/8，通道数为 128，具体细节如表 1 所示。

2.2 多尺度融合模块 (MSFM)

语义分割的解码器阶段需要将低分辨率的特征图逐步恢复到原始图像的分辨率。常见的操作是在恢复的过程中通过 skip connections 融合下采样过程中的特征图，U-net 和 DeeplabV3+使用拼接 (FCM) 的方式融合特征图，如图 3(a) 所示，而 FCN 则是使用相加 (FDM) 的方式融合特征图，如图 3(b)。然而，这些方式忽略了不同分辨率的特征图之间的不对齐，可能会丢失许多的语义信息，导致目标边界上的分割性能较差。在 DBMFnet 中，高分辨分支与低分辨分支分别输出为输入分辨率 1/8 的特征图与 1/64 分辨率的特征图，二者尺寸相差过大。如果用传统线性插值的方式上采样再进行融合，则会丢失许多的语义信息。同理，高分辨分支特征图的分辨率与原图特征图的分辨率也相差过大，直接上采样同样会丢失许多的语义信息。为了提高对目标的检测能力，提出一种多尺度融合模

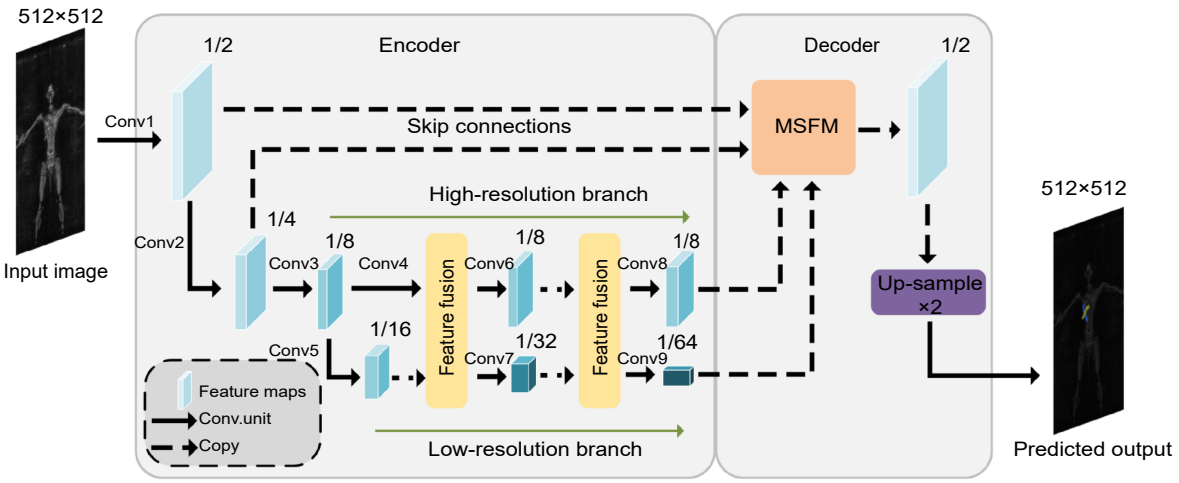


图 1 DBMFnet 网络结构图

Fig. 1 DBMFnet network structure diagram

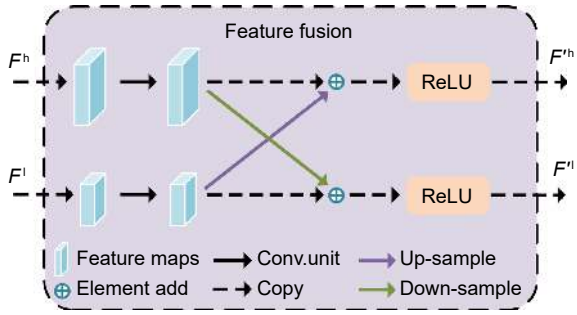


图 2 特征融合过程

Fig. 2 Feature fusion process

表 1 双分支特征提取网络结构

Table 1 Architectures of DBFEN

Stage	Output	DBFEN	Stage	Output	DBFEN
Conv1	256×256	3×3, 64, stride 2	Conv6	64×64	$\begin{pmatrix}3\times3,\ 128\\3\times3,\ 128\end{pmatrix}\times2$
Conv2	128×128	3×3, 64, stride 2	Conv7	16×16	$\begin{pmatrix}3\times3,\ 256\\3\times3,\ 512\end{pmatrix}\times2$
Conv3	64×64	$\begin{pmatrix}3\times3,\ 64\\3\times3,\ 128\end{pmatrix}\times2$	Conv8	64×64	$\begin{pmatrix}3\times3,\ 128\\3\times3,\ 256\end{pmatrix}\times2$
Conv4	64×64	$\begin{pmatrix}3\times3,\ 128\\3\times3,\ 128\end{pmatrix}\times2$	Conv9	8×8	$\begin{pmatrix}3\times3,\ 512\\3\times3,\ 1024\end{pmatrix}\times2$
Conv5	32×32	$\begin{pmatrix}3\times3,\ 128\\3\times3,\ 256\end{pmatrix}\times2$			

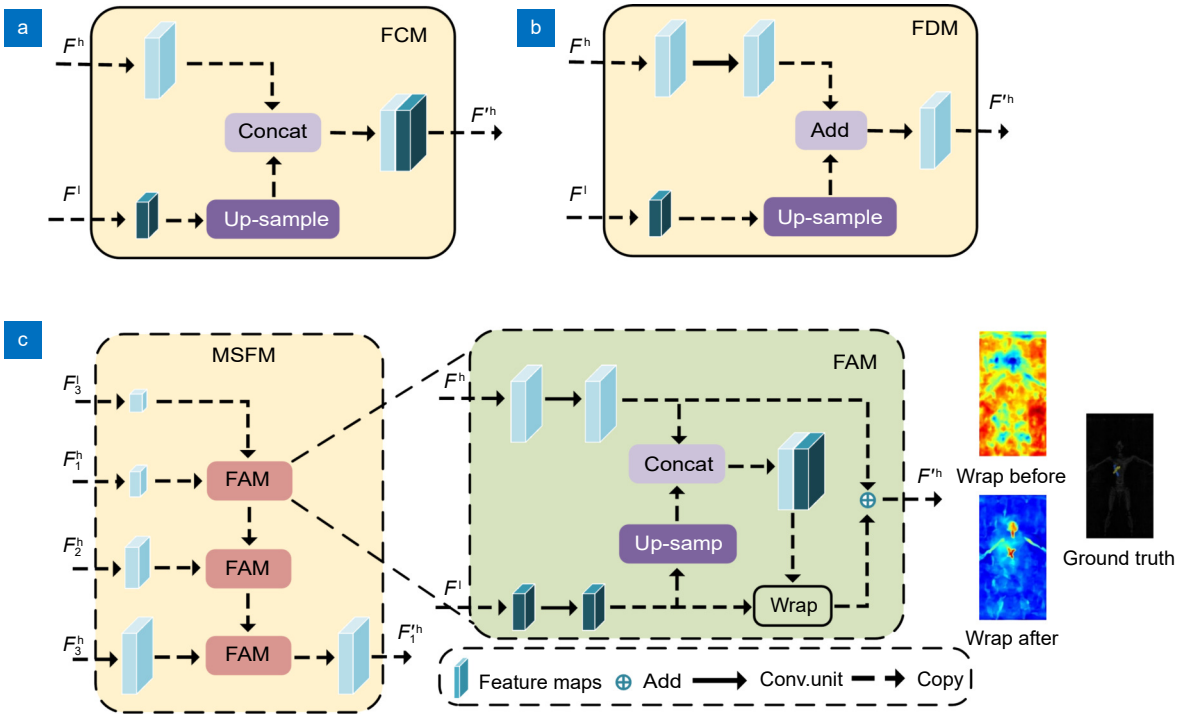


图 3 不同的特征融合方式。(a) FCM; (b) FDM; (c) MSFM;

Fig. 3 Different feature fusion methods. (a) FCM; (b) FDM; (c) MSFM

块 (MSFM)，如图 3(c) 所示。该模块由特征对齐模块 (FAM) 构成[22]，它将高分辨率特征图F^{h}与低分辨率特征图F^{l}作为输入，分别经过一个 1×1 的卷积改变为相同通道数，将高分辨特征图与经过双线性插值上采样的低分辨率特征图进行拼接，然后拼接后的特征图经过一个 3×3 的卷积层得到与F^{h}相同大小的偏移场$\Delta \in R^{2\times H^{\mathrm{h}}\times W^{\mathrm{h}}}$。在数学上，上述步骤可以写为

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

其中：$upsample(.)$ 为双线性插值上采样，$cat(.)$ 代表拼接操作，$conv(.)$ 代表 3×3 的卷积层。在获得偏移场后，低分辨率特征图将被扭曲 (warp) 为高分辨率特征图，偏移场的每个点$(X_{\mathrm{h}}, Y_{\mathrm{h}})$都需要被$F^{\mathrm{l}}$中的点$(X_{\mathrm{l}}, Y_{\mathrm{l}})$映射，公式如:

$$\begin{cases} X_{\mathrm{l}} = \dfrac{(1+\Delta x)}{N} X_{\mathrm{h}} \\ Y_{\mathrm{l}} = \dfrac{(1+\Delta y)}{N} Y_{\mathrm{h}} \end{cases}, \tag{2}$$

其中：Δx和Δy表示点$(X^{\mathrm{h}}, Y^{\mathrm{h}})$可学习的 2-D 变换偏移，$N$ 表示高低分辨率相差的倍数，然后，通过可微分双线性采样机制使用$(X^{\mathrm{l}}, Y^{\mathrm{l}})$的四领域插值来获得扭曲的高分辨率特征图$U(F_{\mathrm{h}}(X_{\mathrm{h}}, Y_{\mathrm{h}}),)$的位置。数学表达式为

$$\begin{aligned} U(F^{\mathrm{l}}(X_{\mathrm{l}}, Y_{\mathrm{l}}), \Delta) = & \sum_{Y_{\mathrm{l}}=1}^{H^{\mathrm{l}}} \sum_{X_{\mathrm{l}}=1}^{W^{\mathrm{l}}} F^{\mathrm{l}}(X_{\mathrm{l}}, Y_{\mathrm{l}}) \\ & \cdot \max(0, 1-|X_{\mathrm{l}} - X_{\mathrm{h}}|) \\ & \cdot \max(0, 1-|Y_{\mathrm{l}} - Y_{\mathrm{h}}|) , \end{aligned} \tag{3}$$

其中：H^{l}和W^{l}表示低分辨率特征图的大小。整个 FAM 的数学表达式如下：

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

在 MSFM 中，引入低分辨率分支、高分辨率分支与通过 skip connections 获得的输入分辨率 1/4 的特征图，分别为F_1^{l}、F_1^{h}、F_2^{h}，引入 skip connections 获得的输入分辨率 1/2 的特征图为F_3^{h}，将多个不同层的高分辨率特征图融入低分辨率特征图F_1^{l}中，得到输入分辨率 1/2 的高分辨率特征图$F_1'^{\mathrm{h}}$。所提出的多尺度融合模块，增加了对目标的检测能力。

3 实验与分析

3.1 数据集

基于实验室的 MIMO-SAR 架构的主动式毫米波人体安检成像系统，构建了毫米波 SAR 人体安检图像数据集，命名为 HM-SAR，数据集共 1100 张图片，其中 90% 的图片用于训练，10% 的图片用于测试。该系统为平扫系统，工作频率为 35 GHz，天线阵列放置于 X 轴，系统沿 Y 轴上下扫描，可同时对人体正面和背面进行扫描，成像效果如图 4(a)、4(b) 所示。扫描成像以 jpg 格式保存，每个图像固定为 200×400 像素。数据集中的图像使用 Labelme 进行标记，对不同的目标赋予不同颜色的标签，剩下未标记的全部被归为背景类。在数据采集过程中，目标物被随机隐藏在人体表面与身体边缘。我们使用了四种违禁品作为识别目标，分别为扳手、锤子、手枪和小刀。

图 4 HM-SAR 安检图片。(a) 背面扫描的人体图片；(b) 正面扫描的人体图片

Fig. 4 HM-SAR security images. (a) Back scanning image of the human body; (b) Frontal scanning image of the human body

3.2 评价指标

为了进行定量评估，测试网络性能参数主要为平均像素准确率 (MPA)、平均交并比 (mIoU)、$F1$。

$$MPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} , \tag{5}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} , \tag{6}$$

$$Precision = \frac{TP}{TP+FP} , \tag{7}$$

$$Recall = \frac{TP}{TP+FN} , \tag{8}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} , \tag{9}$$

上述式中，k 表示像素类总数，p_{ii} 表示原本为 i 类被预测为 i 类的像素点数量，p_{ij} 表示原本为 i 类被预测为 j 类的像素点数量，p_{ji} 表示原本为 j 类被预测

为 i 类的像素点数量。*MPA* 表示对所有类别的 CPA 的平均值。*mIoU* 表示对所有类别的 IoU 的平均值。*mIoU* 的值越高，表示模型的预测值和真实值的重合度更好，说明模型的分割性能越好。*Precision*为精确率，*Recall*为召回率，*F*1 是精确率和召回率的调和平均值，*TP*是将正类预测为正类数，*FP*是将负类预测为正类数，*FN*是将正类预测为负类数。本文中的所提出的模型是在 Pytorch 1.12.0 框架中实现的，Cuda 版本为 11.6，采用 VOC2007 格式的数据集，并使用 Adam 优化器对网络进行端到端的训练，初始学习率为 5×10^{-4}，最小学习率为 5×10^{-6}。所有实验均在具有单张 NVIDIA GeForce RTX 3090 GPU 台式电脑上进行，迭代训练次数为 300，批量大小为 8，输入图片大小为 512×512。

3.3 热力图分析

为了进一步了解 DBMFnet 模型网络结构中各个步骤特征图变化情况，我们对整个网络进行热力图分析，我们将网络中各个步骤输出的特征图提取出来，并进行可视化展示，展示结果如图 5 所示，图中颜色越偏向红色的区域表示网络越关注的区域，即权重值越高，预测为藏匿物的概率也就越大。从图中可以看出，高分辨率分支为浅层特征图，其中语义信息较少，目标位置相对比较准确，图中第一行浅层特征图几乎都能清晰看出人体轮廓。而低分辨率分支为深层特征图，其中语义信息比较丰富，目标位置比较粗略，图中第二行的深层特征图已经几乎看不清楚人体的位置信息，这导致其检测小目标的能力较弱。在 DBFEN 中，我们不断地将浅层特征图和深层特征图进行融合，从而让浅层特征图获得更多的语义信息。在 MSFM 中，让最深的特征图依次融合来自不同层的浅层特征图，从而去获得更多的位置细节信息以增加对目标的检测能力，图中可以看出特征图经过 MSFM 后，网络最关注区域变为藏匿物的存在的区域。

3.4 有效性评估

为了评估 DBMFnet 模型的有效性，选择 U-net、Pspnet、FCN-8s、Deeplabv3+和 HRnet-v2 作为模型性能的对比对象，这些都是语义分割领域的经典模型，且都有很大的影响力。根据表 2 中的数据，我们提出的 DBMFnet 具有最好的分割性能，MPA、mIoU 和 F1 值分别为 85.01%、75.44%、85.21%。模型的

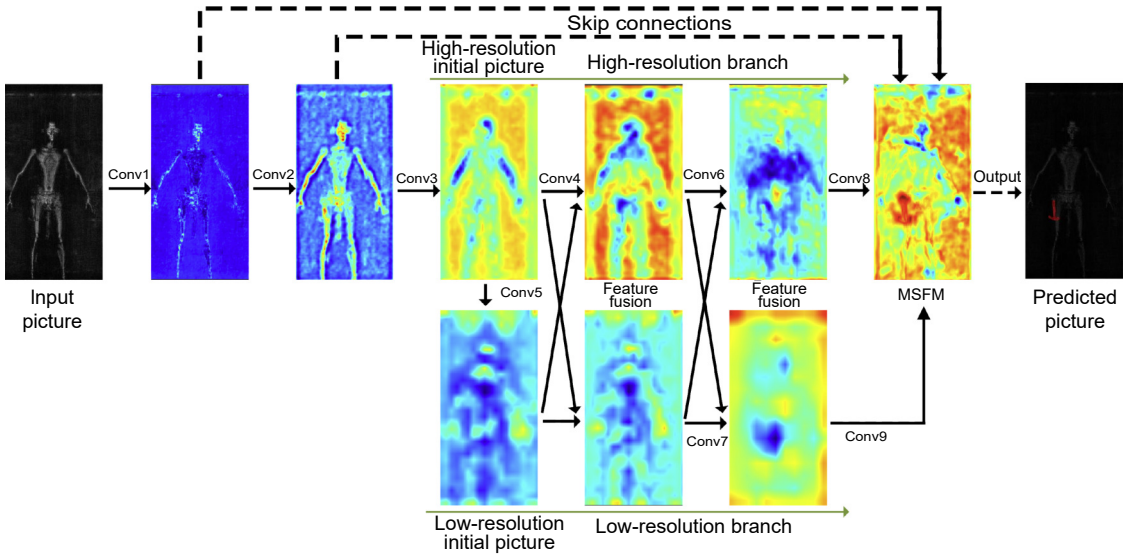


图 5 DBMFnet 热力图

Fig. 5 DBMFnet thermal diagram

表 2 各模型在 HM-SAR 数据集中的分割性能比较

Table 2 Comparisons of the segmentation performance of each model in the HM-SAR dataset

Network model	MPA/%	mIoU/%	F1/%	Network model	MPA/%	mIoU/%	F1/%
U-net	80.29	70.35	81.87	Deeplabv3+	81.05	70.58	82.00
Pspnet	82.98	72.32	83.28	HRnet-v2	82.33	72.90	83.69
FCN-8s	81.29	72.11	83.11	DBMFnet (ours)	**85.01**	**75.44**	**85.21**

表 3 各模型在 HM-SAR 数据集中的目标分割性能比较

Table 3 Comparisons of the objects segmentation performance of each model in the HM-SAR dataset

Class	U-net		Pspnet		Deeplabv3+		HRnet-v2		FCN-8s		DBMFnet (ours)	
	Pre	IoU	Pre	IoU	Pre	IoU	Pre	IoU	Pre	IoU	Pre	IoU
Hammer	80.74	61.98	76.49	63.7	80.15	63.99	79.93	67.35	79.16	65.17	**81.91**	**69.33**
Wrench	82.66	66.78	82.88	71.84	80.61	66.57	78.80	66.15	84.04	69.56	**84.22**	**75.24**
Pistol	75.63	63.77	77.3	64.21	75.45	62.65	85.71	69.47	81.07	65.81	**87.89**	**70.56**
Knife	78.59	59.4	81.36	62.01	78.82	59.84	81.67	61.68	80.06	60.16	**82.55**	**66.15**

图 6 各模型测试结果，每一行代表相同图片测试的结果，每一列代表同一模型的测试结果。黑色像素表示背景，红色像素表示锤头，绿色像素表示扳手，黄色像素表示手枪，蓝色像素表示小刀

Fig. 6 Test results of each model. Each row represents the test results of the same picture, and each column represents the test results of the same model. Black denotes the background, green denotes the wrench, yellow denotes the pistol, red denotes the hammer, and blue denotes the knife

mIoU 值越高，说明目标的预测掩码越贴近 ground truth，模型的 MPA 值越高，说明对目标的定位就越准确，模型的 F1 值越高说明模型对目标识别的精度越好。

表 3 显示了各模型对不同目标的分割结果。我们的模型无论是目标的精确率 (Pre) 还是交并比 (IoU)，都优于其他的模型。锤头的精确率和交并比相对于其他模型分别提高了 1.17% ~ 5.42% 和 1.98% ~ 7.35%，扳手的精确率和交并比相对于其他模型分别提高了 1.34% ~ 5.42% 和 3.40% ~ 9.09%，手枪的精确率和交并比相对于其他模型分别提高了 2.18% ~ 12.44% 和 1.09% ~ 6.79%，小刀的精确率和交并比相对于其他模型分别提高了 0.88% ~ 3.96% 和 4.14% ~ 6.75%。结果表明 DBMFnet 对比与其他模型，不同目标的精确率和交并比均有提升，这是因为多尺度融合模块的引入，提升了我们模型检测小目标和物体轮廓的能力，从而提高了检测性能。

我们选取五张图像的分割结果来更直观地说明各个模型的分割效果。图 6 显示了各个模型在 HM-SAR 数据集中的测试结果，图中第一行，手枪与安检系统扫描平面垂直，只能成像出手枪的握把，部分模型出现了分割错误的情况。图中第二行，小刀完全融入了人体轮廓，部分模型同样出现了分割错误的情况。图中第三行，手枪的握把非常的细，只有 Unet 与 DBMFnet 能正确的将手枪的握把分割出来，并且 DBMFnet 还能够分割出扳手的开口。图中第四行，测试图中的目标之间有重叠的区域，这会增加分割的难度，从分割结果可看出，其他模型都存在错误分割或分割不完整，只有 DBMFnet 的分割结果最接近 Ground truth。图中第五行，锤头几乎看不清，大部分模型出现分割错误或没有检测出来的情况，仅有 DBMFnet 分割出来且类别正确。以上实验结果说明本文模型具有更准确的像素分类能力以及识别小物体轮廓的能力。

除了测试模型精度外，我们还对模型复杂性和推理速度进行测试，如表 4 所示。这里使用参数量、浮点运算 (GFLOPs) 量和 FPS (frames per second) 来评估模型的复杂性。本文中，所有模型均通过单尺度推断进行评估，输入图像分辨率为 512×512。表 4 中，我们提出模型的参数量和浮点运算量最小，因此本文模型对硬件性能的要求较低，可以很容易地部署在各种安检系统中。在推理速度方面，与其他模型相比，本文模型没有优势，这是因为：1) 在特征提取阶段的特征融合过程中，低分辨率分支需要等待高分辨率分支计算完成后才能进行融合，在等待同步需要花费时间；2) 在特征融合阶段多尺度融合模块中，低分辨率特征图依次向上融合，高分辨特征依然要等待下级特征图融合完成后才能继续向上融合，等待过程中也会花费时间。然而，在保证检测精度的前提下，我们模型的检测速度已经可以满足实际的安检需求。

3.5 消融实验

在本文中，双分支特征提取网络用于编码器来提高特征提取性能；多尺度融合模块用于解码器来提高分割精确度。为了验证所提出的模块有效性，我们分别对其进行了消融实验，所有实验均在数据集 HM-SAR 上进行。

我们首先评估双分支特征提取网络的性能，我们使用没有解码器的 DBMFnet，即双分支特征提取网络作为比较基线 (Baseline)，将双分支特征提取网络的高分辨率分支输出与低分辨率分支上采样 8 倍后的结果直接进行拼接，然后再上采样 8 倍得到预测结果，如图 7 所示。随后引入解码器模块逐步恢复边界，将 Baseline 中拼接后的结果上采样 2 倍后与 1/4 输入分辨率的特征图进行融合，再上采样 2 倍与 1/2 输入分辨率的特征图进行融合，再上采样 2 倍得到预测结果。解码器分别以 FCM(图 3(a) 所示)、FDM (图 3(b) 所示) 和 MSFM (图 3(c) 所示) 的方式进行特征融合。结果如表 5 所示，未加入解码器的 Baseline 模型性能已经优于带解码器的 Deeplabv3+和 FCN-8s 网络的性能。在加入解码器之后，各个模型的 mIoU 相对于基线模型均有提升，计算量也有所增加，而 MSFM 方式的特征融合达到了最好的平衡，在增加少许计算量 (GFLOPS) 的同时，性能提升最多，并且参数量还有所降低。

4 总 结

针对安检系统 SAR 图像中违禁品的检测，提出一种语义分割方法 DBMFnet，该方法能够精确定位和识别 SAR 图像中多种小目标的违禁品。该方法包括双分支并行特征提取网络和多尺度融合模块。并行输出结构能够在特征提取过程中不断融合交换高分辨率特征和低分辨率特征的信息，减少下采样过程中的特征损失，有利于小目标和相互重叠目标的识别，同时双分支结构大大减少了模型复杂度，降低了部署的

表 4 各个模型的计算复杂度和推理速度

Table 4 Calculation complexity and inference speed of each model

Network model	Params/M	GFLOPs	Speed/(f/s)
U-net	24.89	452.31	32
Pspnet	46.7	118.43	**33.5**
FCN-8s	32.95	277.74	16
Deeplabv3+	54.71	166.87	21
HRnet	29.55	80.18	11.5
DBMFnet(our)	**19.54**	**47.36**	26

表 5 使用不同解码器模块的模型性能对比

Table 5 Comparisons of models using different decoder modules

Network model	mIoU	Params/M	GFLOPs
Baseline	72.61	23.15	**38.78**
Deeplabv3+(FCM)	70.58	54.71	166.87
FCN-8s(FDM)	72.11	32.95	277.74
Baseline+FCM	74.1	22.44	100.8
Baseline+FDM	73.16	**21.65**	45.27
Baseline+MSFM	**75.44**	23.06	47.86

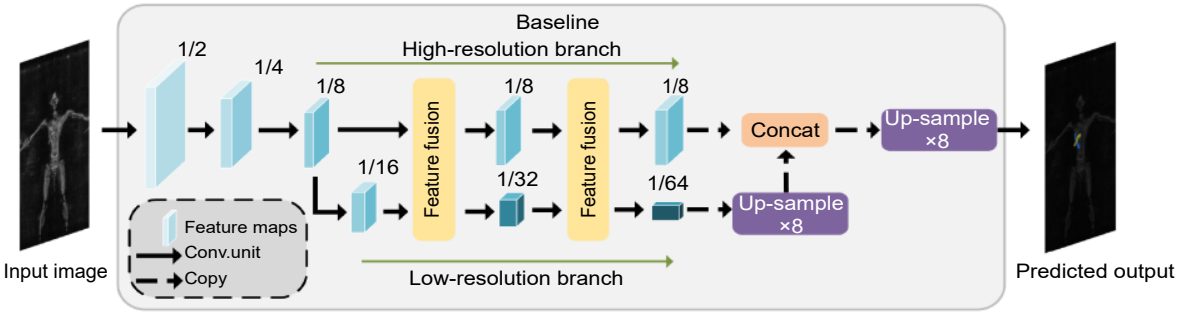


图 7 基线模型

Fig. 7 Baseline model

硬件成本。使用 HM-SAR 数据集测试时，我们所提出的模型与现有的表现最好语义分割模型相比 mIoU 提升了 2.54%。消融实验表明，所提出多尺度融合模块均可有效提升 mIoU 值。我们的方法可以拓展到其他遥感场景，比如检测船舶、土地分类、水体检测等。在未来的工作中，我们期望将实例分割应用到 SAR 图像的检测中，能利用 SAR 图像中尽可能多的信息，促进 SAR 图像目标识别的研究。

参考文献

[1] Saadat M S, Sur S, Nelakuditi S, et al. MilliCam: hand-held millimeter-wave imaging[C]//*Proceedings of 29th International Conference on Computer Communications and Networks*, Honolulu, 2020: 1–9. https://doi.org/10.1109/ICCCN49398.2020.9209710.

[2] Jing H D, Li S Y, Cui X X, et al. Near-field single-frequency millimeter-wave 3-D imaging via multifocus image fusion[J]. *IEEE Antennas Wirel Propag Lett*, 2021, **20**(3): 298−302.

[3] Nozokido T, Noto M, Murai T. Passive millimeter-wave microscopy[J]. *IEEE Microw Wirel Compon Lett*, 2009, **19**(10): 638−640.

[4] Appleby R, Anderton R N. Millimeter-wave and submillimeter-wave imaging for security and surveillance[J]. *Proc IEEE*, 2007, **95**(8): 1683−1690.

[5] Işiker H, Ünal İ, Tekbaş M, et al. An auto - classification procedure for concealed weapon detection in millimeter - wave radiometric imaging systems[J]. *Microw Opt Technol Lett*, 2018, **60**(3): 583−594.

[6] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016: 770–778. https://doi.org/10.1109/CVPR.2016.90.

[7] Chollet F. Xception: deep learning with depthwise separable convolutions[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017: 1251–1258. https://doi.org/10.1109/CVPR.2017.195.

[8] Ren S Q, He K M, Girshick R B, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//*Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, 2015.

[9] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[C]//*Proceedings of the 14th European Conference on Computer Vision*, Amsterdam, 2016: 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

[10] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//*Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 2016: 779–788. https://doi.org/10.1109/CVPR.2016.91.

[11] Xie E Z, Wang W H, Yu Z D, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[C]//*Proceedings of the 35th International Conference on Neural Information Processing Systems*, 2021.

[12] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network[C]//*Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. https://doi.org/10.1109/CVPR.2017.660.

[13] Chen L C, Zhu Y K, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//*Proceedings of the 15th European*

Conference on Computer Vision, Munich, 2018. https://doi.org/10.1007/978-3-030-01234-2_49.

[14] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]// *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, 2019. https://doi.org/10.1109/CVPR.2019.00584.

[15] Pan H H, Hong Y D, Sun W C, et al. Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes[J]. *IEEE Trans Intell Transp Syst*, 2023, **24**(3): 3448−3460.

[16] López-Tapia S, Molina R, de la Blanca N P. Deep CNNs for object detection using passive millimeter sensors[J]. *IEEE Trans Circuits Syst Video Technol*, 2019, **29**(9): 2580−2589.

[17] Liu C Y, Yang M H, Sun X W. Towards robust human millimeter wave imaging inspection system in real time with deep learning[J]. *Prog Electromagn Res*, 2018, **161**: 87−100.

[18] Sun P, Liu T, Chen X T, et al. Multi-source aggregation transformer for concealed object detection in millimeter-wave images[J]. *IEEE Trans Circuits Syst Video Technol*, 2022, **32**(9): 6148−6159.

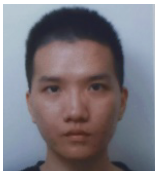
[19] Wang L H, Yuan M H, Huang H, et al. Recognition of edge object of human body in THz security inspection system[J]. *Infrared Laser Eng*, 2017, **46**(11): 1125002.
王林华, 袁明辉, 黄慧, 等. 太赫兹安检系统人体图像边缘物体识别[J]. 红外与激光工程, 2017, **46**(11): 1125002.

[20] Wang C J, Yang K H, Sun X W. Precise localization of concealed objects in millimeter-wave images via semantic segmentation[J]. *IEEE Access*, 2020, **8**: 121246−121256.

[21] Liang D, Pan J X, Yu Y, et al. Concealed object segmentation in terahertz imaging via adversarial learning[J]. *Optik*, 2019, **185**: 1104−1114.

[22] Li X T, You A S, Zhu Z, et al. Semantic flow for fast and accurate scene parsing[C]//*Proceedings of the 16th European Conference on Computer Vision*, Glasgow, 2020: 775–793. https://doi.org/10.1007/978-3-030-58452-8_45.

作者简介



丁俊华 (1998-)，男 ，硕士研究生，从事计算机视觉研究。

E-mail：1483802325@qq.com

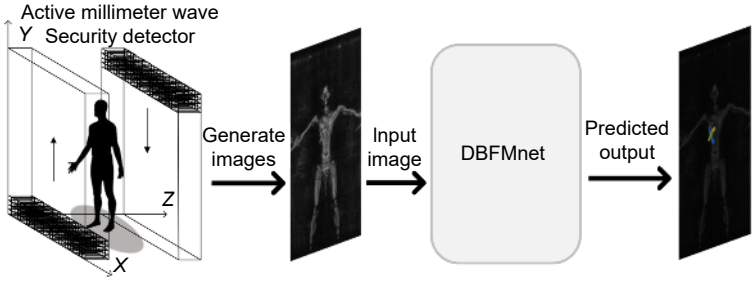
【通信作者】袁明辉 (1974-)，男，博士，副教授，从事太赫兹器件研究。

E-mail: yuanminghui@usst.edu.cn

扫描二维码，获取PDF全文

A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network

Ding Junhua[1,2], Yuan Minghui[1,2*]



System flow chart

Overview: With the advancements of millimeter wave technology, millimeter wave security inspection systems have reached a higher level of maturity. Compared with traditional security inspection technologies such as X-ray, infrared, and metal detectors, millimeter wave security imaging not only enables the detection of the metallic objects hidden under fabrics, but also identifies dangerous items such as plastic firearms, knives, explosives, etc. Significantly, it is crucial to note that millimeter waves are non-ionizing and do not cause harm to the human body. The utilization of millimeter wave security inspection enables the acquisition of precise image information and significantly reduces the occurrence of false alarms, making millimeter wave imaging equipment extensively employed in the security inspection of the human body.

There are several major challenges in the detection and identification of contraband in millimetre-wave synthetic aperture radar (SAR) security imaging: the complexities of small target sizes, partially occluded targets and overlap between multiple targets, which are not conducive to the accurate identification of contraband. To address these problems, a contraband detection method based on Dual Branch Multiscale Fusion Network (DBMFnet) is proposed. The overall architecture of the DBMFnet follows the encoder-decoder framework. In the encoder stage, a dual-branch parallel feature extraction network (DBPFEN) is proposed to enhance the feature extraction. In the feature extraction process of DBMFnet, one branch preserves the high resolution while the other branch extracts the 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 the 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. In the decoder stage, a multi-scale fusion module (MSFM) is proposed to enhance the detection ability of the targets. The module consists of the Feature Alignment Module (FAM), which allows multiple low-resolution feature maps to merge into high-resolution maps. The FAM is inspired by the optical flow for the motion alignment between adjacent video frames, where the feature maps F^h, F^l of different resolutions are used as the 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 up-sampling layer.

The experimental results show that when tested using the HM-SAR dataset, our proposed model improves mIoU by 2.54% compared to the existing best performing semantic segmentation models. The ablation experiment shows that the proposed MSFM can effectively improve the mIoU value.

Ding J H, Yuan M H. A multi-target semantic segmentation method for millimetre wave SAR images based on a dual-branch multi-scale fusion network[J]. *Opto-Electron Eng*, 2023, **50**(12): 230242; DOI: 10.12086/oee.2023.230242

Foundation item: Project supported by National Natural Science Foundation of China (61601291), and Shanghai Committee of Science and Technology (14dz1206602)

[1]Terahertz Technology Innovation Research Institute, University of Shanghai for Science and Technology, Shanghai 200093, China; [2]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

* E-mail: yuanminghui@usst.edu.cn