自适应融合多尺度特征的无锚框遥感图像目标检测算法

王坤[1)], 文武[1)], 铁菊红[1,2)], 吴锡[1)], 符颖[1)*]

[1)] (成都信息工程大学计算机学院 成都 610225)
[2)] (四川大学计算机学院 成都 610065)
(fuying@cuit.edu.cn)

摘 要: 针对遥感图像内容丰富且复杂, 具有目标种类多、密集分布和尺寸变化剧烈等特点, 导致遥感图像中目标多尺度尤其是小目标难以检测的问题, 提出一种基于自适应多尺度特征融合(AMFF)和注意力特征增强(AFE)的无锚框遥感图像目标检测算法. 首先将主干网络提取的图像特征输入 AMFF, 自适应地融合多个尺度的特征, 增加特征复用, 提升网络的多尺度特征表达能力; 然后将 AMFF 输出的特征输入到加入了 AFE 模块的检测头中, AFE 通过结合多分支空洞卷积与注意力机制, 在提高网络对目标尺度的泛化能力的同时增强有效特征信息; 最后进行分类和回归, 得到检测结果. 在 DIOR 和 NWPU VHR-10 公开数据集上, 与多种主流目标检测算法的实验结果表明, 所提算法在 2 个数据集上的平均检测精度分别为 72.4%和 87.4%, 较基线网络分别提升 9.4 和 13.5 个百分点, 比次优结果分别提升 6.3 和 1.7 个百分点; 平均检测精度高于主流目标检测算法, 较基线网络的平均检测精度显著提高, 能够更加准确地检测小尺度目标, 同时有效地提升多尺度目标的检测精度.

关键词: 遥感图像; 目标检测; 无锚框; 特征融合; 注意力机制; 空洞卷积
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023.19673

Anchor-Free Object Detection Method in Remote Sensing Image via Adaptive Multi-Scale Feature Fusion

Wang Kun[1)], Wen Wu[1)], Tie Juhong[1,2)], Wu Xi[1)], and Fu Ying[1)*]

[1)] (*School of Computer Science, Chengdu University of Information Technology, Chengdu* 610225)
[2)] (*School of Computer Science, Sichuan University, Chengdu* 610065)

Abstract: The characteristics of many types, dense distribution and the difference in scales of objects in remote sensing images will result in small objects difficult to be detected. Therefore, a remote sensing image anchor-free object detection method based on an adaptive multi-scale feature fusion (AMFF) and attention feature enhancement (AFE) mechanism is proposed in this paper. Firstly, the image features extracted by the backbone network are input into AMFF, which adopts an adaptive multi-scale feature fusion module to enhance feature reuse, so it can enrich feature information and enhance the multi-scale feature expression ability of the network. Secondly, the output of the features from AMFF is input into the detection head with AFE. AFE combines multi-branch dilated convolution and attention mechanism to enhance both the network's multi-scale generalization ability of the object and the effective feature information. Finally, the detection results are obtained by classification and regression. Experiments with a variety of mainstream object

收稿日期: 2022-04-14; 修回日期: 2022-09-23. **基金项目**: 四川省自然科学基金(2022NSFSC0907); 四川省重点研发项目(2022YFG0042, 2023YFG0025); 四川省科技厅"新一代人工智能平台"重大专项(2019DZX0005). **王坤**(1996—), 男, 硕士研究生, 主要研究方向为计算机视觉、图像处理; **文武**(1979—), 男, 博士, 讲师, 主要研究方向为人工智能在地球科学领域的应用、高性能计算; **铁菊红**(1981—), 女, 博士研究生, 讲师, 主要研究方向为遥感图像分割与融合; **吴锡**(1980—), 男, 博士, 教授, 硕士生导师, 主要研究方向为图像分析与计算成像、高性能与并行分布式计算、智慧气象与数值天气计算; **符颖**(1988—), 女, 博士, 副教授, 硕士生导师, 论文通信作者, 主要研究方向为信号与图像处理、非参贝叶斯方法在图像复原中的应用.

detection algorithms on DIOR and NWPU VHR-10 public datasets show that the average detection accuracy of the proposed algorithm is 72.4% and 87.4%, which is 9.4 percentage points and 13.5 percentage points higher than that of the baseline network and 6.3 percentage points and 1.7 percentage points higher than that of the suboptimal results. The results demonstrate that the average detection accuracy is higher than that of the mainstream object detection algorithms. Meanwhile, the average detection accuracy of the baseline network is significantly improved, which can detect small-scale objects more accurately and effectively improve the detection accuracy of multi-scale objects.

Key words: remote sensing image; object detection; anchor-free; feature fusion; attention mechanism; dilated convolution

随着遥感技术日益成熟, 传感器获取的超高分辨率遥感图像可以提供直观、清晰的地球表面信息, 对地表观测具有重大意义. 遥感图像目标检测作为遥感图像解译分析领域中的重要任务, 可以从广阔的视野范围中定位并标记出感兴趣目标的位置和类别, 是城市规划、土地利用、交通疏导和军事监测等领域的重要基础[1]. 近年来, 深度学习发展迅猛, 图像经过神经网络进行多层卷积操作而获取具有更强语义表达能力的特征, 使得目标检测的性能得到显著提高.

目前, 根据不同的划分标准, 基于深度学习的目标检测算法分为 2 类: (1) 根据是否采用生成区域建议方法, 划分为双阶段检测算法和单阶段检测算法; (2) 根据是否采用预设锚点框, 划分为基于锚点框和无锚框的检测算法.

双阶段检测算法分为2个阶段: 首先在得到预测结果前生成候选区域建议, 然后针对每个区域进行分类和回归, 代表性的算法有基于候选区域的卷积神经网络(region-convolutional neural networks, R-CNN)[2], Fast R-CNN[3], Faster R-CNN[4], Mask R-CNN[5]等. Yang 等[6]提出一种多尺度特征融合结构, 在浅层特征增加更多的特征金字塔网络(feature pyramid networks, FPN), 并通过稠密连接增加特征之间的信息交流来提升船舶的检测精度. 为了提升遥感图像中小目标的检测效果, Zhang 等[7]在 Faster R-CNN 的基础上, 通过对第 1 阶段提取的候选区域上采样来扩大特征图的尺寸, 进而提升网络的检测精度. 考虑到光学遥感图像中的目标具有不同的形状和大小, Dong 等[8]提出基于注意力的多级特征融合模块来自适应地融合 FPN 的多级输出, 以更好地检测具有各种形状和大小的目标.

单阶段检测算法无需生成候选区域, 经过单次检测就能得到最终结果, 所以检测速度比双阶段检测算法快, 但精度一般会有所下降, 代表性的算法有 YOLO(you only look once)[9], SSD(single-shot multibox detector)[10]和 RetinaNet[11]. Fu 等[12]在 FPN 自顶向下融合后又加入自底向上的连接, 该结构增加了特征的复用, 进一步增强特征的表达能力, 有效地提高网络对遥感目标的检测精度. 为了消除背景噪声的影响, Li 等[13]通过对不同尺度的特征分别增加注意力机制模块来增强特征, 提升网络的检测精度. 为了提高小尺寸物体的检测精度. Zhang 等[14]提出基于 YOLOv3 的深度可分离注意力引导网络, 将多级特征级联和通道特征注意机制相结合, 使得网络的特征表达能力得到显著提升, 遥感图像中小尺寸车辆的检测精度得到提高.

大多数检测算法需要预设不同大小和比例的锚点框作为待检测物体的初始检测框, 经过网络回归真实框和锚点框之间的差异, 如其中心点的偏移以及锚点框宽、高的缩放比例, 再对锚点框做进一步的微调, 使其逐渐逼近真实框, 这类算法的检测性能在一定程度上取决于针对不同场景下检测任务的锚点框设计的好坏. 近期兴起的无锚框检测算法摒弃了之前采用锚点框的思想, 直接基于像素层面对目标进行回归和分类来完成检测任务, 代表性的算法有 ConerNet[15], ExtremeNet[16], CenterNet[17]等, 近年来, 这类算法被广泛地应用于遥感图像目标检测领域. Shi 等[18]提出一种检测任意方向的车辆的单阶段无锚框算法, 采用 FPN 融合策略连接多个尺度的特征, 直接用于车辆的回归和识别. 考虑到遥感图像中物体的尺度差异和小物体类别众多, Li 等[19]提出基于密集路径聚合 FPN 的无锚框检测算法, 充分利用遥感图像中的高层次语义信息和低层次位置信息, 在一定程度上避免了浅层特征在传播过程中的信息丢失.

Liu 等[20]提出一个中心和边界的双注意力模块, 利用双重注意机制提取对象中心和边界区域的注意特征, 以减少复杂背景的干扰. Song 等[21]提出自适应融合 FPN, 首先融合来自相邻层的特征, 或融合不相邻的深、浅层特征, 获得更精细的特征表示; 然后对每层引入可学习参数, 实现自适应融合; 结合注意力机制提出前景增强模块来抑制遥感图像复杂背景所带来的噪声, 减少误检和漏检. 杨治佩等[22]在 CenterNet 的基础上改进了原有的检测结构, 提出一个非对称卷积的特征增强模块来增强特征的表达能力, 提高网络在遥感图像中密集目标上的检测精度.

上述算法在遥感图像目标检测领域有着卓越的性能, 但由于遥感图像中小尺寸目标数量庞大(如飞机、汽车和船舶等), 仍然面临小目标较难检测的问题. 虽然大多数算法通过融合深层特征与浅层特征来提高对小尺寸目标的检测准确度, 但是在进行融合时没有充分利用主干网络提取的多尺度特征图, 也未考虑不同尺度特征对融合后特征的贡献度的问题, 导致融合后特征的信息冗杂, 削弱了小目标的特征信息. 不仅如此, 遥感图像中的目标还存在尺度变化较大的问题, 只有极少的算法能做到兼顾小尺寸目标和较大尺寸目标的检测, 网络对目标尺度的泛化能力仍需进一步提升. 此外, 遥感图像中的物体大多分布在复杂背景之中, 可能出现密集排布的现象(如轮船和储罐等), 图像中同一区域存在多个感兴趣目标, 导致漏检、误检的问题, 降低检测精度.

针对上述问题, 本文以 CenterNet 为基础, 提出基于自适应多尺度特征融合(adaptive multi-scale feature fusion, AMFF)模块和注意力特征增强(attention feature enhancement, AFE)模块的目标检测算法, 分为 3 部分: (1) 主干网络; (2) AMFF 模块; (3) 加入了 AFE 模块的检测头. 为了充分利用主干网络提取到的多尺度特征图, AMFF 模块在自上而下进行特征融合时, 采用更多的横向连接以增加相邻特征之间的交流, 得到更丰富的特征信息, 提升网络的多尺度特征表达能力; 由于不同分辨特征对融合后特征的贡献度不同, AMFF 模块中又引入了可学习的权重系数来改善特征的比例不变性; 当遥感图像中存在大小不同的物体时, AFE 模块采用不同扩张率的空洞卷积, 可以同时提取不同尺度目标的特征, 提升网络对目标尺度的泛化能力, 最后利用注意力机制弱化背景和噪声, 增强有效的特征信息. 在 DIOR 数据集[23]和 NWPU VHR-10 数据集[24]上进行大量实验的结果表明, 本文算法可以有效地提升小目标检测精度, 并在多尺度目标检测性能上也有所提升.

1　本文算法

本文算法的结构如图 1 所示. 输入图像首先通

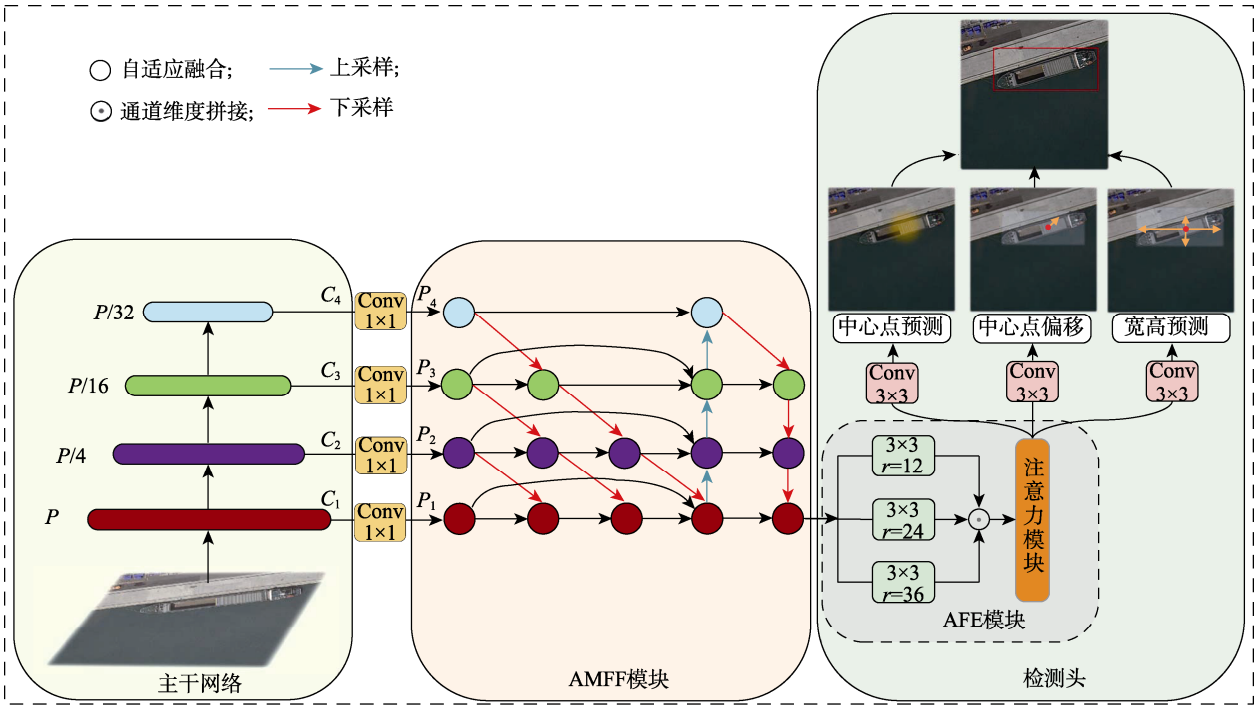


图 1　本文算法结构

过 ResNet[25]主干网络提取不同分辨率的特征；然后通过 AMFF 模块将提取的特征进行自适应地融合；再利用检测头中的 AFE 模块将融合后的特征增强；最后经过 3 个并行的卷积分支分别得到中心点预测、中心点偏移预测和目标宽高预测，得到最终的预测结果.

1.1 AMFF 模块

卷积神经网络提取到的浅层特征具有较高的分辨率，包含更多小尺寸目标的细节和位置信息. 随着网络的加深，经过多次卷积和池化操作之后提取到的深层特征有着更强的语义信息，但其分辨率较低，会丢失目标的位置和细节信息，尤其是小目标信息更是被弱化. 如何有效地表示和利用多尺度特征，一直是目标检测的难点之一. 为此，本文设计了可以充分利用主干网络来提取深、浅层特征信息进行高效融合的 AMFF 模块，其结构如图 2 所示，图中的节点既为融合后特征的名称，又表示该次融合操作.

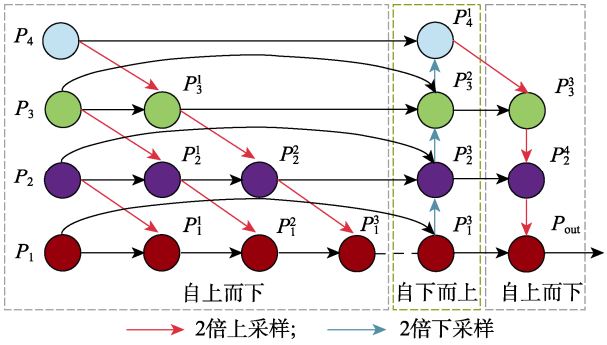


图 2 AMFF 结构

为了方便进行特征融合，首先将主干网络提取到的特征 $\{C_1,C_2,C_3,C_4\}$ 分别进行一次 1×1 的卷积，将所有特征的通道数调整到与 C_1 一致，得到 AMFF 的输入特征 $\{P_1,P_2,P_3,P_4\}$；然后从输入的最深层特征开始，自上而下融合得到 $\{P_1^3,P_2^2,P_3^1,P_4\}$，并且为了充分利用多尺度特征，在融合过程中增加更多的横向连接，既可以确保特征信息的完整性，又能够实现将深层特征的语义信息有效地引入到浅层特征中，增加浅层高分辨特征的语义信息；为了避免因自上而下路径中上采样次数频繁而丢失目标信息，又引入自下而上路径，通过增加跳跃连接使得原始输入特征 $\{P_1,P_2,P_3,P_4\}$ 参与逐层融合的过程，得到 $\{P_1^3,P_2^3,P_3^2,P_4^1\}$；最后再增加一条自上而下路径，通过上采样逐层相加得到较高分辨率 $(P/4)$ 的特征图 P_{out}，在细节和位置信息更加丰富的同时语义信息更强，能够提升网络的多尺度特征表达能力.

在进行多尺度特征融合时，不同分辨率特征图对融合后特征图的贡献度会有所差异[26]. 本文借鉴文献[26]中双向 FPN 的思想引入可学习的权重系数，实现自适应融合多个尺度的特征，改善融合过程中各尺度特征的比例不变性，使得融合后的特征包含更多对检测任务有帮助的信息. 本文首先将待融合的多个尺度特征图的分辨率调整到一致：(1) 在自上而下路径中，深层特征图通过最近邻插值法进行 2 倍上采样；(2) 在自下而上路径中，浅层特征图通过最大池化进行 2 倍下采样. 然后将调整后的特征图分别与其对应的权重系数相乘后，再逐元素相加. 最后经过 Swish 激活函数、卷积和批归一化(batch normalization, BN)来融合特征. 待融合节点在融合过程的简略表达式为

$$F = F_1 w_1 + \cdots + F_i w_i + \cdots + F_n w_n \tag{1}$$

其中，F 表示当前进行融合的特征节点；F_i 表示指向当前节点 F 的某个特征节点；n 表示指向 F 节点的总数；w_i 表示该节点对应于 F 节点的权重系数，公式为

$$w_i = \frac{w_i}{w_1 + w_2 + \cdots + w_n + \delta} \tag{2}$$

w_i 定义为 F_i 占所有指向 F 的 n 个节点的比例，其值为 0~1；δ 取 0.0001 是为了防止学习到的权重系数不稳定而设置的一个极小的值. 在计算 w_i 之前采用 ReLU 激活函数处理，确保 w_i 不为负值.

以 P_{out} 节点为例，自适应特征融合过程如图 3 所示. 其中，上采样表示采用最近邻线性插值法进行 2 倍上采样，使待融合的特征分辨保持一致；在 AMFF 模块中采用深度可分离卷积(depthwise separable convolution, DSC)[27-28]，与常规卷积不同，DSC 先对独立的通道层进行卷积，再使用 1×1 卷积进行深度扩充完成卷积操作，可以有效地缩减卷积计算的耗时与参数量，提高网络检测效率.

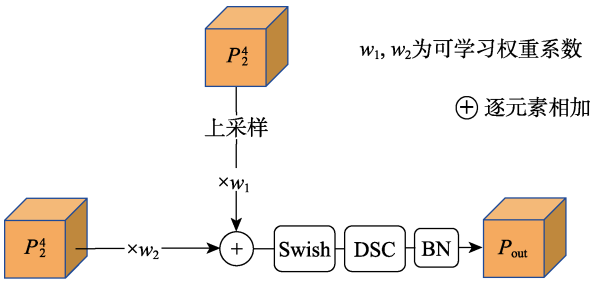


图 3 自适应特征融合示例

1.2　AFE 模块

在卷积神经网络中，固定大小的感受野不利于检测不同尺寸的物体，特别是对于那些遥感图像领域中存在尺度变化很大的目标．在语义分割任务中，通常使用空洞卷积来扩大感受野，对图像中每个像素点都能更好地分类，以实现更精准的分割．如 DeepLab[29]中的空洞空间金字塔池化(atrous spatial pyramid pooling, ASPP)方法中，由于遥感图像中的目标尺度比自然图像变化得更加剧烈，因此该方法并不直接适用于遥感图像目标检测．本文借鉴 ASPP 方法，采用跨度更大的扩张率的空洞卷积分支来获取不同大小的感受野，提升网络对不同尺度目标的检测性能．不同扩张率的空洞卷积的感受野如图 4 所示．

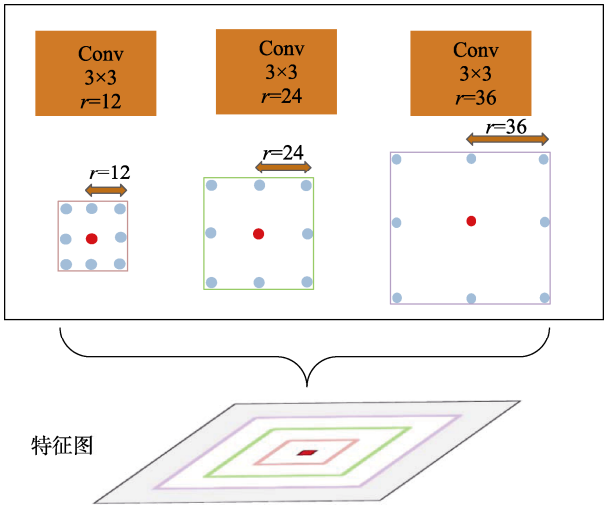


图 4　不同扩张率的空洞卷积的感受野示意图

遥感图像场景复杂，可能出现目标密集排布的现象，导致目标漏检、误检，并且经过特征融合和多分支空洞卷积之后，可能会引入大量的噪声．为此，本文使用混合域注意力机制的卷积块的注意力模块(convolutional block attention module, CBAM)[30]抑制背景和噪声．空间域的注意力根据输入的特征计算得到保存关键空间信息的注意力掩码，将该掩码附加在原来的特征上，增强有用的特征，弱化无用特征，起到特征筛选和增强的效果．通道域注意力单独为每个特征通道增加一个可学习的权重，表示与关键信息的相关程度，加强对当前任务有用的通道；反之，则抑制．CBAM 串行组合更有效地增强了模型对特征的提取能力，提高模型的检测效果．

AFE 模块的结构如图 5 所示．首先，对输入的特征分别取扩张率为 $r=12$，$r=24$ 和 $r=36$，采用卷积核大小为 3×3 的 3 条空洞卷积分支获取不同大小的感受野，并将 3 条分支的输出特征在通道维度拼接；然后经过一个 1×1 的卷积融合特征与输入的原始特征逐元素相加；最后利用 CBAM 抑制背景和噪声，增强特征信息．

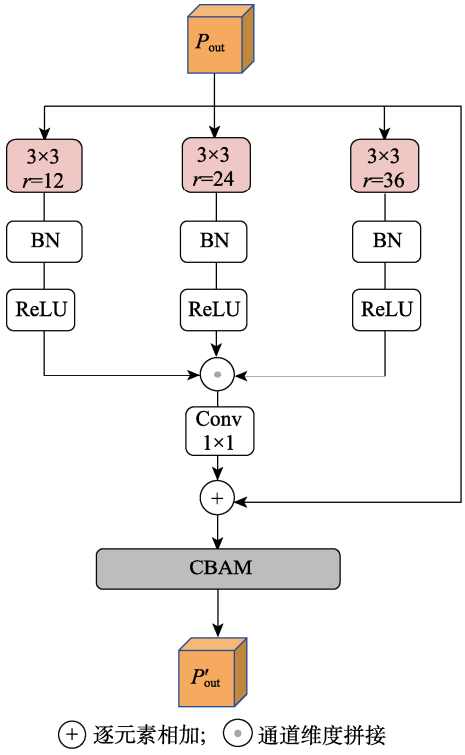


图 5　AFE 模块结构

1.3　损失函数

本文采用与 CenterNet 一致的损失计算方式，分别为中心点损失、目标中心偏移损失和目标尺寸的损失．假设输入一幅宽为 W，高为 H 的图像 $I\in\mathbb{R}^{W\times H\times 3}$，利用本文算法生成一幅关键点的热图 $\hat{Y}\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$．其中，$R$ 表示输出的步长，本文设置为 4；C 表示检测点类别数．当 $\hat{Y}_{x,y,c}=1$ 时，表示当前 (x,y) 处检测到一个类别为 C 的物体；当 $\hat{Y}_{x,y,c}=0$ 时，则表示背景．对于 C 类的每个真实标签中心点 $p\in\mathbb{R}^2$，其相对于输出之后的中心为 $\tilde{p}=\left\lfloor\frac{p}{R}\right\rfloor$，使用高斯核计算中心点附近指定范围内的点相对于中心的距离程度值的分布图，计算公式为

$$H_{xyc}=\exp\left(\frac{-\left(x-\tilde{p}_x\right)^2+\left(y-\tilde{p}_y\right)^2}{2\sigma_p^2}\right)\tag{3}$$

其中，σ_p 表示距离程度值覆盖半径的标准差．

中心点损失函数采用改进的 Focal Loss，计算

公式为

$$L_k=\frac{-1}{N}\sum_{xyc}\begin{cases}\left(1-\hat{Y}_{xyc}\right)^{\alpha}\ln\left(\hat{Y}_{xyc}\right), & Y_{xyc}=1\\ \left(\hat{Y}_{xyc}\right)^{\alpha}\left(1-Y_{xyc}\right)^{\beta}\ln\left(1-\hat{Y}_{xyc}\right), & Y_{xyc}\neq 1\end{cases} \tag{4}$$

其中，α 和 β 表示 Focal Loss 的超参数，本文分别设置为 2 和 4；N 表示某幅图像的中心点数量；$\hat{Y}_{xyc}$ 为输出的概率热图；Y_{xyc} 表示 $\hat{Y}_{xyc}$ 对应的距离程度值.

由于计算 $\tilde{p}$ 时向下取整可能会产生误差，因此增加了中心点偏移 $\hat{O}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$ 预测，且所有类别 C 共享相同的偏移预测，采用 L_1 损失函数，计算公式为

$$L_{\text{off}}=\frac{1}{K}\sum_{p}\left|\hat{O}_{\tilde{p}}-\left(\frac{p}{R}-\tilde{p}\right)\right| \tag{5}$$

假设 C_k 类的目标 k 的边界框标签为 $\left(x_1^{(k)},y_1^{(k)},x_2^{(k)},y_2^{(k)}\right)$，则其中心点坐标 p_k 和实际尺寸 s_k 的计算公式分别为

$$p_k=\left(\frac{x_1^{(k)}+x_2^{(k)}}{2},\frac{y_1^{(k)}+y_2^{(k)}}{2}\right) \tag{6}$$

$$s_k=\left(x_2^{(k)}-x_1^{(k)},y_2^{(k)}-y_1^{(k)}\right) \tag{7}$$

由式(6)(7)可知，采用 L_1 损失函数得到尺寸损失的计算公式为

$$L_{\text{size}}=\frac{1}{N}\sum_{k=1}^{N}\left|\hat{S}_{p_k}-s_k\right| \tag{8}$$

其中，$\hat{S}\in\mathbb{R}^{\frac{W}{R}\times\frac{H}{R}\times 2}$ 表示预测的尺寸.

整体损失 L_{det} 为中心点损失、中心偏移损失以及尺寸之和，计算公式为

$$L_{\text{det}}=L_k+\lambda_{\text{size}}L_{\text{size}}+\lambda_{\text{off}}L_{\text{off}} \tag{9}$$

其中，λ_{size} 和 λ_{off} 表示尺寸损失和偏移损失的权重，本文分别设置为 0.1 和 1.0.

2 实验结果与分析

2.1 数据集

为了验证本文算法的有效性，在 DIOR 数据集和 NWPU VHR-10 数据集上进行实验. 这 2 个数据集都是遥感图像中多类别目标检测广泛使用的公共数据集.

DIOR 数据集是遥感图像目标检测最具挑战性的大规模基准数据集之一，该数据集从 Google Earth 获取 23 463 幅图像，标注了包括飞机、机场、高尔夫球场和车辆等 20 个类别的 192 472 个实例；与其他数据集相比，该数据集的图像和对象实例具有更高的类内变化性和类别之间相似性. 因此，DIOR 数据集被认为非常适用于遥感图像目标检测算法的训练和评估. 本文实验中，数据集的划分遵循官方标准，包括训练集、验证集和测试集，图像数量分别为 5 862, 5 863 和 11 738.

NWPU VHR-10 数据集包括来自 Google Earth 和部分来自 Vaihingen 数据集[31]的 800 幅分辨率非常高的遥感图像，该数据集分为一个带标注的正图像集和一个负图像集. 其中，负图像集中的 150 幅图像不包含任何实例，用于探索半监督和弱监督检测算法；正图像集中的 650 幅图像标注了 10 种物体类别：飞机、轮船、油罐、棒球场、网球场、篮球场、田径场、港口、桥梁和车辆. 本文实验中，将数据集随机划分为包括 260 幅图像的训练集、130 幅图像的验证集和 260 幅图像的测试集.

2.2 实验设置与评价指标

实验平台和系统设置如下：AMD Ryzen 5 3600X 6-Core CPU, NVIDIA GeForce RTX 3090 GPU; Ubuntu 20.04 操作系统，使用 pytorch1.8 深度学习框架, Python 版本为 3.8. 实验中，选用随机梯度下降作为优化算法，动量设置为 0.9，初始学习率设置为 0.01，衰减系数为 0.000 1，并采用 step 策略来调整学习率，批大小设为 18，一共训练 46 轮，分别在第 36 轮和第 42 轮降低学习率；采用有监督数据增强方法，包括色域增强、仿射变换、随机裁剪、随机旋转、随机尺度变换和随机翻转等.

本文采用平均精度(mean average precision, mAP)作为评价指标，其代表各个类别平均精度的平均值；AP 表示某个类别根据准确率 P 和召回率 R 在 0~1 所绘制曲线面积的大小. 计算公式为

$$P=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$R=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{AP}=\int_0^1 P(R)\text{d}(R) \tag{12}$$

$$\text{mAP}=\frac{1}{C}\sum_{i=1}^{C}\text{AP}_i \tag{13}$$

其中，真阳例(true positive, TP)表示模型判断为真，且实际也为真的目标数量；假阴例(false negative, FN)表示模型判断为假，但实际为真的目标数量；假阳例(false positive, FP)表示模型判断为真，但实际为假的目标数量；通过式(12)可以计算出单个类

别的 AP 值; C 表示参与计算的类别数.

2.3　不同扩张率组合的空洞卷积分支分析

扩张率的组合方式不同, 其感受野的大小也会有所差异. 为了探究更适合遥感图像领域的组合方式, 以 ASPP 的扩张率 $r=(6,12,18)$ 为基准, 将其降低为 $r_1=(1,1,1)$, 扩大到 $r_2=(12,18,24)$ 和 $r_3=(12,24,36)$, 分别在 DIOR 和 NWPU VHR-10 数据集上进行对比. 实验中, 主干网络采用 RestNet50, 并且都使用相同的配置, 结果如表 1 和表 2 所示.

表 1　DIOR 数据集上不同扩张率对比　%

地物类别	扩张率组合			
	r_1	r	r_2	r_3
立交桥	53.4	52.2	**54.4**	51.6
高速公路服务区	63.9	63.7	**68.7**	65.5
烟囱	72.2	72.5	72.2	**72.5**
车辆	56.0	55.0	55.2	**56.8**
田径场	**71.8**	68.5	71.2	65.8
港口	35.3	38.0	**40.3**	38.8
桥梁	39.9	40.8	40.3	**40.9**
篮球场	**88.6**	88.5	88.3	88.5
轮船	71.1	**71.3**	71.1	71.0
机场	59.5	59.8	**60.9**	60.5
水坝	41.6	42.4	51.0	**51.2**
储罐	**75.9**	70.6	70.6	70.9
高尔夫球场	40.5	54.6	54.2	**64.2**
高速公路收费站	71.3	75.6	74.5	**75.8**
火车站	**54.0**	48.4	53.6	49.1
飞机	80.1	79.8	80.7	**80.8**
棒球场	78.6	**79.8**	79.5	79.1
网球场	89.7	89.5	89.7	**89.9**
体育场	**70.0**	68.2	62.9	68.4
风车	77.1	78.3	78.5	**85.5**
mAP	64.5	64.9	65.9	**66.4**

注. 粗体表示最优值.

表 2　NWPU VHR-10 数据集上不同扩张率对比　%

地物类别	扩张率组合			
	r_1	r	r_2	r_3
飞机	**100.0**	99.9	90.7	90.9
轮船	**91.0**	90.9	90.4	90.3
油罐	90.2	**90.4**	89.0	90.3
棒球场	**90.3**	90.3	90.3	89.9
网球场	90.5	**90.7**	90.2	90.3
篮球场	90.8	**91.6**	90.9	90.4
田径场	88.2	82.9	**90.6**	89.4
港口	79.6	83.1	76.3	**89.5**
桥梁	51.0	58.4	**67.6**	63.2
车辆	89.1	89.8	89.6	**89.9**
mAP	86.1	86.8	86.6	**87.4**

注. 粗体表示最优值.

从表 1 和表 2 可以看出, 不同的扩张率组合均会影响 2 个数据集中不同尺度物体的检测效果; DIOR 数据集中, 随着扩张率的增大, 水坝、港口、高尔夫球场等纵横比较大的物体类别的检测精度逐渐增大; NWPU VHR-10 数据集中, 随着扩张率的增大, 飞机和轮船等尺度较小的物体类别的检测精度逐渐降低; 由于遥感图像中目标具有多尺度和尺度变化剧烈的特点, 在 2 个数据集中使用扩张率组合较大时, 多数类别的检测效果较好. 综上所述, 本文选择 mAP 值最高的 r_3 组合方式作为空洞卷积分支的扩张率.

2.4　对比实验

2.4.1　在 DIOR 数据集上对比

为了验证本文算法的有效性, 将其与目前一些经典目标检测算法进行对比. 在 DIOR 数据集上, 与文献[23]给出的不同网络在测试集上的基准分数进行比较, 包括 SSD[10]和 RetinaNet[11], 加入了 FPN 的 Faster R-CNN(faster R-CNN with FPN, FR-F)[32], 加入了 FPN 的 Mask R-CNN(mask R-CNN with FPN, MR-F)[32], YOLOv3[33], PANet[34]和 CenterNet[17], 结果如表 3 所示. 其中, 地物类别 C_1~C_{20} 分别对应立交桥、高速公路服务区、烟囱、车辆、田径场、港口、桥梁、篮球场、轮船、机场、水坝、储罐、高尔夫球场、高速公路收费站、火车站、飞机、棒球场、网球场、体育场和风车.

从表 3 可以看出, 本文算法在主干网络使用 ResNet50 时, mAP 值比 CenterNet 提高 10.1 个百分点, 且优于其他算法; 使用提取特征能力更强的 ResNet101 时, mAP 值最高, 达到 72.4%, 比次优算法 mAP 值提升 6.3 个百分点, 并且在飞机、棒球场、篮球场、水坝和烟囱等大部分地物类别的检测上 AP 值最高, 尤其在飞机、储罐、车辆、风车等小尺寸和密集排布目标的平均检测精度提升显著, 比次优结果分别提升 3.0, 10.0, 10.4 和 2.4 个百分点; 对于桥梁、水坝、体育场和高速公路收费站等具有较大纵横比的目标, 网络检测效果较好, 比次优 AP 值分别提升 2.4, 3.4, 2.6 和 9.2 个百分点.

2.4.2　在 NWPU VHR-10 数据集上对比

为了进一步验证本文算法的有效性, 本文算法使用第 2.2 节的实验设置, 其他算法采用对应的参考文献给出的训练策略、参数等进行实验. 在 NWPU VHR-10 数据集上与 YOLOv3[33], FR-F[32], Cascade R-CNN[35], RetinaNet[11]和 CenterNet[17]进行对比, 结果如表 4 所示. 可以看出, 本文算法的 mAP 值最高, 达到 87.4%, 比 CenterNet 提升 13.5 个百

表 3 不同算法在 DIOR 数据集上检测结果对比 %

算法	主干网络	地物类别																			mAP
		C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	
SSD[10]	VGG16	48.1	63.5	65.8	27.4	68.6	49.4	29.7	75.7	59.2	72.7	56.6	46.6	65.3	53.1	55.1	59.5	7.24	76.3	61.0	65.7	58.6
YOLOv3[33]	DarkNet53	49.7	48.6	69.7	48.3	61.1	44.9	31.2	78.6	**87.4**	29.2	26.9	68.7	31.1	54.4	29.4	72.2	74.0	87.3	70.6	78.7	57.1
FR-F[32]	ResNet50	56.0	68.7	72.5	43.1	76.5	42.8	42.6	81.0	71.8	71.4	57.5	53.5	73.1	62.1	53.0	54.1	63.3	81.2	57.0	80.9	63.1
	ResNet101	57.2	75.6	72.5	43.1	**76.8**	46.4	44.8	80.7	71.8	74.5	60.0	53.8	76.0	62.3	59.5	54.0	63.3	81.1	68.3	81.2	65.1
MR-F[32]	ResNet50	56.5	71.6	72.6	43.1	75.8	44.2	38.7	81.0	71.9	72.3	55.9	53.6	73.0	67.0	54.0	53.8	63.2	81.1	58.6	81.1	63.5
	ResNet101	57.4	76.3	72.5	43.0	75.9	46.5	40.2	80.9	71.8	76.6	60.4	53.7	76.0	62.5	62.3	53.9	63.2	81.0	68.3	81.0	65.2
PANet[34]	ResNet50	55.8	68.4	72.5	48.2	74.6	41.6	38.9	80.4	71.7	70.4	56.6	62.3	69.0	60.0	54.6	61.9	71.0	81.2	72.9	86.7	63.8
	ResNet101	56.9	72.1	72.3	47.2	73.4	45.3	43.6	80.5	71.7	72.0	61.4	62.0	72.0	66.7	57.0	60.2	70.6	80.9	70.4	84.5	66.1
RetinaNet[11]	ResNet50	59.6	76.2	72.3	45.1	74.2	**50.7**	44.1	81.3	71.2	**77.3**	62.5	44.8	77.7	66.0	54.2	53.7	69.0	81.3	69.3	83.4	65.7
	ResNet101	59.6	78.6	73.2	44.4	76.6	49.9	44.1	85.0	71.1	77.0	62.4	45.8	**78.6**	62.8	55.2	53.3	69.3	81.3	68.4	85.5	66.1
CenterNet[17]	ResNet50	45.0	39.1	69.2	47.5	48.7	24.5	36.3	88.1	71.0	42.2	36.8	61.6	23.6	63.3	42.8	78.3	76.2	88.9	63.1	79.6	56.3
	ResNet101	50.9	65.8	72.6	48.5	63.2	35.9	41.0	89.0	71.2	54.2	45.3	61.9	53.5	69.7	51.6	71.3	77.1	89.2	68.6	79.5	63.0
本文	ResNet50	51.6	65.5	72.5	56.8	65.8	38.8	40.9	88.5	71.0	60.5	51.2	70.9	64.2	75.8	49.1	80.8	79.1	89.9	68.4	85.5	66.4
	ResNet101	**60.2**	**79.7**	**79.5**	**58.9**	70.0	45.2	**47.2**	**89.3**	71.4	71.7	**65.9**	**78.7**	69.3	**78.9**	**64.6**	**81.3**	**80.8**	**90.1**	**75.5**	**89.1**	**72.4**

注. 粗体表示当前列最优值, 下划线表示当前列次优值.

表 4 不同算法在 NWPU VHR-10 数据集上检测结果对比 %

算法	主干网络	地物类别										mAP
		飞机	轮船	油罐	棒球场	网球场	篮球场	田径场	港口	桥梁	车辆	
YOLOv3[33]	DarkNet50	89.5	81.8	87.3	89.3	84.0	90.2	93.6	88.1	57.0	82.7	84.4
FR-F[32]	ResNet50	90.5	86.6	87.1	90.3	86.4	87.8	90.3	77.0	59.5	78.6	83.4
Cascade R-CNN[35]	ResNet50	90.9	88.4	88.8	90.3	87.5	90.4	**95.8**	83.0	62.3	79.6	85.7
RetinaNet[11]	ResNet50	**92.7**	82.1	89.9	89.6	80.8	84.9	94.4	76.7	57.7	76.9	82.6
CenterNet[17]	ResNet50	90.9	55.4	90.3	**90.5**	73.3	81.9	87.7	41.0	41.4	87.0	73.9
本文	ResNet50	90.9	**90.3**	**90.3**	89.9	**90.3**	**90.4**	89.4	**89.5**	**63.2**	**89.9**	**87.4**

注. 粗体表示当前列最优值, 下划线表示当前列次优值.

分点, 比次优值提升 1.7 个百分点; 在大部分类别检测上 AP 值最高, 对车辆、轮船、油罐这些目标尺度较小的目标检测效果较好, AP 值较次优结果分别提升 2.9, 1.9 和 0.4 个百分点, 对网球场和篮球场等密集排布的目标检测的效果仍有不同程度的提升, 并且对港口和桥梁这类纵横比较大的目标检测仍然有效, 较次优 AP 值分别提升 1.4 和 0.9 个百分点.

综合上述在 2 个数据集上的实验结果表明, 本文算法对车辆、飞机、储罐等物体尺度较小和密集排布的目标可以实现较为精准的检测, 这得益于 AMFF 模块自适应融合主干网络提取到多尺度后得到的更加丰富的特征信息; 在检测高速公路收费站、体育场、桥梁和水坝这类尺度变化大的目标时表现仍然较好, 这是由于 AFE 模块中的并行空洞卷积分支提取不同尺度目标特征, 可获得不同大小的感受野, 再结合注意力模块抑制背景和噪声, 增强了对检测有帮助的特征信息.

2.5 消融实验

进一步, 分析本文算法中各个模块的作用. 主干网络采用 ResNet50, 并按照第 2.2 节的实验设置, 分别在 DIOR 和 NWPU VHR-10 数据集上进行消融实验, 结果如表 5 和表 6 所示. 可以看出: (1) 仅加入 AMFF 模块时, 在 2 个数据集上较基线网络的 mAP 值分别提升 4.6 和 6.4 个百分点, 有效地提高了网络对多尺度特征表达能力, 验证了特征融合策略的可行性与有效性; (2) 仅加入 AFE 模块时, 在 2 个数据集上的 mAP 值分别提升 2.0 和 4.3 个百分点, 表明 AFE 模块能有效地加强特征; (3) 将 AMFF 与 AFE 模块结合, 在 2 个数据集上较基线网络的 mAP 值提升显著, 验证了本文算法的有效性; (4) 当 AMFF 模块采用自适应融合的策略时, 在 2 个数据集上的 mAP 值均有提升, 验证了在多尺度融合时引入可学习参数的有效性.

表 5　在 DIOR 数据集上消融实验结果

算法	AMFF	AFE	mAP/%
基线			56.3
基线+AMFF	√		60.9
基线+AFE		√	58.3
本文*	√	√	65.4
本文	√	√	66.4

注. “√”表示基线网络中加入该模块; *表示未引入可学习参数.

表 6　在 NWPU VHR-10 数据集上消融实验结果

算法	AMFF	AFE	mAP/%
基线			73.9
基线+AMFF	√		80.3
基线+AFE		√	78.2
本文*	√	√	86.8
本文	√	√	87.4

注. “√”表示基线网络中加入该模块; *表示未引入可学习参数.

2.6　计算复杂度和推理速度对比

为了评估本文算法的计算效率, 在 DIOR 和 NWPU VHR-10 数据集上与 YOLOv3[33], FR-F[32], RetinaNet[11] 以及 CenterNet[17] 对比浮点数运算(floating-point operations, FLOP)和推理速度. 计算 FLOP 时, 在 2 个数据集的输入图像大小均为 (800,800), 结果如表 7 所示. 可以看出, 本文算法的推理速度高于 FR-F; FLOP 大于 CenterNet, 但优于表中的其他算法, 这是由于本文算法在 AMFF 和 AFE 模块中引入了较为复杂的计算, 推理速度有待提升.

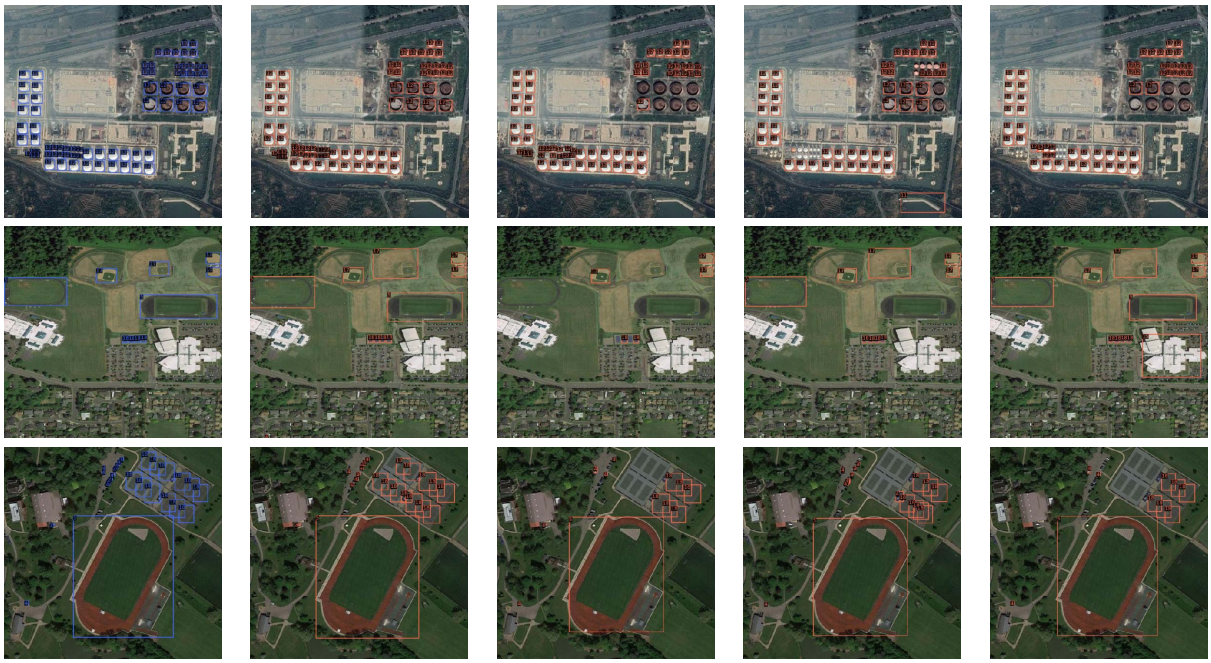
表 7　计算复杂度和推理速度对比

算法	DIOR		NWPU VHR-10	
	FLOP	速度/(帧·s^{-1})	FLOP	速度/(帧·s^{-1})
YOLOv3[33]	121.4G	35.5	121.3G	38.4
FR-F[32]	134.5G	18.6	134.4G	21.3
RetinaNet[11]	130.3G	28.6	133.1G	28.9
CenterNet[17]	63.5G	32.1	63.4G	34.8
本文	102.7G	22.9	102.6G	23.8

2.7　检测示例

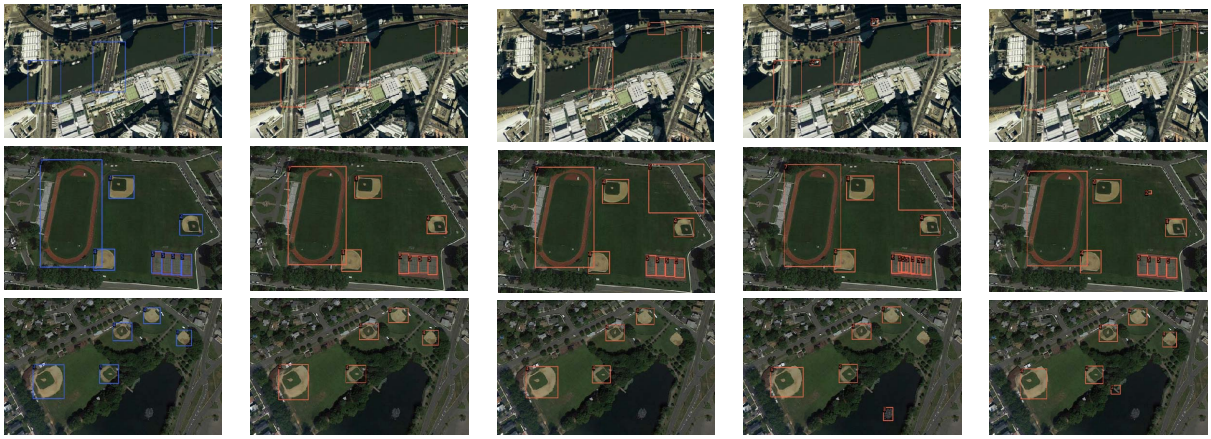
本文算法与 CenterNet, RetinaNet 和 FR-F 分别在 DIOR 和 NWPU VHR-10 数据集中的检测结果如图 6 和图 7 所示. 可以看出, 与 CenterNet 相比, 本文算法在车辆、储罐等小目标和桥梁等纵横比较大的目标检测上效果提升明显; 与其他算法相比, 本文算法的小目标漏检和误检明显减少, 并且在同一幅图像中出现尺度差异较大的物体时, 能够更准确地定位目标.

从图 8 和图 9 可以看出, 在不同分辨率图像中的同一种类尺度差异较大的目标(如飞机、储罐和田径场等)上, 本文算法不仅能够实现较高精度检测, 也能对同一幅图像中类间物体尺度差异较大的物体准确定位, 如同时存在立交桥和车辆的情况; 尤其在飞机、轮船和储罐等尺度较小和密集排布的目标检测上效果显著.



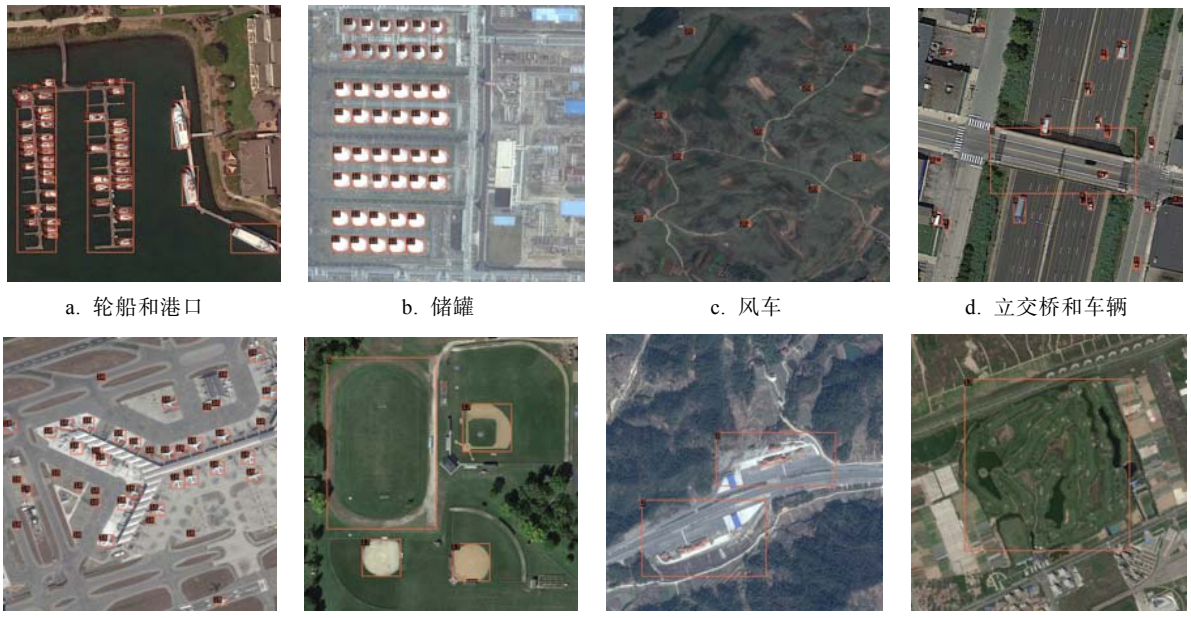
a. 标签　b. 本文　c. CenterNet[17]　d. RetinaNet[11]　e. FR-F[32]

图 6　在 DIOR 数据集上检测结果对比



a. 标签 b. 本文 c. CenterNet[17] d. RetinaNet[11] e. FR-F[32]

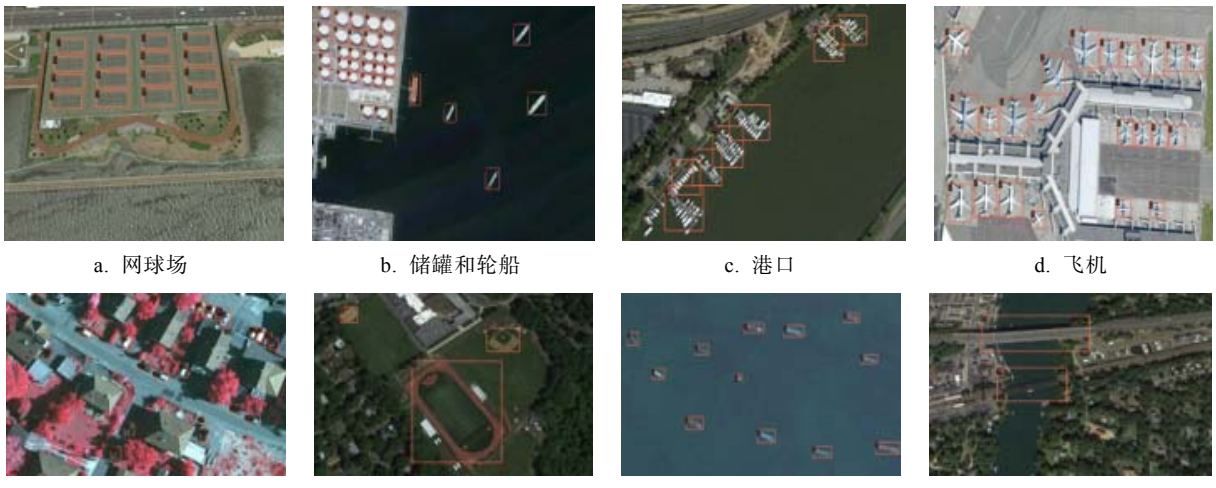
图 7 在 NWPU VHR-10 数据集上检测结果对比



a. 轮船和港口 b. 储罐 c. 风车 d. 立交桥和车辆

e. 飞机 f. 田径场和棒球场 g. 高速公路服务区 h. 高尔夫球场

图 8 本文算法在 DIOR 数据集上检测实例



a. 网球场 b. 储罐和轮船 c. 港口 d. 飞机

e. 车辆 f. 田径场和棒球场 g. 轮船 h. 桥梁

图 9 本文算法在 NWPU VHR-10 数据集上检测实例

3 结　语

针对遥感图像内容丰富且复杂, 目标种类多、密集分布和尺度变化剧烈, 导致遥感图像目标多尺度尤其小目标难以检测的问题, 本文提出一种基于 AMFF 模块和 AFE 模块的无锚框遥感图像目标检测算法. AMFF 模块有效地融合主干网络提取的多尺度特征, 丰富特征信息, 增强了网络对多尺度特征的表达能力; AFE 模块使用空洞卷积分支提取不同尺寸目标的特征, 并通过混合注意力抑制背景和噪声, 增强有效特征信息. 实验结果表明, 本文算法可以实现对尺寸较小和密集分布目标的高精度检测, 对类间和类内尺寸变化大的目标也能实现比较准确的定位, 验证了该算法的可行性与有效性.

参考文献(References):

[1] Nie Guangtao, Huang Hua. A survey of object detection in optical remote sensing images[J]. Acta Automatica Sinica, 2021, 47(8): 1749-1768(in Chinese)
(聂光涛, 黄华. 光学遥感图像目标检测算法综述[J]. 自动化学报, 2021, 47(8): 1749-1768)

[2] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 580-587

[3] Girshick R. Fast R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448

[4] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

[5] He K M, Gkioxari G, Dollár P, *et al*. Mask R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2980-2988

[6] Yang X, Sun H, Sun X, *et al*. Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network[J]. IEEE Access, 2018, 6: 50839-50849

[7] Zhang W, Wang S H, Thachan S, *et al*. Deconv R-CNN for small object detection on remote sensing images[C] //Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Los Alamitos: IEEE Computer Society Press, 2018: 2483-2486

[8] Dong X H, Qin Y, Gao Y H, *et al*. Attention-based multi-level feature fusion for object detection in remote sensing images[J]. Remote Sensing, 2022, 14(15): 3735

[9] Redmon J, Divvala S, Girshick R, *et al*. You only look once: unified, real-time object detection[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 779-788

[10] Liu W, Anguelov D, Erhan D, *et al*. SSD: single shot MultiBox detector[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 21-37

[11] 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

[12] Fu K, Chang Z, Zhang Y, *et al*. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 161: 294-308

[13] Li Q P, Mou L C, Jiang K Y, *et al*. Hierarchical region based convolution neural network for multiscale object detection in remote sensing images[C] //Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. Los Alamitos: IEEE Computer Society Press, 2018: 4355-4358

[14] Zhang Z Y, Liu Y P, Liu T C, *et al*. DAGN: a real-time UAV remote sensing image vehicle detection framework[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(11): 1884-1888

[15] Law H, Deng J. CornerNet: detecting objects as paired keypoints[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 765-781

[16] Zhou X Y, Zhuo J C, Krähenbühl P. Bottom-up object detection by grouping extreme and center points[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 850-859

[17] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[OL]. [2022-04-14]. https://arxiv.org/abs/1904.07850

[18] Shi F R, Zhang T, Zhang T. Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(6): 5221-5233

[19] Li Y Y, Pei X, Huang Q, *et al*. Anchor-free single stage detector in remote sensing images based on multiscale dense path aggregation feature pyramid network[J]. IEEE Access, 2020, 8: 63121-63133

[20] Liu S, Zhang L, Lu H C, *et al*. Center-boundary dual attention for oriented object detection in remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-14

[21] Song K, Huang P M, Lin Z P, *et al*. An oriented anchor-free object detector including feature fusion and foreground enhancement for remote sensing images[J]. Remote Sensing Letters, 2021, 12(4): 397-407

[22] Yang Zhipei, Ding Sheng, Zhang Li, *et al*. Anchor-free remote sensing image detection method for dense objects with rotation[J]. Journal of Computer Applications, 2022, 42(6): 1965-1971(in Chinese)
(杨治佩, 丁胜, 张莉, 等. 无锚点的遥感图像任意角度密集目标检测方法[J]. 计算机应用, 2022, 42(6): 1965-1971)

[23] Li K, Wan G, Cheng G, *et al*. Object detection in optical remote sensing images: a survey and a new benchmark[J]. ISPRS

Journal of Photogrammetry and Remote Sensing, 2020, 159: 296-307

[24] Cheng G, Han J W, Zhou P C, *et al*. Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2014, 98: 119-132

[25] 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

[26] Tan M X, Pang R M, Le Q V. EfficientDet: scalable and efficient object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 10778-10787

[27] Chollet F. Xception: deep learning with depthwise separable convolutions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 1800-1807

[28] Howard A G, Zhu M L, Chen B, *et al*. MobileNets: efficient convolutional neural networks for mobile vision applications[OL]. [2022-04-14]. https://arxiv.org/abs/1704.04861

[29] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

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

[31] Han X B, Zhong Y F, Zhang L P. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery[J]. Remote Sensing, 2017, 9(7): 666

[32] Lin T Y, Dollár 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

[33] Redmon J, Farhadi A. YOLOv3: an incremental improvement[OL]. [2022-04-14]. https://arxiv.org/abs/1804.02767

[34] Wang K X, Liew J H, Zou Y T, *et al*. PANet: few-shot image semantic segmentation with prototype alignment[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 9196-9205

[35] Cai Z W, Vasconcelos N. Cascade R-CNN: high quality object detection and instance segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1483-1498