智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于MobileViT和多尺度特征聚合的遥感图像目标检测

梁礼明, 冯耀, 龙鹏威, 李仁杰

引用本文:
梁礼明, 冯耀, 龙鹏威, 李仁杰. 基于MobileViT和多尺度特征聚合的遥感图像目标检测[J]. 智能系统学报, 2024, 19(5): 1168-1177.
LIANG Liming, FENG Yao, LONG Pengwei, et al. Remote sensing image object detection based on MobileViT and multiscale feature aggregation[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(5): 1168-1177.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202310022

DOI: 10.11992/tis.202310022
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240828.1041.022

基于 MobileViT 和多尺度特征聚合的遥感图像目标检测

梁礼明, 冯耀, 龙鹏威, 李仁杰
(江西理工大学 电气工程与自动化学院, 江西 赣州 341000)

摘　要: 针对遥感图像目标检测存在复杂背景干扰、微小目标提取难和目标多尺度差异问题, 提出一种基于 MobileViT 和多尺度特征聚合的遥感图像目标检测算法 (FWM-YOLOv7t)。首先设计多尺度特征聚合模块, 建立遥感目标上下文依赖关系, 提升多尺度目标和小目标检测精度; 然后利用 MobileViT 模块, 融合卷积神经网络和视觉 Transformer 优点, 有效编码局部和全局信息, 抑制非目标噪声干扰; 最后引入 Wise-IoU 损失函数, 重点关注普通质量锚框, 提高算法检测性能。在公共数据集 RSOD 和 NWPU VHR-10 上的实验结果表明, FWM-YOLOv7t 能够显著提升遥感图像目标检测的平均准确率。与其他目标检测算法相比, FWM-YOLOv7t 对复杂背景目标、小目标和多尺度目标的检测更有效。

关键词: 深度学习; 遥感图像; 目标检测; YOLOv7-tiny; MobileViT 模块; 多尺度特征融合; 上下文信息; Wise-IoU
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)05−1168−10

中文引用格式: 梁礼明, 冯耀, 龙鹏威, 等. 基于 MobileViT 和多尺度特征聚合的遥感图像目标检测 [J]. 智能系统学报, 2024, 19(5): 1168–1177.
英文引用格式: LIANG Liming, FENG Yao, LONG Pengwei, et al. Remote sensing image object detection based on MobileViT and multiscale feature aggregation[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1168–1177.

Remote sensing image object detection based on MobileViT and multiscale feature aggregation

LIANG Liming, FENG Yao, LONG Pengwei, LI Renjie
(School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou 341000, China)

Abstract: A new algorithm is proposed based on MobileViT and multi-scale feature aggregation (referred to as FWM-YOLOv7t) to address problems such as complex background interference, difficulty in extracting small objects, and object multi-scale differences in remote sensing image object detection. First, we design a multi-scale feature aggregation module to establish context dependencies for remote sensing targets, which improves the accuracy of detecting multi-scale and small targets. Then, we utilize the MobileViT module to fuse the advantages of convolutional neural networks and vision transformers for effective local and global information encoding to suppress non-target noise interference. Finally, we introduce the Wise-IoU loss function, which focuses on ordinary quality anchor boxes to enhance the detection performance of the algorithm. Experimental evaluations on the public RSOD and NWPU VHR-10 dataset demonstrate that FWM-YOLOv7t can significantly improve the average accuracy of remote sensing image target detection. Furthermore, compared with other object detection algorithms, the FWM-YOLOv7t algorithm exhibits superior effectiveness in detecting complex, small, and multiscale objects in remote sensing imagery.

Keywords: deep learning; remote sensing image; object detection; YOLOv7-tiny; MobileViT module; multi-scale feature fusion; contextual information; Wise-IoU

收稿日期: 2023−10−17.　**网络出版日期**: 2024−08−28.
基金项目: 国家自然科学基金项目 (51365017, 61463018); 江西省自然科学基金面上项目 (20192BAB205084); 江西省教育厅科学技术研究重点项目 (GJJ170491).
通信作者: 梁礼明. E-mail: lianglm67@163.com.

遥感图像的目标检测是一项重要的计算机视觉技术, 旨在对图像中的物体进行识别和定位, 在环境监测、土地规划和军事侦察等领域有着广

阔的应用前景[1]。由于遥感图像场景变化剧烈、背景杂乱、物体尺度多变和环境因素等特点会导致各种不可预测的干扰, 目标检测难度较大。因此, 亟待设计一种精准且高效的遥感图像目标检测算法[2-3]。

传统的目标检测算法常采用滑动窗口进行区域选择, 进而构建目标的特征表征, 构建方法有尺度不变特征变换[4] 和定向梯度直方图[5] 等, 然后使用分类器完成目标的检测。虽然传统目标检测算法理论较为完善, 但其面对复杂场景仍有不足之处: 一是滑动窗口区域选择缺乏针对性, 造成窗口冗余严重, 算法效率低下。二是基于手工设计的特征难以适应多样化的任务场景。随着深度学习快速发展, 卷积神经网络 (convolutional neural network, CNN) 被广泛应用于计算机视觉等领域[6-7]。当前, 以 CNN 为基础的目标检测算法分为一阶段算法和二阶段算法两大类。二阶段算法包括 R-CNN(region-based convolutional neural network)[8]、Fast R-CNN[9] 和 Faster R-CNN[10] 等, 对目标检测精度高但速度慢。针对其缺点, Liu 等学者提出 SSD(single shot multibox detector)[11]、YOLO (you only look once)[12-14] 系列等一阶段算法, 删除生成候选区域阶段, 直接完成目标位置和类别预测, 提高检测实时性。吴萌萌等[15] 在 YOLOv5 算法基础上设计自适应双向特征金字塔网络, 引入可学习的特征融合因子动态调整各阶段特征图的融合权重, 以增强模型的特征表示能力。但改进后的算法损失函数没有充分考虑数据集目标尺寸的长尾分布特点, 导致模型对不同尺寸目标的检测能力存在差异, 特别是对大目标的检测精度提高不明显。文献 [16] 改进 YOLOX 网络, 引入自适应空间特征融合结构应对不同尺度目标检测问题, 设计高效通道注意力模块, 抑制无关背景信息, 该网络虽然取得较好的检测效果, 但仍存在微小目标特征信息丢失严重等问题。Wang 等[17] 构建 YOLOX_w 算法, 使用切片辅助超推理 (slicing aided hyper inference, SAHI)[18] 对图像进行预处理以及数据增强, 将包含丰富空间信息的浅层特征图引入特征融合结构, 有效地提高检测小型物体的能力, 但也导致模型实时性降低。上述算法虽然有不错的检测表现, 但仍未较好地解决检测中复杂背景信息干扰、小尺寸目标检测难和目标尺度多变问题。

卷积操作的局部性限制了 CNN 获取全局上下文信息的能力。相反, Transformer 可以全局关注不同区域图像特征之间的关系, 并通过自注意力机制来提取充分的特征信息[19]。因此, 研究人员尝试将 Transformer 应用于遥感图像目标检测。Zhu 等[20] 在 YOLOv5 算法基础上将原卷积预测头替换为 Transformer 预测头, 集成卷积块注意力模块 (convolutional block attention module, CBAM) 以提高网络检测能力。Sahin 等[21] 增加检测层数量并在颈部网络头部引入视觉 Transformer(vision Transformer, ViT)[22] 解决小目标检测难的问题。上述模型在目标检测领域虽然有不错表现和贡献, 但计算复杂度过高, 网络实时性较差。因此, 本文受此启发, 引入轻量化 MobileViT 模块[23] 嵌入算法主干网络底部, 结合 CNN 和 ViT 的优点, 捕获图像全局信息, 减少计算复杂度。

综上所述, 针对遥感图像检测中存在的问题, 本文提出一种基于 MobileViT 和多尺度特征聚合的遥感图像目标检测算法 (FWM-YOLOv7t)。首先设计多尺度特征聚合模块替换原模型中模块, 以更好地学习不同尺度特征信息, 增加网络对目标尺度多变的鲁棒性和微小目标的检测精度。然后为解决图像中复杂背景信息对检测的干扰, 引入轻量化 MobileViT 模块, 提高模型全局信息感知力的同时, 抑制背景信息噪声。最后使用 Wise-IoU 损失函数以提高算法整体检测性能。

1　YOLOv7-tiny 算法

2022 年 Wang 等[14] 提出 YOLOv7 算法, 其检测速度和精度超越以往 YOLO 系列算法。根据应用场景与需求的不同, YOLOv7 算法变体为 tiny、x、d6、e6、w6 等版本。其中, YOLOv7-tiny 算法在 YOLOv7 基础上进行精简, 相比于其他变体具有参数量低、检测速度快、硬件兼容性高等特点。鉴于遥感图像目标检测对资源占用、实时性能以及检测精度等多方面要求, 本文选用轻量级算法 YOLOv7-tiny 作为基线模型, 并对其进行改进。YOLOv7-tiny 算法主要由输入端 (input)、主干网络 (backbone)、颈部网络 (neck) 和头部网络 (head) 4 部分组成, 算法结构如图 1 所示。

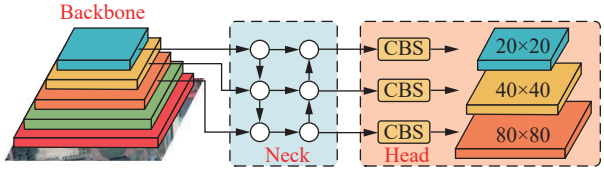

图 1　YOLOv7-tiny 网络结构

Fig. 1　YOLOv7-tiny network structure

首先输入端通过马赛克、混合数据增强技术和自适应锚框计算等方法对输入图像进行预处理。然后主干网络采用若干卷积计算单元、高效

层聚合网络 (efficient layer aggregation networks, ELAN) 和最大池化层交替组合提取图像特征信息。接着颈部网络利用路径聚合网络结构对主干提取的多尺度特征信息进行融合以提高检测精度和鲁棒性。最后头部网络对图像特征层预设不同大小的先验框，根据先验框中是否包含目标及其类别进行评分，筛选出满足置信度要求的预测框，并通过非极大抑制得到最终的预测框，输出目标检测结果。

2　FWM-YOLOv7t 算法

2.1　算法设计

针对遥感图像中目标尺度变化大、微小目标提取难和非目标噪声干扰等问题，本文以 YOLOv7-tiny 为基础模型，提出基于 MobileViT 和多尺度特征聚合算法 (FWM-YOLOv7t)，算法结构如图 2 所示。首先设计多尺度特征聚合模块 (multi-scale feature aggregation module, MFAM) 替换 ELAN，获取不同尺度显著性特征信息，增强对目标尺度变化的鲁棒性，提高算法对微小目标的敏感度。其次主干网络网络中引入融合 CNN 和 Transformer 优点的轻量化模型 MobileViT 替代尾部 MFAM，抑制复杂背景信息干扰，扩展模型专注于不同位置能力，提升全局表达能力。最后通过 Wise-IoU 损失函数重点专注普通质量锚框，提高网络检测性能。

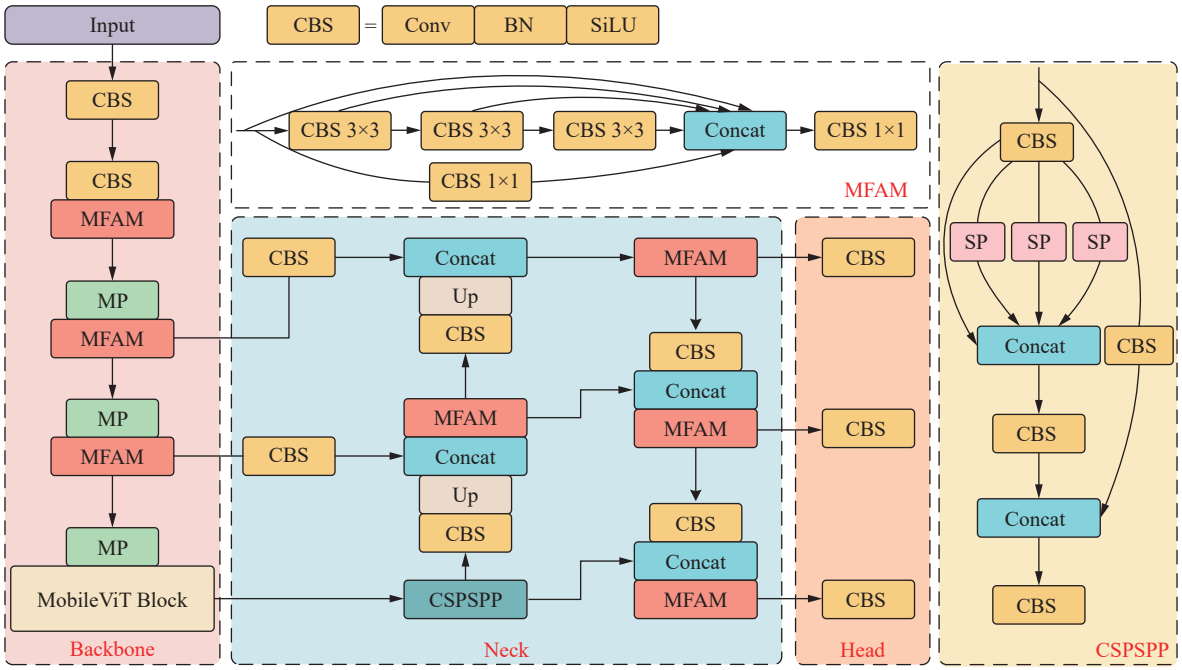

图 2　FWM-YOLOv7t 网络结构

Fig. 2　FWM-YOLOv7t network structure

2.2　多尺度特征聚合模块

在计算机视觉任务中，利用不同尺度卷积核对图片进行特征捕获来获得多尺度信息具有良好效果，针对不同尺度特征之间上下文信息易丢失、遥感图像小目标特征提取难的问题，本文设计一种多尺度特征聚合模块 (MFAM)，其结构如图 3 所示。

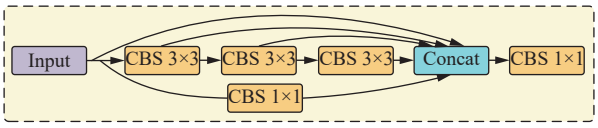

图 3　多尺度特征聚合模块

Fig. 3　Multi-scale feature aggregation module

MFAM 通过不同尺寸的卷积核进行显著性特征信息提取。其中，为降低计算成本，有效提取遥感目标特征，利用多个连续的3×3卷积替代5×5卷积和7×7卷积。同时，针对网络层数堆叠易出现网络退化和梯度爆炸等问题，引入跳跃连接，输出初始特征图，实现早期特征重用。输入遥感图像经 5 条并行支路进行特征提取后，可捕获多尺度特征信息。再通过 Concat 和1×1卷积操作对不同支路所包含的特征信息进行交互以实现跨支路的信息聚合，可进一步加强网络特征学习能力。其具体表达式为

$$C_{i\times i}(\boldsymbol{P}_{\text{in}})=\text{Cat}[C_{1\times1}(\boldsymbol{P}_{\text{in}}),C_{3\times3}(\boldsymbol{P}_{\text{in}}),$$
$$C_{3\times3}(C_{3\times3}(\boldsymbol{P}_{\text{in}})),C_{3\times3}(C_{3\times3}(C_{3\times3}(\boldsymbol{P}_{\text{in}})))]$$
$$\boldsymbol{M}_{\text{out}}=C_{1\times1}(\text{Cat}[\boldsymbol{P}_{\text{in}},C_{i\times i}(\boldsymbol{P}_{\text{in}})])$$

式中：$\boldsymbol{P}_{\text{in}}$表示多尺度特征聚合模块的输入，$C_{3\times3}(\cdot)$表示卷积核尺寸3×3的卷积操作，$\boldsymbol{M}_{\text{out}}$表示多尺度特征聚合模块的输出。

2.3　MobileViT 模块

在遥感目标检测任务中, 能否提取到更多的准确目标特征信息对检测结果非常重要。CNN 使用固定尺寸的卷积核对图像进行局部特征建模, 但随着网络层数增加, 容易出现网络退化和梯度爆炸等问题。ViT 通过 Self-Attention 编码图像全局特征, 但具有较高的计算复杂性, 导致模型实时性较差。

为捕获更多遥感图像目标像素, 抑制背景噪声干扰, 引入 MobileViT(MViT) 模块[23], 如图 4 所示。MViT 模块结合 CNN 和 ViT 的优点, 同时使用卷积和自注意力机制, 可以有效地将局部和全局信息进行编码, 聚合和处理图像中特征信息, 扩展模型专注于不同位置能力, 具有全局感受野, 提高目标检测能力。MViT 模块由局部特征提取块、全局特征提取块和特征融合 3 部分组成。首先输入特征块 $\boldsymbol{X}_{\mathrm{I}} \in \mathbf{R}^{H\times W\times C}$ 通过局部特征提取块进行特征建模, 内部经过尺寸为 $n\times n$ 的卷积核进行局部建模, 再通过 1×1 大小的卷积核将输入特征投影到高维空间得到特征块 $\boldsymbol{X}_{\mathrm{L}} \in \mathbf{R}^{H\times W\times d}$。在全局特征提取块部分, 为学习具有空间归纳偏置, 将局部特征块 $\boldsymbol{X}_{\mathrm{L}} \in \mathbf{R}^{H\times W\times d}$ 展开为 N 个序列块 $\boldsymbol{X}_{\mathrm{U}} \in \mathbf{R}^{P\times N\times d}$ 输入到 L 组 Transformer 中编码图像全局特征, 具体公式为

$$\boldsymbol{X}_{\mathrm{G}}(p) = L\times \mathrm{Trans}(\boldsymbol{X}_{\mathrm{U}}(p)), p\in[1,P]$$

式中: $\boldsymbol{X}_{\mathrm{G}}$ 表示全局特征块, p 表示每个序列块中位置为 p 的像素信息, L 设置为 3, 详见 3.4 节实验分析。然后将输出的全局特征块 $\boldsymbol{X}_{\mathrm{G}} \in \mathbf{R}^{P\times N\times d}$ 折叠还原得到 $\boldsymbol{X}_{\mathrm{F}} \in \mathbf{R}^{H\times W\times d}$。最后通过逐点卷积对通道进行投影还原, 经通道拼接操作建立输入特征块和全局特征块的上下文依赖关系, 最终通过 $n\times n$ 大小卷积核融合全局和局部特征实现对遥感目标特征的提取, 其数学表达式为

$$\boldsymbol{X}_{\mathrm{O}} = C_{n\times n}(\mathrm{Cat}(\boldsymbol{X}_{\mathrm{I}}, C_{1\times 1}(\boldsymbol{X}_{\mathrm{F}})))$$

式中: $\boldsymbol{X}_{\mathrm{O}}$ 表示输出融合特征块, $C_{1\times 1}(\cdot)$ 表示卷积核尺寸 1×1 的卷积操作, $\mathrm{Cat}(\cdot)$ 表示特征块通道维度进行拼接操作, $C_{n\times n}(\cdot)$ 表示卷积核尺寸 $n\times n$ 的卷积操作。

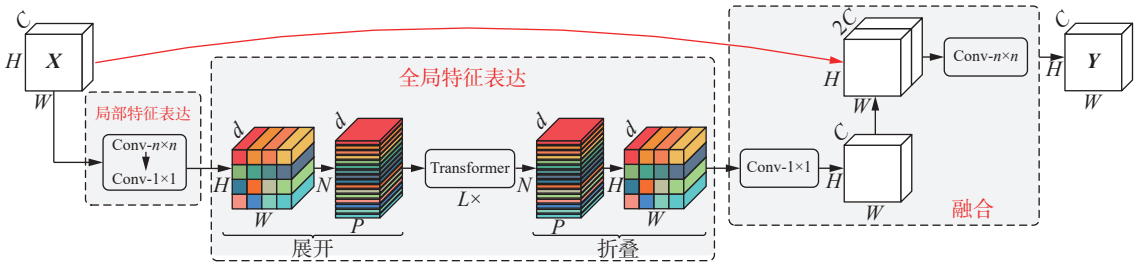

图 4　MobileViT 模块

Fig. 4　MobileViT module

图 4 中的 Transformer 由多头自注意力模块和全连接前馈网络 2 部分构成, 每一部分均进行归一化处理, 并采用残差结构, 形成跳跃连接, 提升模型检测效果。多头自注意力机制的运算公式为

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}\Big/\sqrt{d_k}\right)\boldsymbol{V}$$

$$H_i = A\left(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}\right), i = 1,2,\cdots,h$$

$$M_{\mathrm{MultiHead}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Cat}(H_1,H_2,\cdots,H_h)\boldsymbol{W}^{\mathrm{O}}$$

式中: $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询、键和值向量, d_k 表示缩放因子, $A(\cdot)$ 表示进行自注意力运算, $\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$ 和 $\boldsymbol{W}_i^{\boldsymbol{V}}$ 表示线性变换的权重矩阵, $\boldsymbol{W}^{\mathrm{O}}$ 表示输出的权重矩阵, $M_{\mathrm{MultiHead}}(\cdot)$ 表示进行多头自注意力运算。

2.4　Wise-IOU 损失函数

定位损失函数在目标检测中起到关键作用, 其合理的定义将为网络带来显著的性能提升。YOLOv7-tiny 网络采用 Complete-IoU(CIoU)[24] 计算定位损失, CIoU 考虑了重叠面积、归一化中心点距离和纵横比 3 个几何因素, 但训练数据中不可避免的会包含低质量示例, 归一化中心距离和纵横比等几何因素会加重对低质量示例的惩罚, 从而降低模型的泛化能力。

为使预测框回归过程聚焦普通质量锚框, 增强模型泛化能力, 本文引入 Wise-IoU 损失函数 (WIoU)[25]。WIoU 使用离散度来评估锚框质量, 并根据离散度动态分配梯度增益, 降低高质量锚框的竞争力和低质量示例产生的有害梯度, 重点关注普通质量锚框, 以提高网络的检测性能。WIoU 具体表达式为

$$\mathcal{L}_{\mathrm{WIoU}} = \frac{\beta}{\delta\alpha^{\beta-\delta}}\exp\left(\frac{(x-x_{\mathrm{gt}})^2+(y-y_{\mathrm{gt}})^2}{\left(W_{\mathrm{g}}^2+H_{\mathrm{g}}^2\right)^*}\right)(1-I_{\mathrm{oU}})$$

$$\beta = \mathcal{L}_{\mathrm{IoU}}^*\Big/\overline{\mathcal{L}}_{\mathrm{IoU}} \in [0,+\infty)$$

式中: $\mathcal{L}_{\mathrm{WIoU}}$ 表示定位损失函数值, α 和 δ 表示超参数, β 表示离群度, x 和 y 表示预测框的中心坐标, x_{gt} 和 y_{gt} 表示真实框的中心坐标, W_{g} 和 H_{g} 分别表示包围真实框和预测框最小矩形的宽高, $*$ 表示将

W_g和H_g从梯度计算中分离出来，I_{oU}表示预测框和真实框的交并比，$\mathcal{L}^{*}_{IoU}$表示单调聚焦系数，$\overline{\mathcal{L}}_{IoU}$表示动量为 m 的滑动平均值。m 的数学表达为

$$m = 1 - \sqrt[n]{0.5}$$

式中：t 为平均准确率均值提升明显减慢的轮次，n 为批次数量，$t \times n$为 7 000。

3　实验

3.1　实验环境与参数设定

实验基于 PyTorch2.0.0 框架和 Python3.8.16 语言进行编程实现，处理器为 Intel(R) Core(TM) i9-13900H @2.60 GHz、显卡为 NVIDIA GeForce RTX 4060、显存为 8 GB。实验设置输入图像分辨率为 640 像素×640 像素，初始学习率为 0.01，动量参数为 0.937，权重衰减系数为 0.000 5，训练轮次为 300，批量大小为 16，WIoU 损失函数超参数α和δ通过迭代最优解分别设为 1.9 和 3。

3.2　实验数据集

实验使用武汉大学标注的数据集 RSOD[26] 进行训练验证，该数据集涵盖 976 张遥感图像，覆盖飞机、油桶、操场、立交桥 4 种目标，共 6 950 个标注样本信息。为验证模型的泛化性，使用西北工业大学标注的数据集 NWPU VHR-10[27] 进行实验，该数据集包含目标的图像共 650 张，涵盖飞机、油罐、船只、田径场、网球场、棒球场、篮球场、桥梁、港口和车辆 10 类目标，共 3 651 个标注样本信息。2 个数据集均采用 8∶2 比例随机划分训练集和验证集。

3.3　评价指标

为实现对遥感目标检测结果的量化评价和分析，本文用准确率 (precision, P)、召回率 (recall, R)、平均准确率 (average precision, AP)、平均准确率均值 (mean average precision, mAP) 和检测速度 (每秒处理图像数量) 等评价指标进行比较分析。AP 和 mAP 分别表示模型在单类和多类目标检测的精确度，mAP@0.5 表示交并比 (intersection over union, IoU) 阈值为 0.5 时的平均准确率均值，其计算式分别为

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

$$A_{AP} = \int_0^1 P(R)\mathrm{d}R$$

$$M_{mAP} = \frac{1}{N}\sum_{i=1}^{N} A^i_{AP}$$

式中：A_{AP}表示平均准确率，M_{mAP}表示平均准确率均值，N_{TP}表示正类样本被检测为正类的数量，即模型检测成功的遥感目标；N_{FN}表示正类样本被检测为负类的数量，即模型漏检的遥感目标；N_{FP}表示负类样本被检测为正类的数量，即模型误检的遥感目标；N 表示数据集中目标类别数。

3.4　消融实验

为探究 MobileViT 模块中 Transformer 层数对实验结果的影响，在 RSOD 数据集上对重叠因子 L 进行对比实验，检测结果如表 1 所示，其中加粗表示该项的最佳指标。M0 表示 YOLOv7-tiny 引入 MFAM，并以此为对比基准；M1(L=k) 表示在 M0 基础上加入包含 k 层 Transformer 的 MobileViT 模块。观察表 1 可知，随着 Transformer 层数的增加，参数量也随之增加，但平均准确率均值并非与之呈正相关。因此，在综合考虑参数量和平均准确率均值的平衡关系，本文选用 L=3 时的 MobileViT 模块作为最优方案。

表 1　重叠因子 L 对比分析

Table 1　Comparative analysis of overlap factor L

模型	准确率/ %	召回率/ %	mAP@0.5/%	参数量/10^6
M0	94.6	94.3	97.0	7.9
M1(L=1)	95.5	94.5	97.0	**7.6**
M1(L=2)	96.2	94.8	97.6	8.1
M1(L=3)	**96.6**	**95.1**	**97.8**	8.6
M1(L=4)	95.5	92.3	97.4	9.1
M1(L=5)	93.1	93.5	96.5	9.6

为更好地检验本文提出模型的有效性，使用 RSOD 数据集进行消融实验对比，检测得到的准确率、召回率、平均准确率和 mAP@0.5，如表 2 所示，其中加粗表示此项的最优指标。A1 表示 YOLOv7-tiny，A2 表示设计 MFAM 替换 A1 模型中 ELAN，A3 表示引入 MViT 模块代替 A2 模型主干网络尾部的 MFAM，A4 表示在 A3 中引入 WIoU 损失函数，构成本文模型。

表 2　消融实验数据

Table 2　Ablation experimental data　%

模型	准确率	召回率	平均准确率				mAP@0.5
			飞机	油桶	立交桥	操场	
A1	90.9	94.1	98.7	97.2	89.2	98.0	95.8
A2	94.6	94.3	98.9	**98.3**	92.3	98.6	97.0
A3	**96.6**	95.1	98.7	97.5	95.6	99.2	97.8
A4	96.0	**95.3**	**98.9**	97.5	**97.0**	**99.2**	**98.1**

由表 2 可知，YOLOv7-tiny 模型对遥感目标具有良好的检测效果，初始模型检测飞机和操场遥

感目标已经达到 98.7% 和 98.0% 的平均准确率, 但其准确率、召回率在模型结构固定的情况下仍然较低, 检测复杂背景下立交桥遥感目标的平均准确率仍有提升空间, 于是本文对初始模型进行改进, 以提高模型性能。通过设计 MFAM 替换 A1 内部 ELAN, 各项指标均有提升, 其中准确率提升 3.7 百分点, 检测飞机和油桶遥感目标的平均准确率达到该项最优。说明 MFAM 充分融合多尺度特征信息, 减少信息丢失, 提升飞机小目标和多尺度检测能力。在 A2 基础上引入 MViT 模块, 准确率、召回率和检测复杂背景下立交桥遥感目标的平均准确率有所提升, 其中准确率达到最优值, 检测立交桥的平均准确率提升 3.3 百分点, 说明 MViT 模块可以有效地捕捉复杂背景下目标, 具有全局敏感性。最后引入 WIoU 损失函数, 召回率、平均准确率均值达到各项最优, 说明模型预测框回归过程聚焦普通质量锚框可提高模型整体性能。

综上所述, 本文提出的模型与基线模型相比准确率、召回率和平均准确率均值均有显著提升, 分别提升 5.1 百分点、1.2 百分点和 2.3 百分点, 证明本文提出模型的有效性和合理性。

3.5　多种模型检测结果与分析

3.5.1　定量分析

为客观评价本文提出模型的优势, 将其他目标检测算法同本文模型 FWM-YOLOv7t 应用于 RSOD 数据集, 实验得到的参数量、检测速度、平均准确率和 mAP@0.5 如表 3 所示, 其中加粗表示此项的最优指标。

表 3　不同算法检测数据对比

Table 3　Comparison of detection data from different algorithms

模型	参数量/M	检测速度/(f/s)	平均准确率/%				mAP@0.5/%
			飞机	油桶	立交桥	操场	
Faster R-CNN[10]	72.0	7	77.4	97.9	94.5	**100.0**	92.5
SSD[11]	24.4	41	68.6	96.5	90.2	99.8	88.8
YOLOv3[12]	61.5	24	96.1	97.8	91.5	95.9	95.3
YOLOv4-tiny[13]	6.1	50	70.7	97.3	61.7	99.1	82.4
YOLOv5l	46.6	29	96.4	96.5	89.5	97.8	95.0
YOLOv5s	7.0	**85**	95.2	96.2	84.8	97.1	93.3
YOLODrone+[21]	153.1	22	97.7	**99.2**	84.8	99.5	95.3
YOLOv7[14]	36.5	27	**99.2**	97.4	94.6	99.5	97.7
YOLOv7-tiny[14]	**6.0**	79	98.7	97.2	89.2	98.0	95.8
FWM-YOLOv7t	8.6	77	98.9	97.5	**97.0**	99.2	**98.1**

观察表 3 可知, 本文提出模型相较于 Faster R-CNN、SSD、YOLOv3、YOLOv5l、YOLOv7 算法, 在参数量、检测速度和 mAP@0.5 这 3 个指标上都具有明显优势。与 YOLOv4-tiny 算法相比, 虽然本文算法参数规模上略有扩增, 但检测速度和 mAP@0.5 显著提高, 其中检测速度增长 27 f/s, mAP@0.5 提升 15.7 百分点。同 YOLOv5s、YOLOv7-tiny 算法比较, 在牺牲较少参数量的情况下, 保持检测速度基本不变, mAP@0.5 指标分别增长 4.8 百分点、2.3 百分点。

YOLODrone+基于 YOLOv5s 算法, 在其颈部网络头部引入 ViT 模块, 同时增加检测头数量。文献 [21] 在 VisDrone 数据集中实验证明 YOLODrone+相比基线算法 YOLOv5s 可有效提升小目标检测精度。为检验该模型的有效性, 本文在 RSOD 数据集上进行实验, 发现各类别准确率 AP 均有提高, 但模型参数量过大, 检测速度明显降低。本文提出模型在参数量、检测速度和 mAP@0.5 指标上都优于 YOLODrone+算法, 其中 mAP@0.5 增长 2.8 百分点。

在各个类别的检测结果中, 本文提出的模型对于具有复杂背景的立交桥遥感目标取得了比其他算法更优的结果, 优于基线 YOLOv7-tiny 模型 7.8 百分点。对于包含多尺度和微小目标的飞机类别, 平均准确率值达到 98.9%, 虽略低于 YOLOv7 算法, 但优于其他算法, 和基线 YOLOv7-tiny 模型比较有小幅提升。

3.5.2　定性分析

为直观展现本文提出模型的检测效果, 将本文模型 FWM-YOLOv7t 同其他目标检测算法在遥感图片上进行检测, 检测结果如图 5 所示。其中, 第 1、4 行表示背景复杂遥感图像, 第 2、5 行表示微小目标遥感图像, 第 3、6 行表示多尺度目标遥感图像。图 5(a)~(j) 分别表示原图、Faster R-CNN

算法、SSD 算法、YOLOv3 算法、YOLOv4-tiny 算法、YOLOv5s 算法、YOLODrone+算法、YOLOv7 算法、YOLOv7-tiny 算法和 FWM-YOLOv7t 算法的检测结果图。

(a) 原图　(b) Faster R-CNN　(c) SSD　(d) YOLOv3　(e) YOLOv4-tiny

(f) YOLOv5s　(g) YOLODrone+　(h) YOLOv7　(i) YOLOv7-tiny　(j) FWM-YOLOv7t

图 5　不同算法遥感目标检测结果

Fig. 5　Remote sensing target detection results of different algorithms

由图 5 分析可知，在第 1、4 行复杂背景下遥感图像目标检测结果中，SSD 算法、YOLOv3 算法、YOLOv4-tiny 算法和 YOLOv5s 算法存在目标漏检现象；YOLODrone+算法出现误检问题；Faster R-CNN 算法、YOLOv7 算法、YOLOv7-tiny 算法和 FWM-YOLOv7t 算法均能有效检测出立交桥目标，其中 FWM-YOLOv7t 算法和 Faster R-CNN 算法在检测立交桥目标时表现出最高的置信度。

在第 2、5 行微小目标检测结果中，Faster R-CNN 算法、YOLOv3 算法和 YOLODrone+算法有较多的重叠框问题；SSD 算法和 YOLOv4-tiny 算法存在着目标漏检状况，其中 SSD 算法对小目标特征信息丢失严重，导致漏检率较高；YOLOv5s 算法和 YOLOv7-tiny 算法出现对无关物体误检现象；而 FWM-YOLOv7t 算法和 YOLOv7 算法能够有效地捕获微小目标特征信息，实现目标的精准检测。

在第 3、6 行多尺度目标检测结果中，Faster R-CNN 算法、SSD 算法、YOLOv4-tiny 算法和 YOLOv7-tiny 算法表现不佳，均出现目标漏检现象；YOLODrone+算法能够较好地检测出多尺度目标，但存在着重叠框问题；YOLOv5s 算法对不相关物体产

生误检; 而 FWM-YOLOv7t 算法、YOLOv3 算法和 YOLOv7 算法准确无误地检测出所有目标, 说明 3 种算法对目标尺度变化更具鲁棒性。

综合分析, 本文提出模型 FWM-YOLOv7t 相较于其他目标检测算法不仅有良好的检测精度, 且具有较低参数量和较快检测速度。此外, 在面对复杂背景、微小目标以及多尺度目标时表现出更卓越的检测能力, 其综合检测结果优于其他对比算法。

3.6　泛化性验证

为进一步验证本文提出模型的泛化能力, 将模型 YOLOv7、YOLOv7-tiny 与本文模型 FWM-YOLOv7t 在 NWPU VHR-10 数据集上进行实验对比, 实验结果如表 4 所示, 其中加粗表示此项的最优指标。

表 4　NWPU VHR-10 数据集上检测结果对比

Table 4　Comparison of detection results on NWPU VHR-10 dataset

模型	准确率/%	召回率/%	参数量/10^6	检测速度/(f/s)	mAP@0.5/%
YOLOv7[14]	91.5	**85.1**	36.5	34	**90.7**
YOLOv7-tiny[14]	92.9	83.3	**6.0**	85	88.9
FWM-YOLOv7t	**95.1**	83.0	8.6	**91**	90.6

通过对表 4 分析可知, 相比于 YOLOv7 模型, FWM-YOLOv7t 在参数量较小的情况下, 尽管召回率略有下降, 但其准确率增长 3.6 百分点, 且 mAP@0.5 值基本持平。FWM-YOLOv7t 相较于基准算法 YOLOv7-tiny 虽参数量略有增加, 召回率基本保持不变, 但准确率提升 2.2 百分点, mAP@0.5 值提升 1.7 百分点。此外, 本文模型检测速度达到 91 f/s, 具有较强的实时性能, 能够适应实际应用场景的需求。

本文模型和 YOLOv7-tiny 算法对不同目标类别检测的平均准确率如图 6 所示。从图 6 可以看出, 在对涵盖复杂背景环境的目标类别, 例如桥梁、港口以及网球场进行检测时, 本文模型取得比 YOLOv7-tiny 算法更显著的结果, 分别实现 4.8 百分点、5.5 百分点和 1.1 百分点的提升。对于包含多尺度目标和微小目标的船只类别, 其检测结果也优于 YOLOv7-tiny 算法 1.5 百分点。

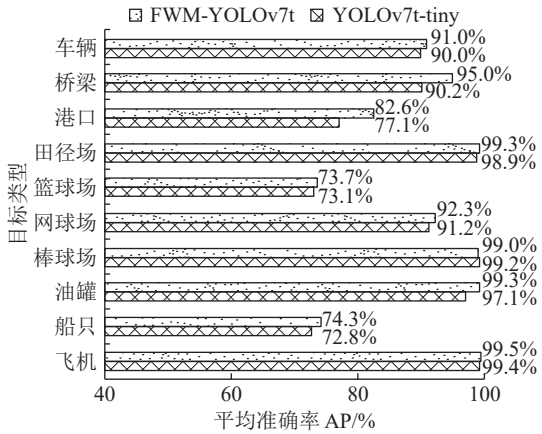

图 6　不同目标类别检测平均准确率结果

Fig. 6　Average accuracy results for detection of different target categories

实验结果表明, FWM-YOLOv7t 不仅在 RSOD 数据集中展现出优异的目标检测表现, 在其他类别的数据集上, 也表现出显著的性能提高, 其泛化性能优异。

4　结束语

针对遥感图像目标检测时存在背景信息干扰、小尺寸目标提取难和目标尺度多变问题, 本文提出 FWM-YOLOv7t 模型。首先使用多尺度特征聚合模块替代网络中 ELAN, 提取图像不同尺度的特征信息, 搭建目标上下文关系, 提高对多尺度目标和微小目标的检测精度。然后在主干网络网络中加入 MobileViT 模块替换尾部 MFAM, 实现对非目标信息的抑制, 关注遥感目标的特征信息。最后引入 Wise-IoU 损失函数以提高模型整体性能。通过 RSOD 数据集实验, FWM-YOLOv7t 相较于基线算法 YOLOv7-tiny, mAP@0.5 提升 2.3 百分点, 说明 FWM-YOLOv7t 对提高遥感目标整体检测效果的有效性; 通过 NWPU VHR-10 数据集实验, mAP@0.5 提升 1.7 百分点, 表明 FWM-YOLOv7t 的泛化能力强; 在包含复杂背景的桥梁、港口、网球场目标类别和具有多尺度目标和微小目标的船只类别上, 平均准确率均显著提高, 得出 FWM-YOLOv7t 可以有效应对遥感图像目标检测问题。但本文提出模型在轻量化方面仍有改进空间。保持模型高速度的同时, 实现高精度和轻量化的平衡, 将是今后的探索方向, 从而更好满足工业场景的需求。

参考文献:

[1]　赵文清, 康怿瑾, 赵振兵, 等. 改进 YOLOv5s 的遥感图像目标检测 [J]. 智能系统学报, 2023, 18(1): 86−95.
ZHAO Wenqing, KANG Yijin, ZHAO Zhenbing, et al. A remote sensing image object detection algorithm with im-

proved YOLOv5s[J]. CAAI transactions on intelligent systems, 2023, 18(1): 86−95.

[2] MING Qi, MIAO Lingjuan, ZHOU Zhiqiang, et al. CFC-net: a critical feature capturing network for arbitrary-oriented object detection in remote-sensing images[J]. IEEE transactions on geoscience and remote sensing, 2022, 60: 5605814.

[3] CONG Runmin, ZHANG Yumo, FANG Leyuan, et al. RRNet: relational reasoning network with parallel multiscale attention for salient object detection in optical remote sensing images[J]. IEEE transactions on geoscience and remote sensing, 2021, 60: 5613311.

[4] SEDAGHAT A, EBADI H. Remote sensing image matching based on adaptive binning SIFT descriptor[J]. IEEE transactions on geoscience and remote sensing, 2015, 53(10): 5283−5293.

[5] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005: 886−893.

[6] 吴珺, 董佳明, 刘欣, 等. 注意力优化的轻量目标检测网络及应用 [J]. 智能系统学报, 2023, 18(3): 506−516.
WU Jun, DONG Jiaming, LIU Xin, et al. Lightweight object detection network and its application based on the attention optimization[J]. CAAI transactions on intelligent systems, 2023, 18(3): 506−516.

[7] 梁礼明, 詹涛, 雷坤, 等. 多分辨率融合输入的 U 型视网膜血管分割算法 [J]. 电子与信息学报, 2023, 45(5): 1795−1806.
LIANG Liming, ZHAN Tao, LEI Kun, et al. Multi-resolution fusion input U-shaped retinal vessel segmentation algorithm[J]. Journal of electronics & information technology, 2023, 45(5): 1795−1806.

[8] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014: 580−587.

[9] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440−1448.

[10] REN Shaoqing, HE Kaiming, 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.

[11] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21−37.

[12] FARHADI A, REDMON J. YOLOv3: an incremental improvement[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1804−2767.

[13] BOCHKOVSKIY A, WANG C Y, LIAO H M, et al. YOLOv4: optimal speed and accuracy of object detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 2−7.

[14] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 7464−7475.

[15] 吴萌萌, 张泽斌, 宋尧哲, 等. 基于自适应特征增强的小目标检测网络 [J]. 激光与光电子学进展, 2023, 60(6): 65−72.
WU Mengmeng, ZHANG Zebin, SONG Yaozhe, et al. Small-target detection network based on adaptive feature enhancement[J]. Laser & optoelectronics progress, 2023, 60(6): 65−72.

[16] 李美霖, 芮杰, 金飞, 等. 基于改进 YOLOX 的遥感影像目标检测算法 [J]. 吉林大学学报 (地球科学版), 2023, 53(4): 1313−1322.
LI Meilin, RUI Jie, JIN Fei, et al. Remote sensing image target detection algorithm based on improved YOLOX[J]. Journal of Jilin University (earth science edition), 2023, 53(4): 1313−1322.

[17] WANG Xin, HE Ning, HONG Chen, et al. Improved YOLOX-X based UAV aerial photography object detection algorithm[J]. Image and vision computing, 2023, 135: 104697.

[18] AKYON F C, ONUR ALTINUC S, TEMIZEL A. Slicing aided hyper inference and fine-tuning for small object detection[C]//2022 IEEE International Conference on Image Processing. Bordeaux: IEEE, 2022: 966−970.

[19] 梁礼明, 何安军, 朱晨锟, 等. 融合 Transformer 和跨级相位感知的结肠息肉分割方法 [J]. 生物医学工程学杂志, 2023, 40(2): 234−243.
LIANG Liming, HE Anjun, ZHU Chenkun, et al. Colorectal polyp segmentation method based on fusion of transformer and cross-level phase awareness[J]. Journal of biomedical engineering, 2023, 40(2): 234−243.

[20] ZHU Xingkui, LYU Shuchang, WANG Xu, et al. TPH-YOLOv5: improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 2778−2788.

[21] SAHIN O, OZER S. YOLODrone: improved YOLO architecture for object detection in UAV images[C]//2022 30th Signal Processing and Communications Applications Conference. Safranbolu: IEEE, 2022: 1−4.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. NewOrleans: ICLR, 2021: 1−22.

[23] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer [EB/OL]. (2021−10−05)[2023−10−17]. https://arxiv.org/abs/2110.02178.

[24] ZHENG Zhaohui, WANG Ping, LIU Wei, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12993−13000.

[25] TONG Zanjia, CHEN Yuhang, XU Zewei, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[EB/OL]. (2023−01−24)[2023−10−17]. https://arxiv.org/abs/2301.10051.

[26] LONG Yang, GONG Yiping, XIAO Zhifeng, et al. Accurate object localization in remote sensing images based on convolutional neural networks[J]. IEEE transactions on geoscience and remote sensing, 2017, 55(5): 2486−2498.

[27] CHENG Gong, ZHOU Peicheng, HAN Junwei. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J]. IEEE transactions on geoscience and remote sensing, 2016, 54(12): 7405−7415.

作者简介:

梁礼明, 教授, 主要研究方向为机器学习、医学影像和系统建模。获得专利授权 6 项, 发表学术论文 100 余篇, 出版专著 1 部。E-mail: lianglm67@163.com。

冯耀, 硕士研究生, 主要研究方向为深度学习与目标检测。E-mail: fybrave@126.com。

龙鹏威, 硕士研究生, 主要研究方向为机器学习、模式识别与图像处理。E-mail: 2637018663@qq.com。