Computer Science and Application 计算机科学与应用, 2023, 13(5), 1157-1165
Published Online May 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.135113

Hans汉斯

STA-YOLOv7：基于Swin-Transformer改进YOLOv7算法用于道路异常病害检测

张冬梅，徐志洁

北京建筑大学理学院，北京

收稿日期：2023年4月27日；录用日期：2023年5月24日；发布日期：2023年5月31日

摘　要

本文提出了一种基于Swin-Transformer改进的YOLOv7道路异常病害检测方法(STA-YOLOv7)，旨在解决道路异常病害图像分辨率较高以及多尺度目标检测不准确等问题。具体地，该方法在YOLOv7的结构中嵌入了Swin-Transformer中基于滑动窗口的设计的编码器，以捕捉不同尺度下病害的上下文信息与全局依赖关系，充分学习目标的语义特征。此外，我们还引入了AlignOTA损失函数来为模型训练提供更精确的标签分配策略，增强分类与回归的一致性。通过与Swin-Transformer、YOLOv7、TPH-YOLOv5等算法进行比较，STA-YOLOv7能够有效检测不同目标，降低漏检率的同时提高了准确率，适用于不同环境下各种尺度病害的检测，达到了实际复杂未知场景中实时性应用的需求。

关键词

深度学习，YOLOv7，Swin-Transformer，AlignOTA，病害异常检测

STA-YOLOv7: Swin-Transformer-Enabled YOLOv7 for Road Damage Detection

Dongmei Zhang, Zhijie Xu

School of Science, Beijing University of Civil Engineering and Architecture, Beijing

Received: Apr. 27th, 2023; accepted: May 24th, 2023; published: May 31st, 2023

Abstract

We propose an improved YOLOv7 method for road damage detection based on Swin-Transformer (STA-YOLOv7), aiming to address the challenges of high-resolution road damage images and inaccurate multi-scale object detection. Specifically, the Swin-Transformer encoder based on a sliding

文章引用: 张冬梅, 徐志洁. STA-YOLOv7: 基于 Swin-Transformer 改进 YOLOv7 算法用于道路异常病害检测[J]. 计算机科学与应用, 2023, 13(5): 1157-1165. DOI: 10.12677/csa.2023.135113

window design is embedded into the YOLOv7 architecture to capture contextual information and global dependencies of damages at different scales, and fully learn the semantic features of the targets. In addition, the AlignOTA loss function is introduced to provide more precise label assignment strategies for model training, enhancing the consistency of classification and regression. Compared with Swin-Transformer, YOLOv7, TPH-YOLOv5 and other algorithms, STA-YOLOv7 can effectively detect different targets, reduce missed positives while improving accuracy, and is applicable for detecting anomalies of various scales in different environments. It meets the requirements of real-time application in complex and unknown scenarios.

Keywords

Deep Learning, YOLOv7, Swin-Transformer, AlignOTA, Road Damage Detection

1. 引言

随着城镇化的迅速发展，道路在使用中出现各种损害如裂纹、坑槽等异常病害，严重影响道路的使用寿命，甚至威胁城镇道路交通安全。实时准确的检测系统有助于及时进行道路养护，完善交通设施。传统的道路病害检测方法，通常采用支持向量机(SVM) [1]、直方图[2]和小波变换[3]等，但这些方法的泛化性较差。近年来，随着计算机视觉领域图像处理技术与深度学习理论的迅速发展，基于深度学习的道路病害异常检测方法表现出显著的优势[4]，逐渐取代传统的方法。在复杂的道路场景下，如何精准地对路面异常病害进行精准识别与定位成为了该行业亟待解决的问题。

道路病害检测(RDD)即检测道路图像或者视频中的病害目标并进行分类和定位。由于不同病害目标的尺度、外形大小不一，以及受到光照、遮挡等影响，目前道路病害异常检测仍存在许多挑战。和一般的目标检测任务一样，道路病害检测主要包括两个阶段，分别是特征提取和分类回归。根据两个过程是否能端到端实现的标准，目前基于深度学习的目标检测算法主要分为两类，分别是两阶段目标检测和单阶段目标检测[5]。两阶段目标检测算法分为两个阶段，首先在第一阶段对输入图像进行卷积提取目标候选区域，第二阶段中对每个候选区域进行标签分配实现分类和回归，代表算法有 R-CNN [6]、Fast R-CNN [7]和 Faster R-CNN [8]等。尽管两阶段算法精度较高，但处理大量的候选区域特征需要庞大的计算量。而单阶段检测算法则直接进行特征提取并生成候选锚框来端到端地预测感兴趣区域的类别和位置，代表算法有 YOLO 系列[9] [10]、SSD [11]和 RetinaNet [12]等。相比于两阶段检测算法，单阶段检测算法凭借浅层的网络结构具有较高的速度，这在应用场景中具有实时性，但简单的卷积采样存在一定的误检率。基于深度卷积神经网络提取图像的语义特征，受卷积核感受野的限制，缺乏对全局信息的感知，因而特征映射未能获得更加完整的表达。

目前单阶段检测器通过对整张图像划分网格的方式进行采样，由于卷积核感受野的限制，使得目标的特征仅仅依赖于高质量的局部卷积。Transformer [13]是一个基于自注意力机制的深度学习模型，由于多头自注意力机制解决了感受野中长距离的依赖问题，凭借其特有的全局操作建模所有像素之间的关系，充分利用图像的上下文特征，弥补了 CNN 局部操作的缺点，并且可以应用于多种视觉任务中，表现出巨大的发展潜力。CNN 与 Transformer 结合既能够保留细粒度的局部特征，又能够获取全局语义信息，有

效提高模型提取特征的表达能力，提升模型的检测性能。MobileViT [14]将轻量级的卷积添加到 ViT [15]模型结构中可实现高效的图像分类；DeiT [16]使用 Convnet 作为教师网络，基于 token 蒸馏的策略进行特征融合，提高了模型的泛化性，在图像分类任务中表现出色。ACmix [17]探索卷积与自注意力之间强大的潜在关系，从机理融合的角度出发，混合不同类别的样本进行表征学习。

目标检测器在训练中通过标签分配划分正负样本，而正样本的数量则直接关系到模型的精度和鲁棒性。静态标签分配策略虽然简单可行，但固定的锚框数量和大小不适用于多样的目标和场景。动态标签策略从优化标签分配方案的角度，逐渐适用于多尺度目标数据集与复杂场景中。YOLOv7 的检测头基于 SimOTA [18]动态标签分配策略为真实框选择不同数量和质量的正样本，引导主检测头和辅助检测头学习。尽管 SimOTA 动态标签分配策略为真实框分配预设锚框提供诸多方案，但仍然存在标签分配不合理的问题，造成训练中分类与回归损失不平衡。基于交并比 IOU 损失和分类的交叉熵损失受到一定的局限性，因此解决分类与回归损失的不一致有助于模型在正确的分类下进行精准的定位。

本文针对以上问题提出了一种改进 YOLOv7 的道路异常病害检测算法，主要贡献分为以下几个方面：

1) 以 YOLOv7 网络结构为基础，设计了一种结合 Swin-Transformer 自注意力机制的 STA-YOLOv7 结构，充分利用全局语义信息，融合不同尺度的特征，提高目标检测的准确率。

2) 针对多尺度目标分类与回归损失不平衡问题，引入 AlignOTA [19]动态标签分配对齐损失函数，优化标签分配策略，提高分类与回归的一致性。

3) 在 RDDBJ 数据集上的实验结果表明，STA-YOLOv7 算法的性能优于其他方法，mAP 达到 62.3%，比 YOLOv7 提高 3.7 个百分点。

2. 相关原理介绍

2.1. YOLOv7 原理

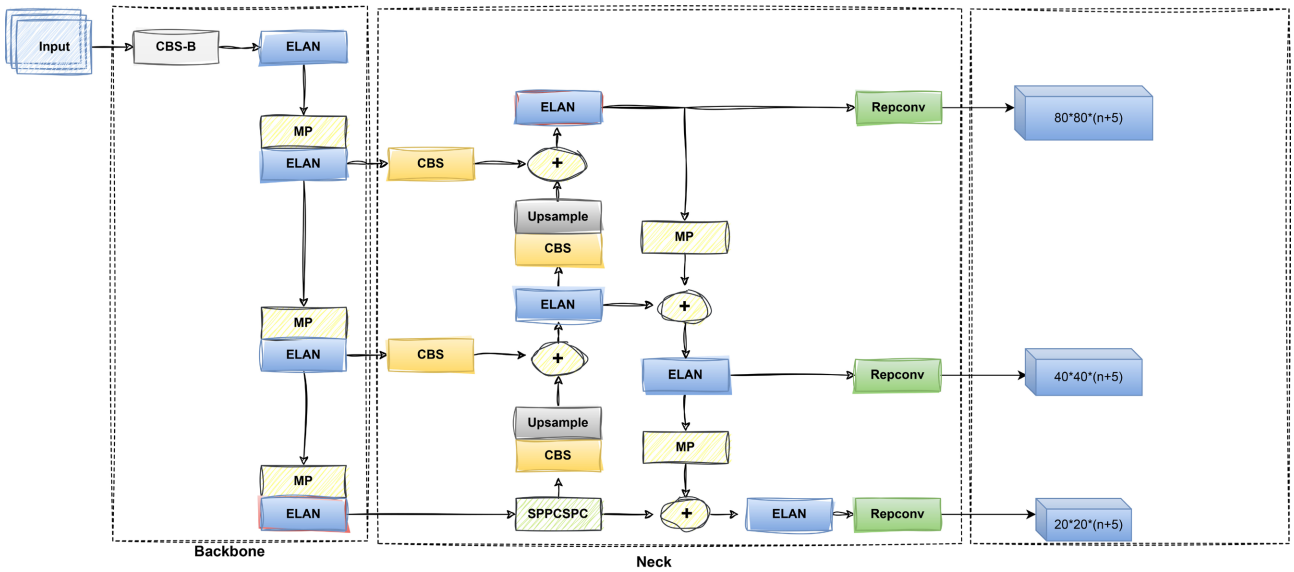

Figure 1. The architecture of YOLOv7 network
图 1. YOLOv7 的网络结构

YOLOv7 网络结构主要分为四部分，如图 1 所示，分别是输入(Input)、骨干网络(Backbone)、颈部网络(Neck)和检测头(IDetect)四部分[20]。输入图像首先经过裁剪、自适应缩放、Mosaic 等一系列数据增强

处理后，被送入到骨干网络中。主干网络在对预处理图像进行特征提取后输出三种不同尺度大小的特征图并送入到颈部网路中。与 YOLOv5 中的颈部网络相同，YOLOv7 的颈部网络主要为 PAFPN 特征金字塔结构，自下而上地将大、中、小三种不同尺寸的特征进行上采样并按通道数融合不同层次的特征。MP 模块与 SPPCSPC 模块中进行多次池化操作以融合不同尺度的特征来避免图像失真问题。经 IDetect 检测头来预测输入图像中目标的类别和位置信息。训练中，YOLOv7 的损失函数包括分类损失、定位损失和置信度损失，其中分类损失与置信度损失函数采用 BCELoss 二值交叉熵损失函数，采用 CIOU 损失函数作为定位损失。同时，YOLOv7 中规划的重参数化卷积 Repconv 在不增加推理成本的基础上优化网络速度。在 5FPS 到 160FPS 范围内，YOLOv7 在速度与精度方面表现优秀[21]，超过了当前主流的目标检测器。

2.2. Swin-Transformer 原理

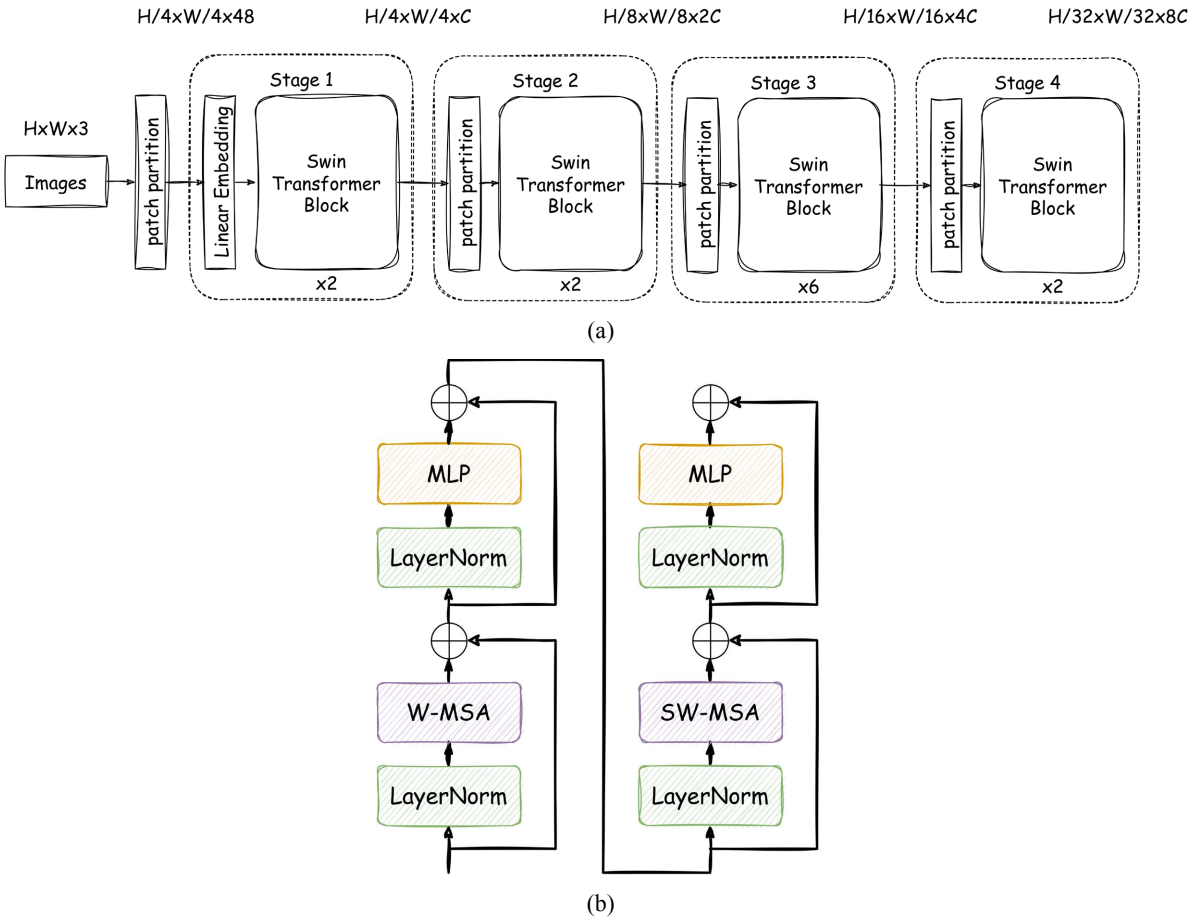

Figure 2. The architecture of Swin-Transformer.
图 2. Swin-Transformer 的网络结构

类似于特征金字塔的层级结构，Swin-Transformer 模型[22]是基于多尺度融合的 Transformer 模型，通过移动不重叠窗口的设计提取不同尺度的特征，同时允许跨窗口连接，实现了局部特征与全局特征的信息交互。如图 2(a)所示，Swin-Transformer 编码器由一个 Patch Partition 模块与四个连续的 stage 组成。Stage 中包含两种注意力模块，分别是窗口多头自注意机制(Windows Multi-Head Self-Attention, W-MSA)模块和滑动窗口多头自注意力机制(Shifted Windows Multi-Head Self-Attention, SW-MSA)模块。W-MSA

模块将特征图划分为多个不重叠的窗口，由多头自注意力机制 MSA 对每个独立窗口中的特征计算注意力分数[23]。但 W-MSA 模块中各窗口之间缺乏全局相关性，因此 SW-MSA 模块改变了 W-MSA 对窗口的划分方式，即通过 Shift Window 循环位移的方式融合多个窗口的特征，同时使用 Mask 机制融合不同尺度的上下文信息，保持特征的相对位置关系。如图 2(b)所示，Stage 中交替使用 W-MSA 模块与 SW-MSA 模块，将分层的局部注意力与全局的自注意力机制相结合，产生不同层级的特征图。Stage1 中含有 Linear Embedding 层对每个像素的通道维度做线性变换将维度映射为 C。其余 Stage 中使用 Patch Merging 层进行下采样合并多个窗口的信息。Swin-Transformer 具有很好的扩展性，适用于处理不同尺度目标与密集目标，在图像分类、检测与分割任务中表现优异。

3. 改进的 STA-YOLOv7 模型

Swin-Transformer 在不同网络层之间进行局部信息与全局信息的交互，提取的特征具有层次性，但模型参数量较大且敏感，训练难度较高。YOLOv7 模型具有较好的实用性，训练速度较快，模型参数量较小，但特征提取能力较弱。本文提出了 STA-YOLOv7 模型，结合 Swin-Transformer 的特征提取优势与 YOLOv7 的实用性，改进 YOLOv7 的特征提取能力，提高多尺度目标检测的精度与速度。

3.1. 网络结构改进

为了解决高分辨率图像中卷积神经网络的语义信息导致的多尺度目标检测不精确的问题，我们在 STA-YOLOv7 模型中将 Swin-Transformer 模型替换 YOLOv7 主干网络顶层的 ELAN 模块，在 ELAN 提取的低分辨率特征映射进行全局像素操作，既可以利用自注意力机制的优势，又可以有效减少计算量，节省内存空间[24]。同时在 FPN 结构中融入 Swin-Transformer 模块捕捉不同区域间的相关性和重要性，有助于提高模型对各种大小目标的适应能力，提高目标检测的准确率，在并行计算的前提下进一步提高目标检测器的速度，在速度与精度之间实现更好的平衡。STA-YOLOv7 模型(图 3)的骨干网络对目标背景、边缘形状等上下文信息具备强大的建模能力，有效利用语义信息指导下游分类与定位任务。同时，该模型具有更好的扩展性和应用价值。

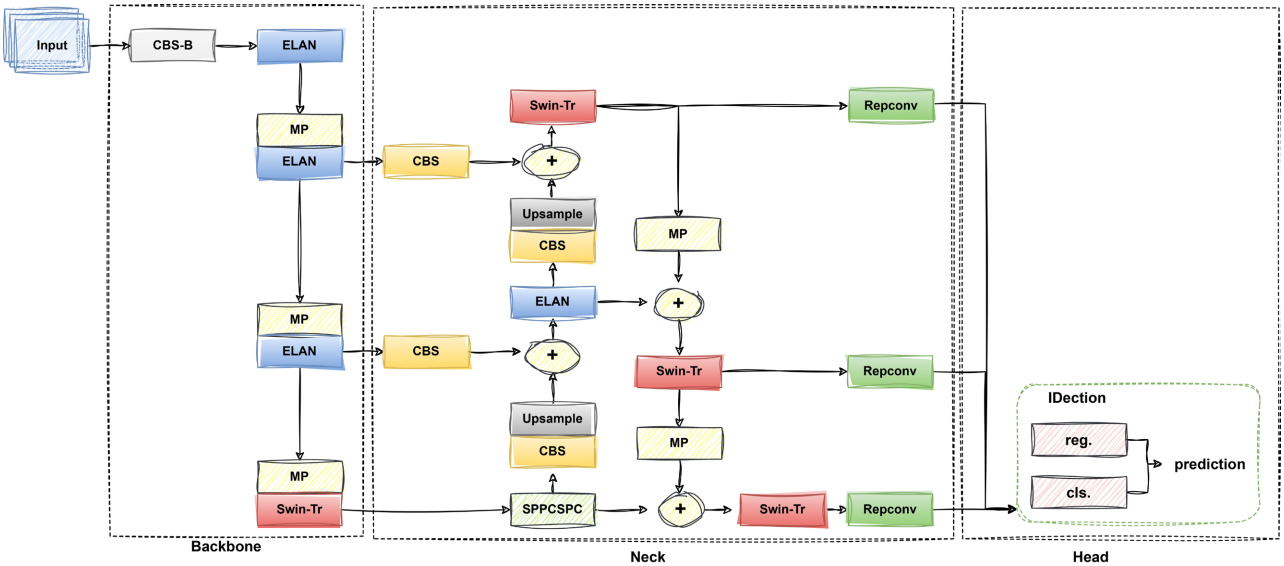

Figure 3. The architecture of STA-YOLOv7.
图 3. STA-YOLOv7 的网络结构.

3.2. AlignOTA 损失函数

准确合理的目标检测器依赖于训练过程中正样本的贡献度。与传统的目标检测损失函数不同，AlignOTA 损失函数[19]在分类损失中引入了 Focal Loss，动态调整正负样本的权重来抵消不平衡产生的影响，缓解了分类与回归不一致的问题，有助于网络学习更准确的定位信息和分类信息。在训练过程中，基于锚框的方法依据预设锚框与真实框的偏移量的标准来划分正负样本，结合在线挖掘困难样本的思想，重点关注预测框与真实框 IOU 较低的困难样本，通过最大化困难样本的目标置信度来惩罚训练过程中产生的错误分类信息。对于不同类型的样本计算与其对应的目标框之间的对齐损失，有助于帮助模型为每个目标选择分类与回归对齐的样本。正确的标签分配能够使模型在训练中学习正确的类别信息，提高模型的性能和应用价值。AlignOTA 在不同尺度大小的目标上具有很强的灵活性。具体来说，AlignOTA 损失函数由回归损失 IOU 损失函数与分类损失交叉熵损失函数构成，IOU 损失衡量预测测框与真实框的定位偏移损失，分类损失衡量样本分类置信度损失。在分类损失中，AlignOTA 引入了 Focal Loss 损失函数，缓解不平衡性产生的影响，使得模型具有更强的鲁棒性。我们在 STA-YOLOv7 算法中引入 AlignOTA 损失函数，缓解标签分配中样本不平衡导致的分类回归不一致的问题。具体公式如下：

$$\text{AlignOTA loss} = L_{reg} + L_{cls} \tag{1}$$

$$\alpha = \text{IOU}\left(reg_{gt}, reg_{pred}\right) \tag{2}$$

$$L_{reg} = -\ln\left(\alpha\right) \tag{3}$$

$$L_{cls} = \left(\alpha - cls_{pred}\right)^2 \times CE\left(cls_{pred}, \alpha\right) \tag{4}$$

4. 实验

4.1. 数据集与评估指标

道路病害检测数据获取与标注是非常困难的任务，由于不同天气、关照、角度等影响，道路病害图像的采集存在较大的挑战。本次实验中的数据集 RDDBJ 是由某道路养护单位提供的，覆盖不同天气、光照及路段等多种场景下的数据。本文使用 LabelImg 工具对 RDDBJ 数据集进行标注，通过 One-hot 编码将样本分为 5 类：线裂、网裂、坑槽、泛油和龟裂。采集约 4100 张图片，通过数据清洗，经过 CutPaste 数据增强策略后共有 5390 张用于构建数据集，并按照 9:1 的比例划分训练集和测试集，其中训练集 4312 张，测试集 539 张。训练集中所有样本参与训练，在测试集上进行评估。

实验中使用了一些性能指标来评估算法的性能，如查准率 P (Presion)、召回率 R (Recall)和平均精度 mAP (mean Average Precision)。此外，在工业场景中，FPS (Frame per Second)作为速度评估指标决定了模型能否实现工程化实施。在本文中，我们主要使用 mAP 和 FPS 的主要结果作为评估指标。其中，查准率、召回率和 mAP 表示为：

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{AP} = \int_0^1 P\left(R\right)\text{d}R \tag{7}$$

$$\text{mAP} = \frac{1}{N}\sum \text{AP} \tag{8}$$

其中 TP 表示正样本被正确预测为正样本的数量，FN 表示将正样本错误预测为负样本的数量，FP 表示将负样本被错误预测为正样本的数量。

4.2. 实验环境与训练

实验环境基于 Ubuntu18.04 操作系统，显卡为 NVIDIA RTX3090，CPU 为 AMD Ryzen 9 5950X 16 核心处理器，内存为 64 GB，存储为 1 TB SSD。实验基本上采用 YOLOv7 官方推荐的参数设置，基于 Python 与 Pytorch 深度学习框架搭建模型，输入图像大小为 640 × 640，使用随机翻转和数据增强等策略进行训练，使用标签平滑，采用 Adam 优化器，训练期间批量大小为 32，初始学习率为 0.01，学习率动量为 0.937，权重衰减率为 0.0005，迭代 200 次。

4.3. 实验结果分析

为验证 STA-YOLOv7 算法的有效性，我们将该模型与 YOLOv7 [10]、Swin-Transformer [22]、TPH-YOLOv5 [25]、YOLOv7 + Transformer 等算法进行比较，实验结果如表 1 所示。结果显示，STA-YOLOv7 模型的 mAP 达到 62.3%，比 YOLOv7 提高 3.7 个百分点，速度达到 72 FPS，在速度与精度之间实现了更好的平衡，验证了 STA-YOLOv7 模型的有效性。Swin-Transformer 编码器将 YOLOv7 的低维特征映射转换为高维空间表示，增强了特征的语义表达能力，有利于后续任务的完成。相比于 TPH-YOLOv5，STA-YOLOv7 更适用于处理高分辨率图像。实验证明 Swin-Transformer 编码器是 STA-YOLOv7 网络性能提升的关键因素，在目标检测任务中发挥着至关重要的作用。我们在图 4 中展示了 STA-YOLOv7 算法在 RDDBJ 数据集中的一些代表性的可视化检测结果。

Table 1. Performance comparison of different algorithms on RDDBJ dataset.
表 1. 不同算法在 RDDBJ 数据集上的性能比较

Method	Size	mAP@.5	FPS	Param.
YOLOv7	640	0.586	156	37.2M
Swin-Transformer	640	0.599	95	37.6M
TPH-YOLOv5	640	0.576	113	38.5M
YOLOv7+Transformer	640	0.609	87	37.3M
STA-YOLOv7	640	0.623	72	37.4M

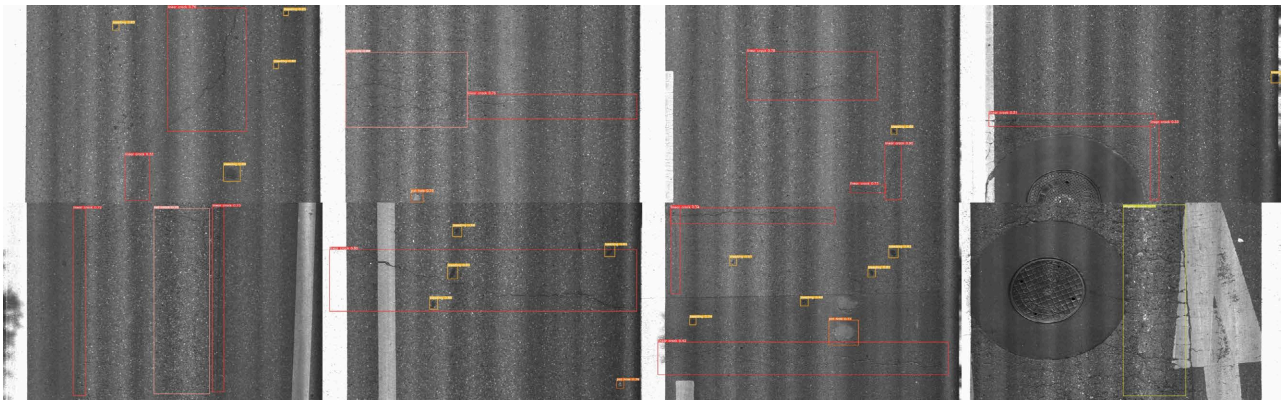

Figure 4. Some representative results of STA-YOLOv7 on RDDBJ dataset
图 4. STA-YOLOv7 在 RDDBJ 数据集上一些代表性结果

4.4. 消融实验

4.4.1. Swin-Transformer 模块消融实验

为验证 Swin-Transformer 模块与 YOLOv7 在融合特征方面对下游检测任务的重要性，在 Swin-Transformer 模块的消融实验中，我们设计了三组实验：1) Swin-Transformer + YOLOv7 (backbone)；2) Swin-Transformer + YOLOv7 (FPN)；3) Swin-Transformer + YOLOv7 (backbone + FPN)，实验结果如表 2 所示。

Table 2. Ablation experiments with different type of component fusion between Swin-Transformer and YOLOv7.
表 2. Swin-Transformer 与 YOLOv7 进行不同组件融合的消融实验

Type	P	R	mAP@.5	FPS
Baseline	0.589	0.552	0.586	145
(1)	0.612	0.561	0.592	126
(2)	0.619	0.583	0.614	108
(3)	0.635	0.590	0.623	72

从上述实验中可以看到将 Swin-Transformer 模块插入 YOLOv7 的不同位置会对模型性能产生不同程度的影响。在 Backbone 中插入 Swin-Transformer 模块时，模型性能稍微有提升，在 FPN 中插入 Swin-Transformer 模块时，模型的 mAP 相较于基线有明显提升。在多个位置比如 Backbone 与 FPN 中插入 Swin-Transformer 模块时，模型的 mAP 进一步提高。

4.4.2. AlignOTA 损失函数消融实验

为进一步验证标签分配对齐损失函数对检测性能的提升，本文设计标签分配对齐损失的消融实验，对比完整的 STA-YOLOv7 模型与去除 AlignOTA 损失函数的模型。

Table 3. Ablation Study of AlignOTA loss introduced in STA-YOLOv7 model.
表 3. STA-YOLOv7 引入 AlignOTA 损失函数的消融实验

Method	P	R	mAP@.5	FPS
Without AlignOTA	0.592	0.561	0.591	85
STA-YOLOv7	0.638	0.615	0.623	72

表 3 中，实验结果显示，移除标签对齐损失后，STA-YOLOv7 的 mAP 由 0.623 下降到 0.591，下降约 3.2%。这证明了 AlignOTA 标签分配对齐损失对检测性能有较大的提升作用，验证了其在 STA-YOLOv7 算法中的重要性。

5. 总结与展望

本文针对高分辨率图像中多尺度病害目标检测精度低的问题提出了 STA-YOLOv7 算法，缓解了标签分配不对齐的问题，同时学习不同尺度目标的全局语义信息与局部细粒度特征，有效提高了模型的准确性与鲁棒性。实验结果表明 STA-YOLOv7 算法能够适应复杂场景下的多尺度病害目标检测[26]，具有较高的应用潜力。虽然我们的方法在实验中取得了令人满意的结果，但仍有许多潜在的改进空间。在未来的工作中，我们将继续探索更高效的模型结构和训练策略，以进一步提高病害检测方法的性能。

基金项目

本研究由北京建筑大学研究生创新项目(No. PG2022145)资助。

参考文献

[1] Cristianinin, Taylorjs. 支持向量机导论[M]. 李国正, 王猛, 曾华军, 译. 北京: 电子工业出版社, 2004: 15-20.

[2] 楼竞. 基于图像分析的路面病害检测方法与系统开发[D]: [硕士学位论文]. 南京: 南京理工大学, 2008.

[3] 马梁, 苟于涛, 雷涛, 等. 基于多尺度特征融合的遥感图像小目标检测[J]. 光电工程, 2022, 49(4): 49-65.

[4] 刘宪明, 辛公锋. 国内基于深度学习的道路路面病害检测研究综述[J]. 电子世界, 2021(8): 96-98.

[5] 谷永立, 宗欣欣. 基于深度学习的目标检测研究综述[J]. 现代信息科技, 2022, 6(11): 76-81.

[6] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[7] Girshick, R.B. (2015) Fast R-CNN. 2015 *IEEE International Conference on Computer Vision* (*ICCV*), Santiago, Chile, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[8] Ren, S., He, K., Girshick, R. and Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[9] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[10] Wang, C.Y., Bochkovskiy, A. and Liao, H.Y.M. (2022) YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. https://arxiv.org/abs/2207.02696

[11] Li, Z. and Zhou, F. (2017) FSSD: Feature Fusion Single Shot Multibox Detector. https://arxiv.org/abs/1712.00960

[12] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324

[13] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017.

[14] Mehta, S. and Rastegari, M. (2021) MobileViT: Light-Weight, General-Purpose, and Mobile-Friendly Vision Transformer. https://arxiv.org/abs/2110.02178

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929

[16] Touvron, H., Cord, M., Douze, M., *et al*. (2020) Training Data-Efficient Image Transformers & Distillation through Attention. https://arxiv.org/abs/2012.12877

[17] Pan, X., Ge, C., Lu, R., *et al*. (2022) On the Integration of Self-Attention and Convolution. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 805-815. https://doi.org/10.1109/CVPR52688.2022.00089

[18] Ge, Z., Liu, S., Wang, F., Li, Z. and Sun, J. (2021) Yolox: Exceeding Yolo Series in 2021. https://arxiv.org/abs/2107.08430

[19] Xu, X., Jiang, Y., Chen, W., *et al*. (2023) DAMO-YOLO: A Report on Real-Time Object Detection Design. https://arxiv.org/abs/2211.15444

[20] 戚玲珑, 高建瓴. 基于改进 YOLOv7 的小目标检测[J]. 计算机工程, 2023, 49(1): 41-48.

[21] 赵元龙, 单玉刚, 袁杰. 改进YOLOv7与DeepSORT的佩戴口罩行人跟踪[J]. 计算机工程与应用, 2023, 59(6): 221-230.

[22] Liu, Z., Lin, Y., Cao, Y., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[23] Lu, S., Liu, X., He, Z., Karkee, M. and Zhang, X. (2022) Swin-Transformer-YOLOv5 for Real-Time Wine Grape-Bunch Detection. https://arxiv.org/abs/2208.14508

[24] Gong, H., Mu, T., Li, Q., *et al*. (2022) Swin-Transformer-Enabled YOLOv5 with Attention Mechanism for Small Object Detection on Satellite Images. *Remote Sensing*, **14**, Article No. 2861. https://doi.org/10.3390/rs14122861

[25] Zhu, X., Lyu, S., Wang, X. and Zhao, Q. (2021) TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios. 2021 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*), Montreal, 11-17 October 2021, 2778-2788. https://doi.org/10.1109/ICCVW54120.2021.00312

[26] 李姗. 基于 YOLOv5 的道路病害检测与分类研究[J]. 现代计算机, 2021, 27(35): 75-79.