中图法分类号： 文献标识码： 文章编号：1006-8961(XXXX)XX-0001-15

论文引用格式：Jin Yuan, Chi Fang'ai, Liu Yingao, Li Yane. XXXX. Study on dougong detection method of mask r-cnn swin with integrated attention mechanism. Journal of Image and Graphics, XX(XX):0001-0015(金远, 池方爱, 刘迎奥, 李颜娥. XXXX. Mask R-CNN Swin融合注意力机制的斗拱检测方法. 中国图象图形学报, XX(XX):0001-0015)[DOI:10.11834/jig.250460]

Mask R-CNN Swin融合注意力机制的斗拱检测方法

金远[1],池方爱[2],刘迎奥[1],李颜娥[1*]

1. 浙江农林大学数学与计算机科学学院,杭州 311300; 2. 长江大学城市建设学院,荆州 434000

摘 要：目的 斗拱作为中国古代木构建筑的重要承重与装饰构件,其形态复杂、方向特征显著,对自动化识别与数字化保护提出了更高要求。为提升斗拱及破损斗拱的识别精度与鲁棒性,研究基于深度学习与Transformer架构构建了一种改进型Mask R-CNN斗拱识别模型。**方法** 以Mask R-CNN_Swin为基础框架,构建包含转角斗拱与非转角斗拱的图像数据集(共3014张样本,来源于灵隐寺、法喜寺等古建筑实景及公开网络资源)。针对斗拱结构的方向性特征与复杂背景干扰问题,提出方向坐标注意力机制,通过在水平与垂直方向上独立建模显著性特征,有效增强模型对方向敏感结构的表征能力;同时引入一维压缩激励机制,以实现通道特征的轻量化重标定,强化全局与局部特征间的依赖建模。**结果** 实验结果表明,改进模型SwinE_AM_DIRCA在非转角斗拱与转角斗拱识别中的精度分别达到94.4%与89.2%,较基础Swin_AM模型分别提升6.7%与8.9%,较Mask R-CNN_ResNet50模型分别提升10.8%与20.4%;在破损斗拱识别任务中亦取得显著性能提升。**结论** 本研究通过引入方向敏感建模与通道压缩激励相结合的注意力机制,显著增强了模型对斗拱结构特征的表达能力与方向判别能力。创新点在于提出了基于方向解耦的坐标注意力机制DIRCA(Directional Coordinate Attention)与一维压缩激励机制1D-SE(1D-Squeeze-and-Excitation)的协同设计框架,实现了方向性特征建模与轻量化通道增强的统一,为古建筑构件的智能识别、自动修复及文化遗产数字化保护提供了可行路径与技术支撑。

关键词： DIRCA;深度学习;文化遗产;Transformer;古建筑

Study on dougong detection method of mask r-cnn swin with integrated attention mechanism

Jin Yuan[1], Chi Fang'ai[2], Liu Yingao[1], Li Yane[1*]

1. College of Mathematics and Computer Science, Zhejiang A&F University, Hangzhou 311300, China; 2. School of Urban Construction, Yangtze University, Jingzhou 434000, China

Abstract: Objective Dougong is a distinctive structural and decorative component of traditional Chinese wooden architecture, playing a critical role in load transfer, seismic resistance, and architectural aesthetics. As a key element connecting columns, beams, and roof structures, Dougong embodies sophisticated construction wisdom and regional architectural characteristics accumulated over centuries. Its structural configuration is composed of multiple interlocking wooden elements arranged in a highly ordered and hierarchical manner, exhibiting strong geometric regularity and pronounced directional

收稿日期：2025-09-22;**修回日期：**2026-01-04

*通信作者：李颜娥 YaneLi@zafu.edu.cn

基金项目：浙江省自然科学基金(MS26F030005,LQ21H180001);浙江农林大学人才发展基金(2019RF065)

Supported by: Natural Science Foundation of Zhejiang Province(MS26F030005,LQ21H180001);Research Development Foundation of Zhejiang A&F University(2019RF065)

dependency. These characteristics make Dougong not only architecturally significant but also visually complex. In real-world heritage environments, the automated recognition of Dougong faces substantial challenges. Variations in spatial orientation, component location (corner and non-corner Dougong), viewing angle, and scale significantly affect visual appearance. Moreover, complex backgrounds, uneven illumination, occlusion by surrounding architectural elements, and long-term material degradation further complicate accurate recognition. In particular, damaged Dougong components often exhibit incomplete structures and blurred boundaries, which severely reduce the effectiveness of conventional image recognition methods. Although deep learning techniques have been increasingly applied to cultural heritage documentation and architectural analysis, many existing approaches rely primarily on convolutional neural networks. Such methods often struggle to capture long-range dependencies and direction-sensitive structural features, limiting their adaptability to ancient architectural components with strong directional organization. Therefore, there is a clear need for a recognition framework that can explicitly model directional structural information while maintaining robustness under complex environmental conditions. The objective of this study is to develop a robust and accurate Dougong recognition model by integrating Transformer-based feature extraction with a novel direction-aware attention mechanism. By enhancing directional feature representation and improving channel-wise feature interaction, this research aims to advance the intelligent recognition of both intact and damaged Dougong components, thereby supporting the digital preservation and systematic documentation of ancient wooden architectural heritage. **Methods** This study proposes an improved Dougong recognition framework based on the Mask R-CNN architecture enhanced with a Swin Transformer backbone. Compared with traditional convolutional backbones, the Swin Transformer enables hierarchical feature extraction and efficient modeling of global contextual information, which is particularly beneficial for complex architectural scenes. A dedicated Dougong dataset was constructed, consisting of 3,014 carefully annotated images that include both corner and non-corner Dougong instances. The dataset was collected from real architectural environments at historically significant sites such as Lingyin Temple and Faxisi Temple, complemented by high-quality images from publicly accessible online repositories. The collected samples cover diverse viewpoints, spatial scales, lighting conditions, and background environments, ensuring strong representativeness and practical applicability. To address the limitations of conventional attention mechanisms in capturing the directional structural characteristics of Dougong, a Directional Coordinate Attention (DIRCA) mechanism is introduced. Unlike traditional spatial or channel attention methods that treat spatial dimensions uniformly, DIRCA decomposes spatial attention into horizontal and vertical directions. This design enables the network to independently encode direction-sensitive information along orthogonal axes, which closely corresponds to the intrinsic structural logic of Dougong, including component stacking order, load transmission paths, and repetitive decorative patterns. By explicitly strengthening the representation of direction-dependent features, DIRCA improves the model's ability to distinguish subtle structural variations under complex background interference. In addition, a one-dimensional squeeze-and-excitation (1D-SE) module is incorporated to achieve lightweight channel recalibration. The proposed 1D-SE mechanism models inter-channel dependencies using a simplified one-dimensional operation, effectively reducing computational cost while preserving channel discrimination capability. This lightweight design facilitates seamless integration with the Transformer-based backbone and enhances the interaction between global contextual features and local structural details. The DIRCA and 1D-SE modules are jointly embedded within the Swin Transformer-enhanced Mask R-CNN framework, forming the proposed SwinE_AM_DIRCA model. Model training and evaluation were conducted under consistent experimental settings. Comparative experiments were performed against representative baseline models, including Mask R-CNN_ResNet50 and Swin_AM, to verify the effectiveness of the proposed architectural improvements. Multiple evaluation metrics were used to assess recognition performance, stability, and robustness, with particular emphasis on corner Dougong and damaged Dougong scenarios. **Results** Experimental results demonstrate that the proposed SwinE_AM_DIRCA model achieves substantial and consistent performance improvements over baseline methods across multiple recognition tasks. The recognition performance for non-corner Dougong reaches 94.4%, while the performance for corner Dougong reaches 89.2%, outperforming the Swin_AM model by 6.7% and 8.9%, respectively. Compared with the conventional Mask R-CNN_ResNet50 model, the proposed framework achieves even more pronounced improvements of 10.8% for non-corner Dougong and 20.4% for corner Dougong recognition, highlighting the effectiveness of Transformer-based feature extraction combined with direction-aware attention. In

damaged Dougong recognition tasks, the proposed model exhibits strong robustness and stability. Despite challenges such as partial occlusion, surface erosion, structural incompleteness, and background clutter commonly observed in ancient architectural heritage sites, SwinE_AM_DIRCA maintains accurate localization and recognition of key structural components. Qualitative visualization results further indicate that the model effectively suppresses irrelevant background features while preserving the integrity of direction-sensitive structural information, which is crucial for reliable heritage documentation. Ablation experiments further confirm the contribution of each proposed module. The DIRCA mechanism significantly enhances the model's sensitivity to directional features, while the 1D-SE module improves channel-wise feature discrimination with minimal computational overhead. When combined, these two modules produce complementary effects, resulting in consistent performance gains and improved model robustness across different Dougong types and damage conditions.

Conclusion This study presents a direction-aware and lightweight attention-enhanced deep learning framework for Dougong recognition in ancient Chinese wooden architecture. By integrating the Directional Coordinate Attention (DIRCA) mechanism and a one-dimensional squeeze-and-excitation (1D-SE) module into a Swin Transformer-based Mask R-CNN architecture, the proposed SwinE_AM_DIRCA model effectively addresses the challenges posed by complex morphology, strong directional characteristics, and real-world damage conditions of Dougong structures. The primary innovation of this work lies in the unified modeling of direction-sensitive structural features and efficient channel recalibration, enabling improved recognition performance and robustness without excessive computational complexity. Experimental results demonstrate that the proposed method significantly outperforms existing baseline models in both standard and damaged Dougong recognition tasks. Overall, this research provides an effective and practical technical solution for the intelligent recognition and digital preservation of ancient architectural components. The proposed framework contributes to the advancement of deep learning applications in cultural heritage analysis and offers valuable methodological support for the systematic documentation and sustainable protection of ancient buildings.

Key words: DIRCA; Deep learning; Cultural heritage; Transformer; Ancient building

0 引　言

斗拱是中国古代建筑的经典构件，承担结构支撑并具有装饰价值(Haiyan 等，2020)根据功能和位置，斗拱可分为柱头斗拱、柱间斗拱和转角斗拱三种类型(Zhao 等，2024)。转角斗拱位于建筑角部，连接墙面并提供支撑；非转角斗拱包括柱头斗拱和柱间斗拱，主要分布在柱子与梁之间(Chai 等，2019)，在多数情况下两者形态相似，有时设计上几乎一致以增强装饰效果(Hao，2014)。斗拱的自动识别与分割对于古建筑的数字化保护、修复及文化遗产传承具有重要意义(Zhou 等，2024)。斗拱结构层次复杂、构件关系多样，给其数字化建模与智能识别带来较大挑战。除结构复杂性外，光照变化、视角差异及复杂背景亦增加了斗拱自动识别与分割的难度。

传统基于形状或纹理的手工特征提取方法在复杂场景下效果有限(Ojala 等，2002)。其中，Ojala 等提出的局部二值模式(Local Binary Pattern，LBP)具有灰度和旋转不变性，为建筑物纹理表征提供了有效特征。但早期建筑物识别方法仍主要依赖边缘与几何特征，局部特征描述子如 SIFT(Scale-Invariant Feature Transform)和 SURF(Speeded Up Robust Features)提高了对旋转和尺度变化的鲁棒性(Lowe，2004)，但仍难以应对复杂背景、遮挡及目标重叠。

近年来，卷积神经网络(Convolutional Neural Networks，CNN)的引入(Bhatt 等，2021)极大改善了建筑物(Belhi 等，2019)和破损建筑物(Duarte 等，2018)的识别精度。Mask R-CNN 通过像素级分割提升了复杂建筑结构(Wang 等，2022)的识别效果(Susetyo 等，2023)。然而，使用标准的 ResNet(He 等，2016)主干网络如 Mask R-CNN_ ResNet50(He 等，2017)在处理细节丰富的建筑图像时仍存在局限(Szegedy 等，2016)。近年来，基于 Transformer(Liu 等，2021)的 Swin Transformer 模型通过局部窗口自注意力机制有效捕捉长距离特征，改善了 CNN 在全局特征提取上的不足(Pramanik 等，2024)。结合 Swin Transformer 与 Mask R-CNN 的多尺度特征提取策略(Gibril 等，2024)，在高分辨率图像(Wei 等，2024)和复杂建筑结构识别中表现优异(Yan 等，2022)。此外，大模型检索增强框架的引入可进一步提升复杂场景下目标识别的鲁棒性(李嘉威等，

2025),为斗拱识别任务中的模型优化提供了新方向。在文化遗产保护背景下,斗拱的目标检测与分类仍具挑战性(Wang 和 Hu,2022)。Guo 等(2022)提出的 Swin-Mask R-CNN 模型通过注意力机制,在斗拱分割与分类中分别达到 90.4% 和 86.2% 的精度,为文化遗产保护提供可视化数据支持。

本研究为传统建筑构件的智能识别与文化遗产保护提供了新的技术思路。徐仕成等(2023)提出的正交约束多头自注意力机制,为优化自注意力的特征提取效率、避免特征冗余提供了重要参考,其设计思路可迁移至斗拱的细粒度特征识别中。基于这种思路在模型设计过程中,比较了 Mask R-CNN_ResNet50、CondInst_ResNet50 和 Swin-Mask R-CNN 等模型,并在 FPN(Feature Pyramid Network)(Lin 等,2017)和 RPN-Head(Region Proposal Network Head)(Ren 等,2016)中引入多种注意力机制,包括压缩与激励注意力机制(Hu 等,2018)、高效通道注意力机制(Wang 等,2020)、卷积块注意力模块(Woo 等,2018)、坐标注意力机制(Hou 等,2021)及方向坐标注意力机制,同时在 Swin 主干网络中加入改进的一维压缩激励机制。实验结果表明,该模型在斗拱识别和分割上优于其他对比模型。

在数据集方面,本研究采集了不同角度、不同光照条件下的转角和非转角斗拱图像,并划分为训练集、验证集和测试集。针对实际场景中的光照不足问题,沈羽翔等(2025)提出生成检测一体化方法,为斗拱图像采集提供了有效技术支持。训练集通过随机遮挡、亮度和对比度调节进行数据增强,以提升模型鲁棒性。研究构建了两套测试集:一是保留原始图像的常规斗拱测试集,二是在原始测试集基础上施加随机裂痕遮挡形成破损斗拱测试集。训练集、验证集和测试集分别占总数据量 81.74 %、8.13 % 和 10.13 %。在相同系统条件下,对 Mask R-CNN_ResNet50、CondInst_ResNet50、SOLOv2_ResNet50 等模型进行训练,并通过平均精度(Average Precision, AP)和平均召回率(Average Recall, AR)指标选出 Mask R-CNN_Swin (Swin-Tiny)作为基础模型。在此基础上,引入 AMP(Automatic Mixed Precision) + MS-Crop(Multi-Scale Cropping),Swin 内部加入 1D-SE 注意力机制,FPN 与 RPN-Head 中引入 SE、ECA、CBAM(Convolutional Block Attention Module)、CA 和 DIRCA 五种注意力机制,最终选出最优模型(见图 1)

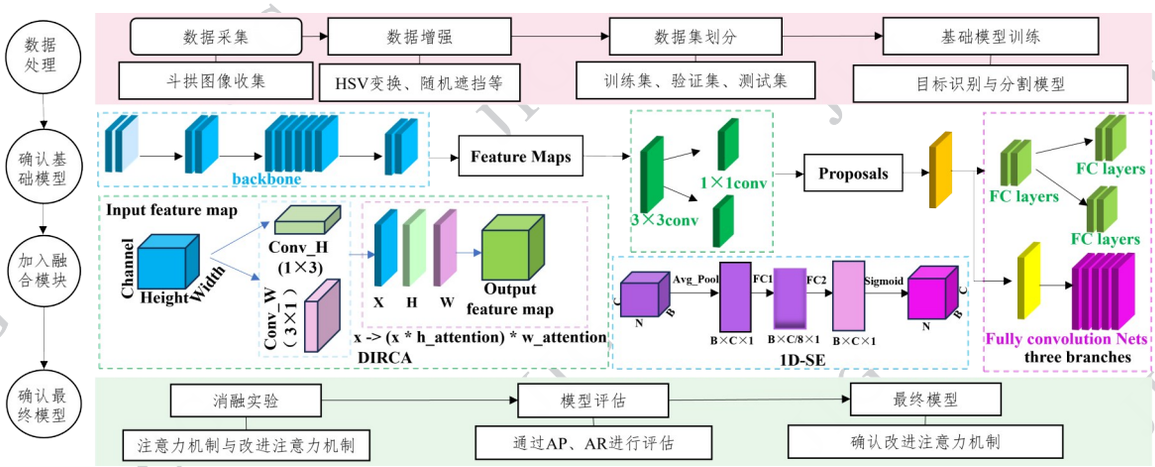


图 1　研究框架图

Fig 1　Framework diagram of study

本研究的主要贡献如下:

1. 构建了涵盖训练集、验证集、常规测试集和破损测试集的完整斗拱图像数据集,训练集通过随机遮挡及光照调节增强模型鲁棒性;根据现有文献和公开资源,尚无可直接使用的同类数据集。

2. 提出改进的注意力机制:方向坐标注意力机制(DIRCA)在捕捉斗拱方向结构上更具优势;1D-SE 机制在压缩激励基础上提取关键特征,提高模型性能。

3. 引入破损斗拱识别,实现对损坏建筑构件的目标导向识别,为后续修复提供数据支持,推动文化遗产智能化保护技术的发展。

1 材料与方法

1.1 数据集

1.1.1 数据采集与增强

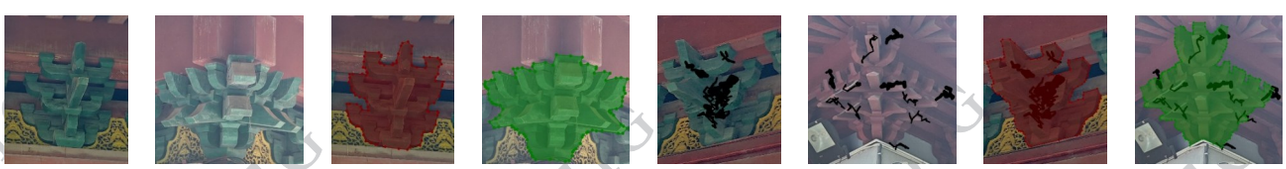
斗拱图像数据采集自灵隐寺、韬光寺、永福寺和法喜寺等场所，并纳入部分公开网络样本作为补充。数据集共包含941张图像、1280个斗拱实例，覆盖多视角及多种光照条件。在此基础上，为提升模型鲁棒性，对491张斗拱图像采用随机遮挡、亮度与对比度调节及HSV变换的数据增强方法，最终扩展至2364张图像，包含3273个非转角斗拱样本和872个转角斗拱样本。

1.1.2 数据预处理

针对光照不均、噪声干扰及结构破损等问题，对部分斗拱图像进行了多阶段预处理。通过非局部均值去噪抑制随机噪声，并结合光照校正与亮度均衡改善成像质量；对于破损区域，采用AOT-GAN网络进行结构化修复，从而提升后续识别与分割效果。

1.1.3 目标区域标注

本研究采用COCO格式构建数据集，并划分为训练集、验证集和测试集，分别包含3388、337和420个样本。其中，测试集进一步分为无遮挡样本与遮挡样本，用于模拟破损斗拱场景。打标后的斗拱示例如图2所示。



(a) (b) (c) (d) (e) (f) (g) (h)

(d) 转角斗拱标注后图像；(e) 破损非转角斗拱原图；(f) 破损转角斗拱原图；(g) 破损非转角斗拱标注后图像；(h) 破损转角斗拱标注后图像；

(c) Labeled non-corner Dougong. (d) Labeled Corner Dougong. (e) Damaged non-corne Dougong. (f) Damaged corner Dougong. (g) Labeled damaged non-corner Dougong. (h) Labeled damaged corner Dougong.

图2 斗拱图像示例与标注结果。(a) 非转角斗拱原图；(b) 转角斗拱原图；(c) 非转角斗拱标注后图像；

Fig 2 Example of a Dougong image with labeling results. (a) Non-corner Dougong. (b) Corner Dougong.

1.2 Swin-Mask R-CNN结合注意力机制模块

1.2.1 Mask R-CNN_Swin网络

Mask R-CNN_Swin将Swin Transformer的多尺度特征提取能力与Mask R-CNN的检测与分割优势相结合。在该模型中，Swin Transformer作为骨干网络，通过层次化窗口自注意力机制提取富含语义和空间信息的多尺度特征图，并经特征金字塔网络(FPN)整合后供后续目标检测使用。模型第一阶段利用RPN对预设锚框进行前景/背景区分及边界框调整，筛选高置信度候选区域；第二阶段通过ROI(Region of Interest)Align对选中ROI进行精确特征对齐与统一尺寸转换，最终在分类、回归和掩码分割分支完成多任务学习。该模型在复杂视觉场景下表现出卓越的检测与分割性能，网络结构如图3所示。

1.2.2 整体网络框架图

本研究以Mask R-CNN搭配Swin Transformer主干网络为baseline，依次引入三项改进模块以提升对复杂结构(如斗拱)的识别与分割能力：

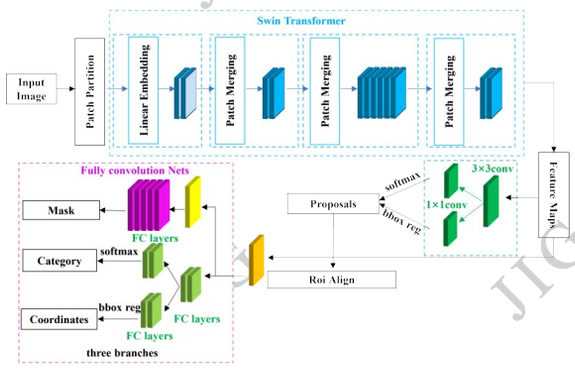


图3 Swin-Mask Rcnn结构图

Fig 3 Swin-Mask RCNN structure diagram

(1)Swin Transformer作为主干网络，其输出特征已具备较强的全局建模能力，但在通道维度上仍存在特征冗余与关键信息抑制的问题。为此，在

Swin Block 之后嵌入一维压缩激励机制(1D-SE),通过沿特征序列方向进行全局平均池化,获得各通道的全局响应描述,再经非线性映射生成权重系数,用以重新标定通道的重要性。该操作能有效增强与斗拱结构相关的特征通道(如层叠纹理、榫卯边缘等)的表达能力,同时抑制无关背景信息,实现特征层面的选择性强化。

(2)特征金字塔网络(FPN)负责将高层语义与低层细节进行自顶向下融合,以实现多尺度特征整合。然而,传统 FPN 在融合过程中对方向信息的保持较弱,不利于捕获斗拱这种具有明显水平与垂直构件分布的目标特征。为此,在 FPN 的特征融合阶段引入方向坐标注意力机制(DIRCA),通过分别沿水平与垂直方向进行卷积特征提取,生成方向敏感的注意力图,从而对特征进行加权增强。该机制能够在多尺度融合过程中强化斗拱梁枋、昂等方向性强的结构区域,提高模型对复杂空间层次的辨识能力。

(3)区域建议网络(RPN)主要负责候选框的生成与筛选,其检测头在滑窗过程中若缺乏方向性信息,容易导致对倾斜或层叠结构的目标定位不准。考虑到斗拱的结构方向具有显著规律性(如横梁延展、垂拱堆叠),在 RPN 检测头中同样引入 DIRCA,以在滑动窗口内捕获水平与垂直方向的显著特征分布。该设计使网络在候选框评分与回归时对方向性结构更加敏感,从而提升候选区域的定位精度与结构匹配度。

(4)整体结构命名为 SwinE_AM_DIRCA(Swin+1D-SE+DIRCA),充分结合空间方向性建模与通道信息压缩,提升网络性能与对复杂背景下关键区域的关注能力。网络框架图如图 4 所示。

SwinE_AM_DIRCA 模型整体流程如下:首先,输入原始斗拱图像并在 Swin 主干网络中逐级提取多尺度特征;并在 Swin Block 的后嵌入一维压缩激励机制来强化关键通道的信息表达;接着,利用特征金字塔网络(FPN)自顶向下融合各尺度特征时,借助方向坐标注意力分别对水平和垂直方向特征进行卷积提取和注意力图生成,对融合结果加权增强;同样地,在区域建议网络(RPN)滑窗检测头中也嵌入 DIRCA,以提高对具有明显方向性结构目标的候选框评分与定位精度;最后,通过 RoIAlign 提取固定尺寸特征,输出目标类别、检测框位置及对应的实例分割掩码。

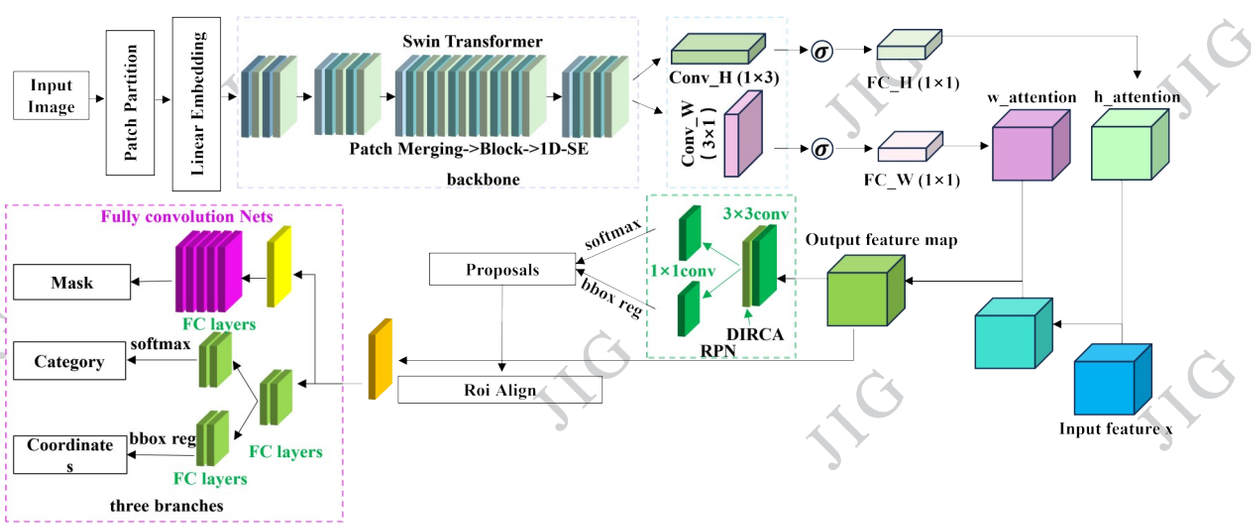


图 4　整体网络框架图

Fig 4　Overall Network Framework Diagram structure diagram

1.2.3　坐标注意力机制及其改进(CA、DIRCA)

(1) 坐标注意力机制

坐标注意力机制(Coordinate Attention, CA)是一种将通道注意力与位置信息相结合的注意力模块(Hou,等,2021)。它通过在水平方向和垂直方向分别进行全局平均池化,生成两个方向感知的通道注意力图,用以捕捉空间坐标信息并对特征图进行加权,显著增强网络对关键区域的响应,抑制背景干扰,提升目标特征的辨识度,尤其适用于背景复杂或细节丰富的图像。然而,原始 CA 模块在面对具有

明显方向性结构(如斗拱这种水平与垂直交织的构件)时,因对两个方向的特征采用统一建模,难以区分横向或纵向上的显著差异,导致对特定方向特征的响应不够敏锐,限制了网络对方向性细节的捕捉能力。其结构如图5所示。公式如下描述:

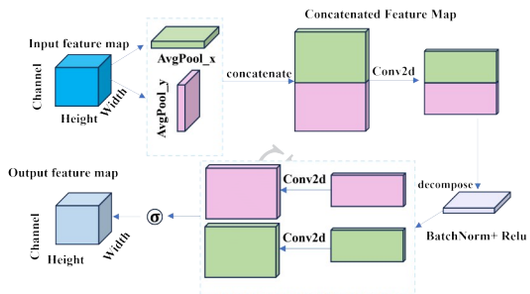


图5　CA结构图

Fig 5　CA structure diagram

CA通过沿H、W方向进行全局平均的池化,生成方向感知的描述向量:

$$z_h(c,x)=\frac{1}{H}\sum_{i=1}^{H}X(c,i,x) \tag{1}$$

$$z_w(c,y)=\frac{1}{W}\sum_{j=1}^{W}X(c,y,j) \tag{2}$$

公式(1)和(2)中,X代表输入特征图,C为通道数,H、W分别为特征图的高度和宽度,$Z_h(c,x)$代表通道c在水平方向x上的全局描述,$Z_w(c,y)$代表通道c在垂直方向y上的全局描述

随后,将所得方向描述向量进行融合,通过1×1卷积映射生成注意力权重,再在通道维度重标定输入特征:

$$Y=X\odot f_{CA}(z_h,z_w) \tag{3}$$

公式(3)中f_{CA}表示通过1×1卷积或全连接映射生成的注意力权重,Y为加权后的输出特征。

(2) 改进的方向坐标注意力制

针对上述不足,提出方向坐标注意力机制(Directional Coordinate Attention, DIRCA)。DIRCA在结构上对水平和垂直方向的空间特征进行独立建模:首先分别采用水平和垂直卷积操作提取对应方向的空间信息;然后通过Sigmoid激活得到归一化的注意力图;最后再经1×1卷积恢复通道数,并将生成的注意力图用于对输入特征图的加权。该设计能够更细致地增强方向性显著区域的特征表达,从而提升模型对具有强方向性结构的目标(如斗拱)的识别与分割能力。其结构如图6所示。公式如下描述:

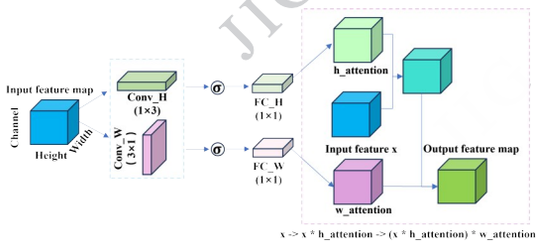


图6　DIRCA结构图

Fig 6　DIRCA structure diagram

水平卷积(H方向):

$$A^H=\sigma(\mathrm{Conv}_{1\times3}(X))\in\mathbb{R}^{B\times C/r\times H\times W} \tag{4}$$

垂直卷积(W方向):

$$A^W=\sigma(\mathrm{Conv}_{3\times1}(X))\in\mathbb{R}^{B\times C/r\times H\times W} \tag{5}$$

公式(4)和(5)中r代表通道压缩比控制中间特征通道数,A^H、A^W表示经过方向卷积并经Sigmoid激活的水平和垂直注意力图。

通道恢复:

$$\hat{A}^H=\mathrm{Conv}_{1\times1}(A^H),\hat{A}^W=\mathrm{Conv}_{1\times1}(A^W)\in\mathbb{R}^{B\times C/r\times H\times W} \tag{6}$$

公式(6)表示A^H、A^W通过1×1卷积恢复到原始通道数的注意力图。

特征加权:

$$Y=X\odot\hat{A}^H\odot\hat{A}^W \tag{7}$$

公式(7)中Y表示加权后的输出特征

计算复杂度对比:

CA:O(B·C·(H+W)+B·C^2/r)

DIRCA:O(B·C·H·W·(2·3/r+1))

DIRCA在H×W大的特征图上增加了空间卷积计算,但避免了全局平均池化带来的空间信息压缩损失;更适合方向性强、局部结构丰富的目标,如斗拱的横梁与垂拱结构。相比CA,DIRCA用卷积直接在空间方向上提取显著性特征,而不是全局平均池化的压缩信息。

1.2.4　一维压缩激励机制(1D-SE)

为了适应一维特征序列场景,将原始的压缩激励注意力机制(Squeeze-and-Excitation, SE)改造成一维压缩激励注意力机制(1D-SE)(Hu等,2018):首先对输入的一维特征沿序列长度进行全局平均池化,生成通道级的全局描述向量;然后通过两层全连接瓶颈结构并经Sigmoid激活,得到每个通道的权重系数;最后用这些系数对原始一维特征进行加权,从而放大关键维度、抑制冗余噪声。与那些省略中间

激活、直接从池化到Sigmoid的轻量化通道注意力不同，1D-SE保留了非线性映射提升表达力；与基于Query-Key-Value的自注意力相比，它仅对通道维度进行重标定，计算复杂度低，因此在保证性能的同时更加轻量且易于部署。

1.3 模型评价指标

研究采用平均精度（Average Precision，AP）和平均召回率（Average Recall，AR）作为主要评价指标，用于衡量模型在斗拱及破损斗拱检测与分割任务中的性能表现。

其中，AP反映了模型在不同召回率下的整体精度表现，mAP（mean Average Precision）表示在所有类别或不同IoU阈值下的平均精度，常用于综合评估模型的检测能力。AR则用于衡量模型在多阈值条件下的平均召回水平，反映检测结果的完整性。此外，MAP75（IoU=0.75时的AP）用于评估模型在高精度匹配条件下的检测性能。

2 结 果

2.1 实验环境和参数设置

该算法采用基于MMDetection 3.3.0框架的Swin Transformer骨干Mask R-CNN模型进行斗拱图像识别，模型训练与推理过程在NVIDIA GeForce GTX 1080 Ti GPU平台上完成。训练图像归一化为800×1300像素，批量大小为2，训练周期为36。使用AdamW优化器，学习率为0.0001，权重衰减为0.05。

表1 实验模型参数

Table 1 Experimental model parameters

训练参数	参数值	训练参数	参数值
输入图片大小	800×1300	学习率	0.0001
批量大小	2	权重衰减	0.05
训练周期	36	一阶矩衰减率(β_1)	0.9
优化器	AdamW	二阶矩衰减率(β_2)	0.999

2.2 识别与分割模型的架构性能分析

为验证Mask R-CNN_Swin（主干网络采用Swin-Tiny）在本研究场景中的可靠性，本实验加入了Mask R-CNN_ResNet50、Condinst_ResNet50、Solov2_ResNet50等模型进行训练并测试，模型均在训练过程完成时达到收敛并进行测试，转角和非转角斗拱识别精度和召回率。

表2 斗拱识别模型对比

Table 2 Comparison of Dougong Recognition Models

模型-识别	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
Mask R-CNN_Swin	**0.875**	**0.986**	**0.734**	**1**	**0.849**
Mask R-CNN_ResNet50	0.836	0.977	0.688	0.801	0.833
Condinst_ResNet50	0.830	0.975	0.576	0.499	0.76
Retinanet_ResNet50	0.844	0.969	0.530	0.455	0.744
Fsaf_ResNet50	0.842	0.979	0.566	0.586	0.788
Retinanet_pvt	0.833	0.976	0.686	0.945	0.803

注：在AP0.75（即IoU阈值为0.75）下，Mask R-CNN_Swin模型在转角斗拱的AP中达到了100%。这是由于在本实验数据中，转角斗拱目标形态特征明显且测试数据量较少，Swin主干网络配合各类注意力机制均能在较高的IoU要求下实现精确匹配，因此后续多组实验在该指标上表现一致

由表2可见，Mask R-CNN_Swin模型在斗拱识别任务中整体表现最佳。其对非转角斗拱的AP为0.875，对转角斗拱的AP为0.734，均显著高于其他模型，且在高阈值条件下（$AP_{0.75}$）几乎达到满分，说明其检测定位精度较高。

相比之下，采用Swin Transformer作为主干网络的模型较基于ResNet50的模型表现更优，说明Swin的层次化自注意力机制能更有效地捕捉斗拱结构的复杂形态和尺度变化，从而提升识别精度。值得注意的是，RetinaNet_PVT由于引入了PVT主干网络，在一定程度上提升了对高复杂度特征的捕捉能力，其$AP_{0.75}$优于其他ResNet50系列模型，但仍不及Mask R-CNN_Swin的综合表现。而CondInst、RetinaNet等单阶段模型在精度与召回率上均略低，表

明其在细粒度目标的特征提取与边界判定方面存在不足。

表3 斗拱分割模型对比

Table 3 Comparison of Dougong Segmentation Models

模型-分割	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
Mask R-CNN_Swin	**0.876**	**0.988**	**0.820**	**1**	**0.865**
Mask R-CNN_ResNet50	0.857	0.976	0.781	0.983	0.860
Condinst_ResNet50	0.757	0.851	0.722	0.888	0.806
Solov2_ResNet50	0.762	0.903	0.582	0.610	0.757
Solo_ResNet50	0.805	0.921	0.711	0.853	0.821

注：关于$AP_{0.75}$为1的现象及原因，见表2注释。

表3可见，Mask R-CNN_Swin模型在斗拱分割任务的各项指标上均表现最佳。其对非转角斗拱和转角斗拱的分割AP分别达到0.876和0.820，整体召回率为0.865，明显优于其他模型。

相比之下，采用Swin Transformer作为主干的模型较ResNet50主干具有更强的特征表达能力，能够更准确地区分斗拱构件间的边界与细节，从而提升分割精度。而CondInst、SOLOv2等模型在复杂结构区域的分割表现相对较弱，说明其在捕捉细粒度结构关系方面存在局限。

表4 破损斗拱识别模型对比

Table 4 Comparison of Damaged Dougong

/Recognition Models

模型-识别	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
Mask R-CNN_Swin	**0.856**	**0.982**	0.667	0.802	0.793
Mask R-CNN_ResNet50	0.824	0.958	0.678	0.852	**0.799**
Condinst_ResNet50	0.781	0.946	0.512	0.481	0.751
Retinanet_ResNet50	0.754	0.932	0.443	0.216	0.661
Fsaf_ResNet50	0.831	0.958	0.566	0.672	0.783
Retinanet_pvt	0.829	0.974	**0.682**	**0.907**	0.799

由表4可见，Mask R-CNN_Swin在破损斗拱识别任务中整体表现最佳，其非转角斗拱AP为0.856，识别精度高，说明模型在结构受损但轮廓仍较完整的情况下具有较强的特征提取与判别能力。相比之下，转角斗拱的AP为0.667，略低于ResNet50主干，表明当破损集中在复杂转角区域时，模型的边界感知与细节恢复能力仍受一定限制。

Swin Transformer的全局自注意力机制在处理特征缺失或纹理不连续的情形下表现出更好的稳定性，但面对形态变化大、遮挡严重的破损部位，其特征聚合仍不足。这说明在破损斗拱识别中，增强局部结构细节建模与跨尺度特征融合仍是进一步提升模型精度的关键方向。

RetinaNet在引入Pyramid Vision Transformer主干后，斗拱的检测精度有所提升，其中RetinaNet_PVT在高阈值条件下（$AP_{0.75}$=0.907）表现优异，显示出PVT在保持整体结构一致性方面的潜力，但其在非转角斗拱破损场景中的稳定性仍不及Mask R-CNN_Swin。

由表5可见，Mask R-CNN_Swin在破损斗拱分割任务中各项指标均明显优于其他模型，非转角斗拱的分割AP为0.854，转角斗拱为0.766，整体表现最为稳定。与ResNet50主干相比，Swin模型在特征提取阶段能更好地保持结构边界与纹理完整性，对局部破损区域的细节恢复能力更强。

这主要得益于Swin Transformer的分层自注意力机制，其能在局部缺损或形态不规则的区域中建立更有效的全局关联，从而提升分割精度。而其他

表5　破损斗拱分割模型对比

Table 5　Comparison of Damaged Dougong

/Segmentation Models

模型-分割	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
Mask R-CNN_Swin	**0.854**	**0.984**	**0.766**	**0.975**	**0.823**
Mask R-CNN_ ResNet50	0.797	0.907	0.752	0.967	0.796
Condinst_ResNet50	0.721	0.786	0.608	0.726	0.755
Solov2_ResNet50	0.708	0.867	0.512	0.526	0.682
Solo_ResNet50	0.759	0.914	0.601	0.735	0.769

基于卷积的模型在特征连续性受破坏时较难重建整体轮廓，导致边界模糊、区域漏分。说明在破损场景下，引入具有全局建模能力的主干网络对于结构类目标的精确分割具有显著优势。

综上所述，在斗拱识别、破损识别及分割等任务中，Mask R-CNN_Swin整体表现最优。因此，在后续研究中以Mask R-CNN_Swin作为基础模型进行拓展与优化。为进一步提升训练效率与模型性能，后续所有基于 Mask R-CNN_Swin 的模型均引入了框架内置的自动混合精度（AMP）与多尺度裁剪（MS-Crop）策略。统称为Swin_AM。

后续所有基于 Mask R-CNN_Swin 的模型均引入了框架内置的自动混合精度（AMP）与多尺度裁剪（MS-Crop）策略。统称为Swin_AM。

2.3　不同注意力机制结合Swin_AM性能比较

为了验证和展示所提出改进算法的效果，分别在Mask R-CNN_Swin模型中的FPN和RPN-Head模块中引入了SE、ECA、CBAM、CA、DIRCA注意机制且在主干网络swin中加入了1D-SE注意力机制，并进行了对比实验。表6、7展示了Mask R-CNN_Swin模型与不同注意机制（SE、ECA、CBAM、CA、DIRCA、1D-SE）相结合后的斗拱识别分割性能对比，其中用SwinE代表在Swin加入1D-SE注意力机制。表8、9展示了破损斗拱识别分割任务中的性能对比。

表6　Swin_AM与不同注意力机制结合下识别斗拱实验对比

Table 6　Comparison of Attention Mechanisms in Swin_AM for Dougong Recognition

模型-识别	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
SwinE_AM_DIRCA	**0.944**	**0.988**	**0.892**	**1**	**0.939**
SwinE_AM_CA	0.903	0.986	0.867	1	0.910
SwinE_AM_CBAM	0.898	0.985	0.802	1	0.886
SwinE_AM_ECA	0.891	0.979	0.831	1	0.909
SwinE_AM_SE	0.904	0.984	0.831	1	0.898
SwinE_AM	0.889	0.986	0.836	1	0.896
Swin_AM	0.877	0.982	0.803	1	0.885

注：关于$AP_{0.75}$为1的现象及原因，见表2注释。

由表6可见，SwinE_AM结合DIRCA注意力机制的模型在斗拱识别任务中表现最佳。其非转角斗拱的AP最高达到0.944，转角斗拱AP也达到0.892，整体召回率0.939，均高于其他注意力机制组合。其他注意力模块虽然在部分指标上略有提升，但整体均低于DIRCA组合，尤其在转角斗拱识别上差距更明显。

Swin提取的全局特征经过DIRCA处理后，不仅保持了整体几何结构信息，也增强了局部纹理和层次关系，使模型在复杂几何或局部破损的斗拱上仍能保持较高精度。相比之下，其他注意力机制多集中于通道特征加权，难以充分捕捉构件间的空间依赖，因此性能略逊。结果表明，全局建模与局部注意力的协同作用对斗拱识别具有显著优势。

表7　Mask R-CNN_Swin_AM与不同注意力机制结合下分割斗拱实验对比

Table 7　Comparison of Attention Mechanisms in Swin_AM for Dougong Segmentation

模型-分割	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
SwinE_AM_DIRCA	**0.904**	**0.988**	**0.862**	**1**	**0.898**
SwinE_AM_CA	0.887	0.982	0.853	1	0.894
SwinE_AM_CBAM	0.902	0.988	0.843	1	0.888
SwinE_AM_ECA	0.896	0.984	0.849	1	0.892
SwinE_AM_SE	0.901	0.987	0.829	1	0.885
SwinE_AM	0.901	0.986	0.856	1	0.895
Swin_AM	0.893	0.985	0.854	1	0.893

注：关于$AP_{0.75}$为1的现象及原因，见表2注释。

由表7可见，SwinE_AM结合DIRCA注意力机制的模型在斗拱分割任务中表现略优，非转角斗拱AP达到0.904，转角斗拱AP为0.862，整体召回率0.898，高于其他注意力机制组合，但差距不大。其他模块在非转角斗拱上的分割表现也接近DIRCA组合，但在转角斗拱细节区域略低。

分割任务本身依赖于FPN多尺度特征，Swin主干已经能较好捕捉全局和局部信息，因此引入注意力机制只能在边缘和纹理细节上带来有限提升。DIRCA在FPN阶段增强了局部空间与通道特征的响应，使模型在复杂区域略有优势，但整体差异不显著，说明对分割任务而言，全局特征的有效提取比注意力机制的优化作用更关键。

表8　Mask R-CNN_Swin_AM与不同注意力机制下识别破损斗拱实验对比

Table 8　Comparison of Attention Mechanisms in Swin_AM for Damaged Dougong Recognition

模型-识别	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
SwinE_AM_DIRCA	**0.936**	**0.990**	**0.873**	**1**	**0.926**
SwinE_AM_CA	0.872	0.965	0.823	1	0.878
SwinE_AM_CBAM	0.899	0.987	0.777	1	0.860
SwinE_AM_ECA	0.887	0.987	0.799	1	0.877
SwinE_AM_SE	0.876	0.988	0.823	0.99	0.875
SwinE_AM	0.884	0.987	0.802	1	0.882
Swin_AM	0.869	0.976	0.771	0.956	0.861

注：关于$AP_{0.75}$为1的现象及原因，见表2注释。

由表8可见，SwinE_AM结合DIRCA注意力机制在破损斗拱识别任务中表现最优。其非转角斗拱AP达到0.936，转角斗拱AP为0.873，整体召回率0.926，高于其他注意力机制组合。其他模块（如SE、ECA、CBAM、CA）在部分指标上有所提升，但整体低于DIRCA组合，尤其在非转角斗拱识别上差距更明显。

Swin主干输出的全局特征经过FPN阶段引入DIRCA后，能够更有效地增强局部纹理与多尺度结构信息，使模型在处理破损或边界不完整的斗拱时仍保持较高的检测精度和稳定性。而其他注意力模块多集中于通道加权，难以充分利用空间依赖关系，因此在破损斗拱识别上表现略逊。结果表明，全局特征提取与局部注意力增强的协同作用对破损斗拱识别尤为重要。

由表9可见，SwinE_AM_DIRCA在破损斗拱分割任务中表现略优，转角斗拱AP较基础Swin模型提升约4.3%，其他指标提升幅度较小，说明注意力机制在分割任务中提升有限。其非转角斗拱和转角斗拱的分割AP分别为0.893和0.852，整体召回率

表9　Mask R-CNN_Swin_AM与不同注意力机制下分割破损斗拱实验对比

Table 9　Comparison of Attention Mechanisms in Swin_AM for Damaged Dougong Segmentation

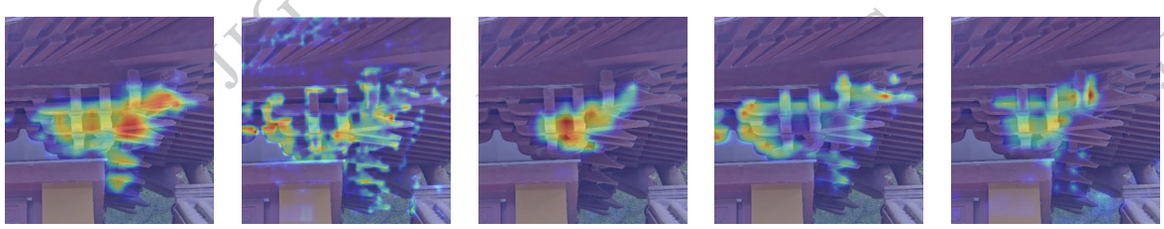
模型-分割	AP for non	AP0.75 for non	AP for cor	AP0.75 for cor	AR
SwinE_AM_DIRCA	**0.893**	**0.987**	**0.852**	**1**	**0.893**
SwinE_AM_CA	0.870	0.973	0.827	1	0.859
SwinE_AM_CBAM	0.879	0.986	0.829	1	0.871
SwinE_AM_ECA	0.873	0.981	0.831	1	0.862
SwinE_AM_SE	0.856	0.979	0.801	0.991	0.857
SwinE_AM	0.866	0.976	0.833	1	0.860
Swin_AM	0.867	0.987	0.809	0.995	0.858

注：关于$AP_{0.75}$为1的现象及原因，见表2注释。

达到0.893，仍高于其他注意力组合。SwinE_AM_DIRCA结合了Swin全局特征与FPN阶段DIRCA注意力，增强了多尺度局部纹理与结构信息的表达能力，使模型在复杂或破损的斗拱上仍保持稳定的识别与分割性能。而注意力机制在分割任务中的提升有限，说明分割效果更多依赖于FPN的多尺度特征融合和Swin全局特征提取。

为进一步验证所提出DIRCA注意力机制在增强特征表征方面的有效性，采用Grad-CAM对不同注意力机制模型生成的中间特征图进行可视化分析。如图7所示，传统注意力机制虽能在一定程度上突出目标相关区域，但其响应仍较为分散，且易受无关背景激活的干扰。

相比之下，DIRCA机制在目标显著区域展现出更为集中的响应，其生成的热力图边界更加清晰，并在空间分布上与真实掩膜保持更高的一致性。上述结果表明，DIRCA在特征选择中具有更强的判别能力，能够生成空间一致性更高、更加聚焦于目标的注意力分布。



(a)　(b)　(c)　(d)　(e)

(c) CBAM注意力机制；(d) ECA注意力机制；(e) SE注意力机制；

图7　SwinE_AM的FPN层Grad-CAM分析。(a) DIRCA注意力机制；(b) CA注意力机制；

Fig 7　Grad-CAM Analysis on the FPN Layers of SwinE_AM. (a) DIRCA attention mechanism; (b) CA attention mechanism; (c) CBAM attention mechanism; (d) ECA attention mechanism; (e) SE attention mechanism.

图8是本研究使用不同方法建立模型的实例分割破损斗拱结果展示，从图中可知Mask RCNN_Swin + DIRCA + 1D-SE的模型分割结果中斗拱边缘清晰，结构基本完整，相比之下原始Mask R-Cnn模型边缘模糊，在细小结构识别分割存在欠缺。本研究一个重要目的是将识别出的斗拱区域取出，作为后续修复的数据集，相较于其他模型Mask RCNN_Swin + DIRCA + 1D-Se的检测框较为准确，尤其是在转角斗拱上其余模型的检测框缺失细节较多。

3　结　论

本文针对斗拱及破损斗拱的目标检测与实例分割任务，提出了一种基于Swin Transformer的改进Mask R-CNN模型SwinE_AM_DIRCA。该方法以

(a)original (b)original (c)labeling (d)labeling (e)Swin (f)Swin

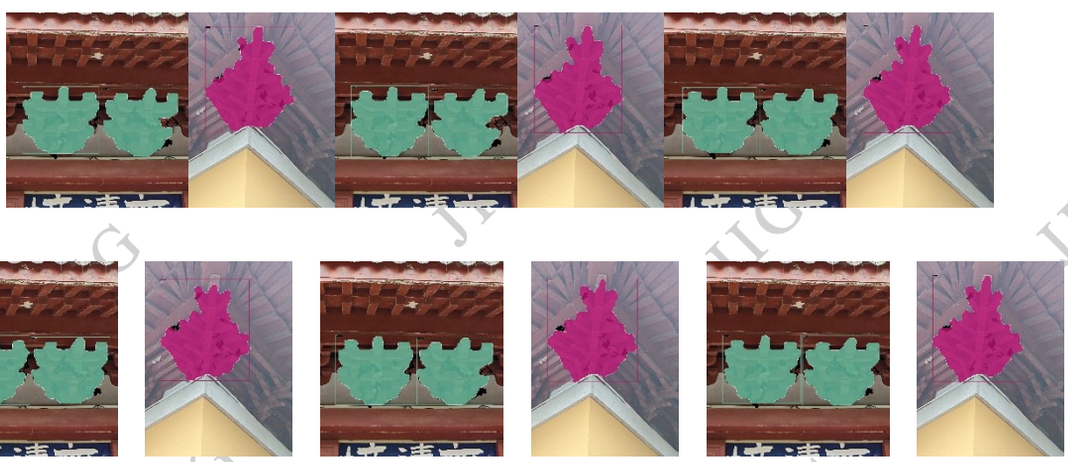
Swin Transformer 作为主干网络，引入一维压缩激励

(g)1D-SE (h)1D-SE (i)SE (j)SE (k)ECA (l)ECA

机制（1D-SE）对通道特征进行重标定，并在 FPN 与 RPN-Head 中融合方向坐标注意力机制（DIRCA），分别对水平与垂直方向特征进行建模，从而增强模型对斗拱方向性结构与局部细节的感知能力。

实验结果表明，所提出的 SwinE_AM_DIRCA 模型在转角与非转角斗拱识别任务中均取得了显著性能提升，识别精度明显优于基于 ResNet50 的传统 Mask R-CNN 模型以及未引入改进注意力机制的 Swin 基础模型；在破损斗拱场景下，模型同样表现出较强的鲁棒性，在结构缺失与噪声干扰条件下仍能保持较高的识别准确率。分割性能提升相对有限，但在复杂背景和遮挡条件下整体优于传统网络结构，具备较好的稳定性。

尽管本文方法在斗拱识别与分割方面取得了较好效果，但仍存在一定不足：一是当前数据集以常见斗拱形态为主，复杂造型与极端破损样本较少，模型的泛化能力仍有待提升；二是本研究尚未将识别与分割结果与图像修复过程进行耦合。未来可进一步扩充数据类型，并结合生成对抗网络等图像生成方法，构建斗拱识别与自动修复相结合的一体化流程。



（m） CBAM （n） CBAM （o） CA （p） CA （q） DIRCA （r） DIRCA

图8 不同模型检测结果示例。

Figure 8 Examples of detection and segmentation results with different models. (a) and (b)Original images of damaged non-corner and corner Dougong. (c) and (d) are the ground truth. (e) and (f) Result of Swin_AM. (g) and(h) Results of SwinE_AM; (i) ~(r) Results of SwinE_AM with the attention mechanisms SE, ECA, CBAM, CA, and DIRCA.

参考文献(References)

Belhi A, Gasmi H, Al-Ali A K, Bouras A, Foufou S, Yu X and Zhang H. 2019. Deep Learning and Cultural Heritage: The CEPROQHA Project Case Study// 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). Maldives: IEEE: 1 - 5. [DOI: 10.1109/SKIMA47702.2019.8982520]

Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K and Ghayvat H. 2021. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. Electronics 10 (20): 2470. [DOI: 10.3390/electronics 10202470]

Chai H, Marino D, So E and Yuan P F. 2019. Design and Fabrication of a Timber Tower Structure through the Reinterpretation of Interlocking Joints// Proceedings of IASS Annual Symposia 2019, no. 20. Barcelona, Spain: IASS: 1 - 8

Duarte D, Nex F, Kerle N and Vosselman G. 2018. Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 4: 89 - 96. [DOI: 10.5194/isprs-annals-IV -2-89-2018]

Gibril M B A, Al-Ruzouq R, Bolcek J, Shanableh A and Jena R. 2024. Building Extraction from Satellite Images Using Mask R-CNN and Swin Transformer// 2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA), Zilina, Slovakia: IEEE: 1 - 5. [DOI: 10.1109/RADIOELEKTRONIKA61599.2024.10524085]

Guo M H, Xu T X, Liu J J, Liu Z N, Jiang P T, Mu T J, Zhang S H, Martin R R, Cheng M M and Hu S M. 2022. Attention Mechanisms in Computer Vision: A Survey. Computational Visual Media 8 (3): 331 - 368. [DOI: 10.1007/s41095-022-0271-y]

Haiyan D, Hong Z and Bin M.2020. Manufacture and Assembly Mechanism of Dougong// IOP Conference Series: Materials Science and Engineering, 960 (2): 022008. Bristol, UK: IOP Publishing. [DOI: 10.1088/1757-899X/960/2/022008]

Hao S. 2014. IDS—Intelligent Dougong System: A Knowledge-Based and Graphical Simulation of Construction Processes of China's Song-Style Dougong System. Columbus: The Ohio State University.

He K, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN// Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy: IEEE: 2961 - 2969. [DOI: 10.1109/ICCV.2017.322]

He K, Zhang X, Ren S and Sun J. 2016. Deep Residual Learning for Image Recognition// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA: IEEE: 770 - 778. [DOI: 10.1109/CVPR.2016.90]

Hou Q, Zhou D and Feng J. 2021. Coordinate Attention for Efficient Mobile Network Design// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA: IEEE: 13713 - 13722. [DOI: 10.1109/CVPR46437.2021.01351]

Hu J, Shen L and Sun G. 2018. Squeeze-and-Excitation Networks// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE: 7132 - 7141. [DOI:10.1109/ CVPR.2018.00745]

Li J W, Yang C Y, Zhang Y C, Sun W L, Meng L, and Meng X X. 2025. Large model retrieval enhancement framework for construction site risk identification. Journal of Image and Graphics, 2025, 1-11. (李嘉威,杨成业,张尧臣,孙玮琳,孟雷,孟祥旭. 2025. 面向工地风险隐患识别的大模型检索增强框架. 中国图象图形学报,2025,1-11).[DOI:10.11834/jig.250333]

Lin T Y, Dollár P, Girshick R, He K, Hariharan B and Belongie S. 2017. Feature Pyramid Networks for Object Detection// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE Computer Society: 2117 - 2125.[DOI: 10.1109/CVPR.2017.106]

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada: IEEE: 10012 - 10022. [DOI: 10.1109/ICCV48922.2021.00986]

Lowe D G. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60: 91 - 110. [DOI: 10.1023/B:VISI. 000002966 4.99615.94]

Pramanik P, Roy A, Cuevas E, Perez-Cisneros M and Sarkar R. 2024. DAU-Net: Dual Attention-Aided U-Net for Segmenting Tumor in Breast Ultrasound Images. PLOS One 19 (5): e0303670. [DOI:10.1371/ journal.pone.0303670]

Ren S, He K, Girshick R and Sun J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (6): 1137 - 1149. [DOI: 10.1109/tpami.2016.2577031]

Shen Y X, Guo X, Wang H, Hu K Y, and Tao C. 2025. Integrated generation-detection approach for pedestrian recognition in dark-light conditions. Journal of Image and Graphics, 30 (10): 3319-3334. (沈羽翔,郭鑫,王昊,胡柯彦,陶超. 2025. 面向暗光条件下行人识别的生成检测一体化方法. 中国图象图形学报,30 (10):3319-3334).[DOI:10.11834/jig.240649]

Susetyo D B, Harintaka H and Rizaldy A. 2023. The Application of Mask R-CNN for Building Extraction//AIP Conference Proceedings, 2941(1). Melville, NY, USA: AIP Publishing. [DOI: 10.1063/5.0139074]

Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z. 2016. Rethinking the Inception Architecture for Computer Vision// Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE: 2818 - 2826. [DOI:10.1109 /CVPR.2016.308]

Wang Q, Wu B, Zhu P, Li P, Zuo W and Hu Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE: 11531 - 11539. [DOI: 10.1109/CVPR42600.2020.01155]

Wang Y, Hu X. 2022. Machine Learning-Based Image Recognition for Rural Architectural Planning and Design. Neural Computing and Applications: 1 - 10. [DOI: 10.1007/s00521-022-07799-w]

Wang Y, Li S, Teng F, Lin Y, Wang M and Cai H. 2022. Improved Mask R-CNN for Rural Building Roof Type Recognition from UAV High-Resolution Images: A Case Study in Hunan Province, China. Remote Sensing 14 (2): 265. [DOI: 10.3390/rs14020265]

Wei S, Zhang T, Yu D, Ji S, Zhang Y and Gong J. 2024. "From Lines to Polygons: Polygonal Building Contour Extraction from High-Resolution Remote Sensing Imagery." ISPRS Journal of Photogrammetry and Remote Sensing 209: 213 - 232.

Woo S, Park J, Lee J Y and Kweon I S. 2018. CBAM: Convolutional Block Attention Module// Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany: Springer Nature Switzerland AG: 3 - 19. [DOI: 10.1007/978-3-030-01252-6_1]

Xu S C, and Zhu Z Q. 2023. Orthogonality-constrained multihead self-attention for scene text recognition. Journal of Image and Graphics, 28 (12): 3855-3869. (徐仕成, 朱子奇. 2023. 正交约束多头自注意力的场景文本识别. 中国图象图形学报, 28 (12): 3855-3869). [DOI:10.11834/jig.221049]

Yan L, Yang J and Zhang Y. 2022. 2022. Building Instance Change Detection from High Spatial Resolution Remote Sensing Images Using Improved Instance Segmentation Architecture. Journal of the Indian Society of Remote Sensing 50 (12): 2317 - 2336. [DOI: 10.1007/s12524-022-01601-z]

Zhao J, Agkathidis A, Lombardi D and Chen H. 2024. Reinterpreting the Dougong Joint: A Systematic Review of Robotic Technologies for the Assembly of Timber Joinery. Architectural Science Review, 65-79 [DOI:10.1080/00038628.2024.2339995]

作者简介

金远，男，硕士研究生，主要研究方向为斗拱图像实例分割与修复。E-mail: 2023611031015@stu.zafu.edu.cn

池方爱，男，教授，主要研究方向为建筑遗产数智化保护。E-mail: 524039@yangtzeu.edu.cn

刘迎奥，男，硕士研究生，主要研究方向为图像分割。E-mail: yingaoliu@stu.zafu.edu.cn

李颜娥，女，副教授，主要研究方向为人工智能与深度学习。E-mail: YaneLi@zafu.edu.cn