Geomatics Science and Technology 测绘科学技术, 2023, 11(2), 122-134
Published Online April 2023 in Hans. https://www.hanspub.org/journal/gst
https://doi.org/10.12677/gst.2023.112014

融合空间结构权重优化注意力机制的建筑物立面元素检测

江　涛[1]，常莉红[2]，魏　征[3,4,5*]，董　震[1]

[1]武汉大学测绘遥感信息工程国家重点实验室，湖北　武汉
[2]武汉大学遥感信息工程学院，湖北　武汉
[3]国家海洋局南海规划与环境研究院，广东　广州
[4]自然资源部海洋环境探测技术与应用重点实验室，广东　广州

收稿日期：2023年3月5日；录用日期：2023年4月18日；发布日期：2023年4月24日

摘　要

本文针对街景图像立面元素检测问题，提出了融合空间结构权重优化注意力机制的立面元素目标检测网络。在主干网络部分使用嵌入基于空间结构优化坐标注意力机制的C3模块，增加横纵坐标权重分支，有效利用空间结构编码信息，提升立面元素定位精度；其次针对立面最主要组成元素窗户、阳台的小目标特性，使用改进的递归门控卷积模块替换原始卷积模块，融合丰富的多尺度上下文信息，并增加小目标检测分支，提升检测精度；最后设计了ECIOU损失同时对检测框的长宽比以及定位中心进行监督，增强网络对立面元素的感知能力，提升网络收敛速度。在FacadeWHU数据集上实验结果表明，本文模型的平均精度比相较于基线网络Yolov5s而言整体平均精度提升了16.4%,窗户目标的平均精度提升了22.4%，阳台目标的平均精度提升了25.5%，可以有效检测立面元素，更好的服务于病害检测、能耗分析等下游任务。

关键词

立面解析，建筑物立面，立面元素检测

Building Facade Element Detection Based on Spatial Structure Weight Optimization Attention Mechanism

Tao Jiang[1], Lihong Chang[2], Zheng Wei[3,4,5*], Zhen Dong[1]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan Hubei

[*]通讯作者。

文章引用: 江涛, 常莉红, 魏征, 董震. 融合空间结构权重优化注意力机制的建筑物立面元素检测[J]. 测绘科学技术, 2023, 11(2): 122-134. DOI: 10.12677/gst.2023.112014

[2]School of Remote Sensing and Information Engineering, Wuhan University, Wuhan Hubei
[3]South China Sea Institute of Planning and Environmental Research, Ministry of Natural Resources, Guangzhou Guangdong
[4]Key Laboratory of Marine Environmental Survey Technology and Application, Ministry of Natural Resource, Guangzhou Guangdong

Received: Mar. 5th, 2023; accepted: Apr. 18th, 2023; published: Apr. 24th, 2023

Abstract

Aiming at the problem of facade element detection in street view image, this paper proposes a facade element object detection network integrating spatial structure weight optimization mechanism. C3 module embedded in the coordinate attention mechanism based on spatial structure optimization is used in the backbone network to increase the weight branches of horizontal and vertical coordinates, effectively use the spatial structure coding information, and improve the positioning accuracy of elevation elements. Secondly, in view of the small target characteristics of Windows and balconies, which are the main components of the facade, an improved recursive gated convolutional module is used to replace the original convolutional module, integrate rich multi-scale context information, and add small target detection branches to improve detection accuracy. Finally, ECIOU loss is designed to supervise the aspect ratio of the detection frame and the positioning center, which enhances the perception ability of the opposite elements of the network and improves the convergence speed of the network. Experimental results on Facade WHU data set show that compared with baseline network yolov5s, the average accuracy of the proposed model is improved by 16.4% overall, 22.4% for window target and 25.5% for balcony target, which can effectively detect facade elements. Better service for disease analysis, energy consumption analysis and other downstream tasks.

Keywords

Facade Parsing, Building Facade, Facade Elements Detection

1. 引言

近年来随着以人工智能、5G、区块链、物联网和量子计算为代表的 5 大关键技术的诞生与发展[1]，数字孪生城市建设迈入新阶段，实景三维中国战略应运而生。在 2021 年颁布的实景三维中国建设大纲[2]中明确指出“构建部件三维模型用于精准表达和按需定制，服务个性化应用。”部件三维模型包括建(构)筑物结构部件、建筑室内部件、道路设施部件、地下空间部件[2]等精细结构信息，在室内外智能导航、自动驾驶、智慧交通、城市管理、城市规划、应急避险、能耗模拟、应急服务[3]等领域具有广泛的应用前景。其中，立面元素作为城市建筑的主要外部结构部件，不仅是实景部件三维模型构建过程中不可或缺的一环，而且是噪声分析、采光分析、能耗分析、遮挡分析、视野分析[3]等一系列下游任务的重要核

心要素，因此如何提高立面元素的位置参数、尺寸参数、语义类别等属性解析的精度仍是国内外学者的研究热点之一。

传统的立面解析方法大多采用基于建筑立面几何先验知识的图像处理方法，得到逐像素的语义分割结果。一类方法从直接推理立面的形式语法树结构，并结合优化算法不断逼近真值标签。Teboul 等人[4]利用基于随机森林算法得到的建筑立面影像初步分割结果，构建形式语法树，并结合随机游走算法得到最优树结构，可同时得到立面的语义与结构信息，但随机游走优化计算过程较慢。为解决这一问题，Teboul 等人[5]对原算法进行改进，结合递归二进制分裂语法，利用层次马尔可夫决策过程表述立面结构的优化过程，使用引入状态聚合的强化学习方法进行求解，速度显著提高。上述算法在推理语法树结构时需要用户或专家提前对形式语法提前预设，然而一套形式语法无法涵盖所有立面风格。为解决此问题，Gadde [6]等人对已有立面的真值解析树结构进行优化，合并重复子树以减少规则数量，进而采用无监督聚类对复杂规则进行合并，经过上述过程可以得到表达能力强且泛化性高的基元语法，可以使得分割步骤快速收敛，同时具有较高的精度。

另一类方法利用立面的对称性、重复性、规则性等典型人工设计风格属性对分割过程施加约束。Pascal Muller 等人[7]利用互信息提取相似立面结构并解析其周期分布规律，进而建立语法树实现立面模型重建。Changchang Wu 等人[8]基于未校正立面影像的主要重复元素沿灭点方向分布且立面元素沿图像中心垂线对称分布的假设，使用灭点检测与 SIFT 描述符对重复元素初始提取，设计重复模式质量评价标准作为对原始结果优化过程的监督。Cohen 等人[9]通过立面的对称性与重复性检测，在改善原有立面分割效果的同时，也可识别、处理与修复被遮挡的立面，取得了一定效果。Xiao [10]等人基于重复结构经常沿水平和垂直方向对齐的观察，通过提取经过立面重复结构边缘的基准线实现重复结构检测。

第三类方法一般将立面解析定义为立面的语义分割任务，借助机器学习方法予以解决。Ali 等人[11]针对窗户检测，提取多尺度哈尔小波特征输入 AdaBoost 分类器检测窗户包围盒，但由于该特征提取器并非专为窗户设计，检测效果一般。高云龙等人[12]提出利用模板匹配自动选择窗户样本训练 JointBoost 分类器以提取窗户，并引入走向线、倾向线、兴趣点以及相似度四个约束构建的规则性模型对原始结果后优化，可以处理复杂背景以及未经校正的立面影像。

尽管以上方法取得了一定效果，但仍然存在一些问题：一方面，基于几何先验知识或规则的方法一般针对特定数据集设计，受数据集影响较大，不具备普适性。另外一方面，计算机视觉传统方法对噪声敏感、鲁棒性较低，适用范围有限。

近年来，随着深度学习在计算机视觉等领域应用的逐渐深入，一些学者尝试在立面解析任务上引入深度学习方法。Hantang Liu [13]等人首次将深度学习方法应用于立面解析任务，基于立面元素为方形的先验假设及对称性约束提出对称损失并使用区域推荐网络对初始目标检测结果精化，取得了较为先进的效果。Hensel 等人采用 Faster R-CNN 检测立面元素，并使用混合整数线性优化思想对初始检测框排布对齐。Yanwei Sun [14]通过在 MaskR-CNN 基础上添加 RPN 区块注意力模块，提升网络的区域感知能力，从而提升立面元素的实例分割效果。CK Li [15]等人借鉴人体姿态检测的思想，将立面元素中的窗户重建定义为关键点检测与节点连接问题，取得了较好的重建效果。

基于深度学习的立面元素解析技术在精度、速度以及鲁棒性等方面都有显著提升，然而现有方法多使用正视角拍摄或者经过校正的立面影像，较少使用街景影像，样本制作流程复杂且获取成本较高。街景影像使用车载相机获取，可低成本快速获取大量建筑物立面样本，且更贴近真实场景，是当前各大场景解析任务的重要数据来源。受采集条件影响，街景影像往往视角多样、光照变化剧烈且背景复杂多样，检测难度较大，现有检测算法存在错检、漏检、定位框不准确的问题。因此本文针对街景建筑立面影像

目标检测问题，提出融合空间结构权重优化坐标注意力机制的立面元素目标检测网络。本文的主要工作概括如下：

1) 提出嵌入基于空间结构权重优化注意力机制的 C3 模块，增加横纵坐标权重调整分支，提升坐标感知能力，提升网络定位精度。

2) 在颈部网络部分，使用改进的递归门控卷积块代替原始卷积，充分融合丰富的多尺度全局上下文信息，同时增加小目标检测分支，改善网络对小型目标的检测效果。

3) 对原始损失函数进行改进，提出 ECIOU 损失，同时对检测框中心点位置以及宽高进行约束，增强网络对立面元素的感知能力，加速网络收敛。

2. 算法设计

本文提出的立面元素检测算法整体流程如图 1 所示，共分为三个部分：输入端、网络主体、预测输出端。输入端采用 Mosaic 与 MixUp 方法进行数据增强，平衡不同尺寸立面目标样本数量，同时增加背景丰富度。网络主体部分如图 2 所示，其中主干网络沿用 CSPDarknet [16]结构进行特征提取，在 C3 模块中嵌入基于空间结构权重优化的坐标注意力机制；颈部网络采用路径增强网络[17]结构，其卷积模块替换为改进的递归门控卷积并增加小目标检测分支以提升小目标检测效果；预测输出端分类目标框类型并回归其坐标后采用非极大值抑制算法消除冗余候选预测框，得到最终结果。

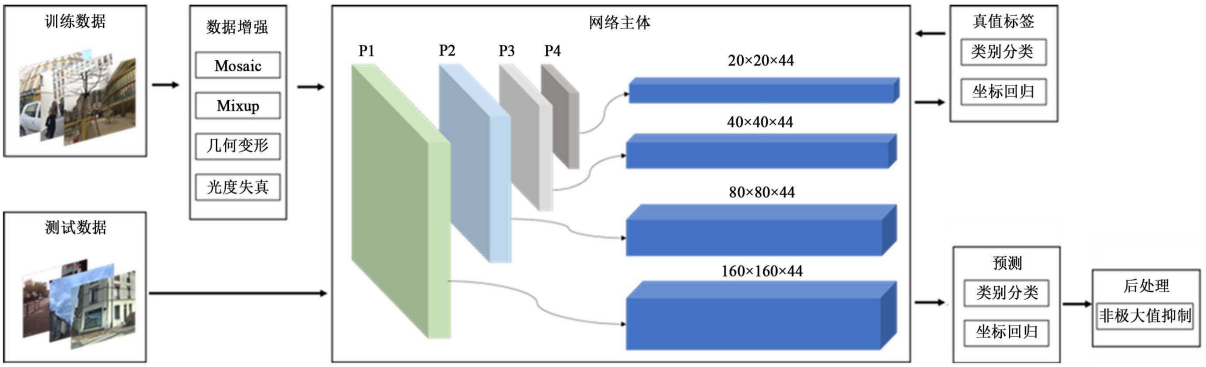

Figure 1. Pipeline of the proposed method
图 1. 本文算法流程

2.1. 基于空间结构权重优化注意力机制的 C3 模块

注意力机制是计算机视觉领域基于人脑注意力机制提出的一种特殊结构，其基本思想是基于原始输入数据提取其关联性，突出重要特征并忽略无关噪声信息，从而提升特征纯度。其中坐标注意力机制[18]是一种典型方法，其核心操作是原始输入特征的水平方向以及垂直方向使用两个一维全局池化分别进行聚合加权，计算得到两个独立的方向感知特征矩阵，用于获取输入特征图在指定方向上的远距离空间依赖权重编码。由于方向信息的嵌入，网络具备更强的空间结构特征感知能力，从而能够精准定位感兴趣目标位置，提升模型检测精度。

然而由于坐标注意力机制获取两个编码分量时仅在单一方向上获取空间感知权重，融合时将二者相乘得到优化特征图，未考虑不同方向空间结构信息权重占比，融合方式较为简单，损失了部分特征感知性能。针对此问题，本文增加方向注意力权重融合分支，通过学习不同方向的空间感知权重，优化原始方向编码，增强坐标感知能力，提升网络特征提取模块性能。

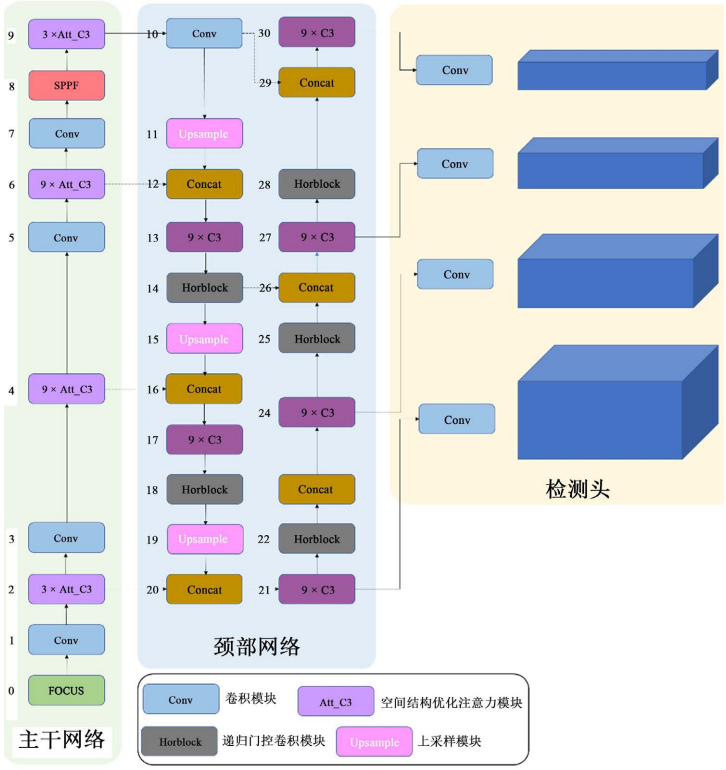

Figure 2. Framework of the proposed network
图 2. 主体网络结构

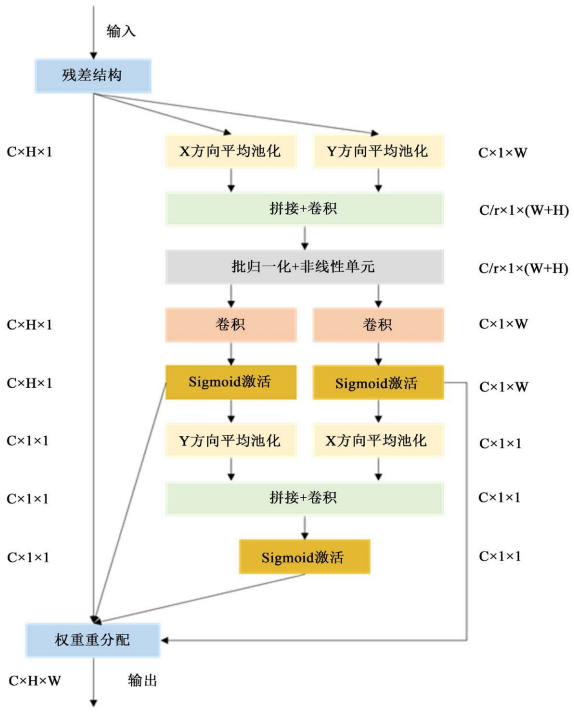

Figure 3. C3 module based on spatial structure weight optimization attention mechanism
图 3. 基于空间结构权重优化注意力机制的 C3 模块

通道注意力机制[19]中，全局池化可编码全局空间信息，但由于其将空间信息压缩为单一通道描述子，难以保持位置信息。因此坐标注意力机制将全局池化拆分为两个一维特征编码操作，以使注意力模块保留精确位置信息，即分别沿水平与垂直坐标方向进行编码：

$$l_c^h(h)=\frac{1}{W}\sum_{0\le i\le W}x_c(h,i)$$

$$l_c^w(w)=\frac{1}{H}\sum_{0\le j\le H}x_c(j,w)$$

获得具有精确编码信息的两个方向特征之后，将其拼接之后送入 1 × 1 卷积 F_2 得到表示空间编码信息的中间特征 f：

$$f=\sigma\left(F_2\left(\left[l^h,l^w\right]\right)\right),\ f\in R^{\frac{c}{r}\times(H+W)}$$

再将中间特征 f 沿空间维度拆分为 $f^h\in R^{\frac{C}{r}\times H}$、$f^w\in R^{\frac{C}{r}\times W}$，并使用 1 × 1 卷积 F_h、F_w 生成注意力分量：

$$g^h=\sigma\left(F_h\left(f^h\right)\right)$$

$$g^w=\sigma\left(F_w\left(f^w\right)\right)$$

以上分量乘积即为原始坐标注意力结果，此处为获得各方向结构信息权重编码，我们将两分量分别沿水平与垂直方向平均池化：

$$p_c^h(h)=\frac{1}{W}\sum_{0\le i\le W}g^h(h,i),h=1$$

$$p_c^w(w)=\frac{1}{H}\sum_{0\le i\le W}g^w(i,w),w=1$$

接着再次使用 1 × 1 卷积 F_2 得到空间结构优化描述子：

$$g^r=\sigma\left(F_2\left(\left[p^h,p^w\right]\right)\right)$$

最终如图 3 所示使用空间结构优化描述子、初始注意力权重以及原始特征图计算注意模块的最终输出：

$$y_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j)\times g_c^r(j)$$

优化之后的特征图带有方向加权编码信息，可使网络对感兴趣目标定位更加准确。

2.2. 基于递归门控卷积的特征融合模块

浅层特征具有较高分辨率且包含更多空间结构细节，但由于感受野限制，语义表征能力不高；深层特征分辨率较低且容易丢失小型目标的关键位置信息，但经过网络多层抽象包含更多的高级语义信息，常用方式为在网络中构建特征金字塔模块融合多层次特征。Yolov5s 沿用 FPN + PAN 结构，FPN 层自顶向下传达强语义特征，而特征金字塔则自底向上传达强定位特征，从而将主干网络不同层参数进行聚合，实现多尺度特征融合。为获得更好的多尺度融合特征，本文针对颈部特征融合网络进行改进:使用递归门控卷积块代替原始卷积并在原始 Yolov5s 模型三个检测头基础上增加一个检测头，即生成四种不同尺度检测头，分别用于街景立面影像中极小目标、小目标、中目标以及大目标检测，增强窗户、阳台等小型目标检测效果。由于窗户、阳台等目标在立面元素中数量占比较大，改进此类目标检测效果可有效提升

整体检测效果。

递归门控卷积[20] (g^nConv)通过门控卷积和递归设计来执行高阶空间交互，具有高度灵活性和可定制性，兼容各种卷积变量，并将自注意力两阶交互扩展到任意阶，而不引入显著的额外计算。门控卷积计算过程第一步将输入特征 f_{in} 经过线性层映射得到中间分量 p_0、q_0：

$$\left[p_0^{HW\times C}, q_0^{HW\times C}\right] = Linear\left(f_{in}\right) \in R^{HW\times 2C}$$

p_0 使用深度可分离卷积 F 后再与 q_0 做点积得到 p_1：

$$p_1 = F\left(q_0\right) \odot p_0 \in R^{HW\times C}$$

p_1 再次经过线性层映射得到门控卷积输出 f_{out}：

$$f_{out} = L\left(p_1\right) \in R^{HW\times C}$$

为提取高层次空间信息，使用递归思想将门控卷积扩展到 N 阶。首先将原始输入特征 f_{in} 经过线性映射得到各中间分量：

$$\left[p_0^{HW\times C_0}, q_0^{HW\times C_0}, \cdots, q_{n-1}^{HW\times C_{n-1}}\right] = Linear\left(f_{in}\right) \in R^{HW\times C_0 + \sum_{0\le k\le n-1} C_k}$$

然后递归执行门控卷积，并将其输出缩放为 $1/\alpha$ 来稳定训练：

$$p_{k+1} = f_k\left(q_k\right) \odot g_k\left(p_k\right)/\alpha, k = 0,1,2,\cdots,n-1$$

其中 f_k 为一系列深度可分离卷积层，g_k 用于按不同顺序匹配输入特征维度：

$$f\left(x\right) = \begin{cases} Identity, & k = 0 \\ Linear\left(C_{k-1}, C_k\right), & 0 \le k \le n-1 \end{cases}$$

同时为确保高阶交互不会引入过多计算开销，将每一阶信道维度设置为：

$$C_k = \frac{C}{2^{n-k-1}}, 0 \le k \le n-1$$

将最后一层输出 q_n 经过线性映射层即可得到递归门控卷积输出。

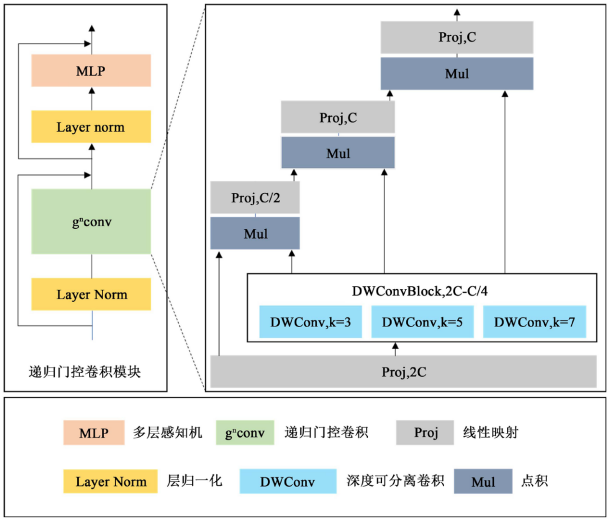

Figure 4. Improved recursive gated convolution module
图 4. 改进的递归门控卷积模块

如图 4 所示，为更有效融合不同层次特征图的多尺度全局上下文信息，我们将原始递归门控卷积中的深度可分离卷积设置为三组模块：每个模块分别使用 k = 3、5、7 三种尺寸的深度可分离卷积核作用于原始输入特征，并以连乘形式串联各核特征计算最终输出。

2.3. ECIOU 损失

Yolov5s 中将预测框与真实框之间的交并比(IOU)拓展得到的 IOU 损失作为位置回归的损失函数，不再使用 smooth-l1 损失函数，其公式为：

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha\upsilon$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$$

$$\alpha = \frac{\upsilon}{\left(1-\mathrm{IOU}\right)+\upsilon}$$

其中 $\rho^2\left(b, b^{gt}\right)$ 为预测框与真实框之间的欧氏距离， c 为能够同时包含预测框与真实框的最小外包矩形区域的对角线距离， w^{gt} 、 h_{gt} 、 w 、 h 分别为预测框与真实框的宽高。

原始 IOU 损失对尺度不敏感，即对于尺寸不一致的预测框，预测框与真实框重叠程度不同但 IOU 值却可能相同，无法加以区分。CIOU 损失[21]通过在附加项中对预测框与真实框的宽高比进行约束，解决了尺度不敏感问题。但由于其计算公式中 v 项仅反映纵横比差异，并非宽高分别与其置信度的真实差异，因此部分情况下会阻碍模型有效优化相似性。针对这一问题，Yi-Fan Zhang [22]等人将 CIOU 中的纵横比拆开，提出了 EIOU 损失，其表达式如下：

$$L_{\mathrm{EIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(C^h\right)^2} + \frac{\rho^2\left(w, w^{gt}\right)}{\left(C^w\right)^2}$$

其中 C_w 与 C_h 为预测框与真实框的最小外包矩形区域的宽高。

EIOU 损失通过将纵横比损失项拆分成宽高预测值分别与最小外接框宽高差值，同时加速了收敛提高了回归精度，此外引入的 Focal Loss 可优化边界框回归任务中的样本不平衡问题，即减少与目标框重叠较少的大量锚框对包围框回归的优化贡献，使回归过程专注于高质量锚框。但由于对于中心点位置约束的降低，造成目标框中心点位置的定位精度下降。针对以上问题，对位置回归损失进行改进，对预测框的中心点位置以及宽高同时进行惩罚以提升网络定位精度，提出 ECIOU 损失，其表达式如下所示：

$$L_{\mathrm{CIOU}} = 1 - \mathrm{IOU} + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha\upsilon + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \frac{\rho^2\left(h, h^{gt}\right)}{\left(C^h\right)^2} + \frac{\rho^2\left(w, w^{gt}\right)}{\left(C^w\right)^2}$$

$$\upsilon = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$$

$$\alpha = \frac{\upsilon}{\left(1-\mathrm{IOU}\right)+\upsilon}$$

网络分类损失与置信度损失保持不变，最终损失函数表达式为：

$$L_{loss} = L_{cls} + L_{obj} + L_{\mathrm{ECIOU}}$$

3. 实验与分析

3.1. 实验环境

本文使用 Windows 10 操作系统，处理器为 Intel(R) Core(TM) i9-10900F CPU@ 2.80 GHz 2.81 GHz，32 G RAM，软件环境为 Python 3.7、VsCode，使用 Pytorch 作为本文深度学习框架，所有模型均在 NVIDIA GeForce RTX 3080ti GPU 中运行，显存为 12 G。模型输入图像尺寸均为 1024 × 1024，批次大小为 20，设置训练 500 轮次，初始学习率为 0.01，终止学习率为 0.2，采用随机梯度下降策略，动量和权重衰减分别为 0.937 和 0.0005，使用早停机制。使用表 1 所示方式进行数据增强：

Table 1. Data augmentation methods
表 1. 数据增强方式

数据增强方式及参数							
混合增强		HSV 空间增强			平移	缩放	水平旋转
Mosaic	Mixup	hsv_h	hsv_s	hsv_v			
1	1	0.02	0.6	0.4	0.2	0.5	0.5

3.2. 实验数据集

本文在武汉大学发布的 Facade WHU [23]街景立面数据集上进行了训练和测试，该数据集包含 900 张街景影像，其中 850 张来自法国巴黎、50 张来自挪威特隆赫姆，训练集验证集测试集的比例为 8:1.5:0.5，可兼容立面语义分割、实例分割以及目标检测等多种任务。Facade WHU 数据集是第一个使用街景影像构建的立面影像数据集，由不同设备在不同时间、角度进行拍摄，包含多种建筑风格的建筑物、多种天气、季节、光照条件等，更加贴近真实场景。

3.3. 评价指标

为评估模型性能，本文选取准确率(Precision)、召回率(Recall)、平均精度(Average Precision, AP)、以及平均精度均值(Mean Average Precision)作衡量本文算法的评价指标。各指标计算方式如下：

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \cdot 100\%$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \cdot 100\%$$

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

$$mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$$

其中，N_{TP} (true positives)表示正确检测的目标数量，N_{FP} (false positives)表示错误检测的目标数量，N_{FN} (number of false negatives)表示漏检目标数量，N_C (number of classes)为类别数量，AP (average precision)为各类别的 P-R 曲线积分值，该指标被用来评估模型对于单个类别的目标检测性能表现，mAP (mean average precision)为所有类别的 AP 平均值，取值范围为[0, 1]，mAP 值越接近于 1 表示模型的性能越好，检测能力越强。本文主要以 mAP^{50}、$mAP^{50:95}$ 以及各个类别的 AP 为测试评估指标，其中 $mAP^{50:95}$ 表示 IoU = 0.5:0.95，步长为 0.05 时的平均精度，mAP^{50} 表示 IoU = 0.5 时的平均精度。

3.4. 实验结果

为验证本文第一小节各算法模块对于网络性能的影响，以 yolov5s 为基线模型在 FacadeWHU 数据集上对各模块进行消融实验，实验结果如表 2 所示。

Table 2. Ablation study results
表 2. 消融实验结果

CA	Gnconv Neck	*ECIOU* loss	mAP^{50}	$mAP^{50:95}$	AP^{50}_{window}
			0.417	0.226	0.578
√			0.531	0.268	0.695
	√		0.428	0.242	0.580
		√	0.435	0.228	0.582
√	√		0.568	0.296	0.792
√		√	0.572	0.307	0.796
	√	√	0.530	0.278	0.687
√	√	√	0.581	0.314	0.802

在 Yolov5s 算法基础上依次添加 AttC3、HorBlock 以及 ECIOU 损失。从表 2 中可以看出，单独添加基于空间结构权重优化坐标注意力机制的 C3 模块， mAP^{50} 和 $mAP^{50:95}$ 分别提升了 11.4%和 4.2%，提升效果显著；单独添加 Gnconv block 和 ECIOU 损失之后，两项指标分别提升 1.1%和 1.6%以及 1.8%和 0.2%，均有一定提升。组合使用任意两项改进策略相较于单一改进策略各项指标均有一定程度提升，而同时使用三种改进模块后相较于原始网络以及单一模块或任意两种模块组合的检测效果大幅提升，三种改进策略可以相辅相承共同促进网络检测效果的提升。

使用 grad-CAM 算法对空间结构权重优化坐标注意力机制添加前后网络对窗户输出热力图可视化，结果如图 5 所示，使用该模块后网络对窗户的关注区域更加集中，定位更加准确，验证了该模块对于网络特征提征提取以及目标定位的促进作用。

(a) 原图　(b) 原始C3模块　(c) 改进C3模块

Figure 5. Heatmap before and after network improvement
图 5. 主干网络改进前后热力图

图 6 为使用 ECIOU loss 前后训练阶段损失曲线，使用 ECIOU loss 时由于网络引入早停机制，为防止出现过拟合训练轮次为 372 时判定网络收敛自动停止，而使用原始损失时未触发早停机制，且根据图 6 可知使用 ECIOU loss 后网络收敛速度明显快于原始网络。

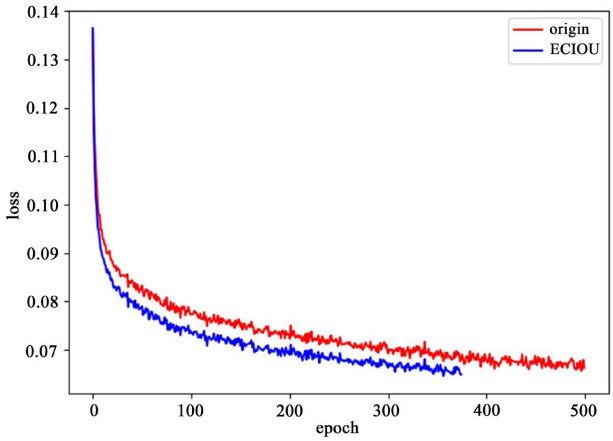

Figure 6. Comparison of training loss
图 6. 训练损失曲线对比图

为定性验证本文算法检测效果，如图 7 所示选取测试集中四种典型外部环境条件下的检测结果进行展示。在强光条件、弱光条件、视角变化剧烈区域等情况下均有较好表现，但在光照变化剧烈区域以及受树木遮挡区域会出现少部分漏检，整体效果良好。

Figure 7. The detection results in different environment conditions
图 7. 不同条件下本文模型检测效果

为测试本文算法立面元素检测效果，将本文算法与目前较为流行的目标检测算法进行定量对比，其中包括两阶段目标检测算法 Faster R-CNN、单阶段目标检测算法 SSD、RetinaNet、Yolov3 以及基线方法 Yolov5s，定量实验结果如表 3 所示。由表 3 可知，本文算法相对于一阶段目标检测算法以及二阶段目标检测算法都表现出了良好性能，对比其余网络本文算法对于立面元素检测所有指标均为最高；对比基线网络 yolov5s，mAP^{50} 提升 16.6%，各类别 *AP* 值均有提升，其中窗户和阳台两类密集小型目标提升最为显著，分别提升 22.4%与 25.5%。

Table 3. The detection results of different algorithms
表 3. 不同算法效果对比

方法	mAP^{50} /%	*AP*					
		Window	Balcony	Door	shop	wall	roof
Faster R-CNN [24]	0.415	0.585	0.389	0.436	0.332	0.401	0.347
SSD [25]	0.386	0.513	0.357	0.416	0.315	0.394	0.326
RetinaNet [26]	0.412	0.581	0.384	0.431	0.330	0.402	0.341
Yolov3 [16]	0.389	0.522	0.362	0.417	0.317	0.397	0.324
Yolov5s	0.417	0.578	0.399	0.453	0.327	0.412	0.332
本文方法	0.581	0.802	0.654	0.682	0.395	0.582	0.369

4. 结束语

针对街景图像立面元素检测问题，本文提出了基于空间结构权重优化注意力机制的建筑物立面元素检测网络，在主干网络中使用基于空间结构权重优化坐标注意力机制的 C3 模块，可有效利用空间信息，学习更多可区分特征，适用于街景影像立面元素检测场景；在颈部网络引入递归门控卷积模块，同时增加小尺度检测分支，充分融合全局多尺度上下文信息，有效提升小型目标的检测效果；引入 ECIOU 损失，同时对检测框的位置以及宽高进行约束，进一步加快模型收敛速度。实验结果表明，本文提出的改进算法 mAP^{50} 可达 58.1%，$mAP^{50:95}$ 可以达到 31.4%，优于当前主流的目标检测算法。不过本文算法在前景遮挡严重、光照剧烈变化等极端条件下会出现漏检，后续研究会对此类情形进行针对性设计。

基金项目

自然资源部超大城市自然资源时空大数据分析应用重点实验室开放基金资助(编节 KFKT-2022-01)。

参考文献

[1] 傅一平. 智慧城市必不可少的五大关键技术[J]. 计算机与网络, 2020, 46(11): 44-45.

[2] 赵玲玲. 《实景三维中国建设技术大纲(2021 版)》印发[J]. 资源导刊, 2021(8): 6.

[3] Klimkowska, A., Cavazzi, S., Leach, R., *et al*. (2022) Detailed Three-Dimensional Building Façade Reconstruction: A Review on Applications, Data and Technologies. *Remote Sensing*, **14**, 2579. https://doi.org/10.3390/rs14112579

[4] Teboul, O., Simon, L., Koutsourakis, P., *et al*. (2010) Segmentation of Building Facades Using Procedural Shape Priors. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13-18 June 2010, 3105-3112. https://doi.org/10.1109/CVPR.2010.5540068

[5] Teboul, O., Kokkinos, I., Simon, L., *et al*. (2011) Shape Grammar Parsing via Reinforcement Learning. *CVPR* 2011, Colorado Springs, 20-25 June 2011, 2273-2280. https://doi.org/10.1109/CVPR.2011.5995319

[6] Gadde, R., Marlet, R. and Paragios, N. (2016) Learning Grammars for Architecture-Specific Facade Parsing. *International Journal of Computer Vision*, **117**, 290-316. https://doi.org/10.1007/s11263-016-0887-4

[7] Müller, P., Zeng, G., Wonka, P., *et al*. (2007) Image-Based Procedural Modeling of Facades. *ACM Transactions on Graphics*, **26**, 85.

[8] Wu, C.C., Frahm, J.-M. and Pollefeys, M. (2010) Detecting Large Repetitive Structures with Salient Boundaries. 11*th European Conference on Computer Vision*, Heraklion, 5-11 September 2010, 142-155. https://doi.org/10.1007/978-3-642-15552-9_11

[9] Cohen, A., Oswald, M.R., Liu, Y., *et al*. (2017) Symmetry-Aware Façade Parsing with Occlusions. 2017 *International Conference on* 3*D Vision* (3*DV*), Qingdao, 10-12 October 2017, 393-401. https://doi.org/10.1109/3DV.2017.00052

[10] Xiao, H., Meng, G., Wang, L., *et al*. (2018) Facade Repetition Detection in a Fronto-Parallel View with Fiducial Lines Extraction. *Neurocomputing*, **273**, 435-447. https://doi.org/10.1016/j.neucom.2017.07.040

[11] Ali, H., Seifert, C., Jindal, N., *et al*. (2007) Window Detection in Facades. 14*th International Conference on Image Analysis and Processing* (*ICIAP* 2007), Modena, 10-14 September 2007, 837-842. https://doi.org/10.1109/ICIAP.2007.4362880

[12] 高云龙, 张帆, 屈孝志, 黄先锋, 崔婷婷. 结合样本自动选择与规则性约束的窗户提取方法[J]. 武汉大学学报(信息科学版), 2018, 43(3): 436-443.

[13] Liu, H., Xu, Y., Zhang, J., *et al*. (2020) DeepFacade: A Deep Learning Approach to Facade Parsing with Symmetric Loss. *IEEE Transactions on Multimedia*, **22**, 3153-3165. https://doi.org/10.1109/TMM.2020.2971431

[14] Sun, Y., Malihi, S., Li, H., *et al*. (2022) DeepWindows: Windows Instance Segmentation through an Improved Mask R-CNN Using Spatial Attention and Relation Modules. *ISPRS International Journal of Geo-Information*, **11**, 162. https://doi.org/10.3390/ijgi11030162

[15] Li, C.K., Zhang, H.X., Liu, J.X., *et al*. (2020) Window Detection in Facades Using Heatmap Fusion. *Journal of Computer Science and Technology*, **35**, 900-912.

[16] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement.

[17] Wang, W., Xie, E., Song, X., *et al*. (2019) Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 8439-8448. https://doi.org/10.1109/ICCV.2019.00853

[18] Hou, Q., Zhou, D. and Feng, J. (2021) Coordinate Attention for Efficient Mobile Network Design. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13708-13717. https://doi.org/10.1109/CVPR46437.2021.01350

[19] Hu, J., Shen, L. and Sun, G. (2017) Squeeze-and-Excitation Networks. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[20] Rao, Y., Zhao, W., Tang, Y., *et al*. (2022) HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. 36*th Conference on Neural Information Processing Systems* (*NeurIPS* 2022), New Orleans, 28 November-9 December 2022, 10353-10366.

[21] Zheng, Z., Wang, P., Ren, D., *et al*. (2022) Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on Cybernetics*, **52**, 8574-8586. https://doi.org/10.1109/TCYB.2021.3095305

[22] Zhang, Y., Ren, W., Zhang, Z., *et al*. (2022) Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *Neurocomputing*, **506**, 146-157. https://doi.org/10.1016/j.neucom.2022.07.042

[23] Kong, G.F. and Fan, H.C. (2021) Enhanced Facade Parsing for Street-Level Images Using Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, **59**, 10519-10531. https://doi.org/10.1109/TGRS.2020.3035878

[24] Ren, S., He, K., Girshick, R., *et al*. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[25] Liu, W., Anguelov, D., Erhan, D., *et al*. (2015) SSD: Single Shot MultiBox Detector. 14th *European Conference Computer Vision*, Amsterdam, 11-14 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[26] Lin, T.Y., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 2999-3007. https://doi.org/10.1109/ICCV.2017.324