计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(7):107−116 [doi: 10.15888/j.cnki.csa.009877] [CSTR: 32024.14.csa.009877]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 RT-DETR 的道路缺陷检测①

朴恒剑, 朱　明

(中国科学技术大学 自动化系, 合肥 230026)
通信作者: 朴恒剑, E-mail: piaohengjian@mail.ustc.edu.cn

摘　要: 道路损坏对道路的使用寿命和安全性构成极大威胁, 及早发现道路损坏有利于进行维护和修理. 传统的道路缺陷检测技术通常依赖于手动视觉检测和车载道路路面监控系统, 然而这些方法在很大程度上受道路维护人员经验的影响. 随着深度学习的发展, 越来越多的研究者将其应用于道路缺陷检测领域, 其中最常见的当属 YOLO 系列目标检测方法及其各种变体. 但这类方法大多需要进行后处理操作, 这会阻碍模型优化、损害鲁棒性并导致检测器延迟推理. 针对这些问题以及道路缺陷检测中存在的多尺度问题, 本文提出了改进后的 RT-DETR 模型, 对主干网络的结构进行了微调, 并提出了 MSaE 注意力机制. 在编码器部分, 使用 GhostConv 卷积和 DySample 模块优化了上采样, 使用 ADown 模块优化了下采样. 本文在公开数据集 SVRDD 中进行了对比实验. 实验结果表明, 本文提出的改进方法在 SVRDD 数据集中的 *mAP*@50 指标达到了 72.5%, 相较于基准的 RT-DETR-R18 提高了 3.8 个百分点, 有效提升了道路缺陷检测能力.

关键词: 目标检测; 注意力机制; 特征融合; 点采样; 多尺度

引用格式: 朴恒剑,朱明.基于改进 RT-DETR 的道路缺陷检测.计算机系统应用,2025,34(7):107–116. http://www.c-s-a.org.cn/1003-3254/9877.html

Road Defect Detection Based on Improved RT-DETR

PIAO Heng-Jian, ZHU Ming

(Department of Automation, University of Science and Technology of China, Hefei 230026, China)

Abstract: Road damage poses a great threat to the service life and safety of roads. Early detection of road defects facilitates maintenance and repair. Traditional road defect detection methods typically rely on manual visual inspection and vehicle-mounted pavement monitoring systems. However, these methods are largely influenced by the experience of road maintenance personnel. With the advancement of deep learning, increasing numbers of researchers have applied it to road defect detection. Among these, the YOLO series of object detection methods and their various variants are the most common. However, most of these methods require post-processing operations, which hinder model optimization, impair robustness, and lead to delayed inference by the detector. To address these issues, as well as the multi-scale challenges in road defect detection, an improved RT-DETR model is proposed. The backbone network is fine-tuned, and the MSaE attention module is introduced. In the encoder, GhostConv convolution and DySample module are used to optimize upsampling, while the ADown module optimizes downsampling. Comparative experiments are conducted on the public SVRDD dataset. Experimental results show that the proposed improved method achieves a 72.5% *mAP*@50 on the SVRDD dataset, 3.8 percentage points higher than the benchmark RT-DETR-R18, significantly enhancing road defect detection performance.

Key words: object detection; attention mechanism; feature fusion; point sampling; multi-scale

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2024-11-18; 修改时间: 2024-12-09; 采用时间: 2025-01-10; csa 在线出版时间: 2025-04-01
CNKI 网络首发时间: 2025-04-02

1 引言

公路交通系统对于促进区域稳定、加速经济发展以及提升人民生活质量具有不可替代的作用. 在公路的使用周期中, 路面会面临多种外部挑战, 包括持续的机械应力 (如车辆反复碾压)、气候波动 (如温度变化和雨水侵蚀), 以及潜在的人为破坏, 这些因素共同作用下, 会导致路面出现裂缝、坑洞、剥落等不同类型的损害, 不仅影响公路的正常使用, 还可能构成交通安全的隐患. 鉴于此, 公路养护工作在当前我国公路交通系统建设中占据了核心地位, 成为保障公路安全、延长公路寿命的关键.

近些年来, 部分研究者采用目标检测方法对道路缺陷展开检测, 主要涵盖两阶段目标检测算法与单阶段目标检测算法. 两阶段目标检测方法的检测精度相对更高, 然而由于需要进行两个阶段的处理, 其检测速度较为缓慢, 具有代表性的方法包括基于 R-CNN[1]的一系列算法. Suh 等人[2]以 ZF-Net 作为特征提取网络, 提升了特征提取的速度, 实现了对桥梁、路面等基础设施的损伤检测. Pang 等人[3]运用 Inception V2 网络进行特征提取, 以获取更为有效的深度特征, 从而实现了道路缺陷的有效检测. 与两阶段网络的流程不同, 单阶段网络采用了一种更为简洁高效的方式: 它们无需生成候选区域, 而是通过一个统一的网络架构直接输出目标的类别和位置信息. 这种设计极大程度地简化了检测流程, 提高了检测速度. 单阶段网络中常用算法有 YOLO (you only look once)[4–7]系列、SSD (single shot multibox detector)[8]等, 这些算法在确保检测精度的同时, 实现了快速实时目标检测, 尤其适用于对速度要求较高的应用场景. Ren 等人[9]在其研究中, 将 YOLOv5 中的路径聚合网络 (PAN) 替换为一种泛化的特征金字塔网络 (GFPN), 即对 YOLOv5 的特征提取部分进行了改进, 利用 GFPN 更好地融合不同层次的特征, 增强模型在物体检测任务中的表现, 特别是提高了对于多尺度目标的检测能力. 高敏等人[10]提出了改进 YOLOv7 的道路坑洼检测算法, 采用 Mosaic+Mixup 进行内置数据增强, 扩充小样本数据集, 增强模型泛化能力, 并引入 CA 注意力机制, 对横纵位置信息进行编码, 在保证计算量的同时又能关注大范围位置信息, 还采用了 BIFPN 双向特征金字塔网络, 通过融合多尺度语义特征提高了检测效率. Wang 等人[11]提出了一种基于 YOLOv8s 的增强型道路缺陷检测算法 BL-YOLOv8. 他们通过融合 BiFPN 概念, 重构其颈部结构, 对 YOLOv8s 模型进行了优化, 减少了模型的参数、计算负载和整体大小. 此外, 为了增强模型的运算性能, 还引入 SimSPPF 模块优化了特征金字塔层, 提高了模型的速度. He 等人[12]提出了一种名为 LMFE-RDD 的道路缺陷检测器, 用于平衡速度和精度, 该检测器构建了轻量级多特征提取网络 (LMFE-Net) 作为骨干网络, 并构建了高效语义融合网络 (ESF-Net) 进行多尺度特征融合.

然而, 上述算法均需在后续处理中采用非极大值抑制 (NMS) 等后处理操作, 这会对模型优化造成阻碍、损害鲁棒性并致使检测器延迟推理. 鉴于此, 研究人员将目光投向在自然语言处理领域表现卓越的 Transformer 架构. Dosovitskiy 等人[13]提出了 Vision Transformer (ViT), 这种应用于计算机视觉的纯 Transformer 结构模型在拥有大量数据以及足够的训练开销的情况下, 能够接近当时最为先进的 CNN 模型的性能. 随后, DETR[14]首次将目标检测任务重新定义为图像到集合的问题, 借助匈牙利算法确定目标分类预测帧与实际帧的最佳匹配, 避免了传统卷积神经网络中的后处理操作, 如 NMS 和锚帧生成. 然而, 它的编码器设计复杂, 导致网络收敛缓慢, 并且对小目标的检测效果不佳. Deformable DETR[15]考虑到多尺度信息对目标检测性能的影响, 采用可变形注意力模块解决了收敛缓慢以及特征分辨率有限的问题, 提升了训练效率和效果, 但同时也增加了模型的复杂性以及计算和内存开销. RT-DETR[16]考虑了特征层输入的重要性以及分类分数与位置置信度对检测的影响, 设计了高效混合编码器来处理多尺度特征, 减少了计算开销, 还提出了 IoU 感知查询选择用来改善初始目标查询. 但是, RT-DETR 仍然不能在道路缺陷检测任务中与同尺寸的 YOLO 系列目标检测器拉开差距, 尤其是在这种多尺度场景下难以识别到小目标, 且对一些大目标的识别也不尽如人意.

因此, 本文在 RT-DETR 的基础上提出了一种改进后的 RT-DETR 模型, 主要贡献如下: 1) 去掉了骨干网络中所有 BasicBlock 中的平均池化层, 并基于 SaE 注意力机制[17], 提出了 MSaE (multi-scale SaE) 注意力模块, 并应用到了骨干网络中的前两个 Block 中, 提高了对于不同尺寸对象的检测能力. 2) 使用 ADown 模块[18]进行下采样, 在减少模型参数量的同时进一步细化了特征图的分辨率. 3) 使用 DySample 模块[19]进行上采样, 并使用 GhostNet[20]中的 GhostConv 替换了两处上

采样前的卷积, 实现了模型精度的提升.

2 RT-DETR 介绍

RT-DETR 是由百度与北京大学电子与计算机工程学院的团队共同开发而成. RT-DETR 主要由主干网络、高效混合编码器 (efficient hybrid encoder) 以及带有辅助预测头的 Transformer 解码器组成.

本文以 RT-DETR-R18 为基准模型, 即主干网络使用 ResNet18[21], RT-DETR-R18 的结构如图 1 所示, 主干网络将最后 3 个阶段的特征输入到编码器中, 接着高效混合编码器将多尺度特征转化为图像特征序列, 再运用不确定性最小查询选择方法 (uncertainty-minimal query selection) 来挑选固定数量的编码器特征, 以此作为解码器的初始目标查询. 最后, 带有辅助预测头的解码器对目标查询进行迭代优化, 进而生成类别和边界框.

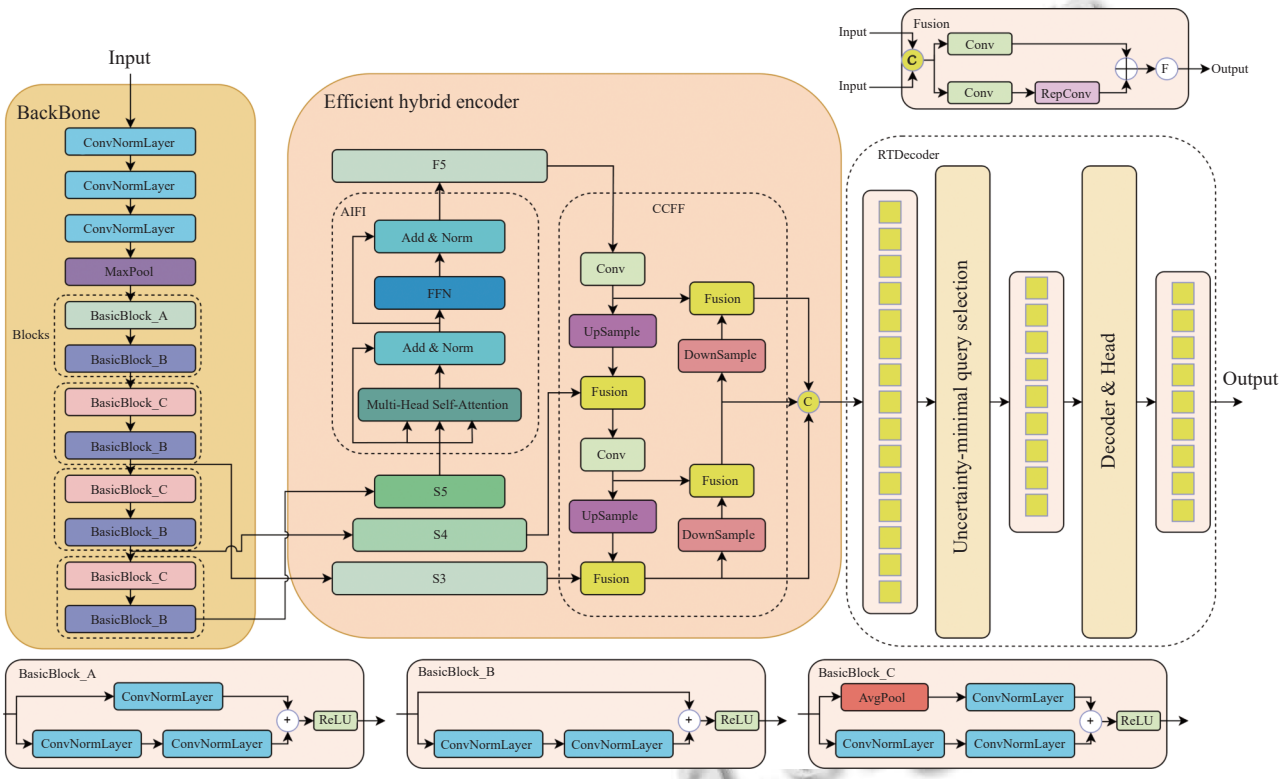


图 1 RT-DETR-R18 结构图

RT-DETR 的主干网络主要承担提取多尺度图像特征的任务, 从主干网络最后 3 个 Block 中提取特征并送入高效混合编码器. 高效混合编码器是 RT-DETR 的关键创新部分之一, 主要由两个组件构成: 基于注意力的尺度内特征交互 (attention-based intra-scale feature interaction) 以及基于 CNN 的跨尺度特征融合 (CNN-based cross-scale feature fusion). 尺度内特征交互借助自注意力机制处理同一尺度内的特征, 而跨尺度特征融合则运用轻量级的跨尺度注意力机制融合不同尺度的特征, 从而提升了模型对不同大小目标的检测能力. 通过这样的方式, 既保留了 CNN 的局部特征提取能力, 又充分利用了 Transformer 的全局建模能力. 随后是 RT-DETR 的另一个主要创新点: IoU 感知查询选择, 其作用是选择初始目标查询. 它从编码器输出的特征中挑选固定数量的特征作为初始查询, 选择过程依据不确定性最小化原则, 以确保选出最有可能包含目标的查询. 最后, 将跨尺度特征融合后的特征拼接在一起并送入 Transformer 解码器 (RTDecoder), 通过迭代优化目标查询, 逐步细化目标的位置和类别信息. 在解码器的每一层都添加了辅助预测头, 这些预测头能够生成中间预测结果, 对模型的训练和收敛起到促进作用, 最终的预测结果则来自最后一层的预测头.

总的来说, RT-DETR 通过创新的编码器设计和感知查询选择机制, 成功地将 Transformer 应用于实时目标检测任务, 在保持高精度的同时实现了较高的检测速度, 为目标检测领域提供了一个强有力的新选择.

3 算法设计和实现

本文在 RT-DETR 原结构的基础上做了些改进, 其整体结构如图 2 所示. 首先是主干网络部分, 去掉了所有 BasicBlock_C 中的平均池化层, 在前两个 Block 中的 BasicBlock_A, BasicBlock_B 以及去掉了平均池化层的 BasicBlock_C 中添加了基于 SaE 注意力模块改进的多尺度 SaE 注意力模块 MSaE, 提高了对于不同尺寸目标的检测能力. 随后, 在高效混合编码器部分做了较多改动, 在 CCFF 模块中的两个上采样前的卷积层替换为了轻量化的 GhostConv 卷积, 同时使用基于点采样的 DySample 模块优化了上采样操作, 最后, 将两个下采样模块替换为了轻量且高效的 ADown 模块, 这些改动整体上虽然会使模型的复杂度略有提高, 但也使得模型的检测精度有了明显的提升, 整体的改进思路也是针对道路缺陷检测这种多尺度场景进行有目的性的优化.

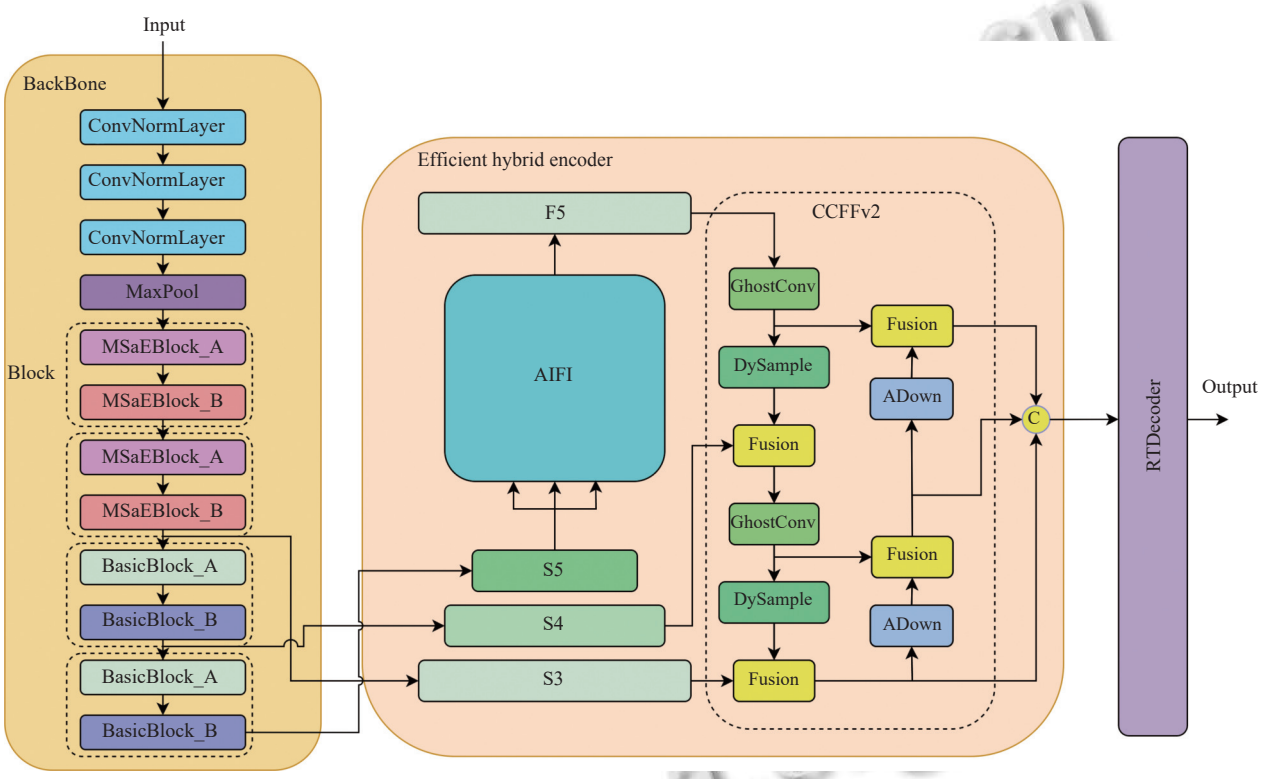


图 2 改进后的 RT-DETR 结构

3.1 增加多尺度 SaE 注意力模块 MSaE

SaE 模块是在 SENet (squeeze and excitation network) 的基础上进行优化的. 它结合了挤压 (squeeze) 和激励 (excitation) 操作, 引入了多分支的全连接层结构, 能够让模型学习到更广泛的全局信息, 进一步增强网络对不同特征模式的捕捉能力, 同时对模型的尺寸不会有太大影响.

SaE 模块结构如图 3 所示, 首先是挤压操作, 模块的输入经过全局平均池化层, 得到通道方向的统计信息, 这一步骤能够将输入的空间信息进行压缩, 聚焦于通道层面的特征. 接着, 通过一个全连接层对全局平均池化层的输出进行处理. 这个全连接层进行了一定的缩减尺寸, 能够在保留关键特征的同时, 减少数据维度. 在挤压操作之后, 紧接着是激励操作, 最后将多路输出拼接成原始形状. 然后, 将激励层的输出与输入层的特征图进行乘法操作, 这能够根据激励层所学习到的权重, 对原始特征图进行重新校准, 突出关键特征的作用.

本文在 SaE 模块的基础上针对道路缺陷检测中存在的多尺度问题提出了改进, 其结构如图 4 所示.

改进后的结构保留了原有的挤压和激励操作, 同时在其头部添加了两层, 第 1 层是一个并行的多尺度特征提取层, 分别使用 1×1 的卷积核与 3×3 的卷积核提取不同尺度的特征, 第 2 层则是对第 1 层的输出进行融合. 新的注意力模块通过这种方式保留了原有的全局信息学习的能力, 同时强化了对不同尺寸目标的特征捕捉能力.

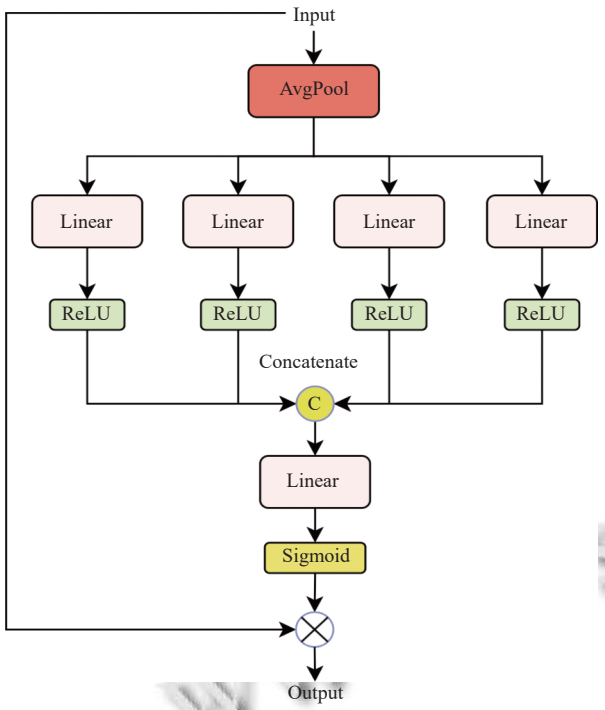


图 3 SaE 模块结构

如图 5 所示, 本文将原主干网络中所有 Block 中的平均池化层全部去掉, 原始的 ResNet 中本来就是没有这一层的, RT-DETR 的作者在添加这一层后取得了更好的效果, 然而在 SVRDD 数据集上的表现并不佳, 同时在前两个 Block 中的 BasicBlock_A, BasicBlock_B 以及去掉平均池化层的 BasicBlock_C 中添加了 MSaE 注意力模块, 构成了新的 MSaEBlock_A 和 MSaEBlock_B 模块. 之所以仅在骨干网络的浅层中添加 MSaE 注意力模块, 主要是为了提高小目标的检测效果, 同时还不会过多地提高模型的复杂度, 并且 MSaE 本身也是面向多尺度场景的, 因此还能兼顾到较大尺寸的目标的检测效果.

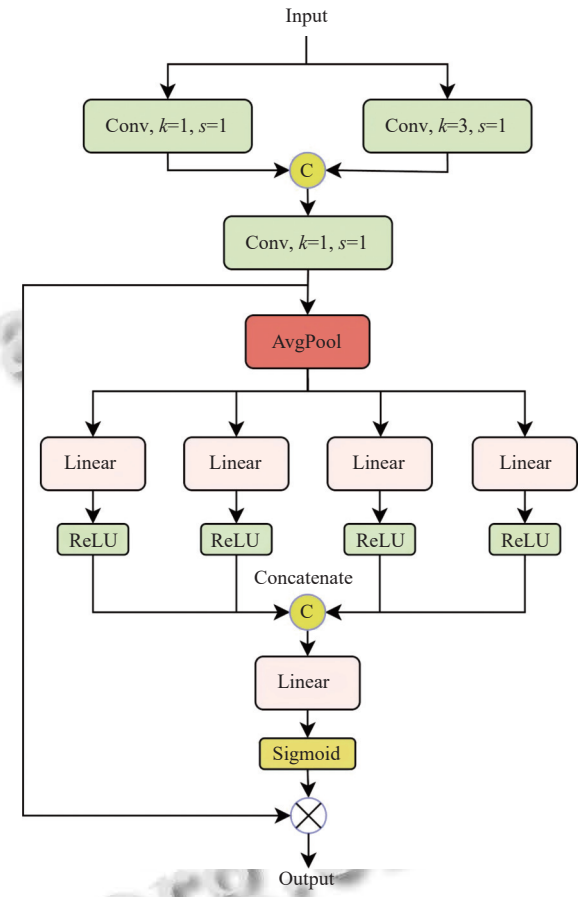
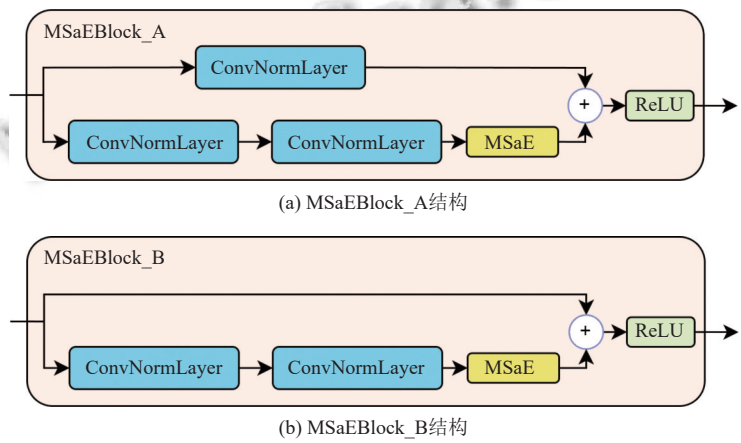


图 4 MSaE 模块结构



(a) MSaEBlock_A结构

(b) MSaEBlock_B结构

图 5 加入 MSaE 注意力机制的 MSaEBlcok 结构

3.2 使用 ADown 模块进行下采样

在深度学习模型中, 特征下采样是一项关键操作, 对于降低计算量、扩大感受野以及提取多尺度特征具有重要意义.

ADown 模块是在 YOLOv9 中提出并使用的一个轻量化的下采样模块, ADown 模块结构如图 6 所示. 采用了独特的下采样策略, 先进行平均池化, 然后进行拆分, 随后对拆分后的特征图分别进行不同的操作再拼接, 这种方式避免了对整个特征图进行复杂的单一操作, 因此能够在降低特征图空间分辨率的同时, 较好地保留图像的特征信息. 尤其是与一些简单的下采样方法 (如直接使用大步长卷积) 相比, 它可以避免过度丢失信息, 更好地捕捉特征图中的关键信息, 使得模型能够更准确地识别目标的位置和类别.

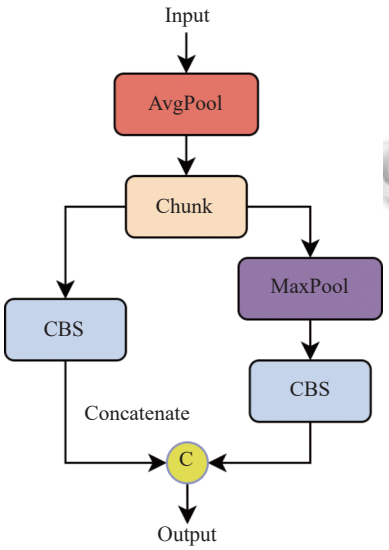


图 6 ADown 结构

ADown 模块在实际应用中展现出显著优势, 在特征下采样过程中, 它能够在保持较高精度的同时, 提高计算效率, 降低资源消耗. 在需要多尺度特征处理的任务中, ADown 模块为模型提供了强大的下采样能力. 考虑到所提 MSaE 注意力模块会提高模型复杂度, 因此本文将原 CCFF 模块中的下采样模块替换为 ADown 模块.

3.3 使用 GhostConv 和 DySample 模块进行上采样

本文结合使用 GhostConv 卷积和 DySample 模块进行上采样, 主要是为了在上采样前保留更多的特征, 并且在上采样时能够根据输入特征自适应地调整采样位置, 从而更好地捕捉图像的细节和结构信息, 更好地应对道路缺陷检测中的多尺度问题.

GhostConv 卷积是一种针对卷积神经网络的创新设计, 特别适用于嵌入式设备, 这些设备通常具有有限的内存和计算资源. Ghost 模块的核心思想是利用已有的特征图通过低成本的线性变换生成更多的“幽灵”特征图 (ghost feature maps), 从而提高网络的计算效率.

本文使用 GhostConv 结构如图 7 所示, 首先进行一次常规卷积, 对输入特征图进行初步的特征提取, 产生了一部分输出特征图. 之后, 数据被分为两路, 一路保持不变, 另一路经过一个卷积核大小为 5, 步长为 1 的卷积层. 最后, 将两路特征进行融合, 得到与原始卷积操作相同维度的输出特征图, 在上采样前尽可能多地保留和生成有用的特征, 同时降低了模型的复杂度和计算成本.

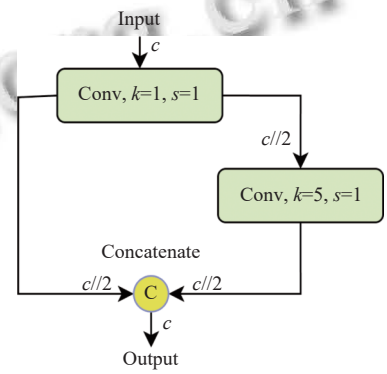


图 7 GhostConv 结构

接下来是 DySample 上采样, 它是一种新型的轻量且高效的上采样技术. DySample 不需要引入高分辨率特征. 它提出了一种用于上采样操作的点采样方法, 该方法在连续特征图上生成具有感知内容的采样点, 然后对这些点进行重新采样. 这种方法可以节省基于核的动态卷积所带来的计算开销, 并进一步提高计算效率.

DySample 结构如图 8(a) 所示, 给定一个大小为 $C\times H_1\times W_1$ 特征图 X, 以及一个由采样点生成器 (sampling point generator) 产生的大小为 $2\times H_2\times W_2$ 的采样集 S, 其中第 1 维的 2 表示 x 和 y 坐标. *grid_sample* 函数使用 S 中的位置对假设的双线性插值 X 进行重新采样, 生成大小为 $C\times H_2\times W_2$ 的 X', 这个过程定义为:

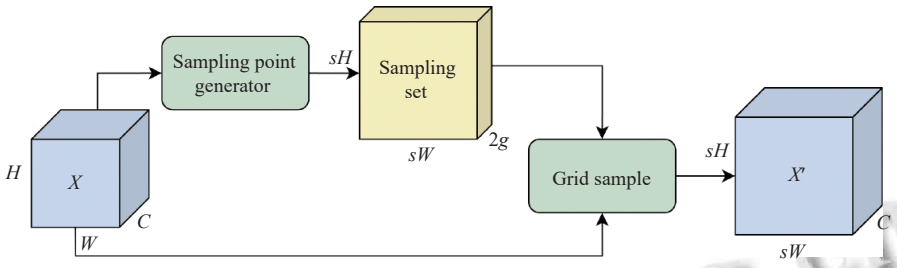
$$X' = grid\_sample(X, S) \tag{1}$$

本文中使用的 DySample 的采样点生成器使用了静态范围因子, 如图 8(b) 所示, 给定一个上采样尺度因子 s 和一个大小为 $C\times H\times W$ 的特征图 X, 使用一个线性层, 其输入和输出通道数分别为 C 和 $2gs^2$, 来生成大小为 $2gs^2\times H\times W$ 的偏移量 O, 其中 g 为超参数. 然后, 通过随机排序将其重塑为 $2g\times sH\times sW$, 采样集 S 是偏移量 O 和原始采样网络 G 的和, 即式 (2) 和式 (3).

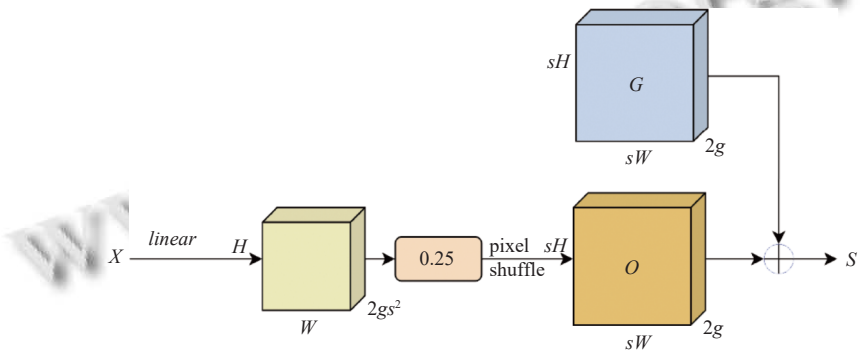
$$O = linear(X) \tag{2}$$

$$S = G + O \tag{3}$$

本文将原 CCFF 模块中的上采样模块替换为了 DySample 模块, 这一改进不仅减少了模型参数和计算量, 还提高了模型的精度.



(a) 基于采样的动态上采样



(b) 静态范围因子

图 8 DySample 结构

4 实验分析

4.1 数据集介绍

SVRDD[22]数据集由北京大学遥感与地理信息系统研究所的 Ren 等人共同开发. 他们从百度地图获取了 8 000 张街景图像, 并对其中 20 804 个损坏实例进行了标注. 此数据集共有 7 种标签, 主要包括纵向裂纹、横向裂纹、鳄鱼裂纹、坑洼、纵向斑块、横向斑块, 同时考虑到坑洼和井盖在分类时可能出现错误的情况, 特意添加了井盖这一类别. 这些图像的背景极为复杂, 涵盖了行人、车辆、建筑物、高架桥、树木及其阴影等多种元素, 并且图像是在不同的季节、天气以及照明条件下收集而来的. SVRDD 是首个基于街景图像的道路缺陷公共数据集, 在相关研究领域具有重大的开创意义和价值.

4.2 实验细节

在本文实验中, 将最大训练次数设定为 300. 在训练的前 50 个 epoch 被设置为 warmup 阶段, 而在 RT-DETR 源码中此设置为 2 000, 但这样的设置在较小数据集的情况下可能会导致模型不收敛. 初始学习率设置为 0.000 5, 并采用 AdamW 优化策略进行学习率调整, 在最后一轮时学习率降为 0.000 05. 训练过程中, batch-size 设置为 32, 输入图片的尺寸均统一设置为 640×640. 在训练本文改进后的 RT-DETR 模型以及其他模型时均不加载预训练模型. 本实验中使用的硬件配置如表 1 所示.

表 1 训练所用机器配置表

类型	参数
系统	Ubuntu 20.04.2 LTS
CPU	Intel(R) Xeon(R) Platinum 8362 CPU @ 2.80 GHz
GPU	Nvidia GeForce GTX3090, 24 GB
内存	45 GB

4.3 实验结果

为了验证实验结果, 本文使用准确率、召回率和平均精度均值 (mean average precision, *mAP*) 来评估模型的检测性能. 准确率与召回率定义如下:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

其中, *FP* 是假阳性, *FN* 是假阴性, *TP* 是真阳性.

平均精度均值的定义如下:

$$mAP = \frac{\sum P_A}{N_C} \tag{6}$$

其中, N_C 为类别数量, P_A 为各类别的平均精度. 在实验中采用了两种主要的评估指标: *mAP*@50 和 *mAP*@50:95. *mAP*@50 表示当交并比 (intersection over union, IoU) 阈值设定为 0.5 时, 所有类别的平均精度. 而 *mAP*@50:95 则是一个更为全面的指标, 它计算 IoU 阈值从 0.5 到 0.95 (以 0.05 为步长) 范围内的多个 *mAP* 值, 然后取这些值的平均. 这种方法可以更全面地评估模型在不同 IoU 阈值下的性能. *mAP* 的取值范围是 0 到 1 之间. 值越接近 1, 表明模型的整体性能越好, 在各类目标的检测能力上越强.

除了精度指标, 本文还关注模型的效率. 检测速度通常用每秒处理的图片数量 (frames per second, FPS) 来衡量. FPS 越高, 则表示模型的实时性能越好. 而模型大小的评估主要采用两个指标: 计算量和参数量. 计算量反映了模型的计算复杂度, 通常以浮点运算次数 (FLOPs) 为单位, 本文使用 GFLOPs (10^9 FLOPs) 来表示. 参数量则是模型所有参数的总数, 通常以百万 (M) 为单位. 这两个指标对于评估模型在不同设备上的部署可行性和运行效率至关重要.

4.3.1 对比实验

为了验证本文算法改进的有效性, 本文在 SVRDD 数据集上分别验证了 RT-DETR-R18, RT-DETR-R34 以及 YOLO 系列从 YOLOv8 到 YOLO11 相近尺寸的模型, 并与本文算法进行对比, YOLO 系列采用 Ultralytics 默认的超参数, RT-DETR 系列采用与本文算法相同的超参数. 对比结果如表 2 所示. 其中每个模型的 FPS 都是在 TensorRT FP32 精度下测量获得, 从中可以看出, 本文的算法相较于基准的 RT-DETR-R18 性能有所提升, *Precision* 由 0.745 提升至 0.76, *Recall* 由 0.647 提升至 0.675, *mAP*@50 和 *mAP*@50:95 也均提升了 3.8 个百分点, 模型参数量也略有降低, 不过模型计算量略有提升, 检测速度也降低了 5.1%. 对比更大尺寸的 RT-DETR-R34, 本文算法除了 *Precision* 略低, 其他指标均优于 RT-DETR-R34. 对比 YOLO 系列检测器时, 本文的算法也表现出了优秀的性能. 在与尺寸相近的 YOLO 系列模型对比时, 本文算法不仅做到了更高的精度, 同时带来了更快的检测速度.

表 2 SVRDD 数据集对比实验结果

模型	*Precision*	*Recall*	*mAP*@50 (%)	*mAP*@50:95 (%)	参数量 (M)	计算量 (GFLOPs)	FPS
YOLOv8m	0.727	0.645	69.2	41.6	23.22	67.9	172
YOLOv9m	0.717	0.635	67.3	40.3	16.73	61.1	155
YOLOv9c	0.747	0.645	69.8	43.0	21.36	84.1	139
YOLOv10m	0.708	0.619	66.2	40.0	**16.49**	64.0	**212**
YOLOv10b	0.762	0.619	68.8	42.2	20.46	98.7	162
YOLO11m	0.761	0.622	68.7	41.4	20.03	67.7	164
RT-DETR-R18	0.745	0.647	68.7	40.0	20.1	**58.3**	197
RT-DETR-R34	**0.790**	0.643	71.2	41.8	30.21	88.6	163
本文算法	0.76	**0.675**	**72.5**	**43.8**	19.71	66.6	187

4.3.2 消融实验

为了分析各模块组合对模型的影响, 本文设计了消融实验, 采用同样的硬件配置和超参数, 训练 300 个 epoch, 并在 SVRDD 数据集上训练, 对比了在验证集上的检测结果, 结果如表 3 所示. 在加入 GhostConv 卷积后检测精度就获得了较多提升, 不过模型参数量和计算量只有轻微浮动. 随后加入 DySample 上采样模块, *mAP*@50:95 又有了进一步的提升, 说明 DySample 模块有助于提高模型在较高 IoU 阈值下的检测效果. 单独加入 MSaE 注意力模块时, 模型精度也有一些提升, 说明 MSaE 在多尺度场景下能够发挥出不错的效果. 当同时加入 MSaE 模块, GhostConv 卷积和 DySample 上采样时, *mAP*@50 和 *mAP*50:95 分别提升了 2.3 个百分点和 2.7 个百分点, 说明这几个模块的组合使用取得了良好的反应. 最后有了 ADown 下采样的加入, 使得整体性能又有了再一次的提升, 但是没有 MSaE 模块的加入时, DySample 模块并不能与 Adown 模块形成良好的配合, 是因为 DySample 模块上采样后的特征分布和语义信息无法很好地被 ADown 模块理解和利用, 然而加入 MSaE 注意力模块之后, 提高了主干网络的特征提取能力, 对后续的上采样和下采样提供了助力, 恰好解决了这一问题, 实现了检测精度的进一步提升.

为了进一步验证本文算法在多尺度场景下的检测效果, 分别使用改进前后的模型在 SVRDD 数据集上做了进一步的分析, 将目标框面积小于 32×32 像素的为小目标, 大于 96×96 的为大目标, 之间的为中等目标

进行实验.

实验结果如表 4 所示, 表中 *mAP*_s, *mAP*_m, *mAP*_l 分别代表小, 中, 大目标的 *mAP*@50:95 指标, 可以看出改进后的算法对于小目标和大目标提升幅度尤为明显, 说明本文提出的 MSaE 注意力模块以及其他结构上的改进卓有成效.

表 3 SVRDD 数据集消融实验结果

序号	Ghost-Conv	MSaE	ADown	DySample	*mAP*@50 (%)	*mAP*@50:95 (%)	参数量 (M)	计算量(GFLOPs)
1	—	—	—	—	68.7	40.0	20.1	58.3
2	√	—	—	—	70.7	41.6	20.04	58.2
3	—	√	—	—	70.5	42.3	20.61	68.4
4	—	—	√	—	70.0	41.3	19.33	**56.6**
5	—	—	—	√	69.7	42.0	20.11	58.3
6	√	—	—	√	70.7	42.3	20.05	58.2
7	√	—	√	√	68.8	41.1	**19.2**	**56.5**
8	√	√	—	√	71.0	42.7	20.57	68.3
9	√	√	√	√	**72.2**	**43.7**	19.71	66.6

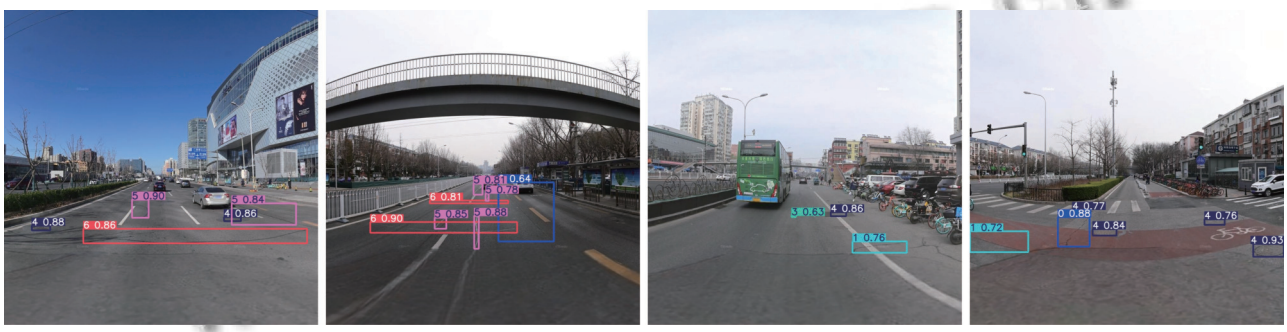
4.4 实验结果对比

为了体现本文算法的有效性, 从测试集中选取了图像进行检测, 检测结果如图 9 所示. 为确保识别的准确度, 置信度阈值设置为了 0.6, 即过滤掉置信度低于 0.6 的检测对象. 第 1 张图中, 改进前的模型并未能检测出细长的横向补丁, 第 2 张图中, 未能检测出右侧的纵向裂纹, 这两个都是大尺寸对象, 而在第 3 张和第 4 张图像中分别有一个很小的坑洼和一个较小的井盖没有被检测到, 改进后的模型则都成功识别到了上述的目标, 说明提出的 MSaE 模块以及其他的改进对于不同尺寸目标的检测效果有明显的提升.

表 4 不同尺寸目标检测精度对比

模型	*mAP*_s	*mAP*_m	*mAP*_l
RT-DETR-R18	0.199	0.346	0.458
本文算法	0.241	0.374	0.630

(a) 原RT_DETR检测结果



(b) 改进RT-DETR检测结果

图 9 街景图像道路缺陷检测

5 结论与展望

针对道路缺陷检测任务, RT-DETR 模型本身作为基于 Transformer 的目标检测模型, 避免了 NMS 等后处理问题, 减少了中间步骤的复杂性和人工干预. 本文在此基础上做了一系列的改进, 在主干网络中, 先是去掉了 4 个 Block 中的所有平均池化层, 同时针对多尺度场景进行改进, 提出新的 MSaE 注意力模块. 并在前两个 Block 中的 BasicBlock_A, BasicBlock_B 以及去掉平均池化层的 BasicBlock_C 中添加了 MSaE 注意力模块, 使得模型能够更多地关注到不同尺度的特征. 在高效混合编码器部分, 重点改进了 CCFF 模块, 使用 ADown 模块优化了下采样, 使用 GhostConv 卷积和 DySample 模块优化了上采样. 这一系列改进使得模型的复杂度略有提高, 但明显提高了检测精度. 不过改进

后的模型仍然存在一些不足, RT-DETR 本身存在的一些问题并没有得到有效解决. 由于使用了 Transformer, 因此在硬件资源受限的情况下, 它的速度会受到很大影响. 后续的工作将着重于简化模型结构, 尤其是编码器和解码器部分, 保持端到端优势的同时, 减小对硬件性能的依赖.

参考文献

1 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

2 Suh G, Cha YJ. Deep faster R-CNN-based automated detection and localization of multiple types of damage. Proceedings of the 2018 SPIE 10598 Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems. Denver: SPIE, 2018. 105980T.

3 Pang J, Zhang H, Feng CC, *et al*. Research on crack segmentation method of hydro-junction project based on target detection network. KSCE Journal of Civil Engineering, 2020, 24(9): 2731–2741. [doi: 10.1007/s12205-020-1896-y]

4 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

5 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6517–6525.

6 Farhadi A, Redmon J. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

7 Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal speed and accuracy of object detection. arXiv:2004.10934, 2020.

8 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision (ECCV 2016). Amsterdam: Springer, 2016. 21–37.

9 Ren M, Zhang XF, Chen X, *et al*. YOLOv5s-M: A deep learning network model for road pavement damage detection from urban street-view imagery. International Journal of Applied Earth Observation and Geoinformation, 2023, 120: 103335. [doi: 10.1016/j.jag.2023.103335]

10 高敏, 李元. 基于 YOLOv7-CA-BiFPN 的路面缺陷检测. 计算机测量与控制, 2024, 32(9): 9–14, 23.

11 Wang XQ, Gao HB, Jia ZM, *et al*. BL-YOLOv8: An improved road defect detection model based on YOLOv8. Sensors, 2023, 23(20): 8361. [doi: 10.3390/s23208361]

12 He QH, Li ZX, Yang WY. LMFE-RDD: A road damage detector with a lightweight multi-feature extraction network. Multimedia Systems, 2024, 30(4): 176. [doi: 10.1007/s00530-024-01367-z]

13 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

14 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with Transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.

15 Zhu XZ, Su WJ, Lu LW, *et al*. Deformable DETR: Deformable Transformers for end-to-end object detection. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

16 Zhao Y, Lv WY, Xu SL, *et al*. DETRs beat YOLOs on real-time object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16965–16974.

17 Narayanan M. SENetV2: Aggregated dense layer for channelwise and global representations. arXiv:2311.10807, 2023.

18 Wang CY, Yeh IH, Liao HYM. YOLOv9: Learning what you want to learn using programmable gradient information. Proceedings of the 18th European Conference on Computer Vision. Milan: Springer, 2024. 1–21.

19 Liu WZ, Lu H, Fu HT, *et al*. Learning to upsample by learning to sample. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. 2023. Paris: IEEE, 2023. 6004–6014.

20 Han K, Wang YH, Tian Q, *et al*. GhostNet: More features from cheap operations. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 1577–1586.

21 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Vegas: IEEE, 2016. 770–778.

22 Ren M, Zhang XF, Zhi XB, *et al*. An annotated street view image dataset for automated road damage detection. Scientific Data, 2024, 11(1): 407. [doi: 10.1038/s41597-024-03263-7]

(校对责编: 张重毅)