基于 BoT-YOLOX 的毫米波图像目标检测

李刚, 叶学义*, 蒋甜甜, 李文杰, 应娜

(杭州电子科技大学通信工程学院　杭州　310018)
(xueyiye@hdu.edu.cn)

摘　要: 主动毫米波(active millimeter wave, AMMW)图像具有噪声多、易含伪影、小目标多等特点, 一直是隐匿目标检测的挑战. 为此, 提出了一种基于 BoT-YOLOX 的毫米波图像目标检测方法. 首先, 在模型主干网络中引入瓶颈型 Transformer(bottleneck Transformer, BoT), 加强模型的特征提取能力; 然后, 调整多尺度目标检测层, 并集成全局注意力机制来提高对小目标的检测能力; 最后, 提出一种多视角加权框融合的后处理方法, 用于集成不同视角检测结果, 以提高模型的鲁棒性. 在自行采集的包括 54 000 幅图像的 AMMW 数据集上, 与基准模型(YOLOX)相比, 该模型达到了 93.22%的检出率和 4.46%的误检率, AP 提升了 6.74 个百分点; 在公开 AMMW 数据集上, 与主流方法相比, mAP 提升了 4.07 个百分点. 实验结果表明, 所提方法对 AMMW 图像场景的目标, 小目标检测准确度更加出色.

关键词: 主动毫米波图像; 瓶颈型 Transformer; 小目标检测; 多视角加权框融合
中图法分类号: TP391.41　　**DOI:** 10.3724/SP.J.1089.2023-00241

Object Detection in Millimeter Wave Images Based on BoT-YOLOX

Li Gang, Ye Xueyi*, Jiang Tiantian, Li Wenjie, and Ying Na

(*School of Communication Engineering, Hangzhou Dianzi University, Hangzhou*　310018)

Abstract: Active millimeter wave (AMMW) images are characterized by high noise, artifacts, and small objects, which has always been challenges for concealed object detection. Therefore, a method is proposed for detecting objects in millimeter-wave images based on BoT-YOLOX. Firstly, Bottleneck Transformer (BoT) is introduced into the model backbone network to enhance feature extraction capability of the model. Then, multi-scale object detection layer are adjusted, and global attention mechanism is integrated to improve detection ability of small objects. Finally, a post-processing method of multi-view weighted boxes fusion is proposed to integrate the detection results of different views to improve the robustness of the model. On the self-collected AMMW dataset, which includes 54 000 images, compared with the baseline model (YOLOX), the model achieves a detection rate of 93.22% and a false detection rate of 4.46%, and AP is increased by 6.74 percentage points. On the public AMMW dataset, compared with mainstream methods, the mAP is increased by 4.07 percentage points. The experimental results show that the proposed method is more accurate in detecting small targets in AMMW image scenes.

Key words: active millimeter wave image; bottleneck Transformer; small objects detection; multi-view weighted boxes fusion

收稿日期: 2023-05-31; 修回日期: 2023-11-06. **基金项目**: 国家自然科学基金(U19B2016, 60802047). **李刚**(1998—), 男, 硕士研究生, 主要研究方向为毫米波图像目标检测; **叶学义**(1973—), 男, 博士, 副教授, 硕士生导师, 论文通信作者, 主要研究方向为模式识别、信息安全; **蒋甜甜**(1999—), 女, 硕士研究生, 主要研究方向为图像处理; **李文杰**(2000—), 男, 硕士研究生, 主要研究方向为图像处理; **应娜**(1978—), 女, 博士, 副教授, 硕士生导师, 主要研究方向为信号处理、语音信号处理、图像处理.

1　相关工作

近年来，随着旅客出行人次的不断增加，机场、火车站和地铁站等公共场所面临着巨大的安检压力. 常规的安检包括人体检查和行李检查，对于行李检查通常采用 X 射线，它穿透性很强但对人体有害；对于人体检查通常采用金属探测仪，但金属探测仪对非金属物体不敏感，也存在侵犯人体隐私的弊端. 因此，目前需要一种新型的安检方式来应对以上的不足. 毫米波是频段在 30 GHz~ 300 GHz 的电磁波，其波长范围在 1 mm~10 mm[1]. 毫米波成像设备是一种利用毫米波进行成像的设备，具有高精度、对人体无危害、非接触式等优点，它可以穿透衣物获取人体的三维图像，从而检测出携带危险物品的人员. 根据成像系统是否主动发射毫米波信号，毫米波成像可以分为 2 类：主动毫米波(active millimeter wave，AMMW)成像[2]和被动毫米波成像[3]. 其中，AMMW 成像受环境影响较小，图像质量较高，因此被广泛应用于安全检查.

早期针对 AMMW 图像目标检测的研究工作[4-6]大多基于统计模型或机器学习方法. 但是鉴于 AMMW 图像自身的特点，传统方法通常缺乏鲁棒性且模型泛化能力较差. 得益于深度学习的发展[7-8]，针对可见光图像，基于深度学习的目标检测网络[9-12]发展迅速. 然而如图 1 所示 AMMW 图像与可见光图像存在很大的差异，其中红色框示意目标. 其有以下特点：(1) 图像的质量较差，目标缺乏纹理信息，噪声形成的伪影与目标非常相似，一些目标与人体部位也会发生混淆；(2) 目标尺寸较小，且这些目标的形状难以被准确描述，模型很难提取到具有鉴别力的特征；(3) AMMW 成像设备每次从特定的角度成像，而目标随机出现在人体不同区域，目标有可能因为低雷达截面而造成不完整和扭曲，从而增加误判的概率.

图 1　AMMW 图像的目标检测

针对 AMMW 图像的 3 个特点，需要专门研究相关检测模型和方法，并注重提高方法的鲁棒性和泛化能力. 一些工作致力于研究利用深度学习模型检测 AMMW 图像，其中，师君等[13]分别使用热图检测和 YOLO 算法进行毫米波图像目标检测，其中 YOLO 算法在检测速度和精度上优势明显. 侯冰基等[14]改进 Faster R-CNN[12]以优化毫米波图像检测，通过反卷积恢复下采样的特征与捷径连接，保留了图像的细粒度特征，同时提高了模型的检测速度. Wang 等[15]使用注意力机制融合不同尺度的特征，并在训练时通过自定步长(self-paced)学习来指导模型学习难以检测的难样本，提高了检出率. Liu 等[16]在 Faster R-CNN 中添加了由不同扩张率的空洞卷积组成的上下文嵌入模块，使检测器可以捕获细节信息和上下文信息，提高了对小目标的检测性能.

然而，上述方法并没有充分地关注 AMMW 图像的特点，AMMW 图像仍然受到低信噪比影响，面对不同尺度、形状扭曲的目标，模型的鲁棒性和检测性能有待进一步提高. 为实现 AMMW 图像高精度检测，本文结合其特点，提出了一种改进的 YOLOX[11]多尺度目标检测方法 BoT-YOLOX. 首先，在原模型的主干网络最后一层的 Bottleneck 层引入了 Transformer[17-18]模块，Transformer 具有良好的全局建模能力；而卷积神经网络能够捕捉特征的局部依赖关系，结合两者可以增强模型的特征学习能力，使模型进一步抑制背景噪声并强调前景目标. 其次，为解决隐匿目标小的问题，以利用高分辨率特征预测小目标和微小目标的检测层替代原模型用于预测大目标的检测层. 必须说明

的是，主干网络底层的高分辨率特征由于缺乏语义信息，会降低检测目标的能力，因此模型在特征融合前引入了全局注意力机制(global attention mechanism, GAM)[19]模块，以增强来自主干网络的高分辨率低级特征. 最后，当模型经过良好的训练并将其应用于单幅图像的检测后，通过集成不同视角图像的检测结果，再采用加权框融合(weighted boxes fusion, WBF)将同一目标的多个检测框合并成一个更准确的框. 通过多视角 WBF，可以有效地避免由矩形框的误差而导致的误检和漏检.

2 YOLOX 模型概述

YOLOX 是一种实时性的一阶段目标检测器，如图 2 所示，其主要由主干网络(Backbone)、颈部(Neck)和检测头(Head)组成. 其中，Backbone 包含聚焦(Focus)模块、跨阶段部分连接(cross stage partial, CSP)模块[20]、空间金字塔池化(spatial pyramid pooling, SPP)[21]和标准卷积层. Focus 模块可以有效地减少下采样带来的信息损失，它通过对输入的图像进行切片操作，将图像划分为 4 个尺度大小相同的特征图，然后沿通道维度拼接(Slice)，使通道数扩充为原来的 4 倍，最后通过一个标准卷积层得到下采样倍数为 2 的特征图. CSP 模块由 3 个卷积层和 1 个 Bottleneck 层组成，该结构可以最大化梯度联合的差异，利用梯度流截断避免不同的层学习到重复的梯度信息，使网络更加高效. SPP 模块由不同窗口大小的最大池化(Maxpool)层组成，它通过融合不同尺度的特征得到语义信息更强的特征. 标准卷积层包括卷积(convolution, Conv)、批归一化(batch normalization, BN)和激活(activation, Act). Neck 由特征金字塔网络(feature pyramid network, FPNet)[22]和路径聚合网络(path aggregation network, PANet)[23]组成. 其中，FPNet 通过自顶向下地上采样，将顶层的特征和底层的特征进行融合；而 PANet 通过自底向上地下采样操作进行特征传递，通过这种路径连接，可以形成一个在所有级别上语义都很强的多尺度特征表示. Head 部分包含 3 个检测层，与以往的 YOLO[9-10]架构不同，YOLOX 采用解耦检测头，即检测框的类别分数、置信度分数和坐标位置不再由同一条路径输出，而是分别通过不同的路径输出，通过这种方式可以有效地提升模型的收敛速度.

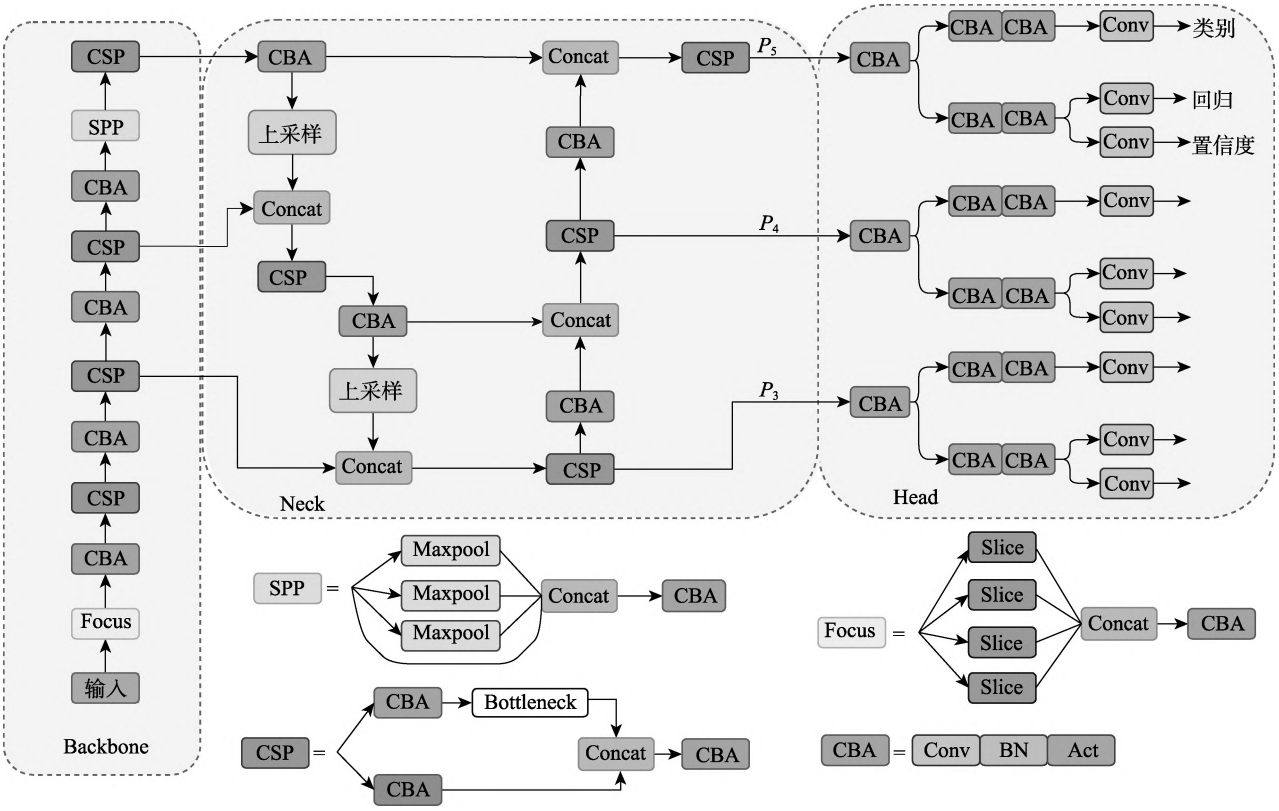


图 2 YOLOX 网络结构

3　BoT-YOLOX

YOLOX 在可见光图像上表现出了不错的检测性能，但却不适用于 AMMW 图像. 为进一步优化整个架构，本文提出了 BoT-YOLOX，其结构如图 3 所示. 遵循原始架构，BoT-YOLOX 由 Backbone，Neck 和 Head 组成. 此外，它在原模型的基础上，加入了 Transformer 模块；调整了检测头输出，并引入了 GAM；在模型后处理阶段，采用多视角 WBF. BoT-YOLOX 比 YOLOX 在 AMMW 图像上表现得更好.

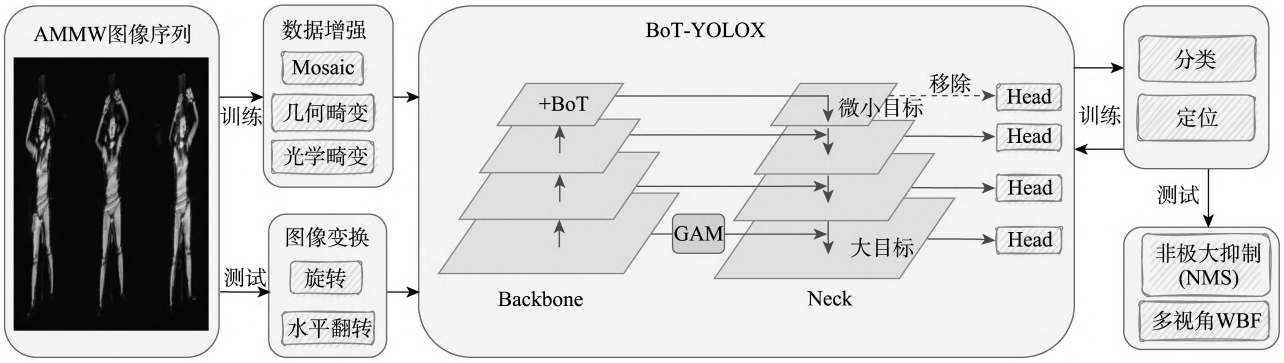


图 3　BoT-YOLOX 模型结构

3.1　BoT

AMMW 图像受低信噪比影响，且目标通常没有足够的纹理信息，导致很难与背景区分开来. 为加强模型的特征提取能力，使网络学习到更独特的特征表示，本文将原始 YOLOX 主干网络中的 CSP 瓶颈块替换为 Transformer 模块.

此外，考虑昂贵的计算和内存成本，Transformer 模块只应用在主干网络的最后一个 CSP 模块中. 如图 4 所示，遵循原始的残差结构，Bottleneck 层中的 3×3 卷积被替换成多头自注意力机制(multi-head self-attention, MHSA)模块. MHSA 模块通过在自注意力机制中引入多个独立的注意力头，使模型可以同时学习到多个不同的注意力表示，捕捉输入特征中的不同关系和重要性，从而提升了模型的表征能力. 其计算公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}(\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{K}+\boldsymbol{P})\boldsymbol{V}\ .$$

其中，Softmax 为激活函数；$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$，$\boldsymbol{P}$ 分别表示查询向量、关键向量、值向量和相对位置编码，它们的计算公式分别为

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{W}_{\boldsymbol{Q}}(\boldsymbol{X})\\ \boldsymbol{K}=\boldsymbol{W}_{\boldsymbol{K}}(\boldsymbol{X})\\ \boldsymbol{V}=\boldsymbol{W}_{\boldsymbol{V}}(\boldsymbol{X})\\ \boldsymbol{P}=(\boldsymbol{R}_{\mathrm{h}}+\boldsymbol{R}_{\mathrm{w}})^{\mathrm{T}}\boldsymbol{Q}\end{cases}.$$

$\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$ 表示输入特征，C，H 和 W 分别表示特征的通道、宽度和高度；$\boldsymbol{W}(\cdot)$ 表示线性变换函数，具体为 1×1 逐点卷积；$\boldsymbol{Q}\in\mathbb{R}^{C\times HW}$，$\boldsymbol{K}\in\mathbb{R}^{C\times HW}$，$\boldsymbol{V}\in\mathbb{R}^{C\times HW}$，$\boldsymbol{P}\in\mathbb{R}^{HW\times HW}$，$\boldsymbol{R}_{\mathrm{h}}\in\mathbb{R}^{C\times H\times 1}$ 和 $\boldsymbol{R}_{\mathrm{w}}\in\mathbb{R}^{C\times 1\times W}$ 表示 2 个可学习参数向量，分别用于高度和宽度的相对位置编码，它们和查询向量 $\boldsymbol{Q}$ 共同形成最终的相对位置编码 $\boldsymbol{P}$. 通过这种编码方式，可以帮助模型更好地理解不同位置特征之间的依赖关系，从而提高模型性能.

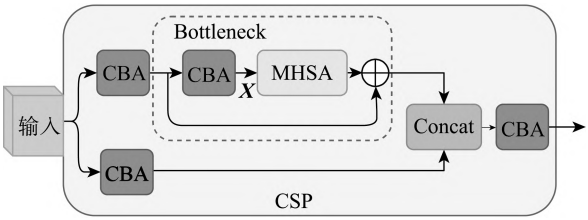


图 4　融合 Transformer 的 CSP 模块

3.2　调整多尺度目标检测层

为了解目标的大小变化和数量分布，在自建 AMMW 图像数据集上对超过 120 000 个目标进行了统计分析，发现它包含许多非常小的目标，如图 5 所示. 大部分目标的面积小于 48^2 像素，面积小于 32^2 像素的目标数量约占总体目标数量的 40%，甚至有目标面积小于 16^2 像素. 即像素面积越小的目标，在特征提取的过程中经过池化或跨步卷积操作更易造成特征信息丢失. 如在 YOLOX 中，最顶层特征的分辨率为原始图像的 1/32，此时面积小于 32^2 像素的目标在顶层的低分辨率特征上小于 1 个特征点.

为解决上述问题，BoT-YOLOX 采用原始 YOLOX 中的 PANet 作为网络的颈部，通过自顶向下和自底向上的路径连接融合语义较强的低分辨率特征和语义较弱的高分辨率特征，形成一个在所有级别上语义都很强的多尺度特征表示，以进

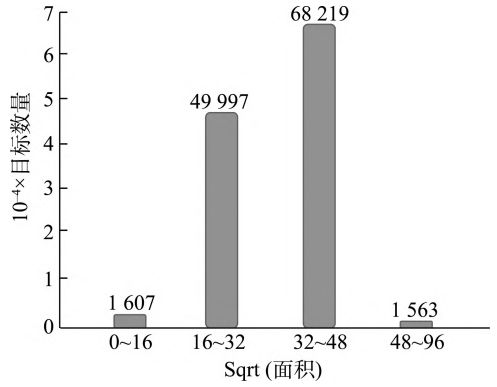


图 5 目标尺寸大小与数量分布

行多尺度预测. 在原始的 YOLOX 中包含了 3 个检测头, 即利用 3 种不同尺度特征进行预测小、中、大 3 种目标, 这 3 种特征的分辨率分别为原始图像的 1/8, 1/16, 1/32. 由于 AMMW 图像中的目标尺寸集中在中目标、小目标和微小目标, 因此, 采用图 6 所示改进后的网络结构, BoT-YOLOX 去掉了原先用于预测大目标的 P_5, 增加了一路由底层下采样倍数为 4 的高分辨率特征生成的检测层 P_2. 由于高分辨率特征含有丰富的空间信息, 因此它对小目标和微小目标更加敏感. 调整后的多尺度目标检测层对 AMMW 图像中的小目标检测能获得更好的效果.

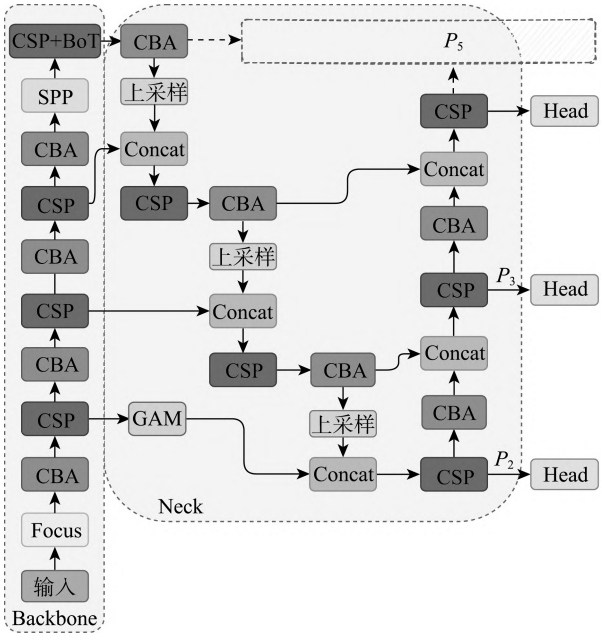


图 6 BoT-YOLOX 网络结构

3.3 GAM

融合空间信息丰富的高分辨率特征有利于小目标的检测, 但考虑底层高分辨率特征只经过浅层的卷积运算, 且 AMMW 图像具有低信噪比的特点, 底层高分辨率特征缺乏语义信息且易受到噪声干扰, 从而会降低目标识别的能力. 因此, 直接融合高分辨率特征和来自顶层的高级特征是不合适的. 为解决这个问题, 在特征融合前引入了 GAM 模块, 通过对上一层的特征进行学习并得到通道和空间上的权重, 对重要信息分配高权重, 对背景噪声等冗余信息分配低权重, 并且不会降低空间信息. 通过这种方式来增强融合前的高分辨率特征, 可以使最终的融合特征在空间信息和语义信息上都得到有效的增强.

GAM 包含了通道注意力模块和空间注意力模块 2 部分, 其结构如图 7 所示. 对于通道注意力模块, 其先对输入特征 $\boldsymbol{F}\in\mathbb{R}^{C\times H\times W}$ 进行维度变换得到 $\boldsymbol{F}'\in\mathbb{R}^{HW\times C}$, C, H 和 W 分别表示特征的通道、宽度和高度; 然后使用多层感知器获取特征通道间的相互关系, 以生成通道注意力特征图 $\boldsymbol{M}_{\mathrm{c}}\in\mathbb{R}^{C\times H\times W}$. 其过程可以表述为

$$\boldsymbol{M}_{\mathrm{c}}(\boldsymbol{F}')=\sigma\left(\boldsymbol{W}_1\left(\boldsymbol{W}_0(\boldsymbol{F}')\right)\right).$$

其中, σ 表示 Sigmoid 函数; $\boldsymbol{W}_0\in\mathbb{R}^{C\times C/r}$ 和 $\boldsymbol{W}_1\in\mathbb{R}^{C/r\times C}$ 表示 2 个全连接层, r 为缩减率. 随后将生成的通道注意力特征图与输入特征 $\boldsymbol{F}$ 相乘, 得到新的特征 $\boldsymbol{F}_{\mathrm{c}}\in\mathbb{R}^{C\times H\times W}$, 即

$$\boldsymbol{F}_{\mathrm{c}}=\boldsymbol{M}_{\mathrm{c}}(\boldsymbol{F}')\otimes\boldsymbol{F}\ .$$

然后将其送入空间注意力模块.

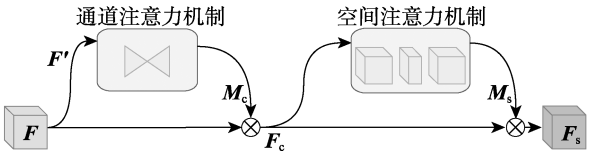


图 7 GAM 模块

区别于一般的空间注意力模块[24], GAM 删除了池化操作, 直接使用 2 个大小为 7×7 的卷积层进行空间信息融合, 以进一步保留特征信息. 但是, 这也会显著增加模型的计算量和参数量. 为了防止增加模型的复杂度, 本文选择采用组卷积来减少复杂度. 其过程可以表示为

$$\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F}_{\mathrm{c}})=\sigma\left(\mathrm{Conv}_2\left(\mathrm{Conv}_1(\boldsymbol{F}_{\mathrm{c}})\right)\right).$$

其中, $\boldsymbol{M}_{\mathrm{s}}\in\mathbb{R}^{C\times H\times W}$ 表示生成的空间注意力特征图; Conv_1和Conv_2表示 2 个大小为 7×7 的组卷积. 最后将空间注意力特征图 $\boldsymbol{M}_{\mathrm{s}}$ 与特征图 $\boldsymbol{F}_{\mathrm{c}}$ 相乘得到

$$\boldsymbol{F}_{\mathrm{s}}=\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F}_{\mathrm{c}})\otimes\boldsymbol{F}_{\mathrm{c}}\ .$$

3.4　多视角 WBF

由于成像角度的原因，目标会发生不完整和扭曲，造成检测困难．因此为提高模型的鲁棒性，使模型对不同尺度和不同角度的目标具有更好的检测效果，在模型的后处理阶段，引入了多视角 WBF．如图 8 所示，对于从人体多个角度进行成像的 AMMW 图像先进行单幅图像检测，然后将其不同视角的检测结果融合到 2 幅图像上(正视图和后视图)，最后采用 WBF 得到最终结果．其中，正确检测被标记为红色框，漏检被标记为绿色框，误检被标记为蓝色框．此外，一些 AMMW 成像设备只在人体正面和背面成像，对其可以采用手动图像变换，即对图像进行水平翻转和三维旋转，以模拟多视角情况；并对变换后的多幅图像进行检测；最后，集成检测结果进行 WBF．

多视角 WBF 具体算法步骤如下．

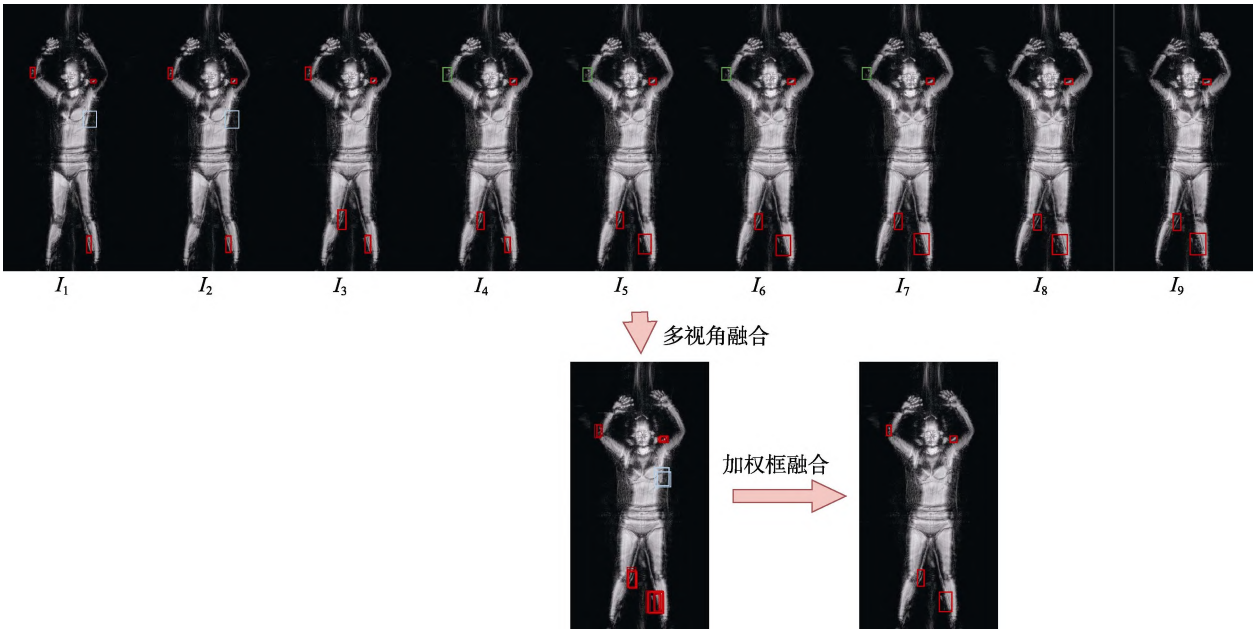


图 8　多视角 WBF 处理过程

输入．一组 AMMW 图像 I_i，$i\in[1,2,\cdots,N]$．

输出．经过多视角加权框融合后的检测框．

Step1．将一组图像送入训练好的模型进行检测，利用非极大抑制去除重复框，将所有的检测框添加到列表 B 中，并按检测框的置信度分数 C 降序排序．

Step2．声明框簇列表 L 和融合框列表 F，L 每个位置包含同一目标的多个检测框，F 每个位置只包含一个框，它是 L 中对应框簇的融合框．

Step3．遍历 B 中的检测框，计算检测框和列表 F 中的融合框的最大交并比(intersection over union, IoU)．

Step4．如果 IoU 大于阈值 t_1，则将该检测框添加到 L 中，添加的位置为 F 中匹配上的融合框的位置 p．

Step5．如果 IoU 小于阈值 t_1，则将该检测框作为新的框簇添加到列表 F 和列表 L 的末尾(在本文实验中，t_1=0.3 接近最佳阈值)．

Step6．当有新的检测框添加到框簇后，需要使用 $L[p]$中累积的 M 个检测框重新计算融合框的坐标位置和置信度分数，计算公式分别为

$$X_{1,2}=\sum_{i=1}^{M}C_i\times X_{1,2}^{i}\Big/\sum_{i=1}^{M}C_i;$$

$$Y_{1,2}=\sum_{i=1}^{M}C_i\times Y_{1,2}^{i}\Big/\sum_{i=1}^{M}C_i;$$

$$C=\sum_{i=1}^{M}C_i\Big/M.$$

其中，$X_{1,2}$ 和 $Y_{1,2}$ 分别为检测框顶点横坐标和纵坐标．

Step7．处理完 B 中所有的检测框后，重新缩放 F 中融合框的置信度分数 C，将其乘以框簇中的检测框数量，再除以多个角度 N；如果框簇中的检测框数量较少，则意味着只有少数视角预测它，因此，需要减少这类情况的置信度分数，处理过程为

$$C=C\times\frac{\min(M,N)}{N}.$$

多视角 WBF 合并所有框，以形成最终结果，其主要思想是利用不同视角同一目标的多个检测框的坐标信息和置信度分数进行加权平均．从图 8 中可以看到，最终的融合框是由多个检测框通过置信度分数构建的平均结果，所以在准确度上，融合框表现更稳定．此外，根据目标是否在多个视角上被检测到，融合框的置信度分数被重新放缩，这样可以有效地改善误漏检问题．

4 实验与分析

4.1 数据集

由于安全隐私以及版权等问题，公开的 AMMW 数据集很少. 因此，分别针对非公开数据集 AMMW1 和公开的数据集 AMMW2 进行实验.

对于 AMMW1，每个人包含 18 幅不同视角的全身图像，正面视角和背面视角分别为 9 幅，其中每幅图像相邻图像之间的角度为 10°，每幅图像的大小为 768×400 像素. 本次实验共收集了 3 000 组(54 000 幅图像)，为了接近真实场景，该数据集中被检人员的性别比例均衡，数据集上有 20 多种不同材料和尺寸的物体，包括如手枪、打火机、匕首、毒品包等，且这些目标随机分布在人体的各个部位. 虽然数据集包含很多的目标种类，但是实验的主要目的是定位人体中的目标，而不是对其进行分类. 因此，所有待检测目标都被标记为同一类别对象. 在实验中，整个数据集被随机划分成训练集和测试集，比例为 8∶2.

AMMW2[25]是一个公开的数据集，其共有 3 157 幅图像，其中一个人只包含正反面的全身图像，待检目标被标记为 11 个类别. 按照原文献的划分方式，其中 2 555 幅图像用于训练，602 幅图像用于测试.

4.2 评估指标

实验主要通过单类别平均精度(average precision, AP)，多类别平均精度(mean average precision, mAP)来评估方法的有效性. 此外，针对实时检测场景，给出单幅图像的平均检测时间，用来衡量模型的检测速度. 其中，AP 综合考量召回率(recall, R)和精准率(precision, P)，衡量模型的在该类别上的好坏. P 表示检测到的目标中真正为目标的比例，R 表示所有真实目标中被正确检测到的比例，它们的计算公式分别为

$$\begin{cases} P=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \\ R=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \end{cases}.$$

其中，T_{P} 为真正例，F_{P} 为假正例，F_{N} 为假反例. P 和 R 是相互制约的关系，不同的置信度分数阈值有不同的结果，这些潜在的截止点可以绘制成一条 P-R 曲线，曲线覆盖的面积则为 AP，面积越大，性能越好. 通常，在不同的 IoU 阈值下计算 AP 值会有很大的差异，IoU 阈值越高，则要求预测框的定位越准确，AP 值就会越低. 因此，为全面评估方法的有效性，在不同的 IoU 下计算 AP，如在 IoU=0.50 时记为 AP@0.5，在 IoU∈[0.1:0.1:0.5]时记为 AP@[0.1, 0.5]，此外，在 IoU∈[0.5:0.05:0.95]时 AP@[0.5, 0.95]简单记为 AP. 最后，实验还给出了各阶段的 F_1 分数、P 和 R. F_1 分数的计算公式为

$$F_1=2\times\frac{P\times R}{P+R}.$$

4.3 实验细节

训练环境为 Ubuntu 22.04 操作系统，CPU 为 AMD Ryzen 93 900X，GPU 为 NVIDIA RTX 3090，并采用开源深度学习框架 Pytorch 完成实验.

在训练和测试时图像大小被调整为 768×416 像素作为网络的输入. 为保证实验的公平性，在训练阶段，不使用预训练权重，一共训练 105 个 epoch；使用随机梯度下降法训练，并使用余弦退火学习率策略[10]，第 1 轮用于预热，初始学习率为 5×10^{-3}，最后一轮衰减到 2.5×10^{-4}. 数据增强包括 Mosaic[10]数据增强、几何失真、光度失真和水平翻转. 此外，综合考虑计算成本和精度要求，实验选择 YOLOX 的 small 版本作为基准网络.

4.4 实验设计与结果分析

在本节中，首先在 AMMW1 数据集上与几种主流目标检测器进行了比较，然后从定性角度与多种 AMMW 图像目标检测方法相比较，最后进行消融实验. 另外，在公开数据集 AMMW2 上进行了额外的实验，以进一步验证本文方法的有效性.

4.4.1 改进方法与其他方法对比

为了验证改进的 BoT-YOLOX 的检测性能，将 YOLOv3[9]，YOLOv5[26]，Faster R-CNN[12]和原始 YOLOX 等主流的目标检测器与 BoT-YOLOX 进行对比，并采用 AP 和 F_1 分数对每个方法进行评估. 结果如表 1 所示，BoT-YOLOX 在 AP@0.5，AP 和 F_1 分数上分别达到了 96.50%，69.35%和 94.37%，其性能优于对比各类主流检测器. 另外，在不采用多视角 WBF 的情况下，BoT-YOLOX 在 AP 上仍然要比其他方法至少高出 2.68 个百分点，这也说明了 AMMW

表 1 本文与其他方法的性能对比 %

方法	AP@0.5	AP	F_1
Faster R-CNN[12]	92.20	55.52	91.56
YOLOv3[9]	94.40	56.90	91.11
YOLOv5[26]	96.10	62.54	92.64
YOLOX[11]	95.07	62.61	92.78
本文	**95.77**|**96.50**	**65.29**|**69.35**	**93.56**|**94.37**

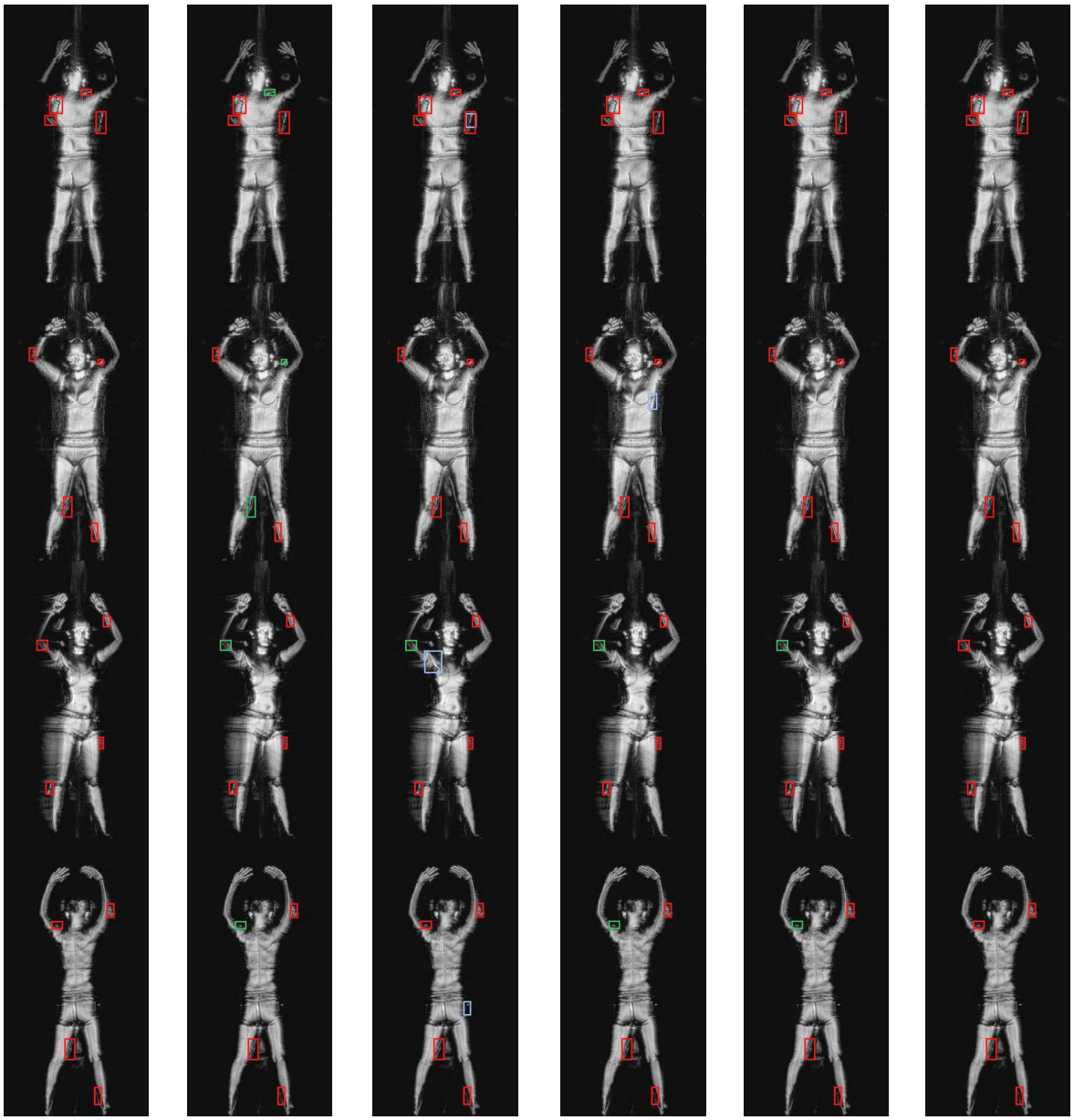
注：粗体表示最优值；“|”左边为单视角图像的检测结果.

图像和可见光图像存在差异，而 BoT-YOLOX 更适用于 AMMW 图像.

图 9 所示为一些具有代表性的可视化结果，其中，正确检测被标记为红色框，漏检被标记为绿色框，误检被标记为蓝色框. 从图 9 中可以看出，未经过改进的检测器无法适用于 AMMW 图像. 其中，一阶段检测器更容易出现漏检，如 YOLO 系列的检测器；二阶段检测器更容易出现误检，如 Faster R-CNN. 相比之下，本文方法缓解了以上的不足，如图 9f 所示.

4.4.2　与其他典型 AMMW 图像检测方法的比较

为了验证方法的优越性，将本文方法与现有的 AMMW 图像检测方法进行比较. 此外，为了比较的客观性，实验直接引用原论文中的结果，并且在检测后处理时不采用多视角 WBF. 如表 2 所示，从定性比较中，本文方法具有显著的改进. 与之前的检测方法相比，本文方法在 AP@[0.1，0.5]上展现了 6.76 个百分点的绝对提高. 结果有力地证明了所提出的方法的综合效果，本文方法优于对比的 AMMW 图像的检测方法.



a. 真实标记　b. YOLOv3[9]　c. Faster R-CNN[12]　d. YOLOv5[26]　e. YOLOX[11]　f. 本文

图 9　4 种主流检测器和本文方法的可视化结果

表 2 不同 AMMW 图像目标检测方法性能对比 %

方法	AP@0.5	AP@[0.1, 0.5]
Liu 等[16]	69.10	83.00
Zhang 等[27]	74.67	80.81
Wang 等[15]	78.55	87.63
Sun 等[28]	82.16	90.61
本文	**95.77**	**97.37**

注. 粗体表示最优值.

4.4.3 消融实验

为了分析所提出的改进对模型性能的影响, 共设计了 5 组实验进行比较, 每组实验使用相同的训练参数, 实验的结果如表 3 所示, 所有的模型都在 NVIDIA RTX 3090 上使用 FP16 和 batch=1 进行测量.

表 3 AMMW1 数据集上的消融实验结果

方法	AP/%	参数量	FLOPs	延迟/ms	FPS
YOLOX[11]	62.61	8.94M	26.64G	8.4	119.0
$-P_5$	62.61	**6.51M**	**24.80G**	**7.4**	**135.1**
$+P_2$	64.01(1.40↑)	7.25M	44.51G	9.9	101.0
+GAM	64.31(0.30↑)	7.28M	45.60G	10.2	98.0
+BoT	65.29(0.98↑)	6.89M	46.36G	10.2	98.0
+WBF	**69.35(4.06↑)**	6.89M	46.36G	12.1	82.6

注. −表示未使用的策略, +表示新增加的改进策略, ↑表示增加, 延迟包括模型前向推理和后处理 2 部分; 粗体为最优值.

从表 3 中实验结果可以看出, 在去掉大目标检测层 P_5 后, AP 没有损失, 说明原模型的 P_5 检测层是冗余的, 移除 P_5 检测层可以有效地降低模型复杂度, 提高对小目标的检测效率; 当在网络中增加用于小目标和微小目标的检测层 P_2 后, 模型的计算量增加, 导致检测速率有所下降, 但 AP 的提升也很高. 从图 9f 可以看出, BoT-YOLOX 在检测小目标时表现良好, 因此, 计算量增加是值得的. 在此基础上, 在融合底层高分辨率特征前引入 GAM 模块, 可以看到 GAM 模块以微小的成本在 AP 上增加了 0.30 个百分点, 表明它可以有效地增强底层高分辨率特征, 从而与顶层高级特征更好地融合. 当主干网络融合 Transformer 模块后, 模型参数量降低, 由于只在低分辨率特征上使用, 计算量略微增加, 但是并不影响模型的检测速率, 且 AP 提高了 0.98 个百分点. 这表明融合 Transformer 模块后模型可以有效地捕获到更具判别能力的特征表示, 以进行更高效的隐匿目标检测.

最后, 模型采用多视角 WBF 对检测结果进行后处理, AP 增加了 4.06 个百分点. AP 的显著提高表明, 多视角 WBF 可以有效地改善检测框精度, 减少模型的不准确预测情况. 不过由于模型需要将不同视角图像的检测结果进行整合, 以得到最终结果, 因此通常也会额外增加 2 ms 左右的后处理时间. 当模型使用全部改进策略后, AP 增加了 6.74 个百分点, 此时模型的检测精度达到最高, 虽然其计算量增加且检测速率有一定下降, 但仍可满足实时检测的要求, 该模型在需要高精度检测任务时具有很大的优势.

对于毫米波图像目标检测而言, 检出率和误检率也是需要关心的重要指标. 如表 4 所示, IoU=0.50 时, 原始 YOLOX 的 F_1 分数为 92.78%, 经过改进后 F_1 分数可以达到 94.37%, R 和 P 分别提高了 0.88 个百分点和 2.38 个百分点; 在不使用多视角 WBF 的情况下, F_1 分数仍能提高 0.78 个百分点, 其中 R 和 P 分别提高了 0.24 个百分点和 1.34 个百分点, 大大地降低了误检率, 这对毫米波图像检测任务来说是意义重大的.

表 4 方法改进前后的 F_1 分数, R 和 P 对比结果 %

方法	F_1	R	P
YOLOX[11]	92.78	92.40	93.16
本文	**93.56|94.37**	**92.64|93.22**	**94.50|95.54**

注. 粗体表示最优值; "|"左边为单视角图像的检测结果.

4.4.4 AMMW2 数据集的实验

为了进一步验证方法的有效性, 将实验扩展到文献[25]使用的公开数据集. 从比较的客观性考虑, 采取与该文献相同的数据集划分比例和相同的评估指标 mAP@0.5, 实验数据直接引用原文献. 此外, 由于 AMMW2 数据集仅包含人体周围的前视图和后视图, 系统必须通过复制原始图像来模拟多视角场景, 但这会大大增加内存开销和降低模型推理速度. 因此, 为了检测速度和精度的平衡, 实验直接采用单视角图像检测.

如表 5 所示, 本文方法比 RetinaNet+P_2[25]中的最佳结果提高了约 10 个百分点, 比基于多源融合 Transformer 的二阶段目标检测器[28]提升了 4.07 个百分点. 改进的结果表明, 所提出的方法具有鲁棒性, 它更适用于 AMMW 图像的目标检测, 并且可以在不同场景中获得较好的检测效果.

表 5　与其他方法在 AMMW2 数据集上的实验结果

方法	Backbone	mAP@0.5/%
YOLOv3[9]	Darknet-53	39.29
YOLOv4[10]	CSPDarknet53	41.39
FRCN-OHEM[29]	VGG16	42.32
RetinaNet[30]	ResNet-50	54.58
RetinaNet+P_2[25]	ResNet-50	60.32
Sun 等[28]	ResNet-50	66.30
本文	CSPDarknet53	**70.37**

注：粗体表示最优值.

5　结　语

本文针对 AMMW 图像与可见光图像的差异及其特点，提出了一种基于 BoT-YOLOX 网络模型的 AMMW 图像目标检测方法. 该方法通过在主干网络添加 Transformer 模块加强网络的特征提取能力；通过调整多尺度目标检测层和引入 GAM 模块来增强对小目标和微小目标的检测；最后，引入了多视角 WBF，使模型对不同尺度、不同角度的目标具有更好的检测效果. 在 2 个 AMMW 数据集上的实验结果表明，BoT-YOLOX 检测性能非常好，且明显优于对比方法，然而在检测速度方面有待提升. 在后续研究中，考虑如何利用轻量化网络技术来控制模型的体积，同时保证检测精度，以使得毫米波安检系统的小型化、高效化.

参考文献(References):

[1] Du Kun, Wang Wei, Nian Feng, *et al*. Concealed objects detection in active millimeter-wave images[J]. Systems Engineering and Electronics, 2016, 38(6): 1462-1469(in Chinese)
(杜琨，王威，年丰，等. 主动毫米波图像的人体携带危险物检测研究[J]. 系统工程与电子技术, 2016, 38(6): 1462-1469)

[2] Grossman E N, Miller A J. Active millimeter-wave imaging for concealed weapons detection[C] //Proceedings of the Passive Millimeter-Wave Imaging Technology VI and Radar Sensor Technology VII. Bellingham: Society of Photo-Optical Instrumentation Engineers, 2003, 5077: 62-70

[3] Clark S E, Lovberg J A, Martin C A, *et al*. Passive millimeter-wave imaging for airborne and security applications[C] //Proceedings of the Passive Millimeter-Wave Imaging Technology VI and Radar Sensor Technology VII. Bellingham: Society of Photo-Optical Instrumentation Engineers, 2003, 5077: 16-21

[4] Ye Jinjing, Zhou Jian, Sun Qianchen, *et al*. A privacy protection algorithm for active millimeter-wave imaging[J]. Journal of Infrared and Millimeter Waves, 2017, 36(4): 505-512(in Chinese)
(叶金晶，周健，孙谦晨，等. 主动毫米波成像隐私保护算法[J]. 红外与毫米波学报, 2017, 36(4): 505-512)

[5] Du Kun, Zhang Lu, Wang Kairang, *et al*. Concealed objects detection on human based on statistical model[J]. Computer Engineering and Design, 2017, 38(10): 2864-2868+2878(in Chinese)
(杜琨，张璐，王凯让，等. 基于统计模型的人体携带危险物检测[J]. 计算机工程与设计, 2017, 38(10): 2864-2868+2878)

[6] Li Z, Jin Y K, Shen Z J, *et al*. A synthetic targets detection method for human millimeter-wave holographic imaging system[C] //Proceedings of the 7th International Conference on Cloud Computing and Big Data. Los Alamitos: IEEE Computer Society Press, 2016: 284-288

[7] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90

[8] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[9] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2023-05-31]. https://arxiv.org/abs/1804.02767

[10] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[OL]. [2023-05-31]. https://arxiv.org/abs/2004.10934

[11] Ge Z, Liu S T, Wang F, *et al*. YOLOX: exceeding YOLO series in 2021[OL]. [2023-05-31]. https://arxiv.org/abs/2107.08430

[12] Ren S Q, He K M, Girshick R, *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 System. Cambridge: MIT Press, 2015: 91-99

[13] Shi Jun, Que Yujia, Zhou Zenan, *et al*. Near-field millimeter wave 3D imaging and object detection method[J]. Journal of Radars, 2019, 8(5): 578-588(in Chinese)
(师君，阙钰佳，周泽南，等. 近场毫米波三维成像与异物检测方法[J]. 雷达学报, 2019, 8(5): 578-588)

[14] Hou Bingji, Yang Minghui, Sun Xiaowei, *et al*. Real-time object detection for millimeter-wave images based on improved faster regions with convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(13): Article No.131009(in Chinese)
(侯冰基，杨明辉，孙晓玮. 基于改进 Faster RCNN 的毫米波图像实时目标检测[J]. 激光与光电子学进展, 2019, 56(13): Article No.131009)

[15] Wang X L, Gou S P, Li J C, *et al*. Self-paced feature attention fusion network for concealed object detection in millimeter-wave image[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(1): 224-239

[16] Liu T, Zhao Y, Wei Y C, *et al*. Concealed object detection for activate millimeter wave image[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9909-9917

[17] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16×16 words: transformers for image recognition at scale[OL]. [2023-05-31]. https://arxiv.org/abs/2010.11929

[18] Srinivas A, Lin T Y, Parmar N, *et al*. Bottleneck Transformers for visual recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los

Alamitos: IEEE Computer Society Press, 2021: 16514-16524

[19] Liu Y C, Shao Z R, Hoffmann N. Global attention mechanism: retain information to enhance channel-spatial interactions[OL]. [2023-05-31]. https://arxiv.org/abs/2112.05561

[20] Wang C Y, Liao H Y M, Wu Y H, *et al*. CSPNet: a new backbone that can enhance learning capability of CNN[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 1571-1580

[21] He K M, Zhang X Y, Ren S Q, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916

[22] Lin T Y, Dollar P, Girshick R, *et al*. Feature pyramid networks for object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 936-944

[23] Liu S, Qi L, Qin H F, *et al*. Path aggregation network for instance segmentation[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 8759-8768

[24] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[25] Liang D, Xue F, Li L. Active terahertz imaging dataset for concealed object detection[OL]. [2023-05-31]. https://arxiv.org/pdf/2105.03677.pdf

[26] Glenn J. YOLOv5[OL]. [2023-05-31]. https://github.com/ultralytics/yolov5

[27] Zhang B, Wang B, Wu X F, *et al*. Domain adaptive detection system for concealed objects using millimeter wave images[J]. Neural Computing and Applications, 2021, 33(18): 11573-11588

[28] Sun P, Liu T, Chen X T, *et al*. Multi-source aggregation transformer for concealed object detection in millimeter-wave images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(9): 6148-6159

[29] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 761-769

[30] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2999-3007