湖南大學
HUNAN UNIVERSITY

硕士学位论文

基于深度学习的城市建筑群快速三维建模与应用研究

作　者　姓　名　　陈祎林

指　导　教　师　　周云教授

　　谈忠坤教授级高工

培　养　单　位　　土木工程学院

专　业　类　别　　土木水利

专　业　领　域　　土木工程

学校代码 10532
学　　号 S2301W0049
密　　级 公开

湖南大学硕士学位论文

基于深度学习的城市建筑群快速三维建模与应用研究

国家自然科学基金（52278306）
湖南省重点研发计划（2022SK2096）
湖南省自然科学基金项目（2023JJ70003）
湖南省应急厅应急管理科技项目（yjtkjxm_202406）
基础设施天-空-地一体化智能安全监测与运维（H202491378193）

作　者　姓　名：陈祎林
指　导　教　师：周云教授
谈忠坤教授级高工
培　养　单　位：土木工程学院
专　业　类　别：土木水利
专　业　领　域：土木工程
论文提交日期：2026 年 6 月 10 日
论文答辩日期：2026 年 6 月 15 日
答辩委员会主席：黄远教授

Research on Rapid 3D Modeling and Application of Urban Building Complexes Based on Deep Learning

by

Yilin Chen

B.E. (Xiangtan University) 2023

A thesis submitted in partial satisfaction of the

Requirements for the degree of

Master of Engineering

in

Civil and Hydraulic

in the

Graduate School

of

Hunan University

Supervisors

Professor Yun Zhou

Senior Engineer Zhongkun Tan

June, 2026

湖 南 大 学
学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

作者签名：陈伟林　　日期：2026 年 6 月 12 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权湖南大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

涉密学位论文在解密后适用本授权书。

作者签名：陈伟林　　日期：2026 年 6 月 12 日

导师签名：　　日期：2026 年 6 月 12 日

摘　要

城市三维建筑信息的快速获取是智慧城市建设与城市安全精细化管理的重要基础。然而，现有城市三维建模方法普遍面临数据获取成本高昂、算力依赖严重、更新周期漫长等瓶颈，同时在建筑轮廓提取中存在边界模糊与复杂背景干扰、高度估计中存在空间信息丢失与跨域泛化困难、三维白模构建流程缺乏系统性且与工程实际需求结合不足等突出问题。如何突破传统现场测绘的局限，实现不到现场即可远程高效获取城市建筑三维信息，为城市数字底座构建、应急管理、规划决策等提供精准可靠的数据支撑，已成为亟待解决的关键科学与工程问题。针对上述挑战，本文提出了一套基于深度学习的遥感影像建筑物轮廓提取、高度估计及城市级三维白模快速构建与工程化应用的完整技术体系，实现了从数据获取到工程落地的全链路贯通。主要研究工作如下：

（1）提出了基于伪多模态结构信息的建筑物轮廓精细提取网络 SPR-Net，解决了遥感影像中建筑物边界模糊、尺度差异显著以及复杂背景干扰等导致提取精度受限的问题。该网络以 Segformer 为主干架构，创新性地设计了编码器端与解码器端两类伪多模态结构增强模块，在仅依赖单一光学影像输入的前提下实现了对建筑轮廓边界的有效强化；同时构建了覆盖长沙市的高分辨率建筑数据集，并设计了复合损失函数以提升分割边界锐度。在公开数据集与自建数据集上的系统对比实验表明，SPR-Net 在 IoU、Recall、Precision 和 F1 值等指标上全面领先多种主流分割网络与 SOTA 模型。消融实验系统验证了各组件的有效性与协同增益，复杂度实验表明模型在保持高精度的同时兼具较高的计算效率。

（2）基于级联动态上采样的建筑物高度估计网络 Segformer-H，解决了遥感影像高度估计在特征融合阶段存在的空间信息丢失、边界模糊以及模型难以应对跨城市形态泛化等核心问题。该网络在保留层次化 Transformer 编码器全局语义建模能力的基础上，引入基于可学习偏移量的级联动态上采样模块替代传统双线性插值，有效克服了固定插值核在建筑物边界与高度突变区域引入的模糊效应；同时自建了涵盖大跨度高度范围的长沙建筑高度数据集。在多个公开数据集与自建数据集上的系统对比实验表明，Segformer-H 在所有评价指标上均取得最优性能，在建筑边缘刻画锐利度与不同复杂度场景下的预测稳定性方面均具有显著优势。

（3）构建了城市级 LoD1 三维白模自动化生成方法，解决了白模构建流

程系统性缺失及其与工程需求结合不紧密的问题，拓展了三维模型在建筑安全监测中的应用范式。通过预测结果空间拼接融合、建筑轮廓正则化、高度赋值及矢量化拉伸建模等关键技术，实现了完整的城市级三维白模自动化生成。在多类典型城市功能区的大范围实验中，该方法均能稳定生成结构合理、几何规整的三维建筑模型。在工程应用方面，构建了 PS-InSAR 形变监测点与三维白模的空间匹配技术链路，可有效识别城市监测漏检建筑；设计了基于多时相遥感影像与三维白模时序分析的建筑违法加层智能识别方法，取得了较高的查准率与查全率。此外，开发了集成多功能模块的可视化平台，实现了三维白模构建与应用全流程的统一管理与交互操作。

综上所述，本文构建了从遥感影像到城市级三维白模的全流程自动化技术体系，突破了传统建模对现场测绘的依赖，实现了建筑信息提取、三维建模及工程应用的端到端自动化处理。该技术体系已在长沙市完成验证，并成功应用于建筑安全监测与违法加层检测等工程场景，为智慧城市建设与城市安全精细化管理提供了高效可靠的技术支撑。

关键词：深度学习；Transformer 模型；遥感光学影像；语义分割；建筑高度估计；三维重建

Abstract

The rapid acquisition of urban three-dimensional building information is a fundamental prerequisite for smart city development and refined urban safety management. However, existing urban 3D modeling approaches generally suffer from several limitations, including high data acquisition costs, strong dependence on computational resources, and prolonged update cycles. In addition, challenges remain in building footprint extraction, such as blurred boundaries and interference from complex backgrounds; in height estimation, including spatial information loss and poor cross-domain generalization; and in 3D white model construction, where workflows lack systematic integration and insufficiently align with practical engineering demands. Overcoming the constraints of traditional field surveying to enable efficient remote acquisition of urban 3D building information has become a key scientific and engineering problem, essential for providing accurate and reliable data support for digital urban infrastructure, emergency management, and planning decision-making. To address these challenges, this study proposes a deep learning–based technical framework for building footprint extraction, height estimation, and rapid construction and engineering application of city-scale 3D white models from remote sensing imagery, achieving full pipeline integration from data acquisition to deployment. The main contributions are as follows:

(1) A fine-grained building footprint extraction network, SPR-Net, based on pseudo-multimodal structural information is proposed. This network addresses limited extraction accuracy caused by blurred boundaries, significant scale variations, and complex background interference in remote sensing imagery. Adopting Segformer as the backbone, the network innovatively incorporates two types of pseudo-multimodal structural enhancement modules at both the encoder and decoder ends, effectively strengthening boundary features using only single optical imagery. Concurrently, a high-resolution building dataset covering Changsha was constructed, and a composite loss function was designed to sharpen segmentation boundaries. Systematic comparative experiments on public and self-built datasets demonstrate that SPR-Net consistently outperforms mainstream segmentation networks and State-of-the-Art (SOTA) models across IoU, Recall, Precision, and F1-score. Ablation studies verify the effectiveness and synergistic gains

of each component, while complexity analysis indicates that the model maintains high computational efficiency alongside high precision.

(2) The Segformer-H network, based on cascaded dynamic upsampling for building height estimation. This network tackles core issues in height estimation, such as spatial information loss during feature fusion, boundary blurring, and the difficulty of generalizing across diverse urban morphologies. While retaining the global semantic modeling capabilities of the hierarchical Transformer encoder, the network introduces a cascaded dynamic upsampling module based on learnable offsets to replace traditional bilinear interpolation. This effectively overcomes the blurring effects introduced by fixed kernels at building edges and height-discontinuity regions. A Changsha building height dataset covering a wide range of elevations was also developed. Experimental results on multiple datasets show that Segformer-H achieves optimal performance across all evaluation metrics, demonstrating significant advantages in sharp edge delineation and predictive stability in complex scenarios.

(3) An automated generation method for urban-scale LoD1 3D white-box models is established. This method addresses the systematic lack of modeling workflows and their poor alignment with engineering needs, expanding the application paradigm of 3D models in structural safety monitoring. Through key technologies—including spatial stitching of predictions, footprint regularization, height assignment, and vectorized extrusion—a complete automated generation pipeline for urban 3D models is realized. Large-scale experiments in typical urban functional zones show that the method stably generates geometrically regular and structurally sound 3D models. Regarding engineering applications, a spatial matching link between PS-InSAR deformation monitoring points and 3D models was developed to effectively identify undetected buildings; additionally, an intelligent identification method for illegal building additions was designed using multi-temporal imagery and 3D time-series analysis, achieving high precision and recall. Furthermore, an integrated visualization platform was developed to provide unified management and interaction for the entire modeling and application workflow.

In summary, this thesis constructs a full-process automated technical system from remote sensing imagery to urban-scale 3D models, breaking the reliance on traditional field surveying and achieving end-to-end processing for information extraction, modeling, and application. The system has been validated in Changsha

and successfully applied to building safety monitoring and illegal construction detection, providing efficient and reliable technical support for smart city development and refined urban governance.

Key Words: Deep Learning; Transformer; Optical Remote Sensing Imagery; Semantic Segmentation; Building Height Estimation; 3D Reconstruction

目　录

插图索引

附表索引

第1章　绪论

1.1 研究背景及意义

随着全球数字化浪潮的深入演进，以及我国城市发展从大规模增量扩展向存量提质增效的历史性转型，“智慧城市”、“数字孪生”[1]与“韧性城市”建设已成为国家提升宏观治理能力、推动高质量发展的重要战略引擎。在这一演进过程中，构建真实、统一、可更新的城市数字化时空底座，尤其是城市建筑群的三维模型，具有不可替代的基础性意义。近年来，国家层面相继出台多项重大政策，明确了三维城市模型在现代城市治理中的核心地位。《实景三维中国建设总体实施方案（2022-2025 年）》中明确提出，城市级实景三维建设需全面展开基础三维模型（即城市白模）的构建[2]；国家发展改革委等多部委联合发布的《关于深化智慧城市发展，推进城市全域数字化转型的指导意见》也指出，数字化基础设施是推动城市智能化转型的核心任务[3]；此外，《关于推进新型城市基础设施建设打造韧性城市的意见》[4]进一步强调，需强化城市信息模型（CIM）平台与其他基础时空平台的协同，全面赋能政务服务、防灾减灾与城市体检等领域。在此宏观背景下，大范围、高精度的城市三维模型已不再是单纯的视觉展示工具，而是支撑国家数字经济发展和城市全域数字化转型的核心数据基础设施。

随着城市化进程的加速，城市空间结构日益复杂，面临的潜在风险类型逐渐多样化。如何高效、准确地监测并评估这些风险，开展常态化的“城市体检”，成为了现代城市精细化管理中的一项重大课题[5]。特别是房屋定期体检、房屋养老金、房屋安全责任保险这“三项制度”的提出，为城市安全风险防控提供了系统性的制度支撑，标志着我国城市管理正从“被动应急”向“主动预防”的深刻转型。以合成孔径雷达干涉技术（InSAR）为例，其凭借高精度、大范围、全天候的特点，已被广泛应用于城市地表形变与基础设施风险的监测中[6]。然而，InSAR 技术获取的雷达信号由于侧视成像的特性，常伴有叠掩、透视收缩等几何畸变，且其解算出的离散形变点云本身缺乏直观的语义与拓扑结构[7]。此时，三维城市模型的构建尤为重要，它能够为多源遥感观测结果提供关键的三维空间依托载体。将 InSAR 形变数据与高精度的三维建筑模型进行空间配准与融合可视化，不仅能精准定位形变发生的具体建筑与楼层部位，还能大幅提升城市风险识别的精度与效率[8]。此外，在常态化的城市安全排查中，城市三维模型（特别是建筑轮廓与高度信息）发挥着举足轻重的作用。通过快速获取大范围的建筑高度与形态数据，城市管理者能够实现对全市范围内违法加层、违规扩建的自动化[9]、精准化监测，有效打破了老旧城区人工巡查难、监管盲区大的困境[10]。在地震、台风等极端灾

害的应急响应中，依托三维模型亦可迅速进行灾损评估与承载力分析，为生命救援和灾后重建提供不可或缺的科学决策依据[11, 12]。

图 1.1 基于实景三维的城市数字化时空底座

然而面对城市级动辄数百乃至数千平方千米的广袤区域，三维数据的获取与模型重建面临着严峻的算力与效率挑战[13]。当前，城市三维模型的获取途径主要依赖于无人机倾斜摄影测量[14]、地面激光扫描（LiDAR）[15]以及基于视觉的SLAM[16]等传统物理建模方法。这些方法虽然在局部微观尺度上具备极高的几何精度，但在城市级宏观尺度应用时却暴露出致命的短板：第一，数据采集成本高昂、部署周期漫长；第二，算法算力需求大、计算效率低。处理一个中等城市的实景三维数据，往往需要庞大的集群算力（GPU、CPU 集群）进行数周乃至数月的连续运算，伴随的算力开销和时间成本极其惊人。这种效率低下、算力门槛极高的技术路径，导致城市模型的更新周期长，难以满足现代“城市体检”与智慧城市建设对时空数据“高频次更新、大范围覆盖、低成本获取”的迫切要求[17]。

在此技术瓶颈下，光学遥感影像凭借其覆盖范围广、获取周期短、数据一致性强等优势，为城市级大范围的三维信息提取提供了最现实可行的数据源[18]。与传统依赖繁杂人工干预、手工设计特征和重度物理几何解算的重建方法相比，基于深度学习的遥感影像建筑三维重建方法展现出了显著的优势，为突破传统算力与效率瓶颈提供了革命性的解题思路。

深度学习技术，尤其是卷积神经网络（CNN）和变压器（Transformer）等先进架构的引入，能够实现自动化、端到端的特征学习和图像处理，无需过多依赖人工设定的规则或先验知识。在建筑物轮廓提取方面，深度学习方法通过大规模训练数据集，能够识别建筑物的多种形态特征，并凭借像素级别的深层学习，自动克服复杂地理环境中的阴影遮挡等干扰，极大提高了轮廓提取的精度和效率[18,

[19]；在建筑高度估计方面，深度学习能够通过深度网络的层次化学习，对二维影像中的建筑物进行精细化分析，从而精确回归建筑物的高度。同时，深度学习模型通过不断优化和更新，其泛化能力持续增强，可以实现对复杂城市环境的自动适应，从而提供更为准确、可靠的高度估算[20, 21]。

此外，深度学习模型最大的优势在于算力与效率的解耦[22]：深度学习方法将庞大的算力消耗前置于模型的训练阶段，一旦模型训练完成，在推理端（即实际生成城市模型时）仅需极小的算力开销即可实现秒级或分钟级的高效计算。基于这种数据驱动范式，模型能够自适应地从二维光学影像中直接提取复杂的建筑物轮廓，并精确回归建筑高度。这不仅有效克服了复杂城市环境下的阴影遮挡、数据噪声等干扰，更彻底省去了传统方法中冗长的图像密集匹配过程[23]，将大范围城市三维建模的处理时间从“月、周”级别压缩至“天、小时”级别[13]，真正实现了高自动化、低算力依赖的快速城市三维重建。

因此，本研究基于深度学习的城市建筑群快速三维建模与应用，将聚焦于光学遥感影像的高效处理与深度学习技术的创新融合，围绕建筑物轮廓精准提取、建筑高度智能估算、城市建筑群三维快速建模及其在城市安全监测中的应用开展系列研究，力求构建一套高效率、低成本、可更新的城市三维建模技术体系。

1.2 国内外研究现状

综合国内外研究成果，当前研究主要聚焦于三个方面：第一，不断创新骨干网络的特征表征能力，通过改进模型架构，从根源上提升特征提取与识别能力；第二，在建筑轮廓提取上，通过多尺度特征融合与边缘解耦监督，实现从像素掩模向规则化几何多边形的精度跃升；第三，在建筑高度估计上，从传统阴影几何解算转向神经网络端到端预测。上述研究共同指向构建高自动化、低算力依赖的城市级数字化底座生成体系，为数字孪生城市建设与常态化城市监测提供时空数据支撑。

1.2.1 基本网络架构发展综述

基于深度学习的遥感建筑提取精度在很大程度上受限于网络架构的发展。梳理网络基本架构的演变并应用最新框架，是提升提取精度的关键途径。

1. 卷积神经网络架构

语义分割的突破性进展始于全卷积神经网络 FCN 的提出。Long[24]等将传统分类网络中的全连接层替换为卷积层，并引入反卷积上采样操作，首次实现了对任意尺寸图像的端到端像素级分类，从根本上改变了此前依赖滑动窗口进行逐像素预测的低效范式。FCN 的核心贡献在于证明了深度卷积特征可以被直接用于密集预测任务，然而其在恢复空间分辨率时采用的简单双线性插值策略导致细节信

息大量丢失，在建筑边界等高频区域的分割结果较为粗糙。为解决 FCN 在精细化分割方面的不足，Ronneberger 等提出的 UNet[25]创新性地采用了对称的编码器—解码器架构，该网络最初面向医学图像分割设计，其核心创新在于通过跳跃连接（Skip Connections）将编码器各阶段的高分辨率空间细节与解码器对应层级的低分辨率语义特征进行逐级融合，使得解码器在恢复空间分辨率的同时能够保留丰富的上下文语义信息。这一经典的“编码—桥接—解码”设计范式深刻影响了后续几乎所有遥感图像分割网络的架构走向，跳跃连接的思想也成为现代密集预测网络中最为基础和普遍的设计元素之一。在目标检测领域，Ren 等提出的 Faster R-CNN[26]通过引入区域生成网络（RPN）构建了两阶段检测框架，在大幅提升检测效率与精度的同时，也为后续实例分割任务奠定了架构基础。为了进一步优化计算效率，Badrinarayanan 等提出的 SegNet[27]在解码器的上采样策略上进行了巧妙的创新。与 UNet 通过跳跃连接直接传递完整的浅层特征图不同，SegNet 的解码器利用编码器阶段最大池化操作时记录的池化索引（Pooling Indices）进行非线性上采样，仅需传递索引位置信息而无需存储完整的特征图，从而有效降低了模型的内存消耗和计算复杂度。这一设计理念在面向资源受限的嵌入式部署场景时表现出显著优势，为后续轻量化分割网络的研究提供了重要启示。

在全局上下文信息建模方面，Zhao 等提出的 PSPNet[28]引入了金字塔池化模块（Pyramid Pooling Module），通过在四个不同尺度上对特征图进行自适应平均池化并聚合，有效捕获了从局部纹理到全局场景布局的多粒度上下文信息。该方法在当时的多个语义分割基准数据集上取得了最优性能，其核心贡献在于揭示了全局上下文建模对于解决复杂场景中类别混淆问题的关键作用，特别是在遥感影像中大面积地物（如广场、停车场）与建筑屋顶容易混淆的情况下，全局信息的引入能够显著降低误分类率。与 PSPNet 采用池化策略不同，Chen 等提出的 DeepLab 系列[29]则另辟蹊径，创新性地引入空洞卷积（Atrous Convolution，又称膨胀卷积），通过在标准卷积核中插入空洞来扩展采样间隔，在不增加参数量和计算量的前提下有效扩大了网络的感受野，成功克服了传统卷积因多次池化而导致分辨率急剧下降的瓶颈。随后的 DeepLabv3[30]进一步改进了空洞空间金字塔池化（Atrous Spatial Pyramid Pooling，ASPP）模块，通过并行使用多个不同空洞率的卷积核捕获多尺度上下文信息，并增加了全局平均池化分支以编码图像级别的全局特征。DeepLabv3+[31]则在此基础上将 ASPP 与编码器-解码器结构深度整合，通过在解码器中融入低层特征来恢复空间细节，极大增强了模型对多尺度特征的捕获能力，并显著改善了建筑物等目标在边界区域的分割精度。DeepLab 系列的持续演进充分说明，感受野的有效扩展与多尺度特征的高效融合是密集预测任务中两个相辅相成的核心技术要素。

2. 实例分割与轻量化

随着应用场景的复杂化，图像分割逐渐向更精细的实例级别发展。Mask R-CNN[32]在 Faster R-CNN 的检测框架上增加了一个并行的像素级分割分支，首次将目标检测与实例分割完美结合，标志着实例分割时代的正式开启。为进一步细化语义分割的边缘特征，UNet++[33]对传统跳跃连接进行了拓展，设计了密集的嵌套跳跃路径（Nested Skip Pathways），与 UNet 仅在编码器和解码器对应层级之间建立单一跳跃连接不同，UNet++在不同层级之间构建了多个中间卷积节点，形成了稠密的子网络结构，使得底层特征在向高层传递时经过逐层语义增强，有效弥合了编码器浅层特征与解码器深层特征之间的语义鸿沟，在细节恢复与分割精度方面均获得了显著提升。针对自动驾驶、无人机遥感等对推理速度要求极高的实时场景，研究者们提出了 BiSeNet[34]、Fast-SCNN[35]和 FastFCN[36]等轻量化网络。以 BiSeNet 为例，其核心的双路径（Bilateral Path）设计分别专注于捕捉细粒度空间信息和提取全局语义信息，成功实现了分割精度与推理速度的良好平衡。与此同时，注意力机制（Attention Mechanism）开始在 CNN 骨干网络中展现出独特的优势：DANet[37]引入了空间与通道双重注意力模块，从两个维度强化了特征的判别能力；而 ResNeSt[38]则通过引入 Split-Attention 模块，有效捕捉了不同尺度和通道间的特征交互，进一步提升了神经网络的特征表示能力。

3. Transformer 架构

近年来，Transformer 凭借其强大的并行计算能力和卓越的全局依赖建模能力，有效弥补了传统 CNN 因感受野受限而导致长距离特征提取不佳的劣势，引发了计算机视觉领域的架构范式变革。Dosovitskiy 等提出的 Vision Transformer[39]率先将自然语言处理领域的 Transformer 架构引入计算机视觉，其核心思想是将输入图像划分为固定大小的图像块（Patch），将每个图像块线性映射为一维向量后作为序列输入标准 Transformer 编码器进行全局自注意力计算。ViT 的成功不仅在于技术层面的突破，更在于它证明了当训练数据足够充分时，纯注意力架构能够在视觉任务上媲美甚至超越精心设计的 CNN 模型，从根本上动摇了卷积操作在视觉特征提取中的主导地位。为解决标准 Transformer 处理高分辨率图像时计算复杂度激增的瓶颈，Swin Transformer[40]引入了局部注意力与滑动窗口机制，不仅限制了计算量，还通过层级化架构生成多尺度特征图，完美契合了密集预测任务的需求。Strudel 等提出的 Segmenter[41]则将图像分割重新定义为掩码分类问题，其独特的掩码变换器（Mask Transformer）解码器通过学习一组可训练的类别嵌入向量，与编码器输出的图像块特征进行交叉注意力交互，直接预测每个类别的分割掩码。这种完全基于 Transformer 的端到端设计摒弃了传统方法中繁琐的手工设计特征和后处理步骤，同时保持了像素间全局关系的建模能力，在多个基准数据集上取得了具有竞争力的分割性能。Chu 等提出的 Twins[42]则利用金字塔聚类和动态滑动窗口机制，在计算资源受限的情况下依然保持了较高的特征提取精度。

然而，上述基于 ViT 的早期变种模型普遍继承了 ViT 较为庞大的参数量和计算开销，在处理遥感领域常见的超高分辨率影像时仍面临显著的效率瓶颈。针对这一痛点，轻量化与高效化成为 Transformer 应用于视觉分割任务的新研究热点。Xie 等提出的 SegFormer[43]巧妙地结合了层次化 Transformer 编码器与轻量级 MLP（多层感知机）解码器，其编码器采用混合注意力机制（Mix-FFN）替代传统的位置编码，通过在前馈网络中引入 3×3 深度卷积隐式编码位置信息，从而避免了因测试分辨率与训练分辨率不一致时位置编码插值带来的性能衰减。该方法在保证分割精度的同时，将参数量缩减至此前最优方法的五分之一，为 Transformer 在资源受限场景下的实际部署提供了切实可行的解决方案。Cheng 等提出的 Mask2Former[44]则通过创新的掩码注意力机制与跨尺度交互，将全景、实例与语义分割统一到了单一框架中，实现了更高效的边界预测。最新提出的 SegNeXt[45]则反其道而行之，反转了常规的 Transformer 为主要架构卷积为辅助架构的思路，利用多尺度卷积特征唤起空间注意力，实现了从局部到全局的高效信息聚合。

随着深度学习架构的进一步演进，最新的 Mamba 网络[46]作为一种基于状态空间模型（State Space Models, SSMs）的全新基础架构，开始在视觉任务中展现出颠覆性的潜力。Mamba 的核心在于其引入了选择性状态空间机制（Selective State Space Mechanism），成功打破了 Transformer 中自注意力机制带来的二次计算复杂度瓶颈。通过在保持甚至超越全局长距离依赖建模能力的同时实现线性时间复杂度，Mamba 不仅大幅降低了处理高分辨率遥感图像时的显存占用与计算开销，还能更高效地提取复杂的空间上下文特征。然而，也正因为其通过压缩状态表示来换取计算效率的提升，Mamba 在某些需要精细化空间推理的任务上（如小目标的精确分割）表现不及同等规模的 Transformer 模型，特征表达能力的上限有所受限。考虑到未来算力资源的持续增长趋势，在面向高精度遥感建筑提取的应用场景中，Transformer 架构凭借其更强的特征建模能力和更灵活的注意力机制设计空间，仍然具有广阔的发展前景和优化潜力。

1.2.2 基于遥感影像的建筑轮廓提取算法

传统的基于遥感影像的建筑轮廓提取方法主要可以分为三类[18]：（1）基于几何特征提取的方法；（2）基于区域分割与组合的方法；（3）基于分类器的方法。这些传统方法的建筑轮廓提取精度普遍偏低，IoU 指标通常在 0.55-0.68 之间，F1 分数仅为 0.60-0.72，难以满足高精度制图需求[47]。

在第一种类型中，通常先提取建筑物的角和边等几何基元，再将其组装成封闭多边形，但该方法的弊端在于边缘和角点的检测极易受图像噪声、阴影和背景纹理的干扰，导致组装复杂且易出错。第二种方法通过图像分区获得超像素片段来描绘建筑，但过度分割会生成大量冗余或无关的超像素，显著增加了后续处理

的复杂性。第三类方法则依赖于手工特征提取并输入分类器，其局限性在于过于依赖人工设计特征，且最终提取的正确率严重受限于分类器的性能。在过去的十年里，大量基于深度学习的方法被相继提出，并在效率和准确性方面显著优于传统方法。现有的基于深度学习的建筑轮廓提取网络，主要从多尺度特征融合、边界与形状强化、多边形直接矢量化提取，以及多源跨模态数据融合等几个核心方向展开了深入研究。

1. 多尺度掩模与特征融合提取

由于遥感影像中建筑分布尺度差异巨大，从城中村的数十平方米小型自建房到大型商业综合体或工业厂房往往共存于同一幅影像中，多尺度特征的高效提取与融合是该类方法的核心挑战，该方法的基本思想是：通过构建多尺度特征提取与融合机制，使模型能够同时感知不同空间尺度的建筑目标，从而提升对大小差异显著建筑的检测与识别能力。Zhu 等[48]针对提取边界不准确及大型建筑不连续的问题，提出多参与路径神经网络（MAP-Net），与现有的多尺度特征提取策略通常采用单路径递进式下采样不同，MAP-Net 创新性地设计了多条并行路径，每条路径在固定分辨率上提取具有不同感受野的高级语义特征，从而在保留空间定位精度的同时实现了真正意义上的多尺度特征学习。此外，MAP-Net 引入注意力驱动的通道融合模块自适应地优化来自各路径的特征权重，并通过金字塔空间池模块捕获全局依赖性以细化不连续的建筑足迹，在多个公开数据集上取得了领先的提取精度。Zhao 等[49]提出了 MSRF-Net，专门面向多尺度建筑物的精确提取。该网络设计了多尺度自适应膨胀（MSAD）模块以丰富特征的感受野覆盖范围，并通过自适应分辨率恢复（ARI）模块在解码阶段恢复特征的空间分辨率，配合多路径解码器捕获多尺度上下文信息。然而，MSRF-Net 的参数量较大且对完整语义分割标签的依赖性强，标注成本高昂；同时，在山区或森林等复杂地形环境下的小型单体建筑识别效果仍有待提高，暴露出该方法在极端场景下泛化能力的不足。为了应对大规模提取的挑战，Huang 等[50]提出的 BldgNet 结合了大窗口注意力与边缘注意模块，并引入半监督训练方法利用开源数据增强模型的鲁棒性。Chen 等[51]则从影像质量增强的角度出发，提出了一种基于超分辨率（SR）技术辅助实例分割的大规模建筑物提取框架。该框架包含场景分类、颜色归一化与图像超分辨率、建筑实例分割以及场景镶嵌四个主要步骤，通过在分割前利用深度学习超分辨率网络增强输入影像的空间细节，有效改善了中低分辨率遥感影像中小型建筑物的可识别性，为提升下游分割精度提供了高质量的输入数据。此外，考虑到特征融合的冗余问题，Huang 等[52]提出轻量级网络 Easy-Net，主张仅融合前三阶段存在的建筑特有特征，结合 CNN 与 Transformer 优势提升了提取效率。近年来，随着基础模型的爆发，基于预训练大模型的遥感建筑提取方法开始崭露头角。例如，基于 Segment Anything Model（SAM）的遥感微调架构通过提示学

习（Prompt Learning）策略，将 SAM 强大的零样本、少样本泛化能力直接迁移至遥感多尺度建筑提取任务中，在极少量标注甚至无标注的条件下即可获得较为理想的提取效果，极大降低了对大规模精确标注数据的依赖，为遥感建筑提取的范式创新开辟了全新方向。

2. 边界正则化与形状约束提取

边界正则化是指通过算法优化使提取的建筑物边界更加平滑、连续和锐利的过程，旨在消除初始提取结果中常见的锯齿状、断裂等不规则现象。形状约束则是利用建筑物固有的几何特性，如矩形结构、直角特征等作为先验知识，引导模型生成更符合真实建筑物形态的规则轮廓。该类方法的基本思想是将建筑物的几何属性与空间相关性转化为显式的优化目标或约束机制，以弥补卷积神经网络在像素级分类中对边缘感知力不足的缺陷。在边界正则化与形状约束提取方面，针对传统卷积神经网络在像素级预测中容易产生弱边界和粗糙标签的问题，研究者致力于通过解耦、图模型与先验约束恢复锐利且规则的轮廓。Li 等[53]提出了一种集成 CNN 和特征成对条件随机场（FPCRF）的端到端方法，利用图模型处理空间相关性以产生清晰边界，该方法的创新在于将概率图模型与深度特征进行了端到端的联合优化，避免了传统 CRF 后处理方法中特征与推理分离带来的次优问题。Li 等[54]则从特征表示的维度出发，提出了高分辨率解耦网络 HD-Net，通过特征解耦-重新耦合（FDR）模块，巧妙地将特征分解为“主体”和“边界”进行深度监督，确保了建筑的整体连续性及边界精度。为了引入形状先验，Hu 等[55]计算了傅立叶描述符并设计形状感知损失函数来约束建筑形状，从而在优化过程中显式地将建筑的规则几何先验知识注入网络的学习目标中，有效改善了预测轮廓的规则度。Wu 等[56]则从另一个角度引入地形感知损失（TAL）以增强异构建筑特征的边界保留能力。同时，生成式网络也被广泛应用，He 等[57]通过社会几何特征驱动的生成引擎应对图像模糊与遮挡；Li 等[58]则利用生成对抗网络（GAN），通过多尺度判别器促使生成器输出更真实、具有边界正则化效果的建筑足迹。

在基于多边形直接组装的矢量化提取方面，为了直接输出符合地理信息系统（GIS）实际应用需求的矢量数据，端到端的多边形提取（轮廓提取的另一个流派，直接输出建筑轮廓的多边形，而不是建筑的白色掩膜）成为一大重要趋势。Wei 等[59]提出了 Line2Poly，采用结合 CNN 和 Transformer 的两阶段策略，以特征线作为几何基元，直接组装成建筑多边形，有效捕获了远程拓扑信息。与此类似，Zhang 等[60]引入了 P2PFormer 管道，首先对顶点、线和角等几何图元进行分割，并使用 Transformer 预测其序列，无需繁琐的后处理即可生成规则轮廓。在分层优化方面，Xu 等[61]提出的 HiSup 方法以掩模可逆性的新视角，面向建筑边缘线段的中层矢量场引导机制（即学习像素到边缘线段的方向向量），通过由底向上的分层监督信号缩小了分割掩模与多边形之间的差距；Li 等[62]则设计了包含多任务深度网络、

顶点选择模块以及基于图的顶点细化网络的完整管道，能够将多边形顶点坐标调整到更准确的位置，生成具有高精度边和形状的最终矢量建筑。

3. 基于多源与跨模态数据融合的提取

在基于多源与跨模态数据融合的提取方面，针对单一光学影像在遮挡、阴影或复杂背景下的局限性，融合高程模型或合成孔径雷达（SAR）数据成为了提升提取精度的重要手段，该方法的基本思想是通过引入多源、多模态数据增强信息表达能力，提升数据的丰富度与互补性，从而为模型提供更加全面的特征输入；其核心在于设计有效的跨模态特征融合机制，使不同来源的数据在特征层面充分交互与协同，增强模型的判别能力与泛化性能，最终提高提取结果的准确性与稳定性。Huang 等[63]开发了 GRRNet，融合高分辨率航空图像和激光雷达点云，并引入门控特征标记（GFL）单元以减少不必要的特征传输。Hosseinpour 等[64]提出的 CMGFNet 则通过门控融合模块（GFM）融合光学图像与数字高程数据，充分利用了多模态特征。针对 SAR 图像中复杂的斑点噪声和多尺度目标，Sun 等[65]提出 CG-Net，利用建筑足迹先验来标准化特征以预测 SAR 图像中的掩模；Jing 等[66]提出的 SSPD 网络不仅通过多层模块自适应提取关键信息，还专门设计了 L 型加权损失（LWloss）以减少线型建筑物的漏检并缓解类别不平衡。此外，Li 等[67]提出的 MMFNet 将相位作为模态不变量，实现了 SAR 与光学图像深层特征的渐进式动态融合。

综上所述，现有的基于深度学习的遥感建筑物提取网络，大多采用“先进视觉架构作为骨干+遥感先验知识作为辅助模块”的范式进行轮廓提取。尽管取得了丰硕成果，但提出一种不依赖繁琐传统先验信息的纯粹深度学习网络架构仍是非常有必要的。最新的研究已充分证明，CNN 与 Transformer 的有机结合在获取建筑局部细节与全局长距离语义信息方面具有显著优势。同时，由于遥感图像中建筑分布尺度的高度不均匀，多尺度特征提取与融合成为几乎每一个网络的必备模块；然而，传统特征金字塔网络中的简单拼接或相加操作往往未能充分利用多尺度融合的潜力，导致跨层级特征之间的相关性交互不足。此外，针对提取任务的特殊性（如直角、规则边缘），量身定制具有极强针对性的损失函数（如形状约束、拓扑感知损失等），也是突破现有网络分割精度瓶颈的核心方法之一。

1.2.3 基于遥感影像的建筑高度估计算法

1. 传统几何物理建模方法

建筑高度作为刻画城市三维形态与空间异质性的关键参数，其遥感估算方法经历了从手工物理建模到深度学习自动特征提取的演变。早期的建筑高度估计研究主要基于遥感影像的几何投影与物理成像机理进行推算，可归纳为三类代表性方法。而受限于几何建模精度与影像质量，这类传统方法的高度估计误差通常在

6-8 米之间，难以满足精细化城市管理的实际需求[68]。

第一类是基于阴影长度的几何推算法。经典思路是利用光学影像中建筑物投射阴影的长度，结合卫星成像时的太阳高度角和方位角等参数，基于空间三角几何关系反算建筑物高度值[68]，该方法原理直观且无需额外数据源，但其精度高度依赖于阴影分割的精细程度——在建筑密集的城市核心区，相邻建筑的阴影频繁重叠甚至完全遮蔽地面，加之不同季节太阳入射角的变化导致阴影方向和长度的剧烈波动，使得阴影的精确提取和归属判定成为难以逾越的技术障碍，方法的鲁棒性在实际城市场景中严重不足。

第二类是基于楼层识别的间接估算方法。从高分辨率光学影像中通过纹理分析或目标检测技术识别建筑物的楼层数，并按照经验性的单层高度标准（通常约 3 米）进行线性折算以估计总体建筑高度[68]。该方法的局限性在于：一方面，楼层数的准确识别受限于影像空间分辨率和建筑外立面可见程度，特别是当建筑采用玻璃幕墙等连续外立面设计时，楼层间界线难以辨识；另一方面，住宅楼、商业裙楼、工业厂房等不同功能类型的建筑，其单层高度差异显著（从 2.8 米到 5 米不等），采用统一的折算标准会引入系统性偏差，且对于超高层建筑而言，单层高度的细微误差经过数十层的累积后将产生严重的总体偏差。

第三类基于雷达干涉的相位测高法。利用合成孔径雷达（SAR）卫星的侧视成像特性，通过干涉雷达技术（InSAR）获取地表散射体的相位信息，结合参考数字高程模型（DEM）分析建筑立面上不同永久散射体（PS）点之间的高差来推断建筑物高度[69]，该方法在理论上能够实现毫米级的形变测量精度，但在实际应用中面临多重挑战：建筑密集区域中不同建筑立面的 PS 点信号容易杂糅，导致难以准确区分和归属；建筑物顶端的 PS 点因屋顶结构复杂而难以精确定位；此外，SAR 影像固有的透视收缩、叠掩和阴影等几何畸变效应也会严重干扰高度信息的准确提取。

2. 基于深度学习的端到端回归方法

随着卷积神经网络的兴起，基于深度学习的端到端回归模型凭借其强大的非线性拟合能力和自动化特征学习能力，逐渐成为建筑高度估计的主流范式。CNN 通过多层卷积结构自动提取影像中的边缘、纹理、阴影及建筑形态等特征，并在大规模样本训练过程中学习这些视觉线索与实际高度之间的非线性映射关系，从而实现由二维影像到三维高度信息的有效推断，克服了传统方法对人工特征设计和精确几何参数的依赖。为了克服中低分辨率影像中空间细节缺失的问题，Cao 等[70]提出了一种超分辨率（SR）与高度分层估计（HS）协同的深度学习网络，实现了从 Sentinel-1/2 融合数据生成空间分辨率为 2.5 米的建筑物高度图。该方法的核心包括两个功能互补的模块：超分辨率模块通过学习低分辨率到高分辨率的映射关系来恢复影像的空间细节；高度分层估计模块则将建筑物高度值划分为若干

离散层级，通过一个编码器和两个解码器分别估计建筑物高度和高度层级标签，以此缓解因高度值分布严重不平衡而导致的训练偏差问题。针对高度值在空间分布上的非线性特征，Li 等[71]提出了一种创新性的问题建模策略，将传统的连续回归问题转化为序数回归问题。该方法将高度值划分为间距递增的离散区间，并设计序数损失函数进行网络训练，利用相邻高度区间之间的有序关系提供额外的监督信号。同时，为实现多尺度特征提取，该方法进一步集成了空洞空间金字塔池化（ASPP）模块从多个膨胀率的卷积层中聚合特征，有效提升了高度突变区域的预测精度。该研究的意义在于揭示了回归问题的建模方式对高度估计精度的显著影响，为后续工作提供了重要启发。针对 SAR 影像的特殊几何畸变，Sun 等[72]提出将高度检索建模为边界框回归问题，通过 TerraSAR-X 影像预测建筑立面投影的边界框长度来反推高度。这些研究证明了深度网络在处理复杂城市地物特征时的强大非线性拟合能力。

当前的研究前沿更倾向于通过多源多模态数据的深度融合与语义约束来消除单一传感器的局限性。在多模态融合方面，Nascetti 等[73]提出的 MBHR-Net 在下采样阶段对光学与 SAR 特征进行深度耦合；Gong 等[74]则引入最小二乘优化的滚动引导滤波（RGF）实现图像的多尺度分解，利用耦合神经系统融合异源特征以抑制散斑噪声。此外，为了增强物理可解释性，Jabbar 等[21]将经典的立体视差公式嵌入深度架构，而 Lyu 等[75]则通过四季复合影像提取阴影分布的统计特征来辅助层数估算。为了解决人口稠密区的信号干扰，Kakooei 等[76]甚至引入夜间灯光数据作为附加约束信息，有效校正了高层建筑区的多重散射偏差。

随后研究开始聚焦于利用 Transformer 架构和多任务学习范式实现建筑轮廓与高度的协同演化。Chen 等[77]提出了一种语义约束框架（MFTSC），通过 Transformer 模型系统地将多级特征与语义分割任务合并，利用建筑轮廓的形状先验引导高度生成，显著改善了建筑边缘模糊的问题。Sun 等[78]开发的 GABLE 框架则代表了目前大规模建模的高水平，其在 Unet 架构中嵌入 Transformer 进行全局建模，实现了全国尺度的 3D 建筑模型构建。更前沿的研究 HeightFormer[20]，通过自注意力机制捕捉建筑足迹与高度的空间长程依赖，彻底改变了传统 CNN 卷积核感受野受限的问题。这类方法不再依赖复杂的后处理模块或人工预验知识，而是通过高性能的注意力机制自发学习建筑几何形态与影像像素之间的对应关系。

综上所述，遥感建筑高度估计已从单一的几何解析演进为多任务协同、多模态融合的智能化制图模式。虽然现有研究在算法精度上取得了显著进展，但多数模型在面对不同地域、不同建筑风格时的泛化能力仍有待提高。现有的网络结构正逐步淘汰繁琐的先验经验与后处理模块，转向基于最新 Transformer 架构或大尺度预训练模型的内生约束设计。未来的研究重点应在于如何进一步解耦复杂的城市空间结构，利用生成式人工智能或自监督学习技术，在缺乏高精度地面真值

标定的情况下，实现更加稳健、精细化的全球建筑高度三维建模，从而为数字孪生城市与城市气候分析提供更为坚实的数据支撑。

1.3 目前研究的不足

本尽管国内外学者围绕遥感影像建筑物轮廓提取、建筑高度估计以及城市三维建模开展了大量研究，但距离低成本、自动化、可推广、可工程落地的城市级三维建筑信息快速获取目标仍有明显差距。现有研究的不足主要体现在以下三个方面：

1. 单一光学影像条件下建筑结构信息挖掘不足，复杂场景中的轮廓精细提取能力仍需提升。

现有建筑轮廓提取方法虽在多尺度特征融合、边界监督和形状约束等方面取得进展，但在高密度城区、老旧居民区和工业园区等复杂场景中，仍易受目标粘连、背景干扰和尺度差异影响，出现边界断裂、轮廓粘连和细长建筑漏检等问题。相关方法普遍侧重语义信息表达，对边缘、角点和规则结构等几何信息建模不足；而多模态融合虽可提升效果，但成本高、流程复杂，难以满足大范围快速应用需求。因此，如何仅依赖单一光学影像实现兼顾边界精细化、复杂背景鲁棒性与计算效率的建筑轮廓提取，仍是亟待解决的问题。

2. 建筑高度估计在空间细节保持和跨区域泛化方面仍存在不足。

现有建筑高度估计方法已由传统几何物理模型逐步转向深度学习驱动的端到端预测，但仍面临空间细节恢复不足、边界模糊及跨区域适应性弱等问题。传统方法对成像条件和先验参数依赖较强，复杂场景下稳定性有限；深度学习方法虽提升了预测效率，但在上采样和特征融合过程中易造成空间信息损失，影响建筑边缘和高度突变区域的表达。此外，现有模型普遍依赖特定区域数据训练，迁移到不同城市和成像条件时性能下降明显。因此，如何增强空间细节表达并提升跨区域泛化能力，仍是建筑高度估计研究的关键难点。

3. 城市级三维白模自动化构建与工程应用研究仍缺乏系统性。

当前研究多聚焦于轮廓提取或高度估计等单一任务，对“轮廓提取—高度估计—白模生成—业务应用”的完整技术链路关注不足，难以支撑城市级三维建筑信息的批量化、自动化和工程化应用。同时，不少三维建模仍依赖倾斜摄影、激光点云等高成本数据，难以满足大范围、快速更新需求。此外，三维白模与城市安全监测、规划核查、违法建设识别等场景的融合仍不深入，业务闭环尚未形成。因此，亟需构建面向城市级范围、以遥感影像为主要数据源、贯通模型生成与应用落地的全流程自动化技术体系。

1.4 本文研究内容

1.4.1 选题依据

本文选题基于国家自然科学基金项目《基于星载 InSAR 变形测量和信息共享平台数据的大跨度桥梁结构状态评估方法》（52278306）、湖南省重点研发计划项目《基于星载 InSAR 技术的城市桥梁群多灾害条件下状态评估与预警研究》（2022SK2096）、湖南省自然科学基金项目《基于天-空-地一体化的基础设施结构安全监测技术研究》（2023JJ70003）、湖南省应急厅应急管理科技项目《基于星载 InSAR 的“长沙市自建房”安全风险排查研究》（yjtkjxm_202406），旨在研究基于深度学习的城市建筑群快速三维建模方法与应用。首先基于 SPR-Net 网络提取完整建筑轮廓信息，实现光学影像建筑轮廓精准提取，其次，建立 Segformer-H 网络，实现建筑群高度的快速估计。最后，联合所获取的建筑轮廓与高度数据，实现广域建筑群外观形态的快速三维建模，并在此基础上开展城市建筑群的大范围智能安全排查，有效识别违法修建与违规加层等行为，为城市安全风险排查提供技术支撑。

1.4.2 研究内容

本文的目标旨在结合深度学习与遥感技术，提出一种基于深度学习的遥感影像建筑建筑群快速三维建模方法（见图 1.2），突破传统实地测量的局限性，实现建筑群级别的远程三维建模，为城市精细化监测提供高效、准确的数字基础设施。特别聚焦于建筑物的轮廓提取、建筑高度估计及城市建筑群三维建模及其在城市风险排查中的应用。主要研究内容如下：

第一章为绪论，首先阐述了选题的研究背景与现实意义，全面论述了城市建筑群三维快速建模的重要性，深入分析了当前技术存在的局限性及采用深度学习方法的明显优势；随后系统梳理了国内外研究现状，分别阐述了主流语义分割网络架构的发展脉络、基于深度学习的建筑轮廓提取算法研究进展，以及基于深度学习的建筑高度估计算法研究现状；然后总的介绍了目前研究的不足；最后阐述了本文的主要研究内容与创新点。

第二章阐述深度学习算法的理论基础。首先，系统阐述了深度学习的基本原理，包括模型参数化定义、损失函数设计、基于梯度下降与反向传播的优化机制，以及学习率调度、激活函数等关键训练组件。其次，深入剖析了卷积神经网络 CNN 与 Transformer 两类主流网络架构的原理与特性。然后，阐述了语义分割与高度回归等密集预测任务基于编码器—解码器的通用架构范式。最后，介绍了正射摄影与倾斜摄影的技术原理及数据特点，为建筑物三维建模的数据获取与处理奠定了基础。

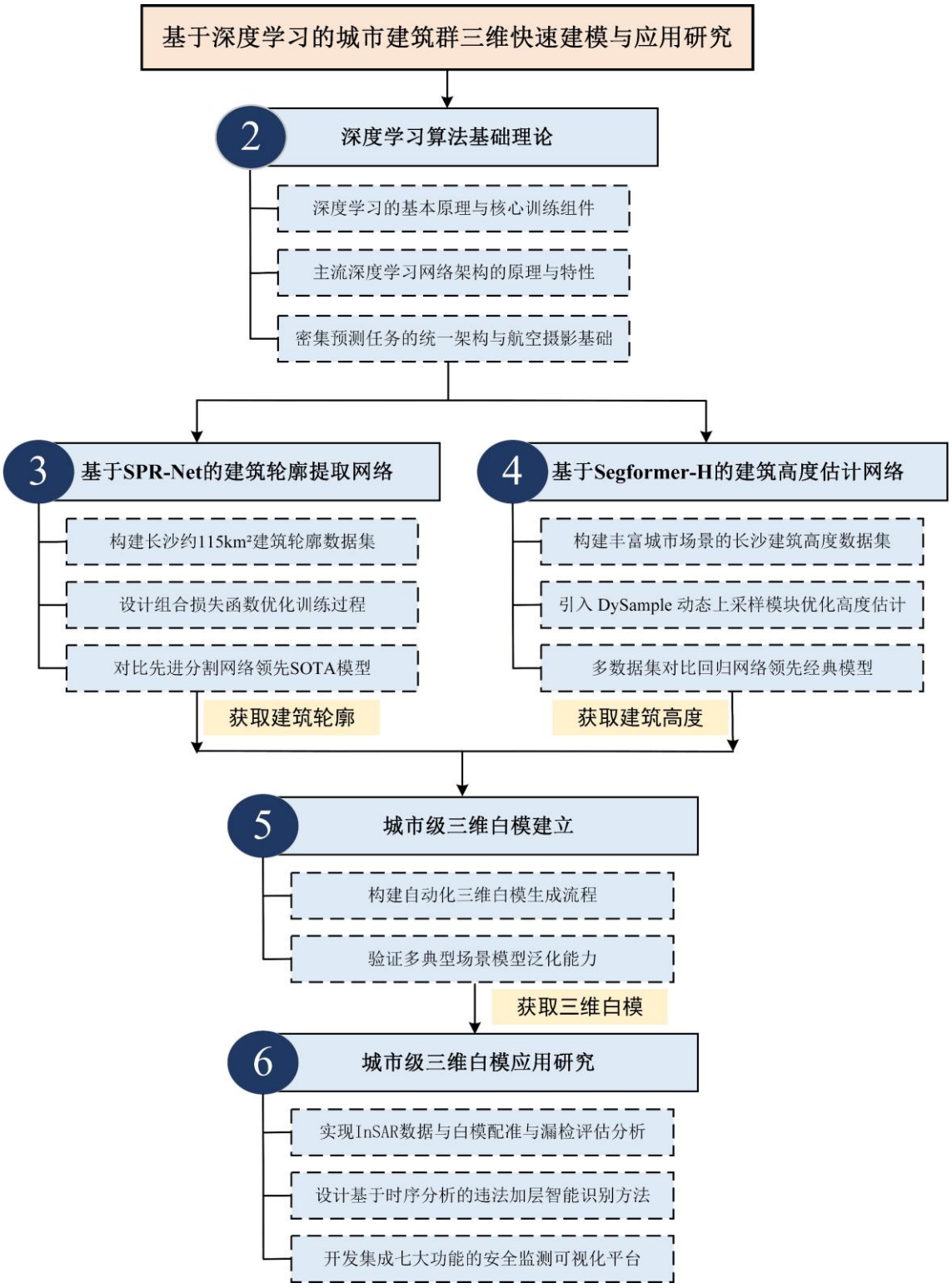


图 1.2 研究内容框架图

第三章提出了一种基于伪多模态结构信息的建筑物精细分割网络 SPR-Net，用于实现城市建筑群轮廓的精确提取。首先，以 SegFormer 作为主干网络，在编码端与解码端分别设计伪多模态模块 PMM_S 和 PMM_C，从高频特征显式构造与预测引导特征重加权两个层面增强边界表达能力，并引入 RFAConv 提升复杂轮廓刻画精度。同时构建复合损失函数，从区域重叠、边界约束与 IoU 优化三个

维度提升训练效果。随后，构建了覆盖长沙市约 115km²、涵盖城乡多样化建筑分布的高分辨率建筑轮廓数据集，并在 WHU 公开数据集与长沙自建数据集上与多种 SOTA 方法进行全面对比，通过消融实验和复杂度分析验证了模型在精度与效率上的优势。

第四章一种面向遥感影像建筑物高度估计的改进 Segformer 网络模型 Segformer-H，用于实现城市建筑群高度的快速估计。首先，在保留 Segformer 层次化 Transformer 编码器高效全局语义建模能力的基础上，将解码器双线性上采样替换为基于 DySample 的级联动态上采样模块，通过可学习偏移自适应优化采样位置，缓解边界与高度突变区域的模糊问题。其次，构建了涵盖 9m 至 250m 高度范围的长沙建筑高度数据集，弥补了现有公开数据集难以反映国内城市高度异质性强、建筑分布密集等特点的不足。最后，在 ISPRS Potsdam、ISPRS Vaihingen 及长沙自建数据集上与等主流方法进行全面对比，并开展跨区域独立泛化测试，验证了 Segformer-H 在多尺度复杂场景下建筑高度估计的精度优势与工程应用可行性。

第五章提出从深度学习预测结果到城市级 LoD1 三维白模的自动化构建流程。通过结果拼接融合、基于 RDP 与几何正交校正的轮廓正则化、中位数高度赋值及矢量拉伸建模，实现三维白模自动生成。在长沙市约 4 万余栋建筑、涵盖多类型城市功能区的实验中验证了方法的泛化能力与工程可行性。

第六章围绕城市建筑安全监测需求，探索了三维白模在 PS-InSAR 形变监测配准、建筑漏检识别以及违法加层智能识别等典型工程场景中的深层次应用。构建“配准—归属—漏检评估”技术链路，提出参数化三维置信空间与两级归属判定机制，实现 PS 点精准归属与漏检识别；同时基于多时相影像与三维白模差分分析，结合自适应阈值与空间一致性检验，实现违法加层自动检测。开发基于 Python 与 PyQt5 的可视化平台，集成三维白模构建与应用功能模块。

最后，总结全文技术成果与创新点，分析方法在数据适用性、模型泛化与工程部署方面的局限，并对未来研究方向进行展望。

第2章 深度学习算法的理论基础

2.1 引言

随着人工智能技术的跨越式发展，深度学习已从早期的模式识别工具演变为驱动计算机视觉、自然语言处理、语音识别乃至科学计算等诸多领域取得突破性进展的核心引擎。特别是在计算机视觉领域，以卷积神经网络为代表的深度学习模型在图像分类、目标检测与语义分割等任务上已全面超越传统方法，展现出强大的特征自动提取与层次化表示学习能力。近年来，Transformer 架构从自然语言处理领域成功迁移至视觉任务，凭借其全局自注意力机制在捕获长距离依赖关系方面表现卓越。这些技术的不断演进，为建筑物遥感影像的精细化语义分割与三维建模提供了坚实的算法基础和广阔的应用前景。本章旨在系统梳理支撑建筑物三维建模任务的深度学习基础理论与核心关键技术，为后续章节所提出的改进算法与建模方案奠定坚实的理论框架。本章首先概述深度学习的基本原理，涵盖模型定义、损失函数、梯度优化、学习率调度策略、激活函数及主流优化算法等核心概念；随后，分别阐述卷积神经网络、Transformer 网络和状态空间模型的基本架构设计思想及其工作原理。

2.2 基本理论

深度学习是机器学习的一个重要分支，其核心思想源于对人工神经网络的深层结构探索[79]。从本质上讲，机器学习的目标在于从数据中自动学习一个从输入空间到输出空间的映射函数，然而许多实际问题所需的映射关系极其复杂，远非人工规则或简单数学表达式所能描述。深度学习正是在此背景下应运而生：它通过构建具有多个处理层的神经网络模型，利用逐层的非线性变换，从原始数据中自动学习出层次化的特征表示，从而逼近高度复杂的目标函数。

与传统机器学习方法需要依赖人工设计特征不同，深度学习的显著优势在于其端到端的学习能力[80]——模型能够直接从原始输入（如图像像素值、文本序列、语音信号等）出发，自动发现对任务最有效的特征表示。深度学习模型的输入形式多样，可以是向量、矩阵（如图像）、时间序列、图结构数据等；经过由大量可训练参数构成的深层网络映射后，其输出可对应于数值回归、类别预测、文本生成、图像分割等多种任务形式。

深度学习的成功依赖于三大要素的协同作用[81]：大规模高质量数据集提供了充足的训练样本，高性能计算硬件提供了必要的算力支撑，以及不断演进的网络架构设计与优化算法则提供了有效的模型表达与训练手段。

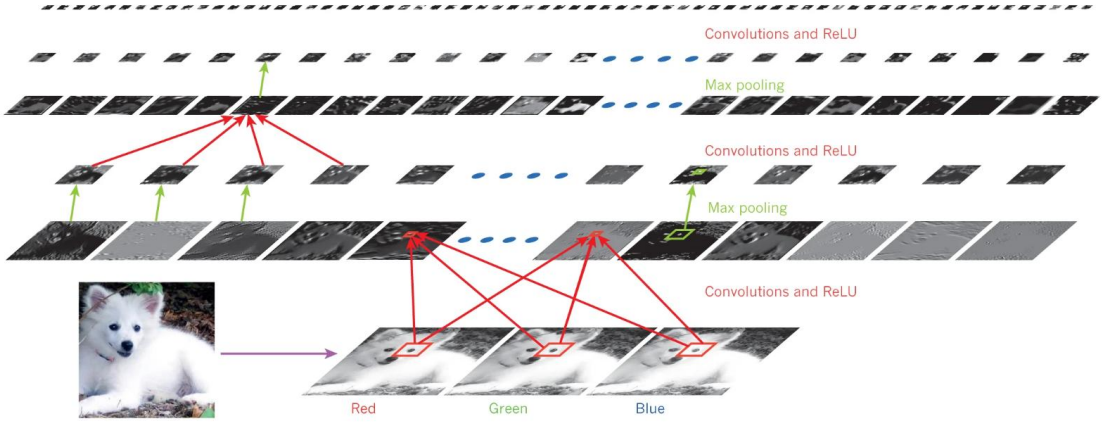


图 2.1 典型卷积神经网络架构应用于萨摩耶犬图像[79]

2.2.1 模型的基本定义

深度学习的语境中，模型（Model）本质上是一个包含大量未知参数的参数化函数。该函数接收输入数据（即特征，Feature），经过一系列数学运算后产生预测输出。以最简单的线性模型为例：

$$y = \omega x + b \tag{2.1}$$

其中，x为输入特征，y为模型预测输出，w为权重（Weight），表征输入特征对输出的影响程度，b为偏置（Bias），用于调整输出的基准值。在该模型中，w与b即为需要通过训练过程从数据中学习得到的未知参数。

然而，现实世界中的任务往往具有高度的非线性和复杂性，单一的线性函数远不足以刻画数据的内在规律。深度学习模型通过将大量简单的非线性计算单元（即神经元）按层次结构组织起来，构建出一个深层的非线性映射函数。一个典型的深度神经网络包含以下基本组成部分：

输入层（Input Layer）：负责接收原始数据，如图像的像素矩阵、文本的词嵌入向量等，并将其传递给后续网络层进行处理。

隐藏层（Hidden Layers）：位于输入层与输出层之间，是深度学习模型的核心组成部分。隐藏层可以包括卷积层（Convolutional Layer）、批归一化层（Batch Normalization Layer）、激活层（Activation Layer）、池化层（Pooling Layer）、注意力层（Attention Layer）等多种类型。网络的“深度”正体现在隐藏层的数量上，层数越多，模型的表达能力和抽象层级越高。

输出层（Output Layer）：根据具体任务的需求，将隐藏层提取的高层抽象特征映射为最终的预测结果。例如，在图像分类任务中，输出层通常采用 Softmax 函数输出各类别的概率分布；在语义分割任务中，输出层需要为输入图像的每个像素预测其对应的语义类别标签。

模型训练的核心目标，便是通过在大规模标注数据集上的迭代学习，不断调

整模型中的所有可训练参数（包括权重和偏置），使模型在给定输入数据的条件下，其输出能够尽可能地逼近真实的标签信息（Ground Truth），从而实现对目标映射关系的有效近似。

2.2.2 损失函数

损失函数（Loss Function）[82]是深度学习训练过程中的核心组件之一，其本质是定义在模型参数空间上的一个标量函数，用于定量度量模型预测输出与真实标签之间的差异程度。损失函数的设计直接决定了模型的优化方向和训练效果。

设训练数据集包含N个样本，对于第n个样本，模型的预测输出为$\hat{y}_n$，对应的真实标签为y_n，则单个样本的损失记为e_n。整个训练集上的总损失通常定义为所有样本损失的平均值：

$$L = \frac{1}{N}\sum_{n=1}^{N} e_n \tag{2.2}$$

损失值L越大，说明模型的预测与真实标签之间的偏差越大，模型性能越差；反之，L越小，模型的预测越准确。训练的核心目标即为寻找使损失函数L取得最小值的一组最优参数。

2.2.3 梯度下降与反向传播

确定了模型结构与损失函数之后，核心问题便转化为如何高效地寻找使损失函数最小化的最优参数组合，这一过程称为优化（Optimization）。深度学习中最基本且最核心的优化方法是梯度下降法（Gradient Descent）[83]，见图 2.2。

梯度下降法的基本思想直观而清晰：在参数空间中，损失函数关于某一参数的梯度（即偏导数）指示了损失函数在该方向上增长最快的方向。因此，沿梯度的反方向更新参数，即可使损失函数值逐步减小。

以包含单个参数ω的简单情形为例，梯度下降的迭代过程如下：

步骤一：参数初始化。随机选取参数的初始值ω^0（通常采用 Xavier 初始化或 He 初始化等策略）。

步骤二：计算梯度。在当前参数值ω^0处计算损失函数L对参数ω的偏导数：

$$\left.\frac{\partial_L}{\partial_\omega}\right|_{\omega=\omega^0} \tag{2.3}$$

步骤三：参数更新。根据梯度方向和预设的学习率η更新参数：

$$\omega^1 = \omega^0 - \eta \cdot \left.\frac{\partial_L}{\partial_\omega}\right|_{\omega=\omega^0} \tag{2.4}$$

其中，学习率η（Learning Rate）是一个至关重要的超参数，它控制着每次参数更新的步幅大小。若梯度为负值，则ω将增大；若梯度为正值，则ω将减小。更

新幅度同时取决于梯度的绝对值和学习率的大小。学习率过大可能导致优化过程在最优点附近剧烈震荡甚至发散；学习率过小则会使收敛速度极为缓慢，且更容易陷入局部最优解。

步骤四：迭代重复。反复执行步骤二和步骤三，直至满足预设的停止条件（如损失值收敛、达到最大迭代轮数等）。

当模型包含多个参数$\theta=\{\omega_1,\omega_2,\ldots,\omega_m,b_1,b_2,\ldots\}$时，需要计算损失函数对所有参数的偏导数，即梯度向量：

$$\nabla_\theta L=\left(\frac{\partial L}{\partial \omega_1},\frac{\partial L}{\partial \omega_2},\cdots,\frac{\partial L}{\partial b_1},\cdots\right) \tag{2.5}$$

参数的整体更新规则为：

$$\theta^{t+1}=\theta^t-\eta\cdot\nabla_\theta L \tag{2.6}$$

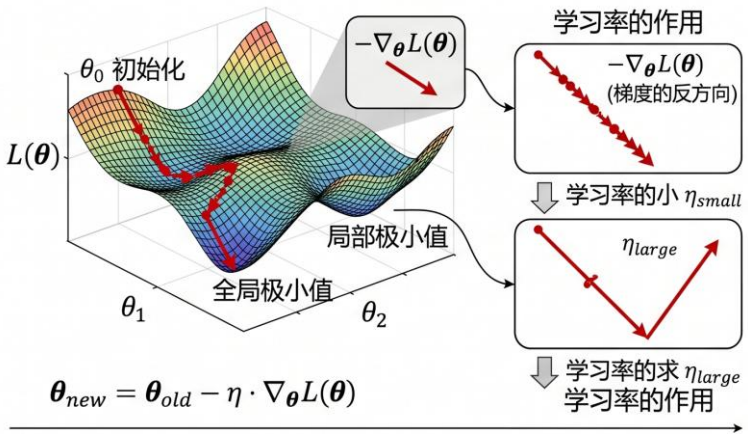


图 2.2 全局损失函数优化

在深度神经网络中，梯度的高效计算依赖于反向传播算法（Backpropagation，BP）。反向传播算法本质上是链式法则（Chain Rule）在多层复合函数中的系统化应用。其过程可分为两个阶段：

前向传播（Forward Pass）：输入数据从输入层出发，依次经过各隐藏层的线性变换和非线性激活，最终在输出层得到预测结果，并据此计算损失函数值。

反向传播（Backward Pass）：损失函数的梯度信号从输出层出发，依据链式法则，沿网络结构逆向逐层传播，计算损失函数对每一层中每一个可训练参数的偏导数（梯度），见图 2.3 所示。

反向传播算法的核心优势在于，它避免了对每个参数单独进行数值微分的巨大计算开销，通过复用中间计算结果，将梯度计算的时间复杂度控制在与一次前向传播相当的量级，从而使得包含数百万乃至数十亿参数的深度网络的训练成为可能。

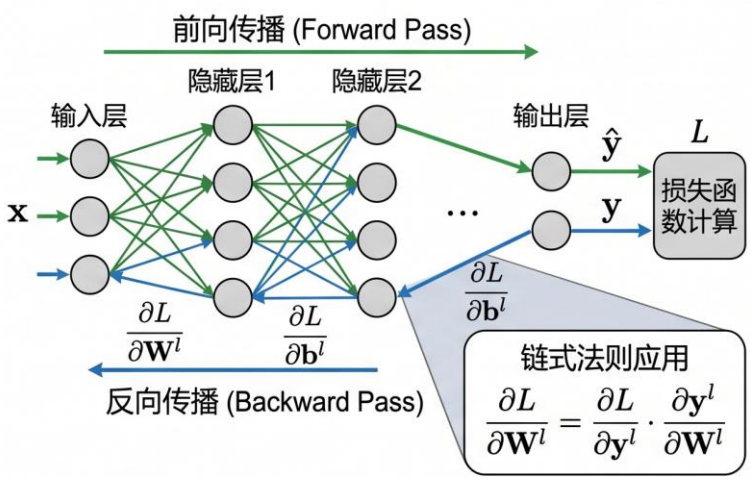


图 2.3 神经网络方向传播算法

在实际训练中，出于计算效率的考量，通常不会在整个训练集上计算梯度（即全批量梯度下降），而是采用小批量随机梯度下降（Mini-batch Stochastic Gradient Descent，Mini-batch SGD）：每次迭代从训练集中随机采样一小批样本（Batch），基于这些样本估计梯度并更新参数。这种做法不仅显著降低了单次迭代的计算量，还因引入的随机性而有助于模型跳出局部最优，提升泛化能力。

然而，随着网络深度的不断增加，梯度在反向传播过程中可能面临梯度消失（Vanishing Gradient）或梯度爆炸（Exploding Gradient）等问题[84]。梯度消失表现为靠近输入层的参数梯度趋近于零，导致浅层参数几乎无法更新；梯度爆炸则表现为梯度值指数级增长，导致参数更新不稳定甚至数值溢出。为应对这些挑战，研究者提出了多种有效的解决方案，包括残差连接（Residual Connection）[85]、批归一化（Batch Normalization）[86]、梯度裁剪（Gradient Clipping）[87]以及合理的参数初始化策略等。

上述模型定义、损失函数计算以及基于梯度下降的参数优化这三个步骤共同构成了深度学习的训练（Training）过程。用于训练的数据称为训练数据（Training Data），模型通过训练学习到数据中的规律；而将训练好的模型应用于训练集之外的新数据并产生输出的过程称为推理或预测（Inference/Prediction）。一个良好的深度学习模型不仅应在训练数据上取得低损失值，更应具备良好的泛化能力（Generalization Ability），即在从未见过的测试数据上同样表现优异。

2.2.4 学习率调度策略

学习率作为控制参数更新步幅的核心超参数，其设置对模型的训练效率和最终性能有着举足轻重的影响。在整个训练过程中使用固定不变的学习率往往难以同时兼顾训练初期的快速收敛与训练后期的精细调优需求。因此，研究者提出了多种学习率调度策略（Learning Rate Schedule），在训练过程中动态调整学习率，以获得更好的优化效果。

1. 分段衰减（Step Decay）[85]

训练进行到预设的里程碑轮次（Milestone Epoch）时，将学习率按固定比例（如乘以 0.1）进行衰减。例如，在总共训练 100 个 epoch 的任务中，可在第 30、60、90 个 epoch 时将学习率降低为原来的十分之一。这种策略简单直观，在许多经典模型的训练中得到了广泛应用。

2. 余弦退火（Cosine Annealing）[88]

受模拟退火算法的启发，学习率按余弦函数的形式从初始值逐步衰减至接近零：

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_{\max} - \eta_{\min}\right)\left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{2.7}$$

其中$\eta_{\max}$和$\eta_{\min}$分别为学习率的上界和下界，T为总训练轮次。余弦退火策略在训练初期保持较大的学习率以探索参数空间，在训练后期逐渐降低学习率以精细调整参数，其平滑的衰减曲线有助于模型跳出尖锐的局部最优并收敛到更为平坦的损失面区域，从而提升模型的泛化性能。该策略在当前主流视觉模型（如 SegFormer、Swin Transformer 等）的训练中被广泛采用。

3. 预热策略（Warm-up）[89]

在训练初始阶段，将学习率从一个极小值线性增加到预设的初始学习率，之后再按上述某种衰减策略进行调整。预热策略可以有效避免训练初期由于参数随机初始化导致的梯度不稳定和损失剧烈波动问题，在大批量训练和 Transformer 类模型的训练中尤为重要。

2.2.5 激活函数

激活函数（Activation Function）是深度神经网络中不可或缺的基本组件，其核心作用是为网络引入非线性变换能力。若神经网络中仅包含线性运算（如矩阵乘法与加法），则无论网络层数多深，其整体功能等价于一个单层的线性变换，无法学习和表达数据中的复杂非线性模式。激活函数的引入打破了这一限制，使深度网络能够逼近任意复杂的连续函数（根据万能近似定理），从而具备强大的学习能力。常见的激活函数包括：

1. Sigmoid 函数[90]

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.8}$$

Sigmoid 函数将输入映射到$(0，1)$区间，输出可解释为概率值，在早期的神经网络和二分类输出层中被广泛使用。然而，Sigmoid 函数存在两个显著缺陷：其一，当输入的绝对值较大时，函数梯度趋近于零，导致反向传播时出现梯度消失问题，严重阻碍深层网络的训练；其二，Sigmoid 函数的输出恒为正值，不是零中

心化的，这会导致后续层的梯度更新方向受限，降低优化效率。

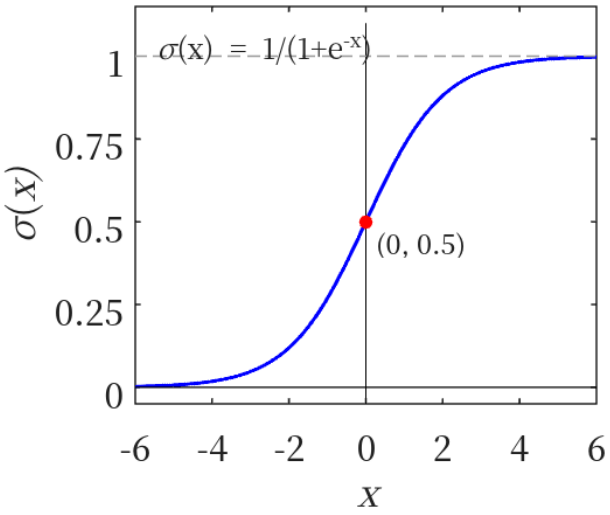


图 2.4 Sigmoid 函数

2. ReLU 函数（修正线性单元）[91]

$$\mathrm{ReLU}(x)=\max(0,x) \tag{2.9}$$

ReLU 函数是目前应用最为广泛的激活函数之一。其结构极为简洁：当输入为正时，直接输出原值；当输入为负时，输出为零。ReLU 具有以下优势：计算效率极高，仅需进行阈值判断操作；在正值区域梯度恒为 1，有效缓解了深层网络中的梯度消失问题，使得训练更深的网络成为可能。然而，ReLU 也存在“死亡神经元问题”：一旦某个神经元的输入在训练过程中始终为负，其梯度将恒为零，该神经元的参数将永远无法更新，相当于该神经元“死亡”。

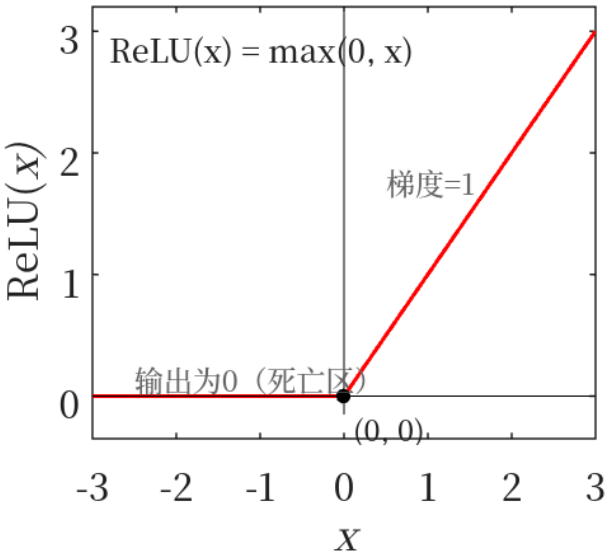


图 2.5 ReLU 函数

2.2.6 优化算法

优化器（Optimizer）是深度学习训练过程中负责基于梯度信息迭代调整模型参数的核心组件。不同的优化算法在收敛速度、训练稳定性和泛化性能等方面各有差异，选择合适的优化器对模型训练至关重要。常用的优化算法包括：

1. 随机梯度下降[83]

SGD 是最基础的优化算法，其参数更新规则为：

$$\theta_{t+1} = \theta_t - \eta \cdot g_t \tag{2.10}$$

其中$g_t=\nabla_\theta L(\theta_t;x_i,y_i)$为基于当前小批量样本计算的梯度估计。SGD 的优势在于实现简单且内存占用低，但由于每次仅基于少量样本估计梯度，更新方向存在较大的随机性和噪声，可能导致优化过程振荡，收敛速度较慢。

2. 动量 SGD[92]

动量 SGD 在标准 SGD 的基础上引入了动量项（Momentum），通过累积历史梯度信息来平滑参数更新方向并加速收敛：

$$v_t = \mu \cdot v_{t-1} + g_t \tag{2.11}$$

$$\theta_{t+1} = \theta_t - \eta \cdot v_t \tag{2.12}$$

其中v_t为动量向量，μ为动量系数（通常取 0.9）。动量项使得参数更新方向在一致的梯度方向上不断加速，而在梯度方向频繁改变的维度上相互抵消，从而有效减少了优化过程中的震荡现象，显著提升了收敛速度。

3. RMSProp（Root Mean Square Propagation）[93]

RMSProp 是一种自适应学习率优化算法，它为每个参数维护一个梯度平方的指数加权移动平均值，并据此独立调整各参数的有效学习率：

$$s_t = \beta \cdot s_{t-1} + (1-\beta) \cdot g_t^2 \tag{2.13}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{s_t + \varepsilon}} \cdot g_t \tag{2.14}$$

其中s_t为梯度平方的移动平均，β为衰减系数，ε为防止除零的极小常数。RMSProp 的核心思想在于：对于梯度较大的参数，其有效学习率会自动降低，防止更新过快；对于梯度较小的参数，其有效学习率会自动增大，促进其充分更新。这种自适应机制使 RMSProp 特别适合处理非平稳目标和稀疏梯度问题。

4. Adam（Adaptive Moment Estimation）[94]

Adam 优化器综合了 Momentum 和 RMSProp 两者的优点，同时维护梯度的一阶矩估计（均值）和二阶矩估计（未中心化方差）的指数加权移动平均：

$$m_t = \beta \cdot m_{t-1} + (1-\beta_1) \cdot g_t \tag{2.15}$$

$$v_t = \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot g_t^2 \tag{2.16}$$

为修正初始阶段因零初始化带来的偏差，引入偏差校正：

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{2.17}$$

最终的参数更新规则为：

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \cdot \hat{m}_t \tag{2.18}$$

其中β_1和β_2的默认值通常分别为 0.9 和 0.999。Adam 优化器凭借其自适应学习率机制和动量加速特性，在绝大多数深度学习任务中均表现出色，已成为当前应用最为广泛的优化器之一。

5. AdamW（Adam with Decoupled Weight Decay）[95]

AdamW 是在 Adam 优化器基础上的重要改进，由 Loshchilov 和 Hutter 于 2019 年提出。其核心创新在于将权重衰减（Weight Decay）与梯度更新过程进行解耦（Decoupling），从而更有效地实现正则化效果。

在传统的 Adam 优化器中，权重衰减通常通过 L2 正则化的方式实现，即将正则化项的梯度直接添加到损失函数的梯度中。然而，由于 Adam 具有自适应缩放梯度的机制，这种耦合方式会导致正则化项的梯度同样被自适应缩放，使得实际施加的权重衰减强度偏离预期，正则化效果大打折扣。AdamW 通过将权重衰减从梯度计算中分离出来，在参数更新步骤中直接对参数施加衰减：

$$\theta_{t+1} = (1-\lambda)\theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \cdot \hat{m}_t \tag{2.19}$$

其中λ为权重衰减系数。这种解耦设计使得权重衰减不受自适应学习率机制的干扰，能够更一致、更有效地控制模型参数的范数，有效抑制过拟合，提升模型的泛化能力。AdamW 已成为训练 Transformer 类视觉模型的标准优化器选择。

2.3 卷积神经网络

卷积神经网络[96]（Convolutional Neural Network，CNN）是深度学习在计算机视觉领域取得突破性成功的奠基性架构。CNN 的设计灵感源自生物视觉系统中神经元的层次化感受野机制，通过引入局部连接、参数共享和平移不变性等归纳偏置（Inductive Bias），使其天然适合处理具有空间结构的图像数据。自 AlexNet 在 2012 年 ImageNet 竞赛中取得划时代的成功以来，CNN 经历了从 VGGNet、GoogLeNet 到 ResNet、DenseNet 等一系列经典架构的快速演进，已成为图像分类、目标检测、语义分割等视觉任务的基础骨干网络。一个典型的卷积神经网络通常由

卷积层、池化层、激活层、归一化层和全连接层等基本模块有机组合而成。

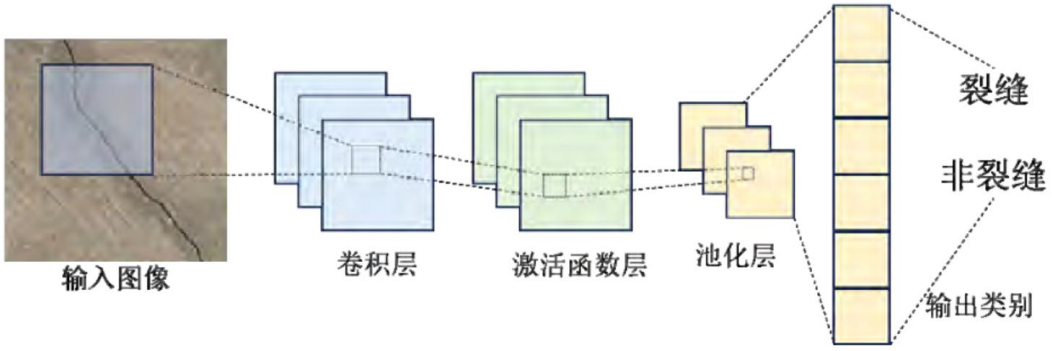


图 2.6 用于图像分类的典型 CNN 架构示意图[97]

2.3.1 卷积层（Convolutional Layer）

卷积层是 CNN 的核心构建单元，负责通过卷积运算从输入特征图中提取局部空间特征。具体而言，卷积操作使用一组可学习的卷积核（Kernel/Filter），在输入特征图上按照预设的步幅（Stride）进行滑动，在每个位置与覆盖的局部区域进行逐元素相乘并求和运算，从而生成输出特征图（Feature Map）。

对于输入特征图$X\in\mathrm{R}^{C_{in}\times H\times W}$，使用$C_{out}$个大小为$k\times k$的卷积核$W\in\mathrm{R}^{C_{out}\times C_{in}\times k\times k}$，输出特征图$Y$中位于空间位置$(i,j)$、第$f$个通道的值计算如下：

$$Y_{f,i,j}=\sum_{c=1}^{C_{in}}\sum_{p=0}^{k-1}\sum_{q=0}^{k-1}W_{f,c,p,q}X_{c,is+p,js+q}+b_f \tag{2.20}$$

其中s为卷积步幅，b_f为第f个卷积核对应的偏置项。

卷积层具有两个极为重要的特性：

局部连接：每个输出神经元仅与输入特征图的一个局部区域（即感受野，Receptive Field）相连接，而非与整个输入全连接。这一设计既符合图像中局部像素之间相关性更强的先验知识，又大幅减少了模型的参数量和计算量。

参数共享：同一个卷积核的参数在输入特征图的所有空间位置上共享使用。这意味着网络在不同位置使用相同的特征检测器，赋予了 CNN 检测平移不变特征的能力，同时进一步降低了参数量。

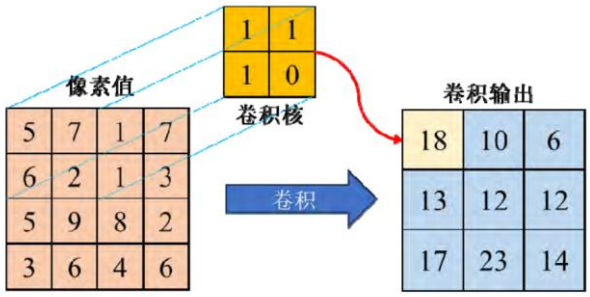


图 2.7 2×2 卷积核计算示例[97]

在深层 CNN 中，浅层卷积层倾向于提取边缘、纹理等低级视觉特征；随着网络层数的增加，卷积层逐步组合低级特征，形成更加抽象和语义丰富的高级特征表示，实现了从底层视觉信号到高层语义概念的层次化特征抽象。

2.3.2 池化层（Pooling Layer）

在图像处理过程中，经卷积层提取的特征图往往包含大量冗余的空间信息。池化层的作用是对特征图进行空间下采样（Downsampling），在保留关键特征信息的同时，减少特征图的空间维度，从而降低后续网络层的计算量和参数量，并增强模型对输入微小平移和形变的不变性（Invariance），提升特征表示的鲁棒性。

池化操作通常在一个固定大小的空间窗口（如2×2）内进行，按照预设的步幅在特征图上滑动，并在每个窗口内进行聚合运算。常见的池化方式包括：

最大池化：取窗口内所有元素的最大值作为输出。最大池化能够突出保留局部区域中最显著、最具判别性的特征响应，在图像分类任务中应用最为广泛。

平均池化：取窗口内所有元素的均值作为输出。平均池化对局部特征响应进行平滑处理，保留了区域内的整体统计信息，常用于网络末端的全局特征聚合（如全局平均池化）。

值得注意的是，池化操作在降低空间分辨率的同时，也不可避免地丢失了部分精细的空间位置信息。这一特性在图像分类任务中通常是有益的（增强不变性），但在对空间精度要求较高的语义分割任务中，则可能导致分割边界模糊、细节丢失等问题，需要通过跳跃连接、特征金字塔等机制加以弥补。

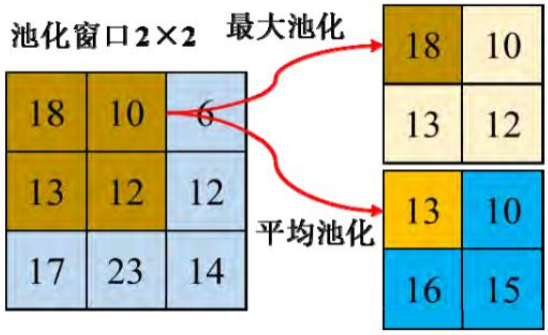


图 2.8 2×2 池化核计算示例[97]

2.3.3 激活层

激活层通常紧跟在卷积层或全连接层之后，对线性变换的输出施加逐元素的非线性变换。如 2.2.5 节所述，激活函数是赋予神经网络非线性表达能力的关键。在 CNN 中，ReLU 及其变体是最常用的激活函数选择。激活层不引入额外的可训练参数，但对网络的学习能力和训练动态具有深远影响。没有激活层的多层网络等价于单层线性变换，无法学习复杂的非线性特征表示。

2.3.4 全连接层（Fully Connected Layer）

在全连接层（也称为密集连接层，Dense Layer）中，每个输出神经元与上一层的所有输入神经元相连接，其计算本质为矩阵乘法加偏置。

在传统 CNN 架构中，全连接层通常位于网络的末端，负责将卷积层和池化层提取的空间特征展平（Flatten）为一维向量，并映射为具体的任务输出（如分类标签的概率分布）。全连接层具有强大的特征整合能力，但由于其参数量与输入维度成正比，在处理高维特征时会带来大量的参数和计算开销，同时也更容易导致过拟合。

在现代深度学习架构中，全局平均池化逐渐取代了全连接层用于分类任务的特征聚合，直接对各通道的特征图取空间平均值后进行分类，极大地减少了参数量。在语义分割任务中，由于需要保留密集的空间信息以实现像素级预测，传统的全连接层较少直接使用，取而代之的是全卷积网络（Fully Convolutional Network，FCN）架构，将全连接层替换为1×1卷积层以保持空间维度。

2.3.5 残差连接（Residual Connection）

随着网络层数的不断增加，深层 CNN 的训练面临严峻的退化问题——网络深度增加到一定程度后，训练精度不升反降，这并非过拟合所致，而是由于深层网络难以学习到恒等映射（因为网络层数太多时，模型在训练时，上一层什么都不改变、直接把输入原样传到下一层的映射关系，即恒等映射。当网络实现这种简单的传递方式时，信息和梯度在反向传播过程中会逐层减弱，导致前面层的参数难以有效更新，从而影响整体性能），梯度在反向传播过程中逐层衰减。

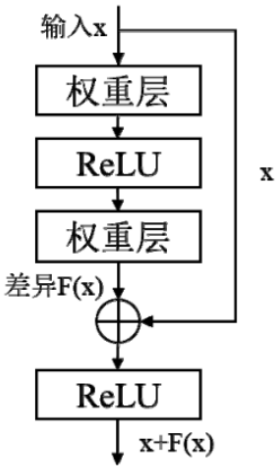


图 2.9 残差连接示意图[97]

何恺明等[85]在 2015 年提出的残差网络（ResNet）通过引入跳跃连接（Skip Connection）/残差连接（Residual Connection），巧妙地解决了这一问题。通过这一设计，网络只需学习输入与输出之间的残差$F(x)=y-x$，而非完整的映射，见图 2.9。

当最优映射接近恒等变换时，学习残差（接近零）远比学习完整映射更容易。残差连接为梯度提供了直通路径，有效缓解了深层网络的梯度消失问题，使得训练超过 100 层乃至 1000 层的极深网络成为可能，是现代深度学习架构设计中最重要的基础组件之一。

2.4 Transformer 网络

Transformer 架构由 Vaswani 等[98]于 2017 年在论文“Attention Is All You Need”中首次提出，最初设计用于机器翻译等自然语言处理（NLP）任务，网络架构见图 2.8。与传统的循环神经网络（RNN）和长短期记忆网络（LSTM）依赖时序递归计算的范式不同，Transformer 完全摒弃了递归结构，完全基于注意力机制（Attention Mechanism）来建模序列中元素之间的全局依赖关系。这一革命性的设计不仅解决了 RNN 难以并行计算和长距离依赖衰减的固有缺陷，还在多项 NLP 基准测试中取得了显著超越前代模型的性能。

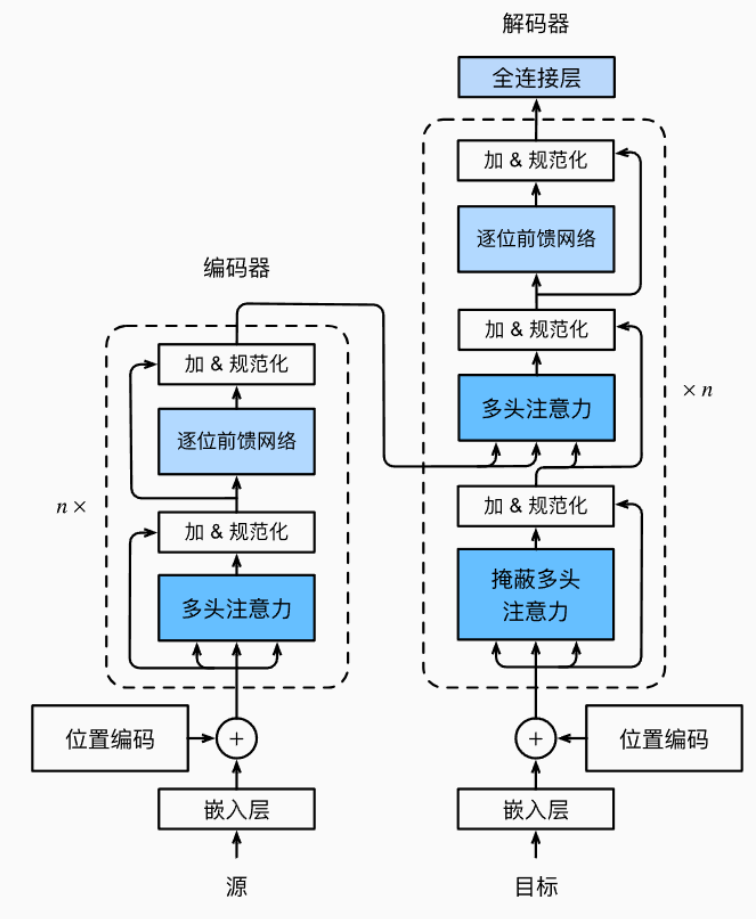


图 2.10 Transformer 网络结构示意图[99]

随后，Dosovitskiy 等[39]于 2020 年提出的 Vision Transformer（ViT）将 Transformer 架构成功引入计算机视觉领域。ViT 将输入图像划分为固定大小的图像块

（Patch），将每个图像块展平为一维向量后作为序列输入 Transformer 编码器，从而将图像识别问题转化为序列建模问题。ViT 的成功证明了 Transformer 在视觉任务中同样具有巨大潜力，其全局自注意力机制能够捕获图像中任意两个位置之间的长距离依赖关系，弥补了 CNN 因感受野有限而在全局上下文建模方面的不足。

2.4.1 自注意力机制（Self-Attention）

自注意力机制是 Transformer 架构的核心组件，其基本思想是：对于序列中的每一个元素，通过计算它与序列中所有其他元素的相关性（注意力权重），来动态地聚合全局上下文信息，从而获得更加丰富的特征表示。

具体而言，对于输入序列的特征矩阵$X\in\mathrm{R}^{N\times d}$（其中$N$为序列长度，$d$为特征维度），自注意力机制首先通过三个可学习的线性投影矩阵$W^Q,W^K,W^V\in\mathrm{R}^{d\times d_k}$，将输入分别映射为查询（Query）、键（Key）和值（Value）三组向量：

$$Q=XW^Q,K=XW^K,V=XW^V \tag{2.21}$$

然后，通过计算查询与键之间的缩放点积来获得注意力权重，并以此对值进行加权求和：

$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.22}$$

其中d_k为键向量的维度，除以$\sqrt{d_k}$的缩放操作是为了防止点积值过大导致 softmax 函数进入梯度饱和区，从而影响训练稳定性。QK^T计算得到的矩阵中，每个元素$\left(i,\ j\right)$表示序列中第i个位置对第j个位置的关注程度；经 softmax 归一化后转化为概率分布，作为注意力权重；最后以这些权重对值矩阵V进行加权求和，得到融合了全局上下文信息的输出特征。

自注意力机制的核心优势在于，它使得序列中任意两个位置之间可以直接建立信息交互通道，而不受距离限制。这种全局建模能力使 Transformer 在捕捉长距离依赖关系方面显著优于 CNN 和 RNN。然而，自注意力机制的计算复杂度和内存占用均为$O(N^2)$，随序列长度二次增长。当应用于高分辨率图像时，序列长度（即图像块数量）极大，二次复杂度将带来巨大的计算负担，这也是后续各种高效注意力变体和层次化视觉 Transformer 被提出的重要动因。

2.4.2 多头注意力机制（Multi-Head Attention）

为了增强模型从不同表示子空间捕获多样化特征模式的能力，Transformer 采用了多头注意力（Multi-Head Attention，MHA）机制，见图 2.11。MHA 将查询、键、值分别投影到h个不同的低维子空间中，在每个子空间独立执行自注意力计算，

最后将各头的输出拼接并经过一次线性变换进行融合：

$$\mathrm{MultiHead}\left(Q,K,V\right)=\mathrm{Concat}\left(\mathrm{head}_1,\mathrm{L},\mathrm{head}_h\right)W^O \tag{2.23}$$

$$\mathrm{head}_i=\mathrm{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right) \tag{2.24}$$

其中$W_i^Q,W_i^K,W_i^V\in\mathrm{R}^{d\times d_k}$为第$i$个注意力头的投影矩阵，$W^O\in\mathrm{R}^{\mathrm{h}d_k\times d}$为输出的线性投影矩阵。多头注意力使模型能够同时关注来自不同位置和不同语义层面的信息，例如某些头可能关注局部纹理特征，另一些头可能关注全局结构关系，从而提供了更为丰富和多元的特征表示。

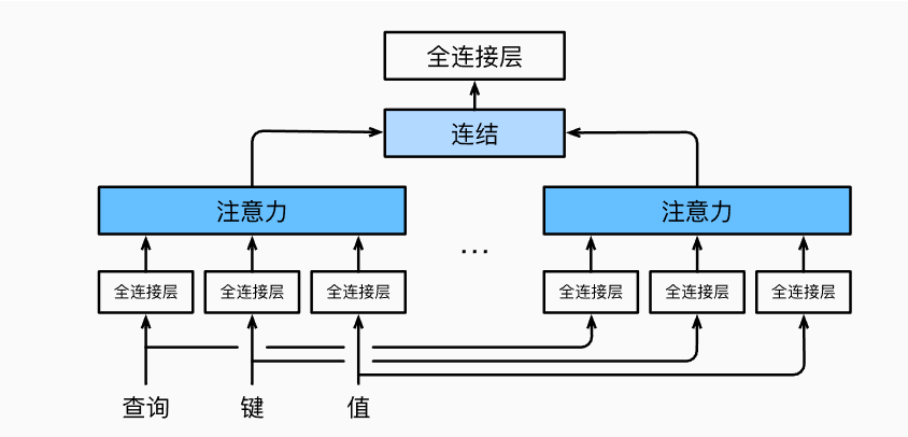


图 2.11 多头注意力机制[99]

2.4.3 位置编码（Positional Encoding）

自注意力机制本身具有置换不变性（Permutation Invariance），即改变输入序列中元素的顺序不会影响注意力计算的结果。然而，在图像和文本等任务中，元素的位置信息至关重要。为了使 Transformer 能够感知序列中元素的位置顺序，需要在输入中额外注入位置编码（Positional Encoding）信息。

原始 Transformer 采用基于正弦和余弦函数的固定位置编码：

$$\mathrm{PE}_{(pos,2i)}=\sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.25}$$

$$\mathrm{PE}_{(pos,2i+1)}=\sin\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.26}$$

其中pos为元素在序列中的位置索引，i为位置编码向量的维度索引。这种设计使得不同位置的编码向量各不相同，且位置之间的关系可以通过线性变换表示。

此外，在视觉 Transformer 中，可学习的位置编码和条件位置编码等方案也被广泛采用。近年来提出的相对位置编码（Relative Positional Encoding）和旋转位置编码等方案进一步增强了位置建模的灵活性和泛化能力。

2.4.4 前馈神经网络（Feed-Forward Network）

Transformer 编码器中的每个层还包含一个前馈神经网络（Feed-Forward Network，FFN），通常由两个线性变换和一个非线性激活函数组成：

$$\mathrm{FFN}(x)=W_2\sigma(W_1x+b_1)+b_2 \tag{2.27}$$

其中σ通常为 GELU 或 ReLU 激活函数。FFN 在每个位置上独立作用，对自注意力层输出的特征进行非线性变换和维度调整，增强模型的表达能力。通常，FFN 的中间隐藏层维度为输入维度的 4 倍，起到特征扩展和压缩的“瓶颈”作用。

2.4.5 Transformer 编码器的整体结构

一个标准的 Transformer 编码器层由以下组件依次串联构成：

$$z'=\mathrm{LN}(x+\mathrm{MHA}(x)) \tag{2.28}$$

$$z=\mathrm{LN}(z'+\mathrm{FFN}(z')) \tag{2.29}$$

其中 LN 表示层归一化。每个子模块（MHA 和 FFN）都伴随有残差连接和层归一化，前者为梯度提供直通路径以稳定深层训练，后者对特征分布进行标准化以加速收敛。多个这样的编码器层堆叠在一起，构成完整的 Transformer 编码器。

2.5 语义分割网络

语义分割（Semantic Segmentation）[24]是计算机视觉领域中一项基础而重要的密集预测任务，其目标是为输入图像中的每一个像素分配一个语义类别标签，从而实现对图像内容的像素级精细化理解。与图像分类仅需判断整幅图像的类别、目标检测仅需定位并分类感兴趣对象不同，语义分割要求模型同时具备精确的空间定位能力和强大的语义理解能力，能够精细地勾勒出不同类别区域的边界轮廓。

在遥感图像分析领域，语义分割具有极为广泛的应用价值。例如，建筑物语义分割可以自动从卫星或航空影像中精确提取建筑物的空间轮廓，为城市规划、灾害评估、变化检测和三维建模等下游应用提供关键的基础数据。

现代语义分割网络大多遵循编码器-解码器（Encoder-Decoder）的架构范式。编码器（Encoder）负责将输入图像逐步进行特征提取和空间下采样，通过多层网络逐步降低特征图的空间分辨率、增加特征通道数，提取出具有丰富语义信息的高层特征表示。编码器通常采用经过大规模数据集预训练的分类网络作为骨干网络（Backbone），如 ResNet、EfficientNet、Swin Transformer、MiT（Mix Transformer）等。

解码器（Decoder）则负责将编码器输出的低分辨率高语义特征图逐步进行上采样，恢复到与输入图像相同的空间分辨率，并在上采样过程中逐步融合编码器各阶段产生的多尺度特征，最终输出逐像素的语义分割预测图。

这种架构设计的核心挑战在于如何有效地平衡语义信息（高层特征，经过多次下采样，语义丰富但空间分辨率低）与空间细节信息（低层特征，空间分辨率高但语义信息有限）之间的关系。不同的语义分割网络在解码器设计和多尺度特征融合策略上各有创新。

2.6 回归任务与语义分割的关系

在基于光学遥感影像的建筑物三维建模任务中，除了获取建筑物的二维平面轮廓（即语义分割），还需要获取建筑物的三维高程信息。基于单视光学遥感影像的建筑物高度估计（Building Height Estimation）本质上属于像素级回归任务（Pixel-wise Regression）。

尽管语义分割旨在预测离散的类别概率，而高度估计旨在预测连续的数值（如高度米数、相对深度），但两者在计算机视觉领域统称为密集预测任务（Dense Prediction Tasks）。这种根本上的同源性决定了两者在深度学习模型架构上具有极高的通用性。

具体而言，无论是执行语义分割还是高度回归，网络的主体结构通常都采用相同的编码器-解码器范式。编码器负责从原始影像中提取多尺度的强语义特征，解码器则负责恢复空间分辨率并融合细节信息。两者在网络结构上唯一的实质性差异仅仅在于网络输出层的设计与损失函数的选择：

语义分割：网络的最后一层通常使用 Softmax 或 Sigmoid 激活函数，将特征映射为各像素属于特定类别的概率分布，并采用交叉熵（Cross-Entropy）等分类损失函数进行参数优化。

高度估计（回归）：网络的最后一层通常不使用非线性激活函数（或仅使用 ReLU 确保输出高程为非负值），直接输出单通道的连续高度数值，并采用均方误差（MSE）、平均绝对误差（MAE）或 Huber Loss 等回归损失函数进行约束。

近年来，大量遥感领域的研究证明了这种架构的通用性及其在特征提取层面的内在联系。光学影像中建筑物的高度分布往往与其二维空间特征（如阴影长度、几何透视形变、屋顶纹理等）高度耦合。文献[24]率先证明了全卷积网络（FCN）可以直接从单张遥感图像中端到端地回归建筑物高度；而后续诸多研究[100]则进一步表明，许多经典的语义分割网络（如 U-Net、DeepLab 等）只需简单替换输出层，即可直接胜任高度回归任务。不仅如此，当前常采用多任务学习（Multi-Task Learning，MTL）框架[101]，利用同一个共享的骨干网络同时提取特征，再通过双分支解码器联合输出分割掩码和高度图。这种联合机制充分证明了分割与回归任务在特征表示上的高度互补性。

2.7 航空航天摄影基本概念

航空航天摄影是指利用搭载在飞行器上的摄影设备对地表进行影像获取的技术手段。根据摄影时相机光轴与地面的相对位置关系，航空摄影主要分为正射摄影和倾斜摄影两类。正射摄影（Vertical Photography）又称为垂直摄影，是指相机光轴与地面保持垂直或接近垂直（倾角小于 3°）时进行的摄影；但由于地形起伏、相机姿态变化、镜头畸变以及地球曲率等因素，即使是垂直拍摄的原始影像也通常存在几何变形，无法直接用于精确量测或与地图进行无缝叠加。为此，需要进行正射校正（Orthorectification）：这是一种基于精确的传感器（相机）参数、飞行姿态数据、地面控制点（GCP）以及数字高程模型（DEM/DSM）对原始航空影像进行几何纠正的过程。其核心思想是：利用高程信息，将影像中的每个像素“投影回”真实地表的水平位置，从而消除因地形起伏和拍摄角度引起的几何畸变，使最终生成的正射影像图（DOM）或正射镶嵌图（Orthomosaic）具有统一的比例尺、真实的地物平面位置关系，并且可直接用于距离、面积、方位等定量测量。倾斜摄影（Oblique Photography）则是指相机光轴与地面呈一定倾斜角度进行的摄影，通常采用多角度（前视、后视、左视、右视及下视）同步获取影像，能够完整展现地物的立面和顶面信息，弥补了正射影像只能表达地物顶部特征的不足。

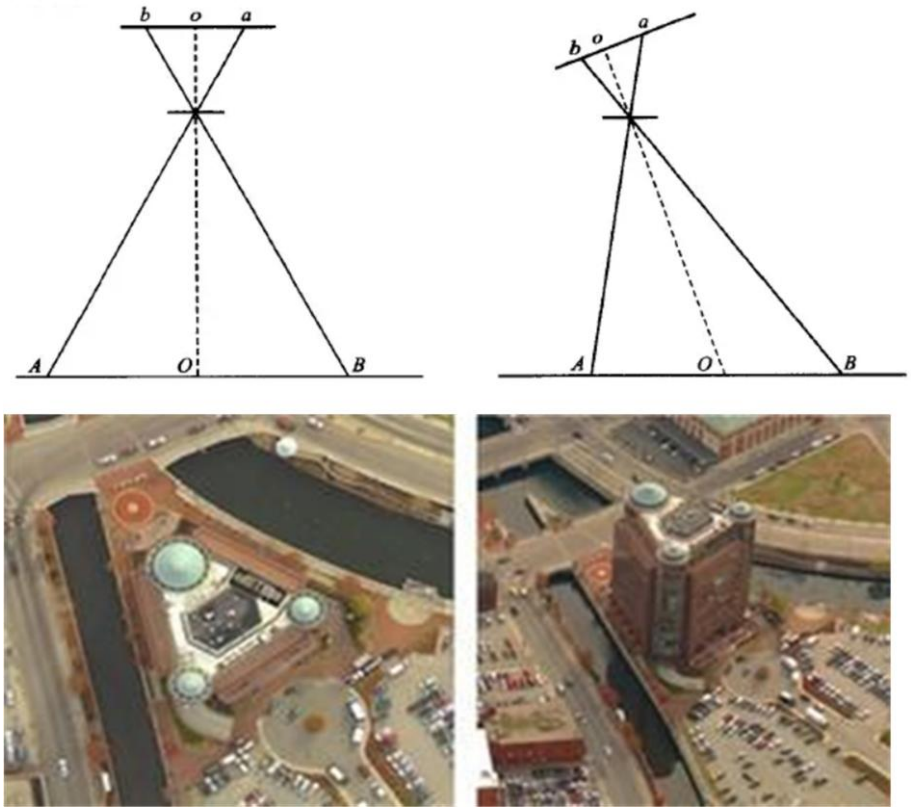


图 2.12 垂直航拍照片（左）及斜向航拍照片（右）[102]

2.8 本章小结

本章围绕高分辨率遥感影像建筑物语义分割与三维建模任务的技术需求，系统阐述了深度学习的基础理论与核心关键技术，为后续章节提出的改进算法与建模方案构建了完整的理论基础。本章主要内容可归纳为以下三个方面：

（1）系统总结了深度学习的基本原理与核心训练组件。本章系统阐述了模型的参数化定义、损失函数的度量机制、基于梯度下降与反向传播的参数优化方法，以及学习率调度策略、激活函数和主流优化算法等核心组件，完成了深度学习训练范式的系统性梳理。

（2）系统归纳了主流深度学习网络架构的原理与特性。本章深入剖析了卷积神经网络和 Transformer 网络两类主流架构。卷积神经网络凭借局部连接与参数共享机制擅长空间特征的层次化提取，但受限于有限感受野；Transformer 架构则通过自注意力机制实现全局依赖建模。两类架构各有所长，其互补特性为后续算法设计提供了重要依据。

（3）回顾了密集预测任务的统一架构，并梳理了航空摄影的数据基础。本章阐述了语义分割与高度回归任务基于编码器—解码器的通用架构范式，论证了两者在网络结构上的内在统一性。此外，介绍了正射摄影与倾斜摄影的技术原理，为建筑物三维建模提供了数据基础。

在本章建立的理论框架之上，后续章节将针对高分辨率遥感影像建筑物分割与三维建模的特定挑战，提出相应的改进算法与具体建模方案。

第3章 基于 SPR-Net 网络的建筑轮廓提取

3.1 引言

目前，现有深度学习分割网络在标准公开数据集上已取得显著进展，但当模型应用于真实遥感场景时，仍面临泛化能力不足与预测稳定性下降等突出问题。其主要原因在于：第一，训练数据与实际场景之间存在显著的地域分布与建筑风格差异；第二，单一光学模态所承载的信息有限；第三，常规网络在逐级下采样过程中难以有效保留建筑物边缘的高频结构信息，导致轮廓模糊与边界粘连问题频发。针对上述局限，引入 SAR、DSM 等辅助模态的多源融合方法虽已展现出改善边界表达的潜力，但受制于数据采集成本高昂与跨模态配准难度大等现实瓶颈，在大范围工程化应用中仍面临诸多限制。为此，本章提出了一种基于伪多模态结构信息的建筑物轮廓提取网络 SPR-Net。该网络以 Segformer 作为主干架构，引入双阶段伪多模态增强策略，在编码器阶段通过拉普拉斯特征变换从单一光学影像中构造高频伪模态信息以增强边缘响应，在解码器阶段利用初步预测结果引导特征重加权以实现边界自适应校准，并在输出端采用感受野注意力卷积进一步精细化局部轮廓建模。同时，构建了覆盖长沙市约 115km²的高分辨率建筑数据集，用于验证方法在国内真实场景下的泛化能力。

3.2 方法介绍

3.2.1 SPR-Net 架构总览

本文提出了一种面向遥感影像建筑物精细分割的网络模型 SPR-Net，旨在有效解决建筑物边界模糊、尺度差异显著以及复杂背景干扰等问题。其整体架构如图 3.1 所示。

SPR-Net 以 Segformer 作为主干网络，采用层次化 Transformer 编码结构，由四个阶段逐级提取多尺度语义特征。Segformer 兼具高效建模与较低计算复杂度的优势，能够为大尺度遥感影像提供稳定的全局语义表示。然而，纯 Transformer 结构在局部边缘与细粒度轮廓建模方面存在不足，尤其在建筑物屋顶边界、拐角等高频区域易出现预测模糊。

为此，本文在 Segformer 框架基础上引入双阶段伪多模态（Pseudo-Multi-Modal，PMM）策略，从特征增强和预测校正两个层面强化边界表达能力。具体网络结构设计体现在以下三个阶段：

1. 编码器阶段：在浅层特征中嵌入单输入伪多模态模块（PMM_S），受MEGANet[103]的思想启发，通过拉普拉斯特征变换自发构造高频“伪模态”，用于增强建筑物边缘与结构轮廓响应；

2. 解码器阶段：引入双输入伪多模态模块（PMM_C），利用初步预测结果引导特征重加权，实现对建筑物边界的自适应校准；

3. 输出阶段：采用感受野注意力卷积（RFAConv）[104]对局部拓扑结构进行精细化建模，进一步提升对直角拐角与复杂轮廓的刻画能力。

通过上述设计，SPR-Net 在保持全局语义一致性的同时，实现了对多尺度建筑边界的精准建模。

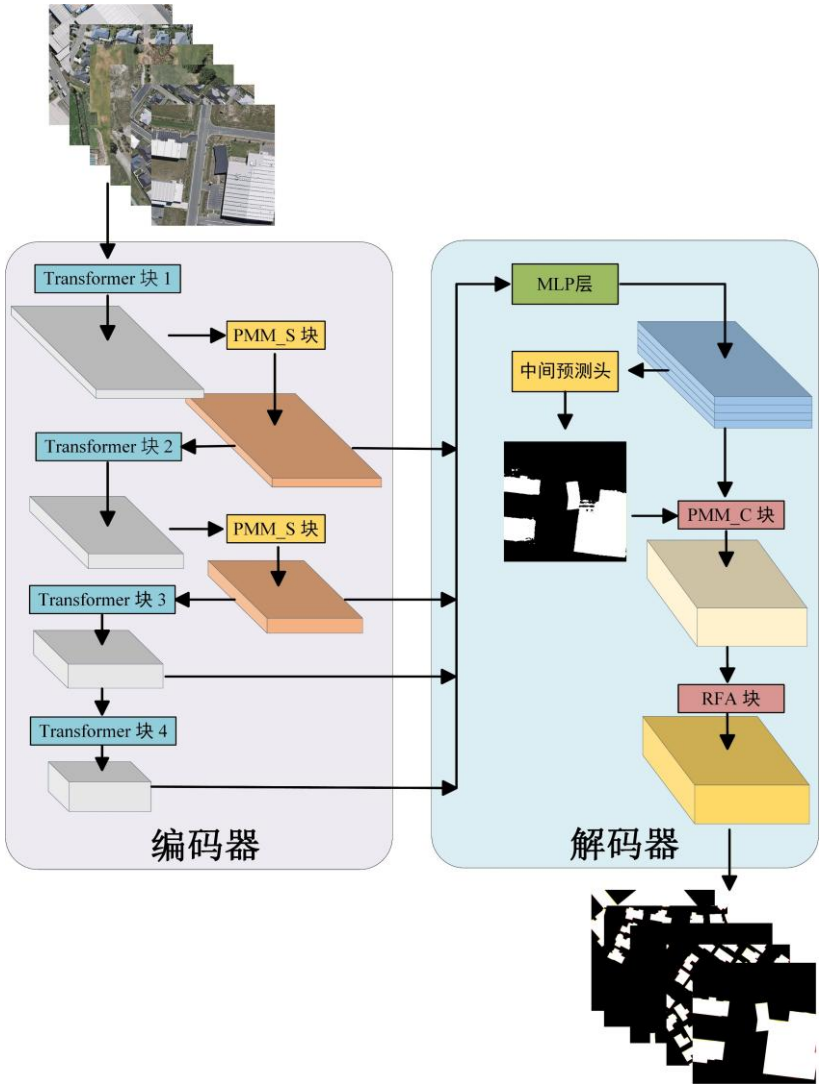


图 3.1 SPR-Net 架构总览

3.2.2 编码器部分

编码器采用 Segformer 的层次化结构，由四个阶段（阶段 1–4）组成，分别对应 1/4、1/8、1/16 和 1/32 的空间分辨率。其中，浅层特征（1/4、1/8 尺度）保留

了丰富的空间细节信息，对建筑物边界和几何结构尤为关键。传统 Transformer 编码器虽然具备强大的全局建模能力，但由于缺乏对高频信息的显式建模，往往难以准确感知建筑物的精细轮廓。针对这一问题，本文在编码器的关键阶段引入 PMM_S 模块，以增强特征对局部边缘的感知能力。编码器遵循层次化结构，数据在每个阶段内按以下流向传递：重叠补丁合并到 Transformer 块再到 PMM_S。

1. 重叠补丁合并层（Overlapping Patch Merging）

在每个阶段开始时，采用重叠补丁合并操作对特征图进行降采样并扩展通道数，其计算形式为：

$$F_{i+1} = \mathrm{Conv}_{k,s,p}\left(F_i\right) \tag{3.1}$$

其中，k，s，p 分别表示卷积核大小、步长和填充。与传统 Patch Embedding 不同，该操作通过重叠感受野保持相邻区域的空间连续性，在实现特征尺度变换的同时，有效减少边界信息和局部结构的丢失。这一特性对于遥感建筑物中连续屋顶结构和相邻建筑区分具有重要意义。

2. Transformer 块处理

每个 Transformer 块由高效自注意力（Efficient Self-Attention）和 Mix-FFN 组成：

为降低标准自注意力在高分辨率特征下的计算复杂度，Segformer 引入了序列缩减率R，其注意力计算形式为：

$$\mathrm{Attention}\left(Q,K,V\right)=\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_{\mathrm{head}}}}\right)V \tag{3.2}$$

通过对键K进行重塑与线性映射：

$$\hat{K}=\mathrm{Reshape}\left(\frac{N}{R},C\cdot R\right)\left(K\right) \tag{3.3}$$

$$K=\mathrm{Liner}\left(C\cdot R,C\right)\left(\hat{K}\right) \tag{3.4}$$

自注意力的计算复杂度由$O(N^2)$显著降低，从而在保证全局依赖建模能力的同时提升了网络效率。

随后，特征输入 Mix-FFN 模块，该模块在前馈网络中引入 3×3 深度可分离卷积，以隐式方式编码空间位置信息：

$$X_{\mathrm{out}}=\mathrm{MLP}\left(\mathrm{GELU}\left(\mathrm{Conv}_{3\times3}\left(\mathrm{MLP}\left(X_{\mathrm{in}}\right)\right)\right)\right)+X_{\mathrm{in}} \tag{3.5}$$

这种设计无需额外的位置编码，即可同时建模全局语义关系和局部空间结构，有利于复杂建筑布局的表达

数据从 Transformer 块输出后进入 PMM_S，如图 3.3 a）。该模块通过拉普拉斯特征变换提取高频“伪模态”成分X_{high}：

$$X_{\text{high}} = X - \text{Upsample}\left(\text{Downsample}\left(\text{Gaussian}\left(X\right)\right)\right) \tag{3.6}$$

其中，Gaussian(·)表示高斯滤波操作，Downsample(·)和 Upsample(·)分别表示降采样和上采样。该过程模拟了拉普拉斯金字塔的构建，通过高斯平滑、降采样再上采样得到低频分量，原始特征减去低频分量即得到高频边缘成分。随后引入 CBAM[105]模块对高频特征进行通道与空间两级注意力重标定，自适应增强与建筑物边界相关的判别性响应并抑制背景干扰，从而提升特征的边界感知能力。

经过 PMM_S 处理后，特征在保持原有语义信息的同时显著增强了边界响应，并分别传递至解码器和下一编码阶段，为后续多尺度融合奠定基础。

3.2.3 解码器部分

经来自四个阶段的增强特征（经过 PMM_S 处理的 1/4 和 1/8 尺度特征，以及未处理的 1/16 和 1/32 尺度特征）首先通过 MLP 层映射至统一的通道维度。该 MLP 层实质上是 1×1 卷积，能够在不改变空间分辨率的情况下实现跨通道的信息交互：随后，所有特征图被上采样至相同的空间尺寸（与最高分辨率特征 1/4 尺度一致），并沿通道维度进行拼接，生成全局融合特征F_{fuse}。这种设计融合了不同尺度的语义信息：高分辨率特征（1/4，1/8）提供精细的空间细节和边缘信息，低分辨率特征（1/16，1/32）提供全局语义和上下文理解，二者的结合使得网络具有强大的多尺度表示能力。

融合特征F_{fuse}结合来自初步预测层的概率图 P 进入 PMM_C，见图 3.2 b）。该模块通过背景注意力、边界注意力和特征边缘三个分支动态修正特征。受 MEGANet[103]启发，其融合过程可表示为：

$$\begin{aligned} F_{\text{refined}} = \text{Conv}([&F_{\text{fuse}} \odot \left(1-\sigma\left(P\right)\right) \cdot F_{\text{fuse}} \odot \text{Laplace}\left(\sigma\left(P\right)\right) \cdot \\ &F_{\text{fuse}} \odot \text{Conv}\left(F_{\text{fuse}}\right)]) \end{aligned} \tag{3.7}$$

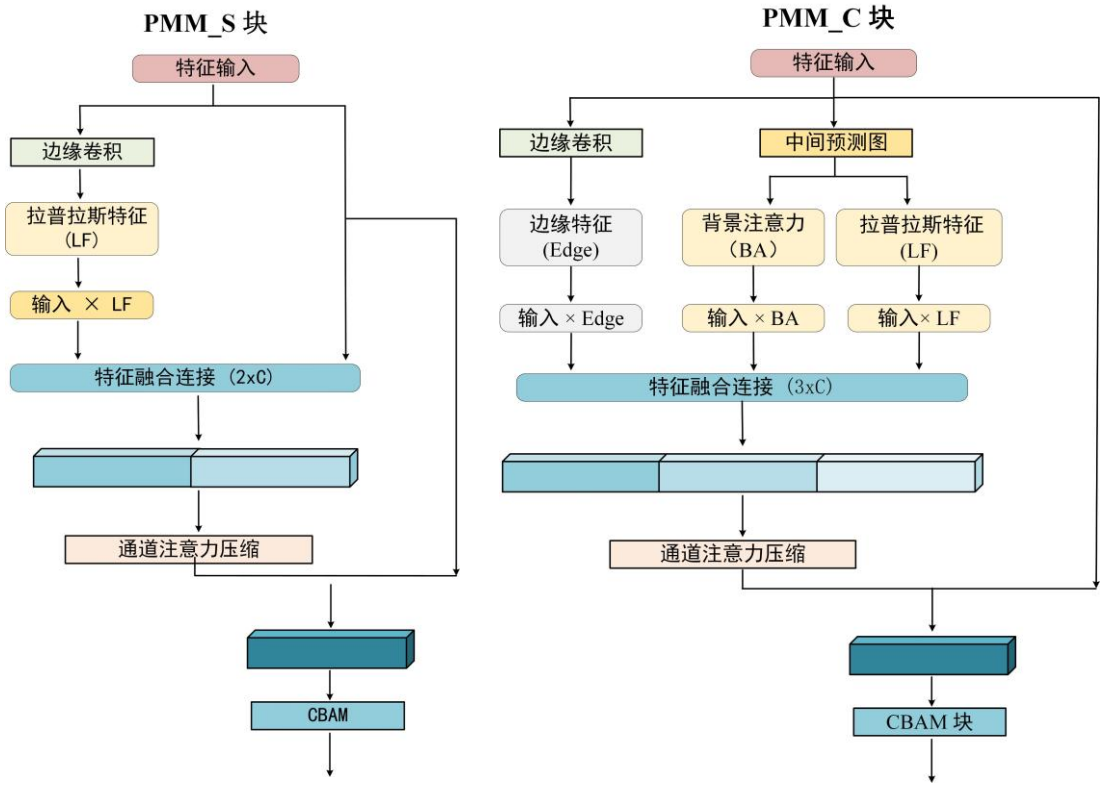
其中$\odot$表示元素级乘法，· 表示通道拼接。

该模块利用预测结果显式引导特征重加权，使网络能够聚焦于建筑物边界不确定区域，有效修正初步分割中存在的边缘偏移问题。

在最终分类前，特征进入 RFA 块，图 3.3 以优化局部拓扑结构。该模块通过感受野注意力权重对局部特征进行重构：

$$Y_{ij} = \sum_{k \in \Omega} W_{\text{spa},k} \cdot W_k \cdot X_{i+k,j+k} \tag{3.8}$$

其中W_{spa}代表 3×3 感受野空间。RFAConv 通过自适应感受野注意力权重，引导网络重点关注建筑物的直角拐角和复杂轮廓区域，从而进一步提升分割结果的几何一致性。



a）PMM_S 模块　　b）PMM_C 模块

图 3.2 PMM 模块

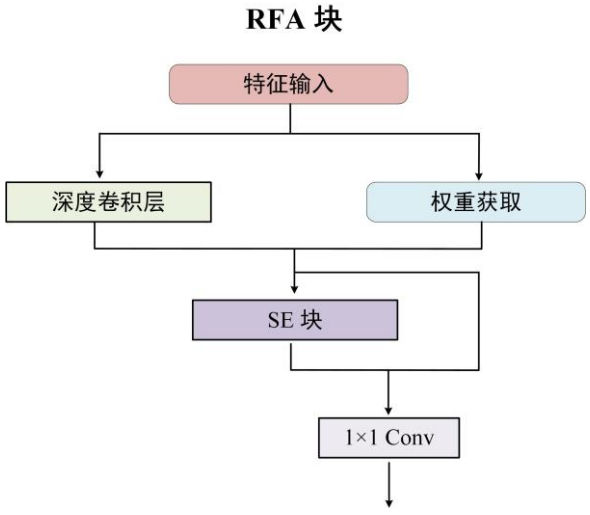


图 3.3 RFAConv 模块

3.2.4 损失函数

为了提高模型的分割精度，本文采用复合损失函数来综合优化模型的区域重叠度、边界精确性和 IoU 性能。复合损失函数融合了 Dice[106]损失$\mathcal{L}_{\text{Dice}}$、边界损失[107]$\mathcal{L}_{\text{Boundary}}$和$\mathcal{L}_{\text{Lovasz}}$损失[107],定义如下：

$$\mathcal{L}=\alpha\mathcal{L}_{\text{Dice}}+\beta\mathcal{L}_{\text{Boundary}}+\gamma\mathcal{L}_{\text{Lovasz}} \tag{3.9}$$

式中，α、β、γ分别表示重叠度损失$\mathcal{L}_{Dice}$、边界损失$\mathcal{L}_{\text{Boundary}}$和 IoU 代理损失$\mathcal{L}_{\text{Lovasz}}$的权重系数；其中：

Dice 损失用于衡量预测分割结果与真实标签之间的区域重叠度，其定义为：

$$\mathcal{L}_{\text{Dice}}=1-\frac{2\sum_{i=1}^{N}p_i g_i+\varepsilon}{\sum_{i=1}^{N}p_i+\sum_{i=1}^{N}g_i+\varepsilon} \tag{3.10}$$

式中，N表示图像中的像素总数；p_i表示模型对像素i预测的概率；g_i表示像素i的真实像素值；ε为平滑项，用于避免数值不稳定；

边界损失通过引入符号距离图来增强模型对分割边界的敏感性，其定义为：

$$\mathcal{L}_{\text{Boundary}}=\frac{1}{N}\sum_{i=1}^{N}s_i p_i \tag{3.11}$$

式中，s_i表示像素i的 SDM 值，该值反映了像素到最近边界的距离；

Lovasz损失作为 IoU 度量的凸代理损失，直接优化 IoU 指标,其定义为：

$$\mathcal{L}_{\text{Lovasz}}=1-\frac{1}{|E|}\sum_{i\in E}\Delta_i(m) \tag{3.12}$$

式中，E表示像素集合；$\Delta_i(\cdot)$表示排序后的 IoU 损失梯度序列；m表示 IoU 损失梯度序列中的任意元素。

综合考虑不同损失项的贡献，本文最终采用的损失函数为：

$$\mathcal{L}=\mathcal{L}_{\text{Dice}}+0.3\mathcal{L}_{\text{Lovasz}}+0.05\mathcal{L}_{\text{Boundary}} \tag{3.13}$$

3.2.5 评价指标

在遥感影像建筑物提取任务中，常用的评价方法大体可以分为两类，即像素级指标与实例级指标，由于本文采用端到端的分类方式，对输入图像中的每一个像素进行建筑与非建筑的二分类，因此我们主要选取像素级评价指标来衡量模型性能。具体指标包括精确率（Precision）、召回率（Recall）、F1 值（F1-score）以及交并比（Intersection over Union，IoU）。

在像素级分类问题中，每个像素的预测结果可归纳为四种情况：将正样本预测为正（TP）、将正样本预测为负（FN）、将负样本预测为正（FP）以及将负样本预测为负（TN）其中，精确率表示所有被预测为建筑的像素中真实为建筑的比

例，召回率表示所有真实建筑像素中被正确识别的比例。F1 值则是精确率与召回率的调和平均数，用于综合反映模型的平衡性能 IoU 指标则衡量预测结果与真实标注之间交集与并集的比值，并在整个测试集上取平均，其能够更加直观地反映分割结果与真实轮廓的一致性。各个指标计算公式如下：

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{3.14}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{3.15}$$

$$\text{F1}=\frac{2\times\text{Precison}\times\text{Recall}}{\text{Precison+Recall}} \tag{3.16}$$

$$\text{IOU}=\frac{\text{Precison}\times\text{Recall}}{\text{Precison+Recall-Precison}\times\text{Recall}} \tag{3.17}$$

此外，为了评估模型在运行效率与复杂度上的表现，本文还采用了浮点运算次数（FLOPs）与可训练参数量（Parameters，Para）作为补充指标 FLOPs 反映了模型在推理过程中所需的计算量，通常用于估计模型执行效率，而参数量表示深度学习模型在训练过程中通过梯度更新的权重总数，可用于衡量模型的存储和计算开销。

3.3 实验与分析

3.3.1 数据集

为全面评估本文所提出方法在遥感影像建筑物提取任务中的有效性与泛化能力，实验选取了一个广泛使用的公开数据集 WHU 航空影像数据集以及一个面向国内实际应用场景自主构建的长沙轮廓数据集开展对比验证实验。上述数据集在地理区域、影像来源及建筑分布特征等方面均呈现显著差异性：WHU 数据集覆盖多个城市区域，建筑类型以规则分布的住宅和商业建筑为主；而长沙自建数据集则包含更为复杂的本土建筑特征，如大面积蓝顶工业厂房、高密度城中村建筑群、新旧建筑混杂区等典型场景。这种数据集的多样性与差异性能够为模型性能验证提供全面且具有代表性的实验基准。

1. WHU 航空影像数据集[19]：

WHU 航空影像数据集是建筑物提取领域中应用最为广泛的公开数据集之一，包含航空与卫星遥感影像及其对应的高精度像素级标注，覆盖新西兰基督城约 450km²区域，见图 3.4。本文选用其中的航空影像子集进行实验评估。该子集原始空间分辨率为 0.075m，公开版本将其下采样至 0.3m，并裁剪为 8139 幅大小为 512×512 的图像块。影像均包含红（R）、绿（G）和蓝（B）三个波段，并提供精

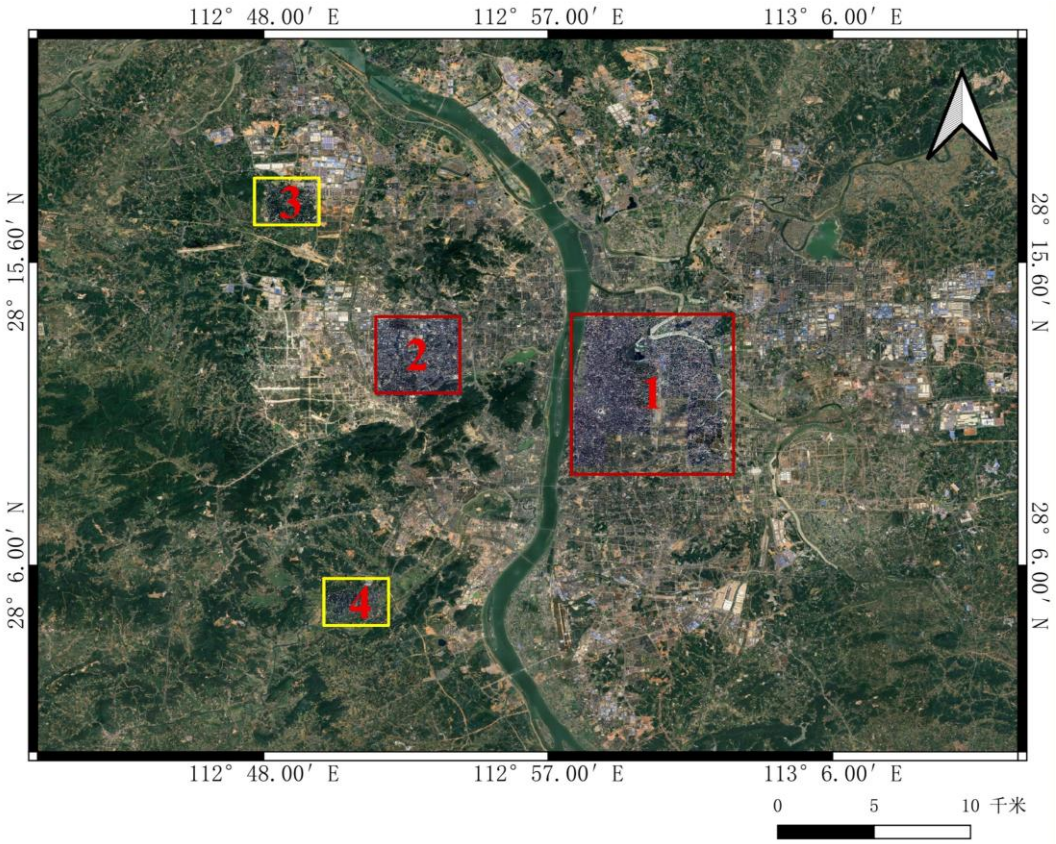
确的建筑物像素级标注。按照官方划分方式，训练集、验证集和测试集分别包含4736、1036 和 2416 幅图像，总计标注建筑物数量约为 18.7 万个。本文严格遵循该数据集的原始划分方案进行实验，以保证结果的可比性。

图 3.4 WHU 航空影像数据集[19]

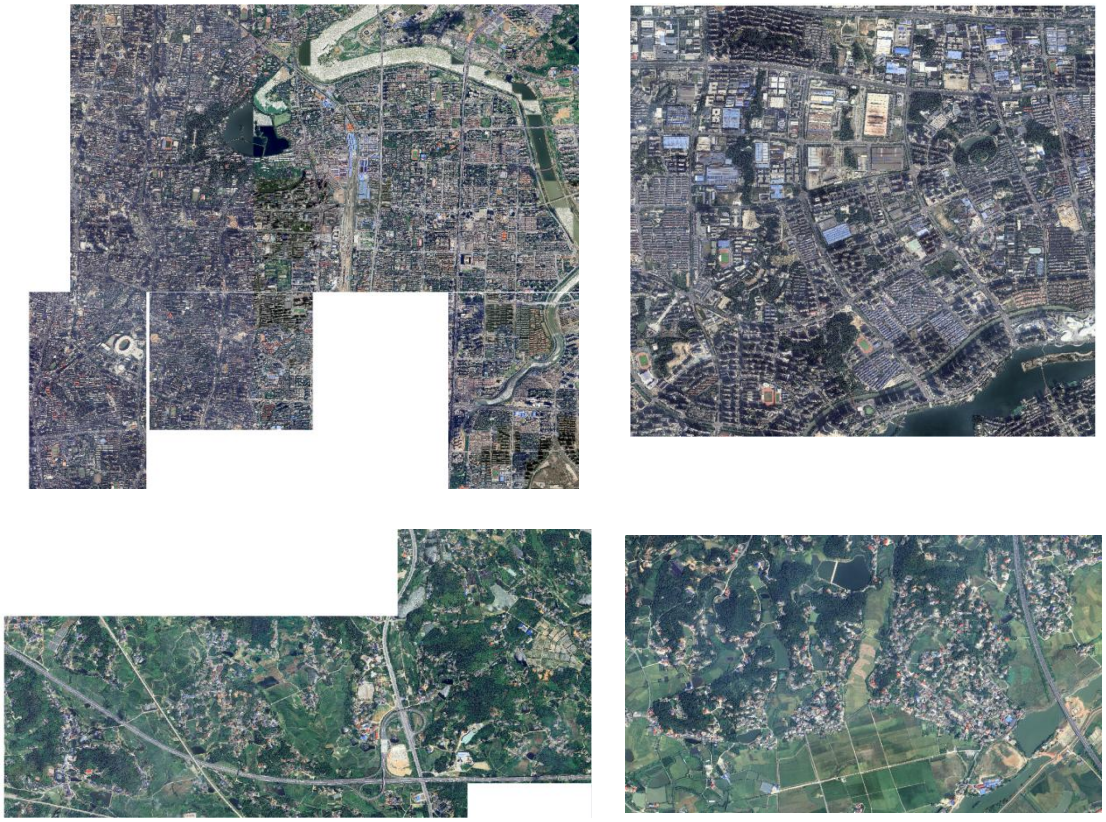
2. 长沙轮廓数据集

鉴于现有公开数据集多来源于国外地区，其建筑风格与国内实际应用场景存在一定差异，直接迁移往往导致分割精度下降。为此，本文针对国内遥感应用需求，构建了覆盖湖南省长沙市的自建数据集，用以进一步验证模型在真实应用环境下的适用性。

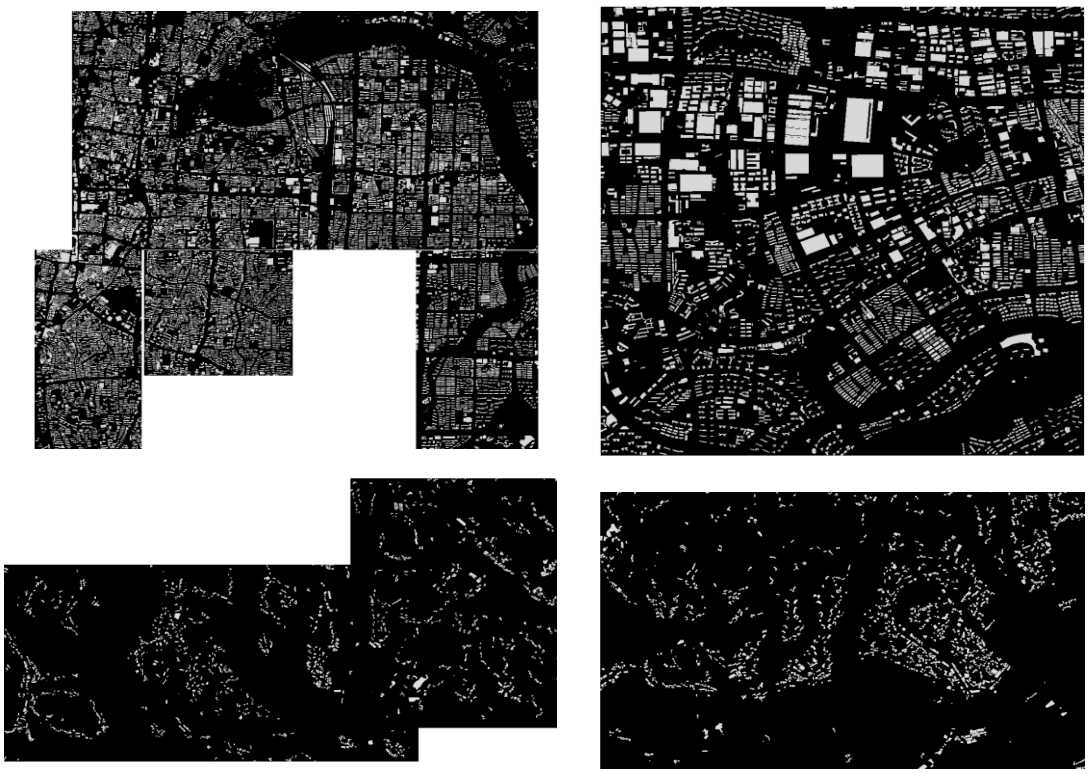
长沙遥感数据集采用 Google Earth 提供的高分辨率卫星影像作为数据源，影像的空间分辨率为 0.3m，影像覆盖长沙市的城市与农村区域，包含多样化的建筑类型与空间分布形态，如图 3.5 a）所示。影像均为 RGB 三波段，标注过程中采用逐栋建筑物人工勾绘的方式，对每一栋建筑物的轮廓边界进行像素级精细标注，累计完成 24995 栋建筑物的标注工作，确保了标签的高质量与高精度。为适配深度学习模型训练，原始影像被裁剪为 512×512 的图像块，空间分辨率保持为 Google Earth 原始分辨率 0.3m。该子集覆盖总面积约 115km²，其中城市区域约 90km²（图 3.5 a）中 1 和 2 部分），农村区域约 25km²（图 3.5 a）中 3 和 4 部分）。影像总数量为 4428 张，城市与农村子集分别按照 8:2 的比例划分训练集与验证集，最终获得 3542 张训练样本和 886 张验证样本。



a）数据集覆盖范围示意图



b）光学影像



c）建筑轮廓标签图

图 3.5 长沙轮廓数据集

3.3.2 实验参数设置

实验分为对比实验与最佳精度实验两类，所有实验均在相同的软硬件环境下完成。具体实验条件和参数设置如下:

对于对比实验，所有模型均基于深度学习框架 PyTorch，在配有 NVIDIA GeForce RTX 4070Ti SUPER 显卡的 Ubuntu20.04 操作系统上进行训练。模型的具体实现基于 1.22 版本的 mmsegmentation 代码库，这是一个基于 PyTorch 深度学习框架构建的开源语义分割工具箱，集成了多种主流算法并提供了统一的标准化评测基准。训练过程中使用在线数据增广来提高模型的范化能力，主要包括 50%到 200%的随机缩放、75%的随机裁剪。采用初始学习率为 6×10^{-5} 的 AdamW 优化器实现模型的收敛，参数β为[0.9,0.999]，权重衰减为 0.01。考虑到两个数据集的规模差异，为确保模型充分训练，WHU 航空影像数据集的迭代训练次数设置为 160000 次，对应 270 个训练轮次；长沙轮廓数据集的迭代训练次数设置为 80000 次，对应 180 个训练轮次。损失函数采用本文提出的复合损失函数，所有网络训练输入的图像大小为 512×512。训练的 batch size 大小为 8，学习率调度策略采用线性预热（Linear Warm-up）与多项式衰减（Polynomial Decay）相结合的方式：在训练初始阶段，学习率通过线性策略从极小值逐步提升至设定的初始学习率，以缓解训练初期的不稳定问题；随后在预热结束后，采用多项式衰减策略逐步降

低学习率直至训练结束，从而保证模型在后期阶段的稳定收敛。

在最佳精度训练实验中，所有模型均使用 mmsegmentation 提供的预训练编码器权重进行初始化，并通过微调获得最优性能。受显存限制，batch size 调整为 4。该策略在加速模型收敛、减少训练时间的同时，有助于进一步提升模型最终分割精度。

3.3.3 WHU 航空影像数据集实验结果分析

为全面验证本文所提出 SPR-Net 的性能，我们选取了 7 种主流语义分割网络作为对比基线，与本文提出的 2 个不同编码器配置的 SPR-Net 变体（SPR-Net-b0 和 SPR-Net-b2），共计 9 个模型进行对比实验。根据网络架构特点，这些对比模型可大致分为四类：（1）轻量化网络，以 Mobilenet-v3-s 为代表，注重计算效率与部署便捷性；（2）经典卷积网络，包括 PSPNet-r50、Deeplabv3plus-r50 和 SegNext-b，代表了传统卷积神经网络在多尺度特征融合、空洞卷积及高效设计等不同发展路径；（3）基于 Transformer 的模型，包括 Swin-t 和 Segformer-b2，代表了利用自注意力机制捕捉全局上下文信息的前沿方向；（4）基于状态空间模型的网络，即 SegMan-s，体现了 Mamba 等新兴架构在语义分割领域的最新探索。这些模型均为近年来学术界广泛认可的主流方法，在实际应用中具有代表性。表 3.1 展示了各网络在 WHU 航空影像数据集上的详细对比结果。

表 3.1 不同网络与 SPR-NET 在 WHU 航空影像数据集上的对比

算法	IoU(%)	Recall(%)	Preci-sion(%)	F1	Flops(G)	参数量(M)	训练时间(h)
SegNext-b[45]	87.01	91.78	94.36	93.06	32.477	27.560	12.16
PSPNet-r50[28]	88.46	93.14	94.62	93.87	183.296	46.602	21.49
Mobilenet-v3-s[108]	82.33	87.21	93.63	90.31	**4.214**	**1.139**	**4.33**
Deeplabv3plus-r50[31]	89.21	93.68	94.92	94.3	180.220	41.216	22.23
Swin-t[40]	88.83	92.87	95.33	94.08	236.000	58.942	20.17
Segformer-b2	88.24	92.34	95.21	93.75	25.245	24.724	10.02
SegMan-s[109]	88.03	92.3	95	93.63	23.779	29.403	14
SPR-Net-b0(ours)	88.2	92.6	94.88	93.73	39.916	5.751	10.13
SPR-Net-b2(ours)	**89.51**	**93.34**	**95.6**	**94.47**	59.326	27.073	16.5

由表 3.1 实验结果可以看出，不同模型在精度与效率方面呈现出明显差异。轻量化网络 Mobilenet-v3-s 虽然在计算开销上优势巨大，但其提取精度 IoU 显著低于其他模型，难以满足高精度要求。经典的语义分割网络 PSPNet-r50 与

Deeplabv3plus-r50 展现了优良的建筑物识别能力，IoU 和 Precision 等指标表现出色。然而，这两者共同的短板在于计算复杂度过高，参数量巨大且训练耗时长，这在一定程度上限制了它们的实际应用部署。SegNext-b 在精度与效率之间取得了更好的平衡，只有中等水平的计算资源消耗，但是其难以获得具有竞争力的分割精度。与之相比的基于 Transformer 的模型普遍展现了强大的特征提取能力。Swin-t 实现了非常高 IoU 和 Recall，证明了 Transformer 结构在捕捉全局上下文信息方面的优势。然而，其高昂的计算成本和庞大的模型规模成为了性能的掣肘。另一款模型 Segformer-b2 则在 Transformer 架构的效率优化方面做出了有效探索，它在保持较高精度的同时，显著降低了计算复杂度和训练时间，展现了更优的效能平衡。而 SegMan-s 取得了 88.03%的 IoU，与 Segformer-b2 性能相近，但其参数量略高（29.403M），训练时间为 14 小时。该网络在精度与效率方面表现中等。

SPR-Net-b0 是 SPR-Net 的轻量化版本，其核心优势在于使用了 Segformer 的 B0 编码器。在如此轻量化的设计下，它依然实现了 88.2%的 IoU，超越了部分更复杂的模型，证明了其架构的高效性与先进性。值得特别指出的是，SPR-Net-b0 的精度与基于 Transformer 的 Segformer-b2（IoU88.24%）几乎持平，但其参数量远低于后者。这一结果极具说服力地证明了我们所提出模块的有效性，它能够在不增加模型深度（层数）的同时，显著提升网络性能，达到了与更深、更复杂模型相当的精度水平。

而 SPR-Net-b2 在使用了 B2 解码器后，在所有对比模型中取得了最优的综合表现。其 IoU（89.51%）、Precision（95.6%）、Recall（93.34%）及 F1 分数（94.47）四项关键精度指标均位列第一，展现了无与伦比的建筑物提取能力。尤为重要的是，这种顶尖的精度并非以牺牲效率为代价。SPR-Net-b2 的模型复杂度和训练时间均控制在合理范围内，显著优于 Swin-t 等精度相近的模型，实现了精度与实用性的完美平衡。

量化的性能指标仅能提供宏观的评估，为了更直观地理解各模型在实际预测中的行为差异和具体优劣，进一步对测试集中的典型样本进行了可视化分析。

如图 3.6 所示，我们精心挑选了覆盖三类常见建筑分布的六个代表性场景（即高密度建筑群、大尺度单体建筑和稀疏零散建筑，每类各两个）。这三种场景分别代表了建筑提取中最具挑战性的典型难题：高密度建筑群因严重的阴影遮挡和边界粘连，通常会出现边界漏检与邻近建筑误判；大尺度单体建筑因内部纹理复杂多变，易产生内部空洞与区域漏提；稀疏零散建筑虽因与周围环境对比度高而相对容易检测，但细小建筑容易产生漏检。为了清晰地展示预测细节，图中以不同颜色对结果进行了解读：白色表示正确提取的像素，红色表示错误识别的像素，而黄色则代表被遗漏的建筑物部分。每个场景的 IoU 值也一并标注在右上角，从而为深入剖析各模型的建筑提取能力提供了直观的视觉证据。

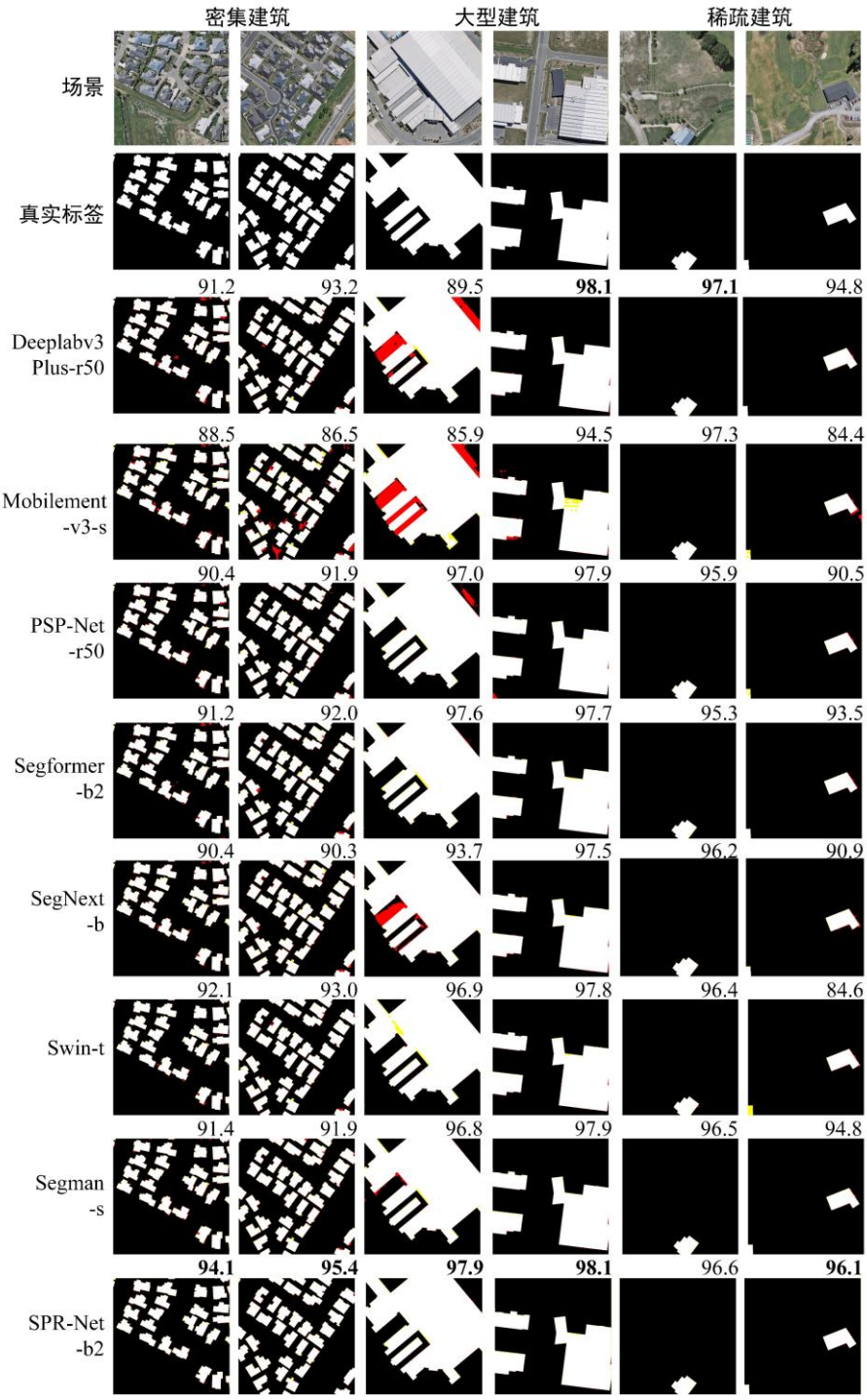


图 3.6 WHU 航空影像数据集预测结果

由图 3.6 可视化结果可知，不同网络在各类场景下的表现差异显著：

（1）密集建筑场景：在密集建筑场景下（如图 3.6 第 1、2 列所示），建筑间距狭窄且伴随严重的阴影遮挡，这对网络的边界划分能力提出了极高挑战。轻量级主干网络 MobileNet-v3-s 在此场景下暴露出明显的局限性，由于特征提取能力有限，出现了较多的边界粘连与误判。相比之下，拥有更深层数及更大参数量的 Deeplabv3plus-r50 和 Swin-t 凭借强大的特征表达能力，展现了更高的分割精度。值得注意的是，SegNext-b、Segformer-b2 和 SegMan-s 等轻量级网络虽然在整

体结构上进行了优化，但在密集边缘的处理上仍逊色于 Deeplabv3plus-r50 和 Swin-t，主要表现为建筑轮廓的模糊。此外，尽管 PSPNet-r50 同样拥有较大的参数量，但在该场景下表现并不突出，这是由于 PSPNet 的金字塔池化模块在聚合全局上下文时，往往会忽略局部的空间细节信息，导致密集小目标的边缘信息丢失。而本文提出的 SPR-Net 通过有效的特征融合策略，在建筑边缘细节上表现最佳，成功区分了紧密相邻的建筑物。

（2）大型建筑场景：在大型建筑场景下（如图 3.6 第 3、4 列所示），建筑物通常具有规则的几何边缘。对于纹理简单的大型建筑（如图中第 4 列），所有网络均能取得较好的预测结果。然而，面对纹理复杂或与地面背景高度相似的建筑（如图中第 3 列），基于 CNN 主干的网络（如 MobileNet，ResNet）普遍出现了不同程度的“空洞”与漏检现象，这是由于卷积操作受限于局部感受野，难以区分相似的纹理特征。相反，以 Transformer 为骨干的网络凭借其强大的全局上下文信息聚合能力，有效克服了局部纹理干扰，保持了大型建筑的内部完整性。同时，以 Mamba 为主干的 SegMan-s 同样展现了不俗的性能，证明了新一代状态空间模型在处理长距离依赖关系方面的巨大潜力。

（3）稀疏建筑场景：在稀疏建筑场景中（如图 3.6 第 5、6 列所示），由于建筑物与周围地形（如植被、道路）的区分度较高，所有网络均能准确地定位建筑区域。此时，性能的差异主要体现在对建筑几何边缘的建模能力上。SPR-Net 在此方面表现优异，其预测结果不仅定位精准，且边缘平滑锐利，仅在极细微的边缘处存在少量像素级的偏差（细线误判），在所有对比网络中展现了最好的形态保真度。

综上所述，SPR-Net 在三个场景中均展现了卓越的鲁棒性，取得了 96.37%的最佳 mIoU 表现，这充分说明了 SPR-Net 在多尺度特征融合与细节恢复方面的优势。同时，以 Mamba 为核心的 SegMan-s 取得了 94.87%的 mIoU，进一步验证了新一代状态空间模型在遥感图像分割任务中的应用潜力。

为了进一步验证 SPR-Net 在现有技术中的先进性，我们在 WHU 航空影像数据集上将其与当前主流的 SOTA（State-of-the-Art）网络进行了全面对比。为了探究模型的性能上限，我们在本次对比实验中采用了特征提取能力更强的 MixTransformer-B5 作为骨干网络（SPR-Net-b5）。结果见表 3.2。

由表 3.2 可以看出，SPR-Net-b5 在 IoU、Recall、Precision 和 F1-score 四项核心指标上均取得了最优表现，展现了显著的性能优势：

综合性能突破：SPR-Net-b5 的 IoU 达到了 91.90%，以 1.04%的显著优势超越了排名第二的 MAP-net（90.86%）。在语义分割任务中，当 IoU 达到 90%的高分段后，性能往往已趋于饱和，每提升 1%都不仅是数值的增长，更是对模型特征提取极限的实质性突破，这有力证明了本文方法的先进性。

表 3.2 SPR-Net 与主流 SOTA 方法在 WHU 航空影像数据集上的性能对比

算法	IoU(%)	Recall(%)	Precision(%)	F1
MAP-net[110]	90.86	94.81	95.62	95.21
MSANet[111]	88.25	94.65	92.88	93.76
HD-Net[54]	90.19	94.68	95.00	94.84
Easy-net[112]	89.16	93.48	94.94	94.20
MSRF-net[49]	90.16	95.00	94.97	94.82
SPR-Net-b5(ours)	**91.90**	**95.31**	**96.26**	**95.78**

高召回率与边缘感知：模型取得了 95.31%的最高召回率，这表明漏检现象得到了极大抑制。这一优势主要归功于本文引入的边缘伪模态，它显式增强了网络对高频信息的敏感度，使得那些在原始 RGB 图像中模糊不清或难以识别的建筑边缘能够被网络精准捕捉。

高精确率与背景抗干扰：同时，模型保持了 96.26%的最高精确率，说明误报率极低，SPR-Net 能够有效区分建筑物与具有相似光谱特征的背景干扰项（如道路、混凝土地面等），从而避免了将非建筑背景错误识别为建筑物的情况。

3.3.4 长沙建筑轮廓数据集实验结果分析

上述在 WHU 公开数据集上的实验初步证实了 SPR-Net 卓越的基础提取能力。然而，为了进一步检验该创新架构在面临更复杂的真实城市建成区时的泛化性能与实际工程应用潜力，本文随后在自主构建的长沙市建筑轮廓数据集上进行了深度评估。该数据集涵盖了更加密集且形态多样的城市建筑群，对模型的抗干扰与边界精细化刻画能力提出了更高要求。表 3.3 展示了各网络在长沙建筑轮廓数据集上的详细对比结果。

表 3.3 不同网络与 SPR-NET 在长沙建筑轮廓提取数据集上的对比

算法	IoU(%)	Recall(%)	Precision(%)	F1
SegNext-b	70.95	80.28	85.93	83.01
PSPNet-r50	75.45	85.47	86.5	86.01
Mobilenet-v3-s	60.75	72.34	79.14	75.59
Deeplabv3plus-r50	74.24	84.26	86.2	85.22
Swin-t	71.15	80.02	86.53	83.14
Segformer-b2	72.74	81.55	87.80	84.22
SegMan-s	71.44	80.76	86.09	83.34
SPR-Net-b2(ours)	**76.93**	**85.8**	**88.15**	**86.96**

由表 3.3 的结果可以看出，从实验结果来看，与 WHU 数据集的结论相符，MobileNet-v3-s 依然是效率最高但精度最低的模型，受限于极低的参数量，其特

征提取能力在不同场景下均显不足。同样，SegNext-b、Segformer-b2 和 SegMan-s 继续保持了在精度与计算成本之间的良好平衡。其中，Segformer-b2 依旧发挥稳定，在较低的计算负载下取得了优于 Swin-t 的结果，再次印证了其作为高效 Transformer 架构的泛化能力。

在长沙数据集上，经典的语义分割网络展现了更强的鲁棒性。PSPNet-r50 和 DeepLabv3plus-r50 虽然计算开销依然巨大，但其精度表现非常出色。特别是 PSPNet-r50，取得了 75.45%的 IoU，反超了在 WHU 数据集中表现亮眼的 Swin-t（IoU 71.15%）。这一现象表明，在某些特定的数据分布下，单纯增加 Transformer 的模型复杂度并不总能带来性能收益，反而是经典的 CNN 结构表现出了更强的适应性。

这一现象充分说明。尽管 Transformer 架构通常在捕捉长距离全局依赖方面具有优势，但在处理本数据集时面临特定挑战。该场景不仅包含布局极度紧凑、边缘粘连严重的“密集城中村”，更显著的特征是存在大量小目标建筑。这些小目标在影像中占比较小，且与背景地物具有极高的光谱相似性。因此，分割任务的瓶颈很大程度上取决于模型对局部高频细节（如微小边界、角点结构）的重构能力。

在此背景下，Swin Transformer 基于窗口的自注意力机制虽然减少了计算量，但在处理窗口内的高度相似特征时，倾向于产生过度平滑的特征表示，导致小目标容易被背景淹没，粘连实例难以分离。相比之下，DeepLabv3plus 和 PSPNet 等经典 CNN 架构凭借卷积操作固有的归纳偏置，对局部纹理和边缘信息更为敏感。特别是 DeepLabv3plus 的空洞卷积能够在保持特征图分辨率的同时有效扩大感受野，而 PSPNet 的金字塔池化模块则显式地融合了多尺度上下文。这种机制使得 CNN 在处理尺度变化大且边界模糊的小目标建筑时，能够比基础的 Transformer 架构生成更锐利、更准确的分割边界。

SPR-Net-b2 的优越性在于它成功打破了这一性能桎梏，在所有对比模型中取得了最优表现。特别是相较于其基准架构 Segformer-b2，SPR-Net-b2 展现了显著的性能飞跃，其 IoU 达到了 76.93%，相比 Segformer-b2（72.74%）大幅提升了 4.19%。这一显著的精度增益有力地证明了本文所提出的基于伪多模态信息的结构增强策略的有效性。针对小目标与粘连建筑丢失高频信息的问题，SPR-Net 在仅依赖光学影像输入的前提下，将拉普拉斯派生的高频结构提示作为伪多模态信息融入网络。这种策略为模型引入了显式的边界先验，直接强化了网络对细粒度结构细节和轮廓边界的感知能力。通过这种方式，SPR-Net 成功克服了原始 Segformer 在下采样过程中丢失小目标细节的瓶颈，实现了对复杂场景下建筑轮廓的精细化刻画。最终，SPR-Net-b2 不仅超越了同类的 Transformer 架构，更在保持轻量级优势的同时，成功击败了计算量巨大的 PSPNet-r50，实现了精度与效率

的双重最优。部分定性可视化结果如图 3.7 所示。

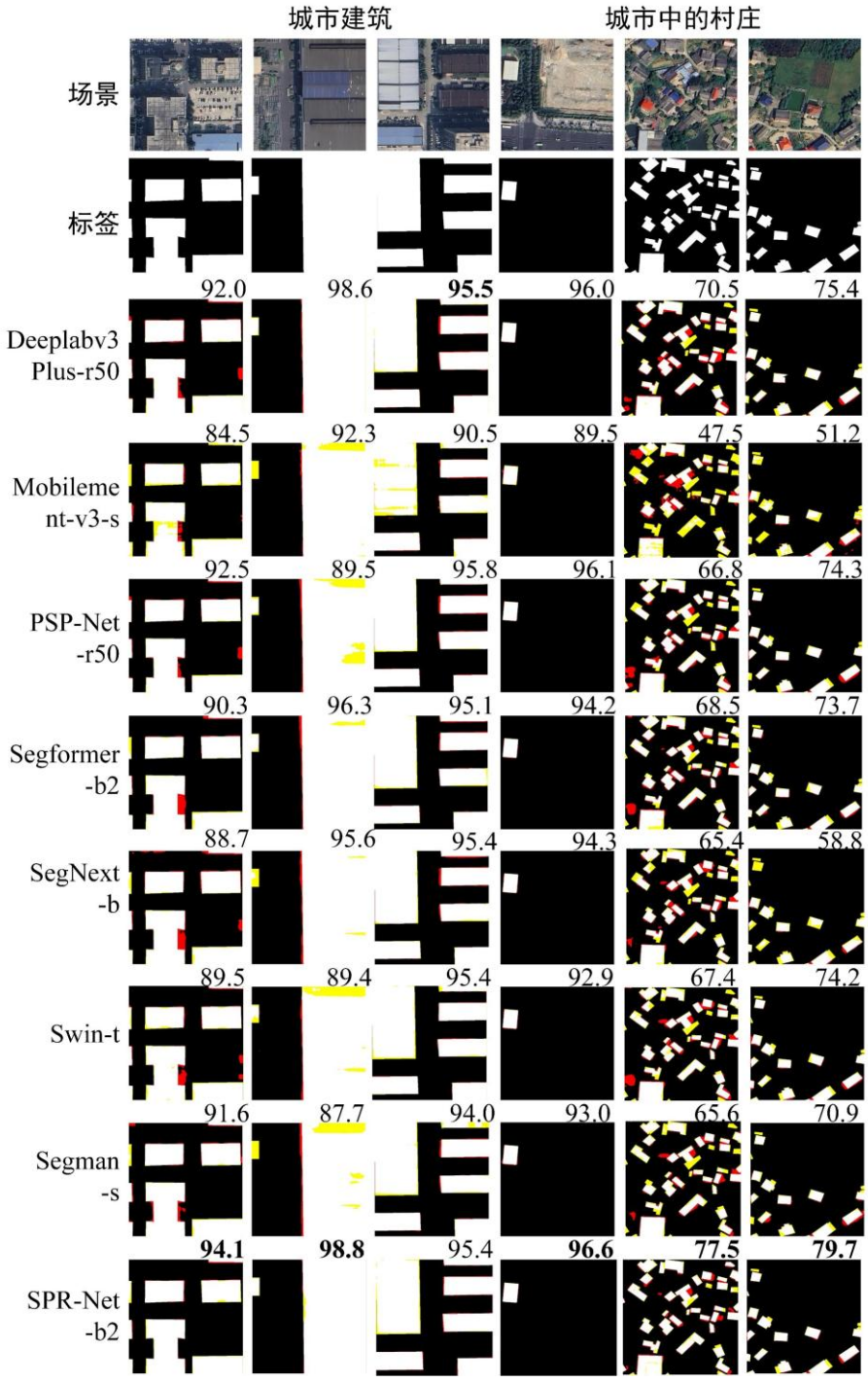


图 3.7 长沙轮廓数据集预测结果

由图 3.7 可视化结果可知，不同网络在各类场景下的表现差异显著：

（1）在城市建筑场景下（如图 3.7 第 1-4 列所示），建筑物通常具有规则的几何结构和较大的尺。MobileNet-v3-s 受限于较浅的网络深度，在处理大型建筑物内部纹理时表现出明显的不稳定性，出现了大量的黄色漏检区域（如图中第 2 列，IoU 仅为 92.3），产生了“空洞”现象。相比之下，DeepLabv3plus-r50 和 PSPNet-r50 凭借成熟的上下文聚合机制，能够较好地保持大型建筑的形态完整性。虽然

Swin-t 和 Segformer-b2 整体表现稳定，但在某些边缘细节的处理上仍略显平滑。SPR-Net（b2）在此类场景中展现了最佳的内部一致性。得益于伪多模态特征增强模块对结构信息的强调，它不仅消除了大型建筑内部的空洞（如图中第 2 列，IoU 达到 98.8），而且在第 1 列存在复杂阴影干扰的情况下，依然保持了 94.1 的高精度，显著优于 Swin（89.5）。这证明了 SPR-Net 在处理多尺度几何特征及抑制环境噪声方面的鲁棒性。

（2）城中村复杂场景：城中村场景（如图 3.7 第 5、6 列所示）是本数据集最具挑战性的部分。该区域建筑物布局极度密集，且由于光谱相似性高，建筑物之间极易发生粘连。观察第 5 列，DeepLabv3plus-r50 取得了 70.5 的 IoU，显著优于 Swin-t 的 65.6。这是因为 Swin 的窗口注意力机制在处理高度相似的相邻目标时产生了过度平滑效应，导致大量红色误检和黄色漏检。相比之下，DeepLabv3plus 的空洞卷积更有效地保留了高频边缘信息。尽管场景极具挑战，SPR-Net-b2 依然取得了压倒性的优势。在第 5 列和第 6 列中，SPR-Net 分别达到了 77.5 和 79.7 的 IoU，远超 DeepLabv3plus 和 Segformer。从图中可以清晰地看到，SPR-Net 生成的预测图中红色与黄色区域最少，成功分割了紧密相邻的独立房屋。

综上所述，SPR-Net 在各类挑战性场景中均展现了卓越的鲁棒性，取得了 90.35%的最佳 mIoU 表现，这充分说明了 SPR-Net 架构的先进性。值得深入分析的是，尽管该总体指标相比 WHU 数据集有所下降，但这主要归因于城中村场景极高的识别难度，显著拉低了整体平均分数。事实上，在常规的城市建筑中，SPR-Net 依然保持了与 WHU 数据集相当的高识别率，预测效果十分出色。然而，即便在这些高精度区域，建筑边缘仍存在少量的误判与漏检。

这主要是由于本数据集中包含大量高层甚至超高层建筑，且所使用的 Google 光学影像存在明显的侧向成像效应，并未进行严格的正射校正。这种成像条件导致建筑立面在影像中大量暴露，由此产生的几何畸变使得部分侧面纹理被模型误判为屋顶，从而影响了边缘分割的极致精度。但即便受此物理条件限制，SPR-Net 仍能通过伪多模态信息有效抑制干扰，准确提取建筑主体轮廓，进一步印证了其在实际复杂遥感应用中的可靠性。

3.3.5 消融实验

为了验证 SPR-Net 中各组件的有效性，我们基于 WHU 航空影像数据集设计了消融实验。实验以 Segformer-b0 作为基线网络（Baseline），采用 IoU、F1-score、Precision 和 Recall 作为评估指标。实验设置如下：

第一步，在基线网络的编码器上加入 PMM_S 模块，在编码器阶段引入伪多模态模块（PMM_S）对特征提取的影响，同时实验对比了仅在浅层（阶段 1、2）加入 PMM_S 与在全阶段（阶段 1-4）加入的效果。第二步，考虑只加入解码器

PMM_C 模块；第三步同时加入 PMM_S 与 PMM_C 模块；最后，再加入 RFA 模块，验证其对处理多尺度目标的能力。

实验结果记录在表 3.4 中，最高值以粗体突出显示。实验表明，我们方法的每一个模块都可以有效提高网络建模能力。进一步的对每一个步骤的机制进行了分析：

表 3.4 SPR-Net 在 WHU 航空影像数据集上的消融实验

算法	IoU(%)	Recall(%)	Precision(%)	F1
Segformer-b0	85.69	90.61	94.05	92.3
Segformer+PMM_S(12)	86.04	90.51	94.57	92.5
Segformer+PMM_C	87.11	91.37	94.91	93.11
Segformer+PMM_S(1234)+PMM_C	87.23	92.13	94.26	93.18
Segformer+PMM_S(12)+PMM_C	87.56	91.77	95.02	93.37
Segformer+PMM_S(12) +PMM_C+RFA(SPR-Net-b0)	**88.2**	**92.6**	**94.88**	**93.73**

结果表明，这种伪模态作为一种结构注意力，迫使编码器在提取语义特征的同时，显式地保留建筑物的几何边界信息。同时仅在浅层加入效果更优。这是因为浅层特征图保留了丰富的空间细节，适合提取高频边缘；而深层特征图主要包含抽象语义，强制引入高频噪声反而可能破坏语义的一致性。因此，我们在最终方案中选择在阶段 1 和 2 部署 PMM_S。解码器中 PMM_C 能够有效抑制背景噪声，并专门针对容易发生粘连的密集建筑边缘进行特征重加权，显著提升了边界的锐度，当同时加入编码器和解码器的增强模块时，网络表现出明显的性能叠加效应。编码器提供的丰富结构信息为解码器的精细化分割提供了更优质的底座，两者形成了“特征保持-边界精修”的互补关系。

再加入 RFAConv 模块，通过注意力机制动态调整感受野权重，使得网络能够自适应地处理多尺度目标。这一改进在几乎不增加计算成本的前提下，进一步提升了模型在复杂尺度变化场景下的 IoU 表现。

为了进一步验证所提出的组合损失函数对网络的性能提升的有效性，进行了消融实验：同样基于 WHU 航空影像数据集，在 Segformer-b0 网络上进行验证，采用 IoU、F1-score、Precision 和 Recall 作为评估指标。实验结果如表 3.5。

实验结果表明，在$\mathcal{L}_{\text{Dice}}$基础上引入$\mathcal{L}_{\text{Lovasz}}$后，IoU 提升至 85.13%，这主要归功于 Lovasz Loss 直接对 IoU 进行凸优化，能够从全局层面更好地处理类间不平衡问题，弥补了 Dice Loss 在非凸优化过程中的局限性，进一步加入$\mathcal{L}_{\text{Boundary}}$后，模型性能获得显著增益，IoU 最终达到 85.69%，且 Recall 指标出现了大幅跃升。这一结果表明，$\mathcal{L}_{\text{Boundary}}$成功迫使模型关注那些在常规像素级损失中容易被忽略

的边界像素，通过显式的几何约束锐化了预测边缘，最终采用的混合损失函数$\mathcal{L}_{\text{Dice}}+0.3\mathcal{L}_{\text{Lovasz}}+0.05\mathcal{L}_{\text{Boundary}}$实现了最佳的精度与召回率平衡，证明了三者结合的必要性。

表 3.5 不同损失函数组合在 WHU 航空影像数据集上的消融实验结果

损失函数	IoU(%)	Recall(%)	Precision(%)	F1
$\mathcal{L}_{\text{Dice}}$	84.31	88.1	93.0	91.1
$\mathcal{L}_{\text{Dice}}+0.3\mathcal{L}_{\text{Lovasz}}$	85.13	88.92	93.51	91.7
$\mathcal{L}_{\text{Dice}}+0.3\mathcal{L}_{\text{Lovasz}}+0.05\mathcal{L}_{\text{Boundary}}$	**85.69**	**90.61**	**94.05**	**92.3**

3.3.6 复杂度实验

为了验证 SPR-Net 的性能与复杂度之间的权衡，我们比较了相关方法的 Flops、训练参数和训练时间，如图 3.8、图 3.9 所示。在这两幅多维气泡图中，横坐标表示模型的参数量（M），数值越小说明模型占用内存越少；纵坐标表示分割精度交并比 IoU（%），数值越大说明模型的准确性越高。图中的气泡大小则直观地反映了模型的运行成本，在图 3.8 中气泡面积代表计算复杂度 Flops（G），在图 3.9 中气泡面积代表训练时间（h）。因此，一个表现优异的模型在图中应当尽可能地靠近左上角，即具备高精度与低参数量，同时拥有尽可能小的气泡面积（即低计算复杂度和较短的训练时间）。

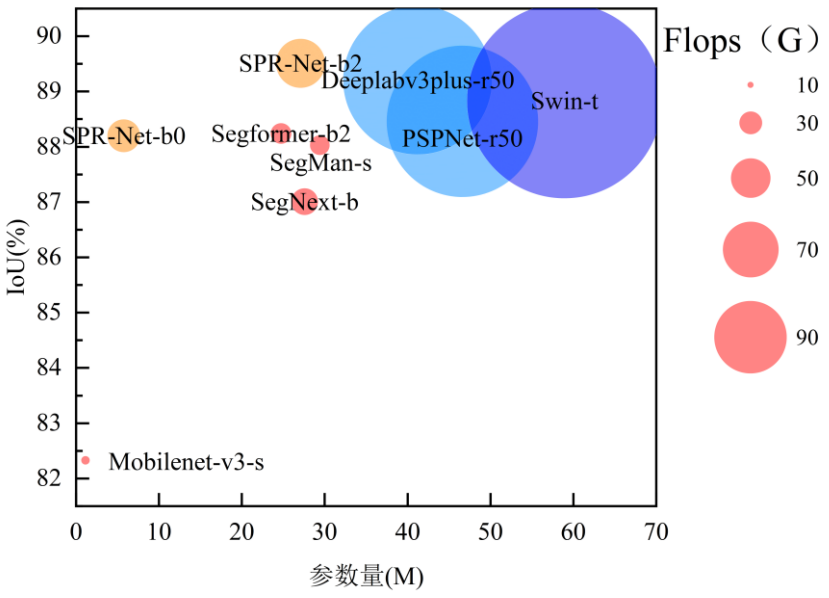


图 3.8 相关方法的复杂度和准确性比较

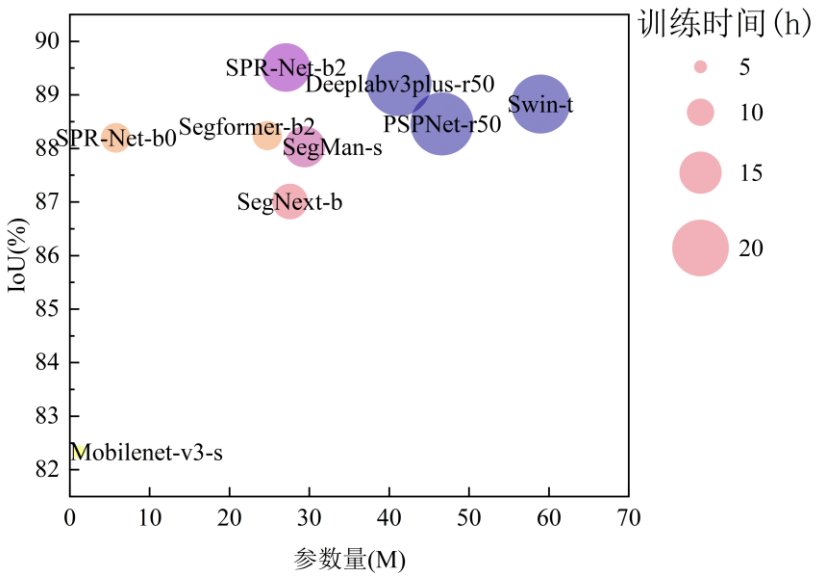


图 3.9 相关方法的训练时间和准确性比较

如图 3.8 所示，SPR-Net-b2 展现了卓越的综合性能，它在参数量仅为约 27M 的情况下，取得了所有对比模型中最高的 IoU 分数（约 89.5%）。与 DeepLabv3plus-r50 和 Swin-t 等重型模型相比，SPR-Net-b2 不仅精度更高，且参数量大幅减少，代表计算量（Flops）的气泡面积也显著更小。此外，SPR-Net-b0 则展示了极致的轻量化优势，以极低的参数量（<10M）实现了超过 88%的 IoU，显著优于 MobileNet-v3-s。

图 3.9 进一步从时间成本角度验证了算法的高效性。得益于高效的网络设计，SPR-Net 在保持 SOTA 精度的同时，其训练耗时仍保持在较低水平，明显优于计算密集的 Swin-t 和 DeepLabv3plus。综上所述，SPR-Net 系列模型在精度、参数量和计算成本之间取得了最佳的平衡。

3.4 本章小结

本章针对光学遥感影像建筑物轮廓提取中存在的边界模糊、尺度差异显著及复杂背景干扰等问题，提出了基于伪多模态结构信息的建筑物精细分割网络 SPR-Net，在仅依赖单一光学影像输入的前提下，通过结构增强与预测校正的协同机制实现了高精度的建筑轮廓提取。主要研究内容与结论如下：

（1）设计了编码器端 PMM_S 与解码器端 PMM_C 两类伪多模态结构增强模块，解决了多模态数据获取成本高且预处理复杂难以满足大范围应用需求的问题。模块分别从高频特征显式构造与预测引导的特征重加权两个层面强化建筑轮廓边界表达；同时构建了覆盖长沙市约 115km²、涵盖城市与农村多样化建筑分布的高分辨率数据集，有效验证了方法在国内真实场景下的泛化能力。

（2）提出了基于伪多模态结构信息的建筑轮廓提取网络 SPR-Net，实现了复杂背景与多尺度条件下建筑轮廓提取精度的显著提升。在 WHU 航空影像数据集上以 91.90%的 IoU 超越 MAP-Net 等当前主流 SOTA 方法，在长沙轮廓数据集上相较基线 Segformer-b2 提升 4.19 个百分点，在复杂背景与多尺度条件下均实现了稳定的轮廓提取效果。

（3）设计了融合 Dice 损失、边界损失与 Lovasz 损失的复合损失函数，解决了常规网络缺乏对建筑边缘高频信息显式建模的问题。该损失函数通过区域重叠度、边界几何约束与 IoU 优化的联合约束，有效提升了分割边界的锐度与整体精度。

第4章 基于 Segformer-H 网络的建筑高度估计

4.1 引言

在遥感影像建筑物高度估计任务中，多尺度特征融合阶段的空间信息丢失与边界模糊，以及模型在不同城市形态下的泛化困难，是当前亟待解决的两大核心痛点。一方面，现有的基于 Transformer 架构的 Segformer 网络虽然在全局语义建模方面展现出巨大优势，但其解码器在特征融合时采用的双线性插值上采样策略存在明显局限。作为一种内容无关的固定核插值方法，其权重仅取决于像素间的空间距离，无法根据特征内容自适应地调整采样策略，导致建筑物边界处的高度预测呈现渐变模糊，且在密集建筑群中极易引起相邻高度信息的相互干扰。此外，低分辨率特征进行大倍率上采样时会产生更为显著的空间失真。另一方面，目前建筑物高度估计领域的主流公开数据集大多基于国外城市特征构建，缺乏面向国内实际应用场景的大规模数据集。由于国内城市建筑高度跨度极大、分布极为密集，依赖现有公开数据集训练的模型在面对国内复杂城市场景时，往往面临严重的泛化困难。为解决上述问题，本章引入 DySample 动态上采样模块改进 Segformer 解码器，并构建长沙建筑高度数据集以增强模型泛化能力。本章将详细阐述网络架构的设计细节，并通过系统的实验验证该方法与自建数据集的有效性。

4.2 方法介绍

4.2.1 Segformer-H 架构总览

本文改进了 Segformer 网络，使其成为一种面向遥感影像建筑物高度估计的网络模型 Segformer-H，旨在有效解决多尺度特征融合过程中传统上采样方法导致的空间信息丢失、建筑物边界高度模糊以及高度突变区域预测失真等问题。其整体架构如图 4.1 所示。

4.2.2 编码器部分

编码器沿用 Segformer 的标准层次化 Transformer 结构，由四个阶段（阶段 1-4）组成，分别输出 1/4、1/8、1/16 和 1/32 空间分辨率的多尺度特征图。每个阶段依次包含重叠补丁合并层、高效自注意力模块和 Mix-FFN 前馈网络，逐级提取从局部纹理到全局语义的多层次特征表示。其具体结构与工作原理已在第三章 3.2.2 节中详细阐述，此处不再赘述。

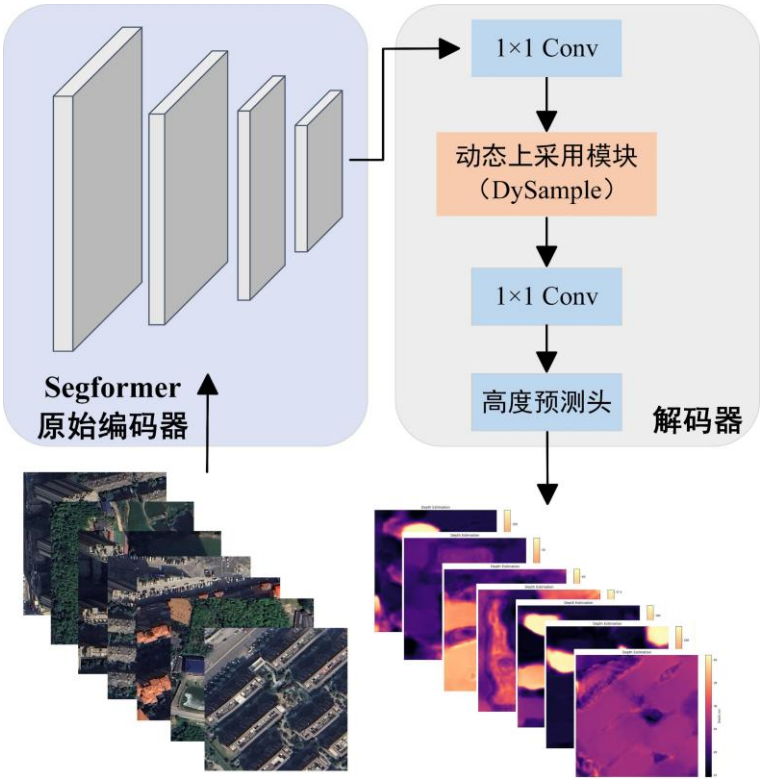


图 4.1 Segformer-H 架构总览

4.2.3 解码器部分

解码器是 Segformer-H 的核心改进所在，负责将编码器输出的四个尺度特征进行融合并生成逐像素的建筑物高度预测值。相较于原始 Segformer 解码器，本文的主要改进在于将双线性插值上采样替换为基于 DySample[113]的级联动态上采样，见图 4.2。

1. 基于 DySample 的级联动态上采样：

来自编码器四个阶段的特征图F_i，首先通过1×1卷积层映射至统一的通道维度C，随后经过 DySample 动态上采样至统一的 1/4 尺度进行融合。对于输入特征图$F_i^{'}$，DySample 首先通过轻量级1×1卷积生成采样位置偏移量：

$$\Delta = \mathrm{Conv}_{1\times1}\left(F_i^{'}\right)\times\alpha + P_0 \tag{4.1}$$

其中P_0为预定义的均匀网格初始采样位置，α为缩放因子，用于约束偏移量的幅度在初始网格间距的合理范围内，防止训练初期因偏移量过大导致梯度不稳定。通过该机制，偏移量生成器能够学习到与特征语义相关的采样策略：在建筑物边界处，偏移量倾向于将采样点向同一高度区域内部偏移，避免跨越边界的混合采样；在纹理平滑区域，偏移量接近于零，保持均匀采样的稳定性。

在获取初始偏移量后，DySample 的核心上采样机制并非在特征层面进行插值，而是对采样坐标进行上采样。具体而言，低分辨率的偏移坐标网格会先通过

像素重排（Pixel Shuffle）操作被放大至目标高分辨率尺度。随后，基于这种高分辨率的偏移采样坐标，DySample 利用可微分的双线性网格采样算子（Grid Sample），直接从低分辨率的输入特征图中“拉取”对应位置的特征。由于采样操作按分组独立进行，不同通道组能够学习到差异化的采样偏好。通过使用高分辨率坐标网格进行采样，该模块一步到位地输出了高分辨率的特征图，避免了传统上采样中大量的冗余计算：

$$\hat{F}_i = \mathrm{PixelShuffle}\left(\mathrm{GridSample}\left(F_i^{'}, P_0 + \Delta\right)\right) \tag{4.2}$$

偏移量生成器的参数通过下游任务的损失函数联合优化，使上采样策略自适应地服务于建筑物高度估计目标。

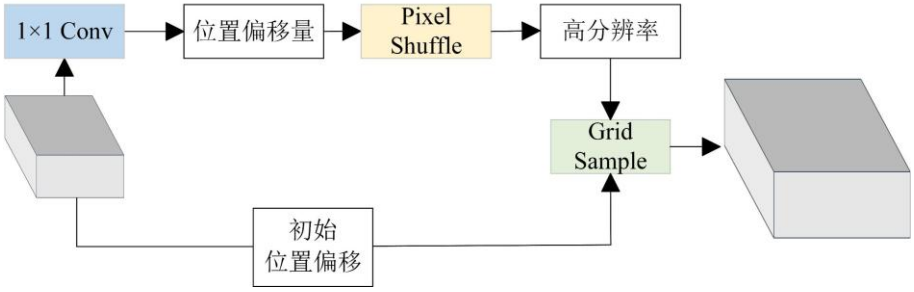


图 4.2 DySample **模块**

对于不同尺度特征到 1/4 分辨率的上采样，本文采用级联策略而非直接大倍率插值：

阶段 1 特征（1/4 尺度）：无需上采样，直接保持原始分辨率；

阶段 2 特征（1/8 尺度）：经 1 个 DySample 模块上采样至 1/4 尺度；

阶段 3 特征（1/16 尺度）：经 2 个级联 DySample 模块，实现4×上采样；

阶段 4 特征（1/32 尺度）：经 3 个级联 DySample，模块实现8×上采样；

级联设计的优势在于：每个模块仅需学习 2 倍上采样偏移量，学习难度显著低于直接学习大倍率偏移量，训练更加稳定；同时后级模块可在前级输出基础上逐步精调采样位置，实现由粗到细的渐进式分辨率恢复。

2. 高度预测头

经 DySample 上采样后的四个尺度特征统一至 1/4 分辨率，沿通道维度拼接并通过1×1融合卷积生成全局融合特征F_{fuse}。融合特征最终通过高度预测头生成逐像素的建筑物高度估计值。该预测头由3×3卷积层（通道数降为$C/2$，附带批归一化与 ReLU 激活）和1×1卷积层（输出单通道）组成：

$$D = d_{\min} + \left(d_{\max} - d_{\min}\right) \cdot \sigma\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{3\times3}\left(F_{\mathrm{fuse}}\right)\right)\right) \tag{4.3}$$

其中$\sigma(\cdot)$为 Sigmoid 函数，$d_{\min}$和$d_{\max}$分别为预设的最小和最大建筑物高度值。Sigmoid 将输出归一化至$[0,\ 1]$区间，再线性映射至$[d_{\min},\ d_{\max}]$的合理高度范围，确保预测值的物理可解释性。3×3卷积在最终预测前提供局部空间上下文聚合，有助于平滑孤立噪声并增强高度预测的空间连贯性。

4.2.4 损失函数

在模型训练阶段，本文选用尺度不变对数损失函数（Scale-Invariant Logarithmic Loss，SI-Log）[114]作为高度估计网络的优化目标。SI-Log 损失函数在对数空间中度量预测高度与真实高度之间的差异，并通过引入尺度不变项，有效削弱不同建筑高度尺度差异对模型训练过程的影响，从而提升模型在整体高度结构建模方面的稳定性与鲁棒性。与传统基于绝对误差或平方误差的损失函数相比，SI-Log 损失更加关注预测结果与真实值之间的相对关系，能够在保证局部高度精度的同时维持全局高度分布的一致性，尤其适用于稀疏高度监督条件下的建筑高度估计任务。其数学表达式如式子 4.1 所示：

$$\mathcal{L}_{\text{SI-Log}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log h_i - \log \hat{h}_i\right)^2 - \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\log h_i - \log \hat{h}_i\right)\right)^2} \tag{4.4}$$

式中，N表示参与损失计算的有效像素或样本点数量；$\hat{h}_i$为第i个像素或建筑实例的预测高度值；h_i为对应的真实高度值。为避免数值不稳定及无效像素对训练过程的干扰，损失计算过程中仅对真实高度大于给定阈值的有效样本进行统计，并在对数运算中引入极小正数以防止对零取对数的情况。

4.2.5 评价指标

本文研究的目标为基于遥感影像的建筑高度估计问题，其本质属于连续数值回归任务。为综合评估本文建筑高度估计模型的整体性能，选取了 6 种在高度回归与单目深度估计任务中广泛采用的评价指标，包括平均绝对相对误差（Abs Rel）、均方根误差（RMSE）、对数均方根误差（$\text{RMSE}_{\log}$）以及三种阈值精度指标δ_1、δ_2和δ_3。上述指标从不同角度对模型的高度预测能力进行定量分析。

Abs Rel 用于衡量预测高度与真实高度之间相对误差的平均水平，能够反映模型在不同高度尺度下的相对偏差情况，见式（4.5），该指标越小越好，较小的 Abs Rel 值表明预测结果与真实高度之间的相对差异较小。

$$\text{Abs Rel} = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|\hat{h}_i - h_i\right|}{h_i} \tag{4.5}$$

式中，N表示参与评价的有效像素或样本点数量；$\hat{h}_i$为第i个像素或建筑实例的预测高度值；h_i为对应的真实高度值。

RMSE 是预测高度与真实高度之间绝对误差的标准度量形式，见式（4.6），通过对误差平方求均值再开方，能够有效反映模型在高度估计任务中的整体预测精度，对较大误差较为敏感，该指标越小越好，RMSE 值越小说明模型预测的高度值越接近真实高度。

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{h}_i-h_i\right)^2} \tag{4.6}$$

$\mathrm{RMSE}_{\log}$通过在对数空间计算均方根误差，见式（4.7），能够降低异常值的影响，更关注预测值与真实值之间的相对关系而非绝对差异。该指标越小越好，较小的$\mathrm{RMSE}_{\log}$值表明模型在对数尺度下具有更好的预测一致性，尤其适用于建筑高度跨度较大的场景。

$$\mathrm{RMSE}_{\log}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\hat{h}_i-\log h_i\right)^2} \tag{4.7}$$

δ_1、δ_2和δ_3是一组基于相对误差阈值的精度指标，见式（4.8），用于统计预测高度与真实高度之比在给定阈值范围内的样本比例，当预测高度与真实高度之间的相对误差满足设定阈值条件时视为预测正确，这三个指标越大越好，较高的δ值表示模型能够在更严格的误差范围内保持较高的预测准确率，从而体现模型的鲁棒性与可靠性。

$$\delta_j:\max\left(\frac{\hat{h}_i}{h_i},\frac{h_i}{\hat{h}}\right)<1.25^j\in\{1,2,3\} \tag{4.8}$$

式中，j表示不同的阈值级别，其中j=1，2，3分别对应δ_1、δ_2和δ_3。

4.3 实验与分析

4.3.1 数据集

为全面评估本文所提出方法在遥感影像建筑物高度估计任务中的有效性与泛化能力，实验选取了两个广泛使用的公开数据集 ISPRS Potsdam 数据集[115]与 ISPRS Vaihingen[116]数据集，以及一个面向国内实际应用场景构建的长沙自建数据集。上述数据集在地理区域、影像来源、空间分辨率及建筑分布特征等方面均具有显著差异，涵盖了欧洲中小城市的典型建筑场景与中国城市的现代化建筑风格，包含不同高度范围、建筑密度、结构类型及城市肌理模式。这种多样性与差异性能够为模型性能验证提供具有代表性的实验基准，通过在不同地域、不同建筑特征的数据集上进行交叉验证，可以有效检验模型对不同场景的适应能力，更全面地体现模型的泛化性能与鲁棒性，确保所提方法在实际应用中具有广泛的适用性。

1. ISPRS Potsdam 数据集：

ISPRS Potsdam 数据集是由国际摄影测量与遥感学会（ISPRS）发布的城市场景语义理解基准数据集之一，广泛应用于建筑物高度估计，见图 4.3。该数据集覆盖德国波茨坦市的城市区域，包含 38 幅大小为 6000×6000 像素的高分辨率航空正射影像，空间分辨率为 0.05m，建筑高度范围为 0-25.5m。影像提供数字表面模

型（DSM）及归一化数字表面模型（nDSM）数据。其中，nDSM 反映了地物相对于地面的归一化高度信息，可直接作为建筑物高度估计任务的真值标签。该数据集所覆盖区域建筑密度较高，包含住宅楼、商业建筑及历史建筑等多种类型，建筑高度分布范围广泛，为模型在密集城市场景下的性能评估提供了可靠基准。为适配深度学习模型的输入要求，原始影像及对应的 nDSM 标签均采用滑动窗口方式裁剪为 512×512 像素的图像块。最终获得 3830 张训练样本、1095 张验证样本以及 547 张测试样本。

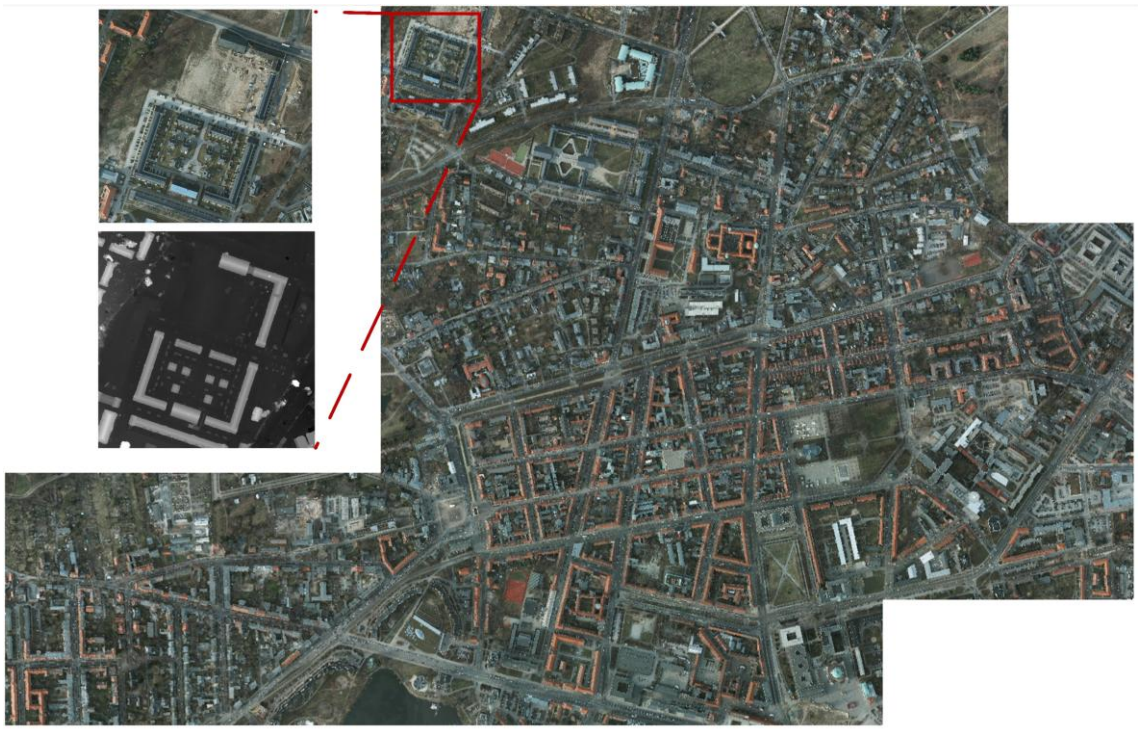


图 4.3 ISPRS Potsdam 数据集示意图

2. ISPRS Vaihingen 数据集

ISPRS Vaihingen 数据集同样由国际摄影测量与遥感学会发布，覆盖德国斐海因根市的城市与郊区区域，是另一个在建筑物高度估计领域被广泛采用的高分辨率航空遥感基准数据集，见图 4.4。该数据集包含 33 幅大小不等的航空正射影像（平均尺寸约为 2500×2000 像素），空间分辨率为 0.09m，建筑高度范围为 0-20m。影像包含近红外（NIR）、红（R）和绿（G）三个波段，同样提供 DSM 与 nDSM 数据作为高度参考信息。与 Potsdam 数据集相比，Vaihingen 数据集的空间分辨率略低，波段配置有所不同，且场景中建筑密度相对较低、以低矮住宅为主，建筑类型与高度分布的差异为评估模型在不同数据条件下的鲁棒性与泛化能力提供了有价值的补充。按照常用的实验划分方案，同样采用滑动窗口方式将影像裁剪为 512×512 像素的图像块，最终获得 348 张训练样本、100 张验证样本以及 48 张测试样本。

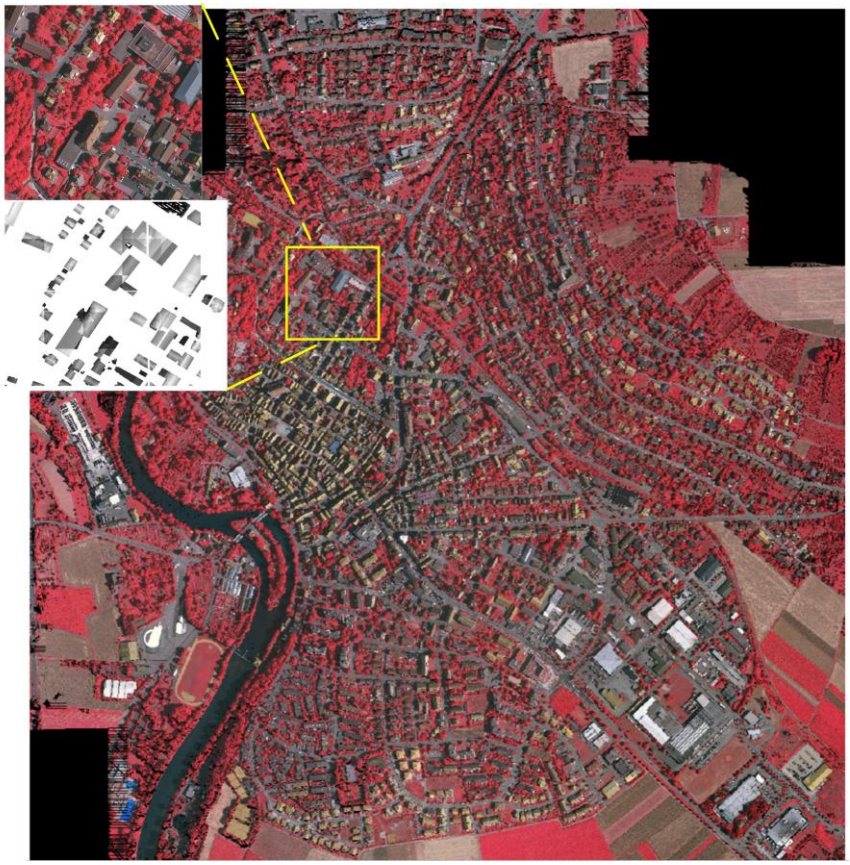
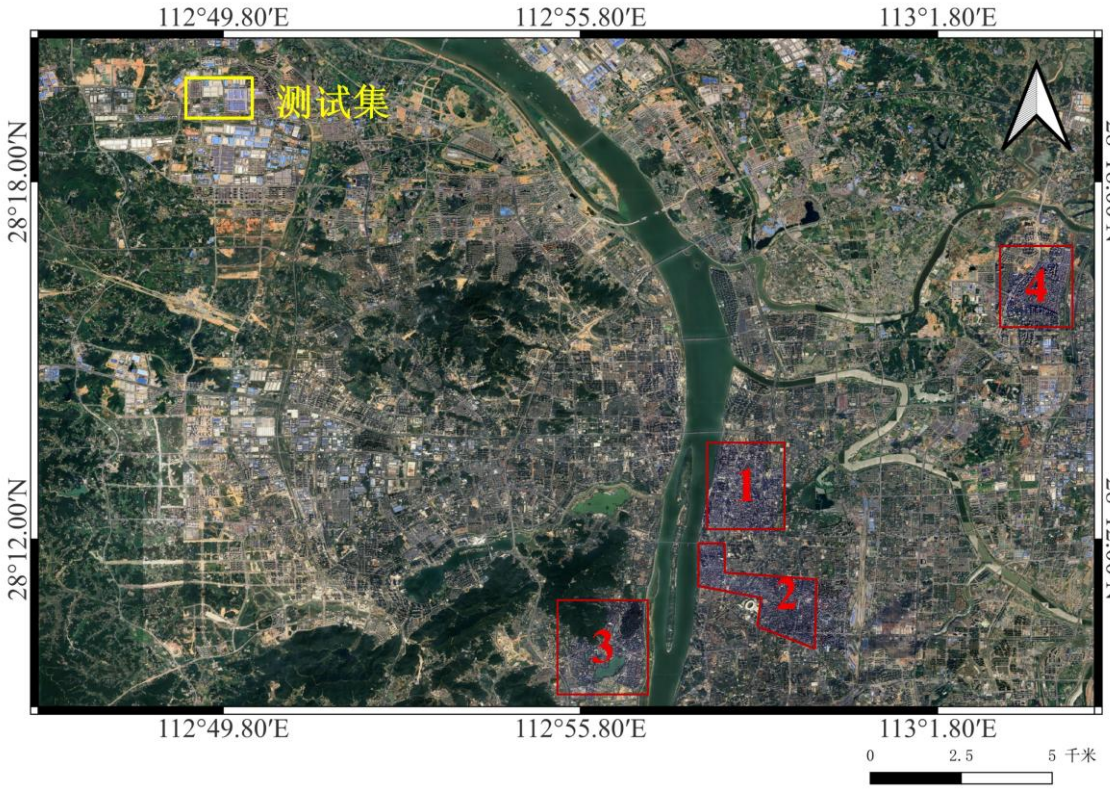


图 4.4 ISPRS Vaihingen 数据集示意图

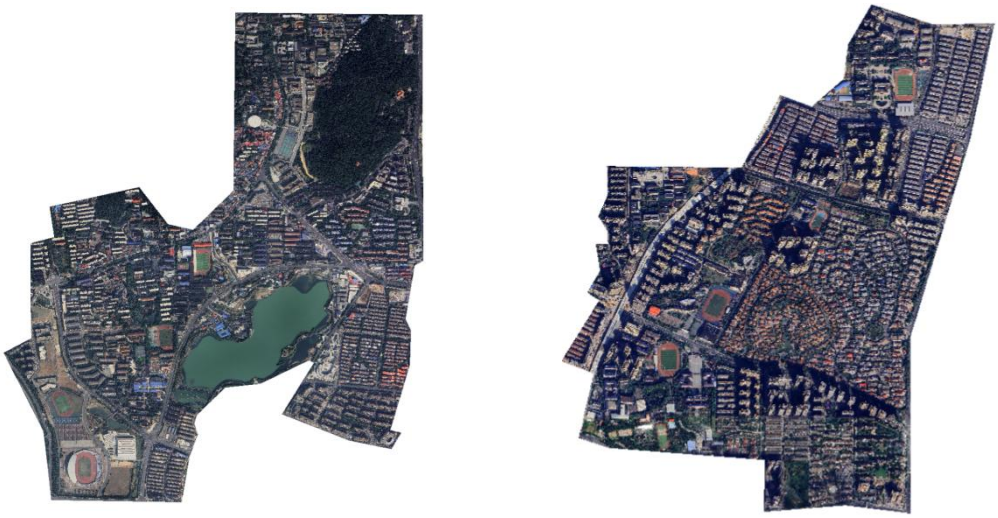
3. 长沙高度数据集

长沙遥感数据集采用 Google Earth 提供的高分辨率卫星影像作为数据源，影像为 RGB 三波段。标签生成方面，为获取高精度的建筑高度真值，本文引入了无人机倾斜摄影生成的实景三维模型作为参考基准。具体做法为：首先将建筑轮廓矢量数据与实景三维模型进行空间配准，提取各建筑轮廓范围内对应的三维点云高度信息；随后，为消除噪点及屋顶结构复杂性带来的干扰，采用区域中位数策略，即取轮廓范围内所有高度值的中位数赋值给该建筑实例。基于上述流程，本阶段累计完成 3930 栋房屋的逐栋精细化高度标注工作。

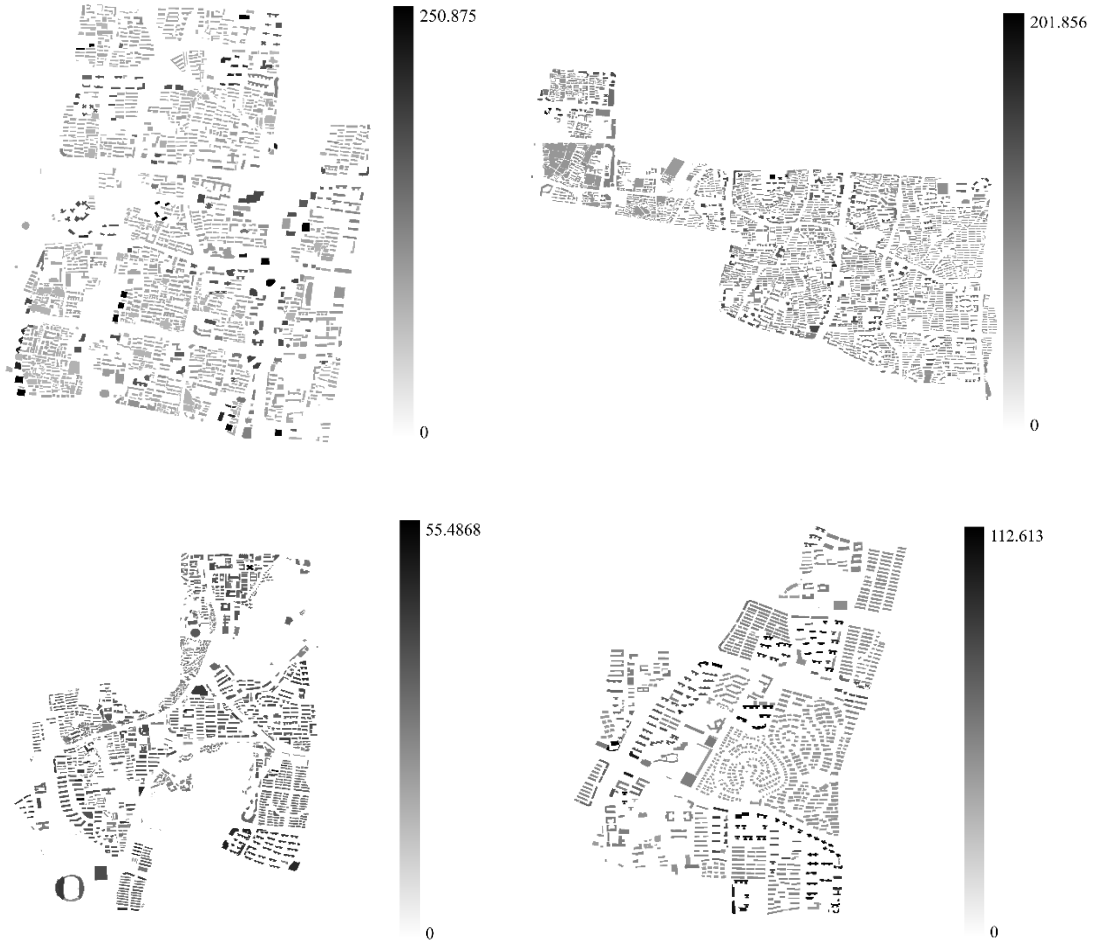
数据集的空间分布如图 4.5 所示。其中，图 4.5 a）展示了数据采集区域的空间范围；图 4.5 b）和图 4.5 c）分别展示了光学影像及其对应的高度标签（单位 m）可视化结果，建筑高度范围为 0-250m。为适配模型的输入要求，原始大幅影像采用滑动窗口方式裁剪为 512×512 像素的图像块。值得说明的是，由于本数据集仅包含建筑轮廓区域内的高度信息，属于稀疏高度监督形式，而非全场景逐像素深度标注，因此在数据清洗阶段剔除了不包含任何建筑目标的纯背景图像块，以避免大量无效样本对模型训练造成干扰。经筛选后，最终获得有效样本共 964 张。数据集按照 8:2 的比例随机划分为训练集与验证集，其中训练集包含 771 张样本，验证集包含 193 张样本。



a）数据集覆盖范围示意图



b) 光学影像



c）建筑高度标签图

图 4.5 长沙建筑高度数据集

4.3.2 实验参数设置

实验仅为对比实验，所有实验均在相同的软硬件环境下完成。具体实验条件

和参数设置如下：

所有模型均基于深度学习框架 PyTorch，在配有 NVIDIA GeForce RTX 4090 显卡的 Ubuntu20.04 操作系统上进行训练。模型的具体实现基于 1.22 版本的 mmsegmentation 代码库。在数据增广方面，本文针对建筑高度估计任务的特殊性，设计了多层次的数据增广策略。几何增广层面，在保持随机水平与垂直翻转操作的基础上，采用随机尺度缩放与裁剪相结合的方式，其中随机缩放的基准尺度设为 768×768，缩放比例范围为 0.8-1.5，随后统一裁剪为 512×512 的固定输入尺寸，从而在保证输入分辨率一致性的同时增强模型对多尺度建筑目标的适应能力。此外，引入了面向高度估计任务的 RandomDepthMix 策略，应用概率设置为 0.25，用于在训练阶段随机混合不同样本的高度信息，以增强模型对高度分布变化的鲁棒性。在光谱与亮度层面，引入基于 Albumentations 库的在线数据增广方法，包括随机亮度与对比度调整、随机伽马变换以及色调-饱和度-明度（HSV）扰动，以提升模型在不同成像条件下的泛化性能。

采用初始学习率为 1×10^{-4} 的 AdamW 优化器实现模型的收敛，参数β为[0.9，0.999]，权重衰减为 0.01，训练过程中 batch size 大小为 16，对于 Potsdam 数据集迭代训练次数为 80000 次，对于 Vaihingen 数据集和长沙高度数据集迭代训练次数为 10000 次。学习率调度策略采用线性预热与多项式衰减相结合的方式：在训练初始阶段，学习率通过线性策略从极小值逐步提升至设定的初始学习率，以缓解训练初期的不稳定问题；随后在预热结束后，采用多项式衰减策略逐步降低学习率直至训练结束，从而保证模型在后期阶段的稳定收敛。

在高度数据处理方面，针对建筑高度数值分布跨度较大的问题，训练与评估阶段均引入了显式的高度取值范围约束。具体而言，通过设置最小与最大有效高度阈值对参与计算的高度样本进行筛选，其中最小高度阈值设为 0.01m，Potsdam 数据集最大高度阈值设为 26m，Vaihingen 数据集最大高度阈值设为 20m，长沙数据集最大高度阈值设为 250.0m。该策略能够有效抑制异常高度值对模型训练与评价结果的干扰，同时为高度归一化处理提供明确的数值边界，有助于提升模型在不同高度区间内的预测稳定性。

4.3.3 ISPRS Potsdam 数据集实验结果分析

本节在 ISPRS Vaihingen 数据集上开展了系统的对比实验。该数据集包含丰富的城市场景信息，建筑物类型多样、分布密集，能够较为全面地检验模型在不同复杂度场景下的高度估计性能。各网络的训练结果见表 4.1。

由表 4.1 实验结果可以看出，不同模型在建筑物高度估计精度方面呈现出显著差异。Segmenter-s 在所有评价指标上均表现最差，表明该模型在缺乏有效局部特征提取机制的情况下，难以充分学习遥感影像中建筑物高度的细粒度回归信息，

预测结果与真实值之间存在较大偏差。FCN-r50 作为经典的全卷积网络，凭借 ResNet-50 骨干网络较强的局部特征提取能力，略高于其他对比模型。然而，其 Abs Rel 为 0.434，δ_1仅为 0.557，表明在高精度要求的逐像素预测中，纯卷积架构对建筑物细节高度信息的捕捉能力仍存在局限。Segnext-b 通过引入多尺度卷积注意力机制，在δ_1指标上达到了 0.571，优于 FCN-r50，但其 RMSE 和 $RMSE_{log}$ 反而有所上升，说明该模型虽然在精细高度感知方面有所改善，但整体回归精度的稳定性不足。Segman-t 的各项指标与 Segnext-b 较为接近，整体性能处于中游水平，未能体现出明显的优势。

表 4.1 不同网络与 SPR-NET 在 ISPRS Potsdam 数据集上的对比

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segmenter-s	0.488	3.375	0.623	0.475	0.682	0.780
FCN-r50	0.434	2.641	0.468	0.557	0.763	0.861
Segnext-b	0.457	2.776	0.500	0.571	0.748	0.837
Segman-t	0.472	2.779	0.518	0.532	0.734	0.831
Segformer-H-b0	**0.400**	**2.586**	**0.462**	**0.591**	**0.774**	**0.862**

本文提出的 Segformer-H-b0 在 Potsdam 数据集上取得了综合最优的性能表现。在误差类指标方面，Segformer-H-b0 的 Abs Rel 为 0.400，较次优的 FCN-r50 降低了 7.8%；RMSE 为 2.586，较 FCN-r50 进一步降低了 2.1%；$RMSE_{log}$ 为 0.462，达到了所有模型中的最低值。在阈值精度指标方面，δ_1为 0.591，较 Segnext-b 提升了 3.5%，较 Segmenter-s 大幅提升了 24.4%；δ_2为 0.774，同样位列所有模型之首。

量化的性能指标仅能提供宏观的评估，为了更直观地理解各模型在实际预测中的行为差异和具体优劣，本文进一步对测试集中的典型样本进行了可视化分析。如图 4.6 所示，展示了五个具有代表性的城市场景的建筑物高度预测热力图。在这些热力图中，模型的预测结果以具体的浮点数表示，并被映射为连续的颜色梯度：颜色越明亮、越偏向黄色，表示预测的建筑高度越大；颜色越暗沉、越偏向黑色，则表示预测的建筑高度越低。此外，每个预测图的右上角标注了该场景对应的 RMSE 值，从而为深入剖析各模型的建筑高度提取能力提供了直观的视觉证据。

由图 4.6 可视化结果可知，Segmenter-s 的预测结果整体误差远高于其他模型，这与其在量化指标上的末位表现完全吻合。FCN-r50 的预测热力图相较于 Segmenter-s 有了明显改善，但在建筑物边界处仍存在较为明显的模糊现象，且部分区域出现高度欠估。Segnext-b 和 Segman-t 的预测结果在视觉上较 FCN-r50 更为清晰，建筑物边界的刻画能力有所增强，但在第三个场景中，Segnext-b 的 RMSE 为 3.57，Segman-t 的预测也出现了局部高度值跳变的现象，说明这些模型在应对

特定复杂场景时的鲁棒性仍有不足。

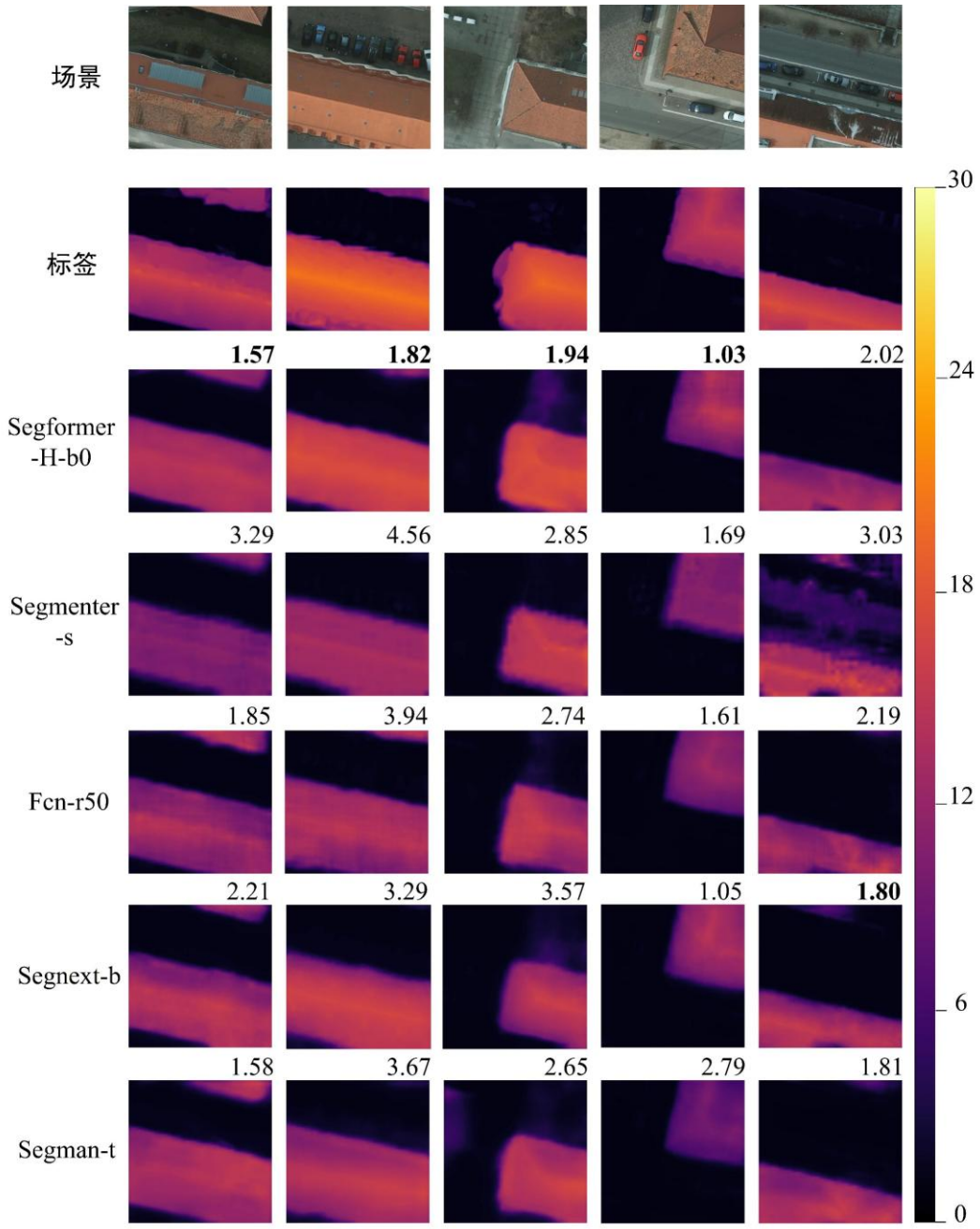


图 4.6 不同网络在 Potsdam 数据集上的建筑高度预测结果

本文提出的 Segformer-H-b0 在前四个场景中均取得了最低的 RMSE 值，分别为 1.57、1.82、1.94 和 1.03，预测热力图与标签之间的吻合度最高。从视觉效果来看，Segformer-H-b0 不仅能够准确还原建筑物的整体高度分布，还在建筑物边缘轮廓的锐利度、相邻不同高度建筑之间的过渡区域以及复杂屋顶结构的高度细节等方面均展现出明显优势。特别是在第四个场景中，该场景包含了建筑物与道路交界的复杂区域，Segformer-H-b0 以 1.03 的 RMSE 值大幅领先于其他模型，充分体现了其在精细化高度估计方面的卓越能力。

然而，在第五个场景中，Segformer-H-b0 的 RMSE 为 2.02，略高于 Segnext-

b（1.80）和 Segman-t（1.81），未能取得最优表现。通过对该场景的原始影像进行深入分析可以发现，第五个场景中建筑物屋顶约有一半区域处于阴影遮蔽状态，光照条件与非阴影区域存在显著差异。阴影区域的像素值整体偏暗，纹理信息和色彩对比度大幅降低，这对依赖光谱与纹理特征进行高度推理的网络模型构成了较大挑战。

Segformer-H-b0 在网络设计中融合了多层次的特征交互与增强模块，其通过丰富的上下文语义信息和多尺度特征融合来实现精细化的高度估计。然而，当建筑物屋顶存在大面积阴影时，阴影区域与非阴影区域之间的特征分布产生了显著不一致，多层次特征融合策略在尝试统一建模这两个特征差异显著的区域时，可能引入了一定程度的特征干扰，导致阴影区域与非阴影区域的高度预测一致性受到影响，从而使整体误差略有增加。

相比之下，Segnext-b 所采用的多尺度卷积注意力机制具有较强的局部特征聚合能力，从而在该特定场景下保持了较好的高度估计稳定性；Segman-t 同样在应对这类阴影干扰时表现出一定的鲁棒性，使其在该场景下取得了与 Segnext-b 相当的精度。但需要强调的是，即便在这一受阴影严重影响的非最优场景中，Segformer-H-b0 的 RMSE 仍显著优于 Segmenter-s 和 FCN-r50，且与最优值之间的差距仅为 0.22，在实际应用中属于可接受的波动范围。综合五个场景的整体表现来看，Segformer-H-b0 的平均 RMSE 为 1.68，远低于 Segnext-b 的 2.38、Segman-t 的 2.50、FCN-r50 的 2.47 和 Segmenter-s 的 3.08，进一步印证了其在建筑物高度估计任务中的综合优越性。

综上所述，无论是从量化评价指标还是可视化预测结果来看，Segformer-H 在 Potsdam 数据集上均展现出最优的综合性能。

4.3.4 ISPRS Vaihingen 数据集实验结果分析

为了进一步验证所提出的 Segformer-H 网络在建筑物高度估计任务中的有效性与泛化能力，本节在 ISPRS Vaihingen 数据集上开展了系统的对比实验。各网络的训练结果见表 4.2。

表 4.2 不同网络与 SPR-NET 在 ISPRS Vaihingen 数据集上的对比

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segmenter-s	0.813	3.336	0.579	0.336	0.601	0.761
FCN-r50	0.620	2.812	0.500	0.425	0.706	0.852
Segnext-b	0.569	2.537	0.459	0.444	0.746	0.891
Segman-t	0.564	2.556	0.465	0.462	0.736	0.892
Segformer-H-b0	**0.508**	**2.354**	**0.435**	**0.478**	**0.785**	**0.915**

由表 4.2 实验结果可以看出，不同模型在建筑物高度估计精度方面呈现出显

著差异。Segmenter-s 在所有评价指标上均表现最差，符合上述 Potsdam 数据集结果，预测结果与真实值之间存在较大偏差。FCN-r50 相较于 Segmenter-s 有了显著提升，Abs Rel 降低至 0.620，RMSE 降至 2.812，δ_1提升至 0.425，这得益于其基于 ResNet-50 骨干网络较强的局部特征提取能力。然而，FCN 解码器结构相对简单，缺乏对多尺度上下文信息的有效聚合机制，使其在精细化高度估计方面仍有不足。

Segnext-b 和 Segman-t，在精度方面取得了较为接近的表现。Segnext-b 的 Abs Rel 为 0.569，RMSE 为 2.537，而 Segman-t 的 Abs Rel 略低至 0.564，RMSE 为 2.556。二者在δ_3阈值精度上几乎持平，表明它们在整体高度趋势的捕捉上具备较好的能力。值得注意的是，Segnext-b 在 RMSE 指标上略优于 Segman-t，而 Segman-t 则在δ_1指标上占优，说明二者在不同误差尺度上各有侧重：Segnext-b 整体误差控制更佳，而 Segman-t 在高精度匹配比例上表现更好。

本文提出的 Segformer-H-b0 在所有评价指标上均取得了最优结果。与排名第二的模型相比，Segformer-H-b0 将 Abs Rel 从 0.564 进一步降低至 0.508，降幅达 9.9%；RMSE 从 2.537 降至 2.354，降低了 7.2%；$RMSE_{log}$ 从 0.459 降至 0.435，降低了 5.2%。在阈值精度方面，δ_1从 0.462 提升至 0.478，δ_2从 0.746 提升至 0.785，δ_3从 0.891 提升至 0.915，三项指标分别提升了 3.5%、5.2%和 2.7%。这些全面性的提升充分表明，Segformer-H-b0 通过引入的改进模块，有效增强了网络对建筑物高度信息的多尺度感知与精细化回归能力，在保持轻量化架构优势的同时，实现了优越的高度估计精度。

进一步对测试集中的典型样本进行了可视化分析。如图 4.7 所示，展示了五个具有代表性的城市场景的建筑物高度预测热力图。需要说明的是，真实标签中除建筑物以外的背景区域（地面、植被等）均设置为无效值，在图中以白色显示，以便更清晰地对比各模型对建筑物与非建筑物区域的区分效果。

由图 4.7 可视化结果可知，Segmenter-s 的预测结果整体呈现为近乎均一的低值色调，几乎无法辨识出建筑物的轮廓与高度分布，五个场景的 RMSE 值均在 5.0 以上，这与其在量化指标上的末位表现一致，进一步证实了该模型在建筑物高度估计任务中的严重不足。FCN-r50 的预测热力图虽然开始呈现出建筑物的大致形态，但高度估计结果较为粗糙，尤其在第一个场景中 RMSE 高达 4.66，表明其对复杂排列的建筑物群高度估计能力较弱。Segnext-b 和 Segman-t 的预测结果有了质的飞跃，建筑物的几何轮廓更加清晰，高度分布与真实标签的一致性显著提升。Segnext-b 在五个场景中的整体表现较为稳定。Segman-t 在第一个场景中 RMSE 为 5.23，明显高于其在其他场景中的表现，观察其预测热力图可以发现，该模型在建筑物与周围地面的区分上出现了明显错误，将部分地面区域错误地赋予了较高的高度值，导致整体误差显著增大，表明该模型在建筑物边界判别与前景-背景分

离方面的鲁棒性有所不足。

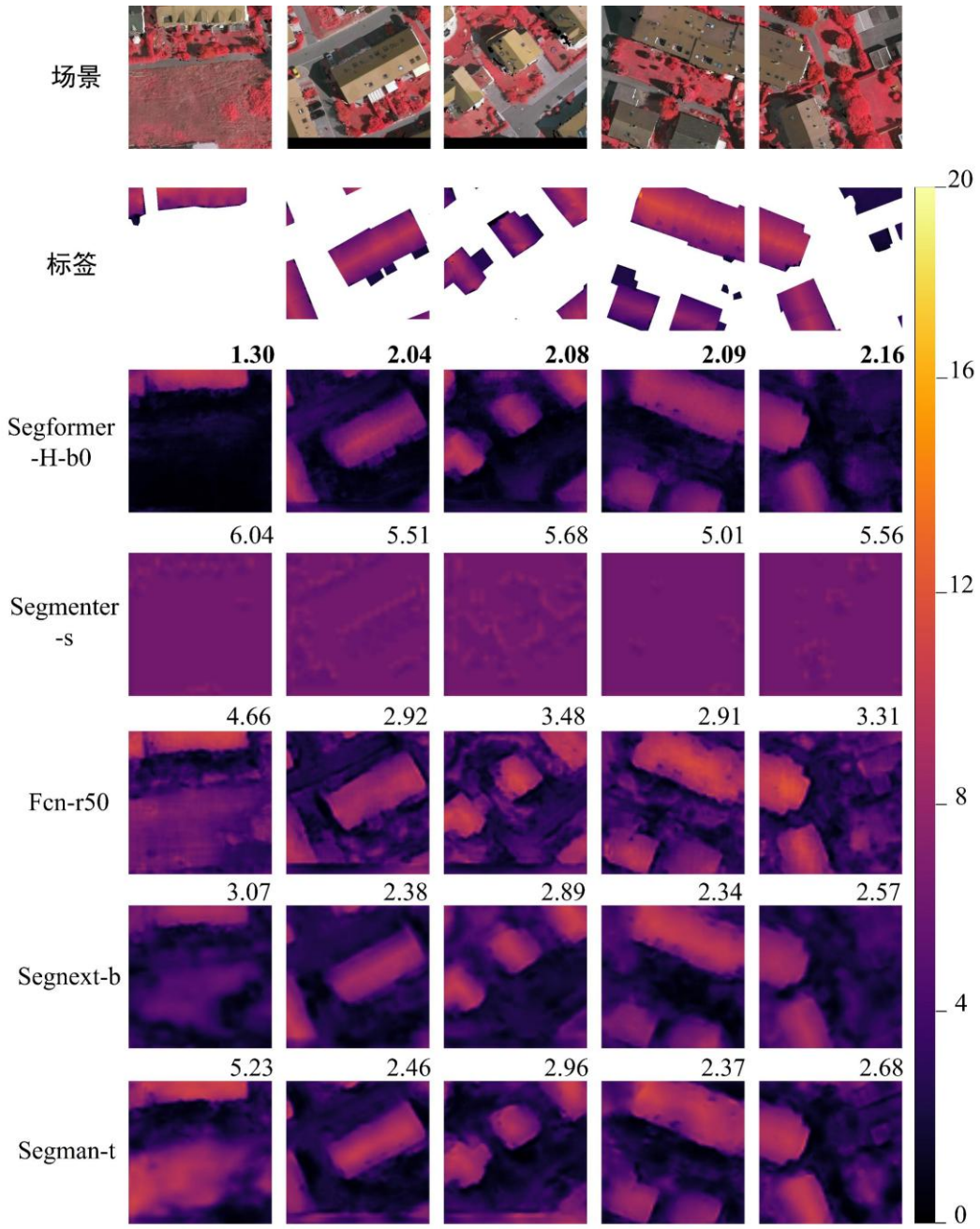


图 4.7 不同网络在 Vaihingen 数据集上的建筑高度预测结果

需要特别指出的是，ISPRS Vaihingen 数据集并非传统的红-绿-蓝（RGB）成像，而是采用近红外-红-绿（NIR-R-G）波段组合。这一特殊设定使得植被在图像中呈现高亮的红色，其光谱响应与部分建筑物屋顶较为接近，同时削弱了建筑物与裸露地面之间的光谱差异，极大增加了模型对建筑物与背景，尤其是与地面进行精准区分的难度。正因如此，上述对比模型均在不同程度上出现了将地面、道路或植被误判为建筑物的现象，热力图中可见明显的“伪高度”区域。

相比之下，本文提出的 Segformer-H-b0 在五个场景中均展现出最优的高度估计性能，不仅绝对数值最低，而且各场景间的波动最小，体现出卓越的稳定性与

鲁棒性。从预测热力图来看，Segformer-H-b0 的结果在建筑物边缘的刻画上更加锐利清晰，建筑物内部的高度分布更加均匀平滑，与真实标签的色彩梯度高度吻合。即使在 NIR-R-G 波段带来的强干扰条件下，Segformer-H-b0 仍能精准地将建筑物与地面完全分离，地面区域的预测值接近于零，几乎没有出现其他模型中普遍存在的虚假高度响应，建筑物区域的高度响应则集中、准确，前景与背景的对比度极为清晰。特别值得关注的是，在第一个场景中，Segformer-H-b0 的 RMSE 仅为 1.30，远低于其他模型。该场景中建筑物与地面交错分布，且由于 NIR 波段影响，地面与建筑屋顶的光谱差异较小，FCN-r50、Segnext-b 和 Segman-t 均在地面区域产生了不同程度的高度误判，而 Segformer-H-b0 能够有效抑制背景干扰，地面部分几乎不产生任何伪高度，建筑物部分的高度估计也与真实标签高度一致，充分说明了本文所提改进方法在增强网络对建筑物与非建筑物区域判别能力方面的显著优势。此外，在第四和第五个场景中，Segformer-H-b0 对不同尺度、不同朝向的建筑物均能给出准确的高度估计，同时保持地面区域的低值响应，进一步验证了其多尺度特征融合机制在提升空间辨别力与高度回归精度方面的有效性。

综上所述，无论是从量化指标还是可视化分析来看，本文提出的 Segformer-H 在 Vaihingen 数据集上均展现出了优于对比方法的建筑物高度估计能力，尤其在建筑物与地面的精准区分、预测精度、边界清晰度和场景适应性方面表现突出，验证了所提出网络改进策略的有效性。

4.3.5 长沙建筑高度数据集实验结果分析

为了进一步验证所提出的 Segformer-H 网络在建筑物高度估计任务中的有效性与泛化能力，本节在自制的长沙建筑高度数据集上开展了系统的对比实验。该数据集包含更加丰富的城市场景信息，建筑物类型多样、分布密集，且存在大量高层建筑与密集低矮建筑交错分布的复杂区域，能够较为全面地检验模型在不同复杂度场景下的高度估计性能。各网络的实验结果见表 4.3。

表 4.3 不同网络与 Segformer-H 在长沙建筑高度提取数据集上的对比

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segmenter-s	0.498	18.160	0.514	0.329	0.586	0.740
FCN-r50	0.393	12.962	0.386	0.470	0.725	0.881
Segnext-b	0.448	13.450	0.404	0.469	0.724	0.872
Segman-t	0.421	12.774	0.397	0.470	0.700	0.880
Segformer-H-b0	**0.332**	**11.004**	**0.343**	**0.555**	**0.780**	**0.912**

从整体结果来看，本文所提出的 Segformer-H-b0 在长沙数据集上的所有评价指标中均取得了最优表现，充分验证了该网络架构在复杂城市场景下的高度估计能力。具体而言，Segformer-H-b0 的 Abs Rel 为 0.332，RMSE 为 11.004m，$RMSE_{log}$

为 0.343，相较于次优模型 FCN-r50，分别降低了 15.5%、15.1%和 11.1%。在阈值准确率指标上，Segformer-H-b0 的δ_1、δ_2、δ_3分别达到了 0.555、0.780 和 0.912，相较于 FCN-r50 分别提升了 18.1%、7.6%和 3.5%。上述结果表明，Segformer-H-b0 不仅在整体误差控制方面具有显著优势，能够更准确地逼近建筑物的真实高度值。

在对比网络中，与前面的数据集结果一致，Segmenter-s 的表现最为不理想，各项指标均大幅落后于其他模型。Segnext-b 作为融合了注意力机制的改进型网络，虽然在δ_1和δ_3等阈值指标上优于 Segmenter-s，但其 Abs Rel 和 RMSE 仍然偏高，说明该网络在全局上下文建模与局部精细特征提取之间的平衡仍有不足。

值得关注的是，FCN-r50 和 Segman-t 在长沙数据集上展现了较为稳健的性能。FCN-r50 凭借 ResNet-50 骨干网络强大的特征提取能力，在 $RMSE_{log}$、δ_2和δ_3等指标上表现出色，表明经典 CNN 架构在处理建筑高度估计任务时仍具有良好的基线性能。Segman-t 则以 12.774m 的 RMSE 略优于 FCN-r50，体现了其在中等尺度建筑高度估计上的竞争力。然而，这两个模型在δ_1指标上均为 0.470，与 Segformer-H-b0 的 0.555 存在明显差距，反映出它们在精确匹配建筑真实高度方面仍存在较大的提升空间。

综合分析可以发现，长沙数据集相较于其他数据集具有更大的场景复杂度和建筑高度分布差异性，这对模型的多尺度特征融合能力和上下文理解能力提出了更高要求。Segformer-H-b0 通过引入的改进模块，有效增强了网络对不同尺度建筑物特征的感知与融合能力，使其在高层建筑密集区域和低矮建筑混合区域均能保持较高的估计精度。特别是在δ_1指标上，Segformer-H-b0 以 0.555 的得分大幅领先所有对比网络，这说明该模型在严格的阈值约束下依然能够对超过半数的建筑物实现高精度的高度估计，充分体现了所提方法在实际应用场景中的可靠性与鲁棒性。

尽管本模型的 RMSE（11.004m）在数值上显得较大，但这主要归因于数据集内巨大的高度异质性。文献[70]指出由于中国城市建筑高度分布差异显著，即便是当前最先进的模型（SOTA）在中国地区的预测 RMSE 也达到了 10.318m。本数据集具有相似的分布特征，建筑高度跨度极大，从约 9m 的低层建筑到 250m 左右的超高层建筑，这种巨大的量级差异导致了均方根误差的自然升高。

总体而言，长沙数据集上的实验结果与前述数据集的结论保持一致，进一步证实了 Segformer-H 网络在建筑物高度估计任务上的有效性和良好的泛化能力。即使在建筑类型更加多样、空间分布更加复杂的城市场景中，Segformer-H-b0 依然能够以较低的模型复杂度实现最优的估计精度，为遥感影像建筑物高度提取提供了一种高效且实用的解决方案。

为了进一步验证 Segformer-H 网络的跨区域泛化能力与实际应用效果，本文

选取了与训练集地理位置相距较远的区域作为独立测试集，如图 4.5 a）所示。在该测试区域中，进一步筛选出四个具有代表性的特征区域（A、B、C、D），如图 4.8 所示，分别涵盖低层建筑群、高层建筑群以及高低层混合建筑群三种典型城市场景，用以更加清晰地区分不同类型建筑群的高度估计误差特征。该测试区域内建筑高度分布跨度较大，最高建筑高度达到 86.2m，涵盖了从低层住宅到中高层建筑的多种类型，能够较为全面地检验模型在未见区域上的预测能力。

各区域的具体量化评价指标见表 4.4。

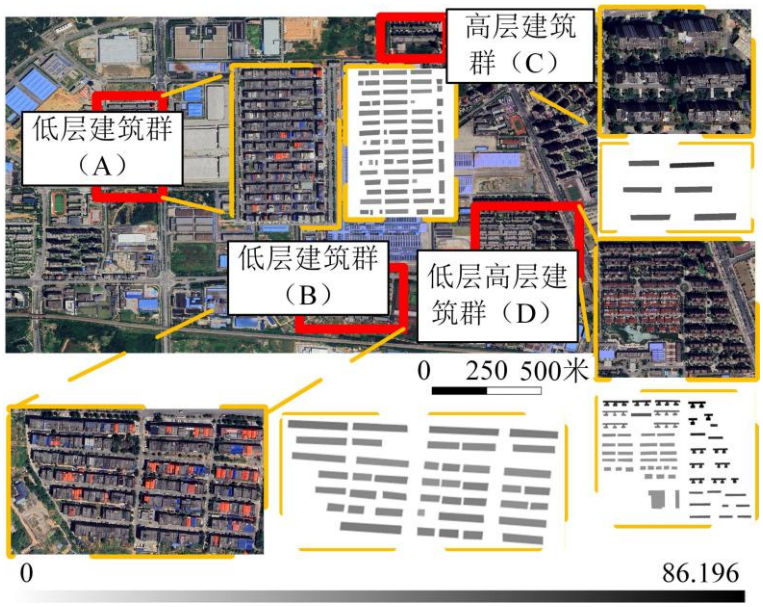


图 4.8 测试集房屋以及其对应标签

表 4.4 测试区域综合指标结果

类别	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
A	0.159	4.054	0.213	0.837	0.944	0.981
B	0.223	4.674	0.247	0.642	0.897	0.999
C	0.345	15.488	0.317	0.149	1.000	1.000
D	0.275	11.525	0.295	0.452	0.902	0.986
合计	0.255	8.935	0.268	0.520	0.935	0.991

结合表 4.4 的统计数据，本文对不同场景下的模型性能进行了深入剖析：

（1）低层建筑群表现优异：在低层建筑群（A 区域）中，模型取得了所有区域中最优的精度表现。Abs Rel 仅为 0.159，RMSE 为 4.054m，δ_1高达 0.837，表明在常规高度范围内，模型具有极高的预测置信度，绝大多数建筑物的高度预测误差被控制在极小范围内。低层建筑群（B 区域）同样表现良好，RMSE 为 4.674m，δ_1达到 0.642，δ_3接近 1.0，说明模型对低层建筑的高度估计整体稳健可靠。A 区域与 B 区域之间的性能差异主要源于建筑密度与布局复杂度的不同，B 区域建筑

排列更为紧密，相互遮挡与阴影干扰更为显著，导致精度略有下降。

（2）高层建筑的尺度效应：在高层建筑群（C 区域）中，RMSE 升高至 15.488m，δ_1降至 0.149。这一现象符合深度学习回归任务中的尺度规律，即绝对误差往往随目标值的增大而同比增加。然而，值得注意的是，该区域的δ_2和δ_3均达到了 1.000，这说明尽管高层建筑的精确预测难度较大，但模型并未产生严重的预测失效，所有预测结果均稳定落在宽松阈值范围内，体现了模型对建筑高度量级判断的正确性与稳定性。此外，计算该区域的归一化相对均方根误差$RMSE_{rel}$为 0.180，相较于其绝对 RMSE 数值而言，相对误差仍处于可接受范围内，进一步佐证了误差主要源于高度量级本身而非模型的系统性偏差。

（3）混合建筑群的整体泛化性能：高低层混合建筑群（D 区域）作为最接近真实城市形态的样本，其δ_2达到 0.902，RMSE 为 11.525m，证明模型能够有效应对建筑高度分布的复杂异质性。尽管该区域同时包含低矮建筑与中高层建筑，高度跨度较大，但模型仍保持了较为均衡的预测表现，δ_3高达 0.986，表明几乎所有建筑物的预测高度均落在合理的误差区间内。

（4）整体泛化测试结果：从合计指标来看，模型在跨区域泛化测试中整体取得了 RMSE8.935m 的预测精度，δ_2和δ_3分别达到 0.935 和 0.991。计算整体归一化相对均方根误差$RMSE_{rel}$为 0.102，表明在考虑测试区域高度范围后，模型的相对预测误差约为 10.4%，具有较好的实用价值。再次印证了模型在对数空间内对不同尺度建筑具有一致性处理能力，不会因建筑高度量级的变化而产生显著的性能退化。

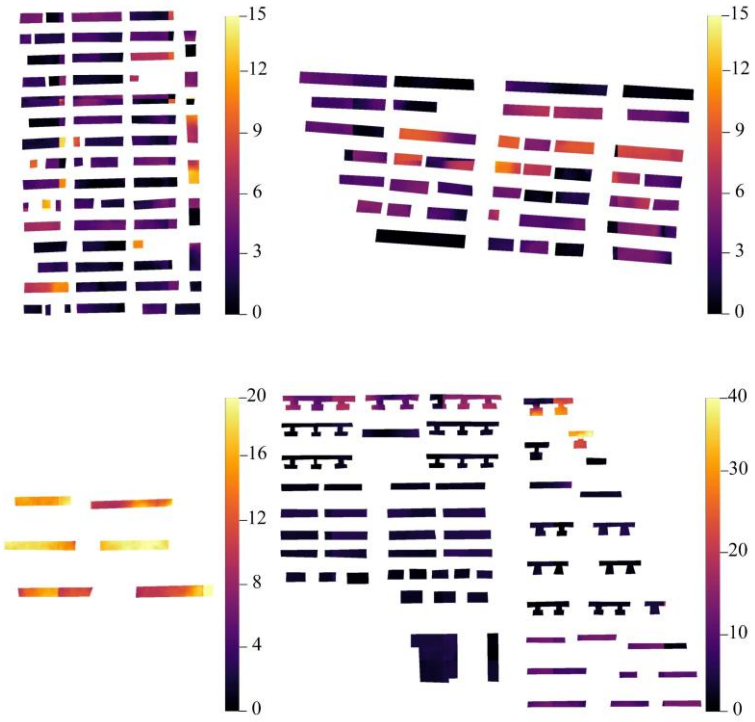


图 4.9 典型建筑群像素级高度预测误差空间分布

进一步地，本文对测试集中的典型样本进行了可视化分析，输出四个建筑群类别的预测绝对误差分布（即图中通过颜色梯度直观映射出单体建筑预测值与真实值之间的绝对误差大小），结果见图 4.9。

从图 4.9 可以直观地观察到，低层建筑群的误差分布较为均匀，绝大多数像素的预测误差集中在较低的数值范围内。混合建筑群的误差分布呈现出与建筑高度相关的梯度特征，低矮建筑的预测误差较小，而中高层建筑的误差相对较大，这与前述定量分析的结论相吻合。

上述结果是基于像素级的绝对预测精度进行分析的。然而，实际建筑屋顶表面的高度并不均匀，如存在坡屋顶、设备层、女儿墙等局部高度变化。为了更好地模拟实际应用场景，本文对预测结果进行了进一步后处理，即提取每个建筑轮廓内所有预测像素的中位数作为该建筑单体的最终估计高度，并与标签高度进行对比分析。结果见表 4.5，对应的误差可视化效果如图 4.10 所示。

表 4.5 测试区域中位数高度值综合指标结果

类别	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
A	0.149	3.742	0.201	0.870	0.945	0.978
B	0.233	4.786	0.251	0.613	0.899	1.000
C	0.356	15.057	0.310	0.161	1.000	1.000
D	0.278	11.781	0.294	0.418	0.911	0.988
合计	0.254	8.842	0.264	0.516	0.939	0.991

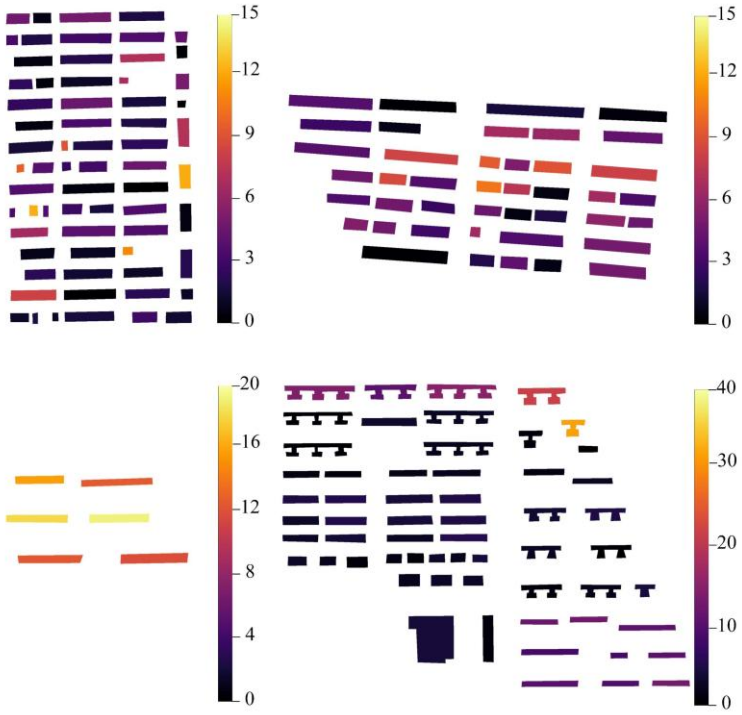


图 4.10 典型建筑群中位数高度预测误差空间分布

由表 4.5 和图 4.10 可知，对比表 4.4（像素级评估）与表 4.5（中位数评估）的数据，整体 RMSE 由 8.935m 降低至 8.842m，降低了 1.0%；δ_2由 0.935 提升至 0.939。这些变化表明，“轮廓内中位数”策略有效地过滤了单目视觉预测中不可避免的像素级噪声与建筑屋顶局部高度波动。

从各区域的分项指标来看，低层建筑群在采用中位数策略后，RMSE 由 4.054m 降低至 3.742m，δ_1由 0.837 提升至 0.870，进一步提升了预测的精确度与可靠性。高层建筑群的 RMSE 由 15.488m 降低至 15.057m，δ_1由 0.149 提升至 0.161，虽然改善幅度相对有限，但δ_2和δ_3均保持在 1.000 的完美水平，表明中位数策略在高层建筑中同样发挥了积极的平滑作用。

考虑到在构建三维城市模型、城市规划分析以及灾害风险评估等实际应用中，通常只需要为每个建筑单体提供一个统一的代表性高度值，而非离散的像素高度场，因此表 4.5 的结果不仅在数值上进一步验证了 Segformer-H 的估计精度，更从工程应用角度证明了本模型结合中位数后处理策略后，能够直接产出满足实际测绘与三维建模需求的高质量建筑高度数据，具有明确的落地应用价值。

此外，为进一步量化分析模型预测白模与真实模型之间的几何偏差，基于第三章获取的建筑轮廓预测结果，从轮廓长度与面积开展系统性对比，结果如表 4.6 所示。由表中数据可知，在四个测试区域内，长度差值整体控制在较低水平，反映出所提模型在大规模建筑场景下能够较为完整地还原建筑边界的整体几何形态。值得注意的是，C 区域因建筑规模相对较小，轮廓基数有限，致使相同绝对误差在比例尺度上被显著放大，其长度差值百分比达到 8.51%；然而该区域对应的面积差值百分比仅为 2.79%，表明尽管局部边界细节存在一定偏差，模型对该区域建筑覆盖范围的整体预测精度仍保持在可接受区间内。综合而言，长度差值、面积差值与 IoU 三类指标在评价结果上相互印证、彼此支撑，充分说明本文所构建的深度学习模型不仅在像素级评价指标上展现出良好的预测性能，同时在建筑轮廓的几何尺度还原方面亦具备较高的精度水平，能够有效满足后续白模生成及三维建模应用对轮廓精度的需求。

表 4.6 测试区域轮廓综合指标结果

类别	轮廓总长度（m）	预测总长度（m）	长度差值（m）	长度差值百分比（%）	面积差值百分比（%）	IoU（%）
A	7069	7243	174	2.46	8.02	0.79
B	5363	5519	156	2.91	8.66	0.76
C	1034	1122	88	8.51	2.79	0.87
D	8749	9034	285	3.26	7.23	0.83
合计	5554	5730	176	3.17	6.68	0.81

4.3.6 消融实验

为了验证 Segformer-H 中 DySample 模块的有效性，基于 Vaihingen 数据集、Potsdam 数据集和长沙建筑高度数据集设计了消融实验。采用 Abs Rel、RMSE、$RMSE_{log}$ 以及三种阈值精度指标δ_1、δ_2和δ_3作为评估指标。实验结果记录在表 4.7、表 4.8 与表 4.9 中。

表 4.7 在 Potsdam 数据集上的消融实验

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segformer	0.512	2.711	0.483	0.499	0.677	0.803
Segformer-H	**0.400**	**2.586**	**0.462**	**0.591**	**0.774**	**0.862**

表 4.8 在 Vaihingen 数据集上的消融实验

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segformer	0.563	2.465	0.459	0.456	0.764	0.893
Segformer-H	**0.508**	**2.354**	**0.435**	**0.478**	**0.785**	**0.915**

表 4.9 在长沙建筑高度数据集上的消融实验

	Abs Rel	RMSE	$RMSE_{log}$	δ_1	δ_2	δ_3
Segformer	0.375	12.528	0.386	0.507	0.752	0.865
Segformer-H	**0.332**	**11.004**	**0.343**	**0.555**	**0.780**	**0.912**

结果表明，在三个数据集上引入 DySample 动态上采样模块后，Segformer-H 相较于基线模型 Segformer 在所有评估指标上均取得了一致性的提升，充分验证了该模块的有效性。DySample 动态上采样模块通过自适应地学习上采样核参数，相较于 Segformer 原始解码器中的固定插值上采样方式，能够更精细地恢复特征图的空间细节信息，从而在多尺度特征融合过程中有效减少高度信息的损失。验证了将 DySample 模块集成到 Segformer 解码器中用于建筑高度估计任务的合理性与有效性。

4.4 本章小结

本章针对遥感影像建筑物高度估计任务中存在的特征融合空间失真以及国内复杂场景泛化困难等问题，改进了 Segformer 网络模型并构建了长沙建筑高度数据集，系统地开展了实验验证与分析。本章的主要工作与结论如下：

（1）改进了 Segformer 网络架构（Segformer-H），解决了固定插值核在建筑物边界与高度突变区域引入模糊效应的问题。基于 DySample 的级联动态上采样模块替换了 Segformer 解码器中的双线性插值，实现了根据输入特征局部语义内容自适应计算采样位置，有效增强了模型对高度异质性区域的敏感度。

（2）构建了涵盖 9m 至 250m 大跨度高度范围的长沙建筑高度数据集，实现了对模型在复杂真实城市场景下跨区域泛化能力的有效验证。该数据集结合高分辨率卫星影像与无人机倾斜摄影实景三维模型，解决了现有公开数据集难以完全反映中国城市高低建筑落差大、分布密集现状的问题，为模型面向实际工程落地应用提供了有力的数据支撑。

（3）在三个数据集上开展了系统的对比实验，验证了模型的优越性。实验结果表明，Segformer-H 在所有评价指标上均取得了最优性能，全面优于 Segmenter-s、FCN-r50、Segnext-b 和 Segman-t 等对比方法。可视化分析进一步表明，Segformer-H 在建筑物边缘刻画的锐利度、建筑物与背景区域的精准区分以及不同复杂度场景下的预测稳定性方面均展现出明显优势。

第5章 长沙市中心城区建筑群三维模型构建

5.1 引言

基于深度学习的建筑信息提取方法虽然在建筑轮廓识别和高度估计等单一任务上取得了显著进展，但现有研究仍存在以下不足：第一，缺乏面向城市级尺度的三维白模快速构建完整流程，尚未系统性地构建起从原始遥感影像到最终三维白模的完整技术链路；第二，受遥感影像中树木遮挡、建筑阴影以及地物混淆等因素干扰，预测轮廓存在边缘锯齿化、形状不规则等问题，影响三维模型的几何精度与视觉质量；第三，目前尚缺乏一套简洁、完整且可复现的自动化处理流程将各环节有机串联，制约了三维白模在实际工程中的规模化应用。针对上述问题，本章提出一套涵盖预测结果拼接融合、建筑轮廓几何正则化、高度属性赋值及矢量化拉伸建模的全流程自动化方法，实现从深度学习预测结果到城市级 LoD1（Level of Detail 1）三维白模的快速生成。LoD1 模型是忽略屋顶结构细节与墙面纹理信息，仅保留建筑物的二维轮廓与统一高度属性，以简洁的块状几何体形式呈现城市建筑群的空间分布与高度格局。本章首先阐述城市级三维白模构建的整体技术路线与核心方法，包括预测结果空间拼接、基于几何约束的轮廓正则化算法、中位数高度赋值策略以及矢量化拉伸建模流程；随后，在长沙市多类型城市功能区开展大范围实验验证，通过对高密度住宅区、复杂地形混合区、工商业建筑区以及农村稀疏建筑区等典型场景的三维重建结果分析，验证方法的有效性与鲁棒性。

5.2 城市级三维白模构建方法

在获取高精度建筑轮廓与高度数据的基础上，本节提出一套面向城市级尺度的 LoD1 三维白模自动化构建方法。整体技术路线如图 5.1 所示，主要包含两个环节：前处理步骤与后处理步骤。

其中前处理步骤分为：光学遥感影像的下载、裁剪与预测；后处理步骤分为：预测结果拼接、建筑轮廓正则化、高度中位数赋值以及栅格矢量化。

由深度学习模型输出的建筑轮廓与高度预测结果以图像块为单位生成，无法直接用于大范围连续场景的三维建模。此外，受限于遥感影像中树木遮挡、建筑阴影以及地物混淆等因素的干扰，预测轮廓往往存在边缘锯齿化、形状不规则等问题，高度预测值也可能包含像素级噪声与局部异常。这些问题若不加处理，将严重影响三维模型的几何精度与视觉表达质量。因此，需要设计系统性的后处理流程，对原始预测结果进行逐步优化，以满足城市级三维建模对拓扑规则性、属性一致性与空间连续性的要求。

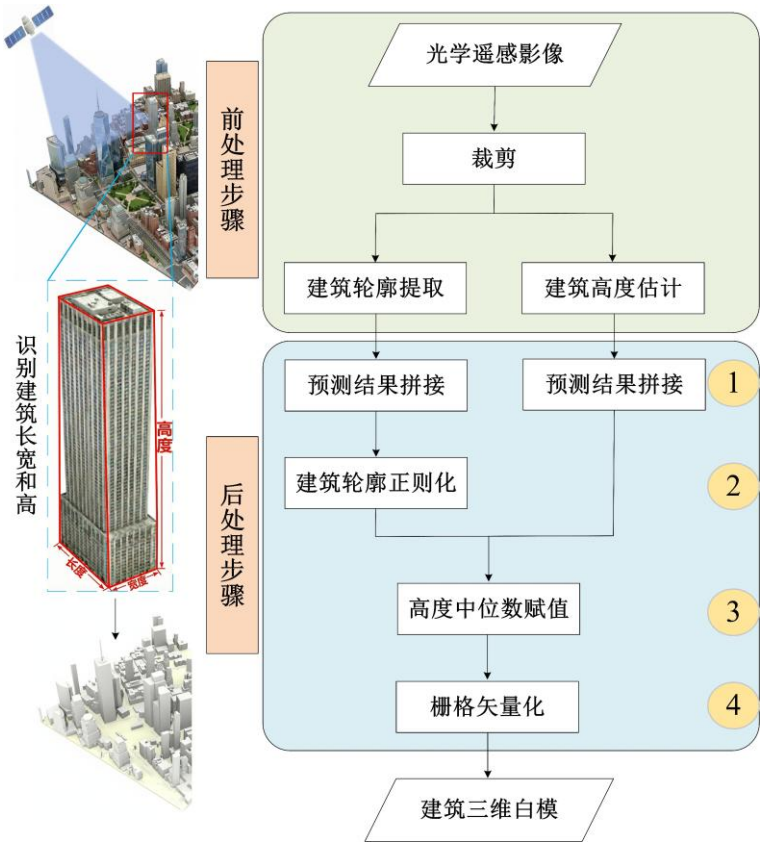


图 5.1 城市级建筑群三维白模建立流程图

1. 预测结果拼接

在利用第三章 SPR-Net 网络生成建筑轮廓图和第四章 Segformer-H 网络完成建筑高度估计后，需要对 512×512 像素的预测图像块进行空间拼接，以生成区域尺度的连续预测结果。拼接过程中，根据各图像块在原始影像中的地理坐标和空间索引关系，分别对轮廓预测图与高度预测图进行无缝镶嵌，最终形成覆盖整个研究区域的完整建筑轮廓图和建筑高度图。该步骤为后续的正则化处理与三维建模提供了数据基础，有效保证了空间信息的连续性与完整性。

2. 建筑轮廓正则化

针对深度学习预测轮廓中普遍存在的边缘锯齿化与形状不规则问题，本文引入基于几何约束的建筑轮廓正则化方法[117]，对建筑边界进行规则化处理。该方法旨在消除轮廓噪声，恢复建筑物的直角特征与规则几何形态，从而提升三维重建的视觉效果与模型精度。具体包含轮廓提取与简化、主方向判定以及几何正交校正三个阶段。

首先是轮廓提取与简化，对二值化的建筑预测图进行连通域分析，提取每个建筑物的原始轮廓点集$P=\{p1，p2，\cdots，pn\}$。由于原始轮廓由像素边缘组成，包含大量冗余顶点，不仅增加后续计算开销，还会引入不必要的几何噪声。为此，采用 Ramer–Douglas–Peucker（RDP）算法进行多边形简化。给定距离阈值ε，RDP 算法通过递归方式剔除到弦距离小于ε的非关键点，从而在保持多边形主要几何特

征的同时显著减少顶点数量，为后续正交化处理提供简洁而有效的轮廓表达。

然后进行主方向判定。绝大多数建筑物具有规则的几何形态，其边缘通常相互垂直或平行。为了恢复这一特征，需要首先确定建筑物的主方向。对于简化后的轮廓点集，计算每条边向量$vi=p(i+1)-pi$，并进一步计算对应的边长L_i以及方位角θ_i。本文假设建筑物最长的边缘代表其主要朝向，因此选取长度最大的边所对应的方位角作为该建筑的主方向θ_{main}，其定义为：

$$\theta_{\min} = \theta_k \tag{5.1}$$

其中 k 为使L_i取得最大值的边索引（即：k 对应于 Li 最大的那一条边）。该主方向将作为后续正交化约束的基准参考。

最后是几何正交校正。基于确定的主方向，对所有边缘进行几何约束校正。算法流程如下：

第一步，方向对齐与边缘分类。遍历轮廓的每一条边，计算其方位角θ_i与主方向θ_{main}的偏差$\Delta\theta$。根据$\Delta\theta$将边缘分类为平行边或垂直边。具体而言，若偏差接近0°或180°，则约束为平行；若接近 90°或 270°，则约束为垂直。

第二步，坐标旋转与重构。为便于计算，将所有轮廓点绕中心点旋转，使得主方向与坐标轴对齐。

第三步，相邻边求交与修正。针对相邻的两条边E_i和E_{i+1}，根据其约束状态进行处理。当E_i和E_{i+1}相互垂直，计算两条直线的解析交点作为新的角点。若E_i和E_{i+1}被判定为同向（即由原来的非直线段被强制约束为平行线），则计算两直线间的距离d。若d小于设定阈值，则通过平移操作合并边缘；否则，插入连接线段以维持拓扑闭合。

第四步，逆变换。将修正后的顶点坐标旋转回原始地理坐标系，完成规则化。

经过上述处理，建筑轮廓的锯齿状边缘被有效消除，边界形态恢复为符合建筑实际几何特征的规则多边形，为高质量三维建模奠定了基础。

3. 高度中位数赋值

在获得正则化建筑轮廓后，需要为每个单体建筑赋予统一的高度属性。通过空间叠加分析，提取每个独立建筑轮廓范围内所有像素对应的高度预测值。考虑到原始像素级高度预测中可能存在由遮挡、阴影或模型不确定性引起的噪声与异常值，直接采用均值统计容易受极端值影响。因此，本文选用中位数统计法确定每个建筑对象的统一高度值，该统计量对离群值具有更强的鲁棒性，能够更稳定地反映建筑的实际高度水平。中位数统计法的有效性已在前述高度估计精度验证中得到证实。

4. 矢量化拉伸与 LoD1 建模

采用栅格转矢量算法（Raster-to-Vector Conversion）[118]，将带有高度属性的栅格轮廓转换为矢量多边形要素。在转换过程中，保持正则化后的规则几何边界特征不被破坏，并将中位数高度值作为属性字段赋予对应的矢量多边形。

最终，基于高度属性对每个矢量建筑多边形执行垂直拉伸操作，即沿竖直方向将二维平面多边形拉伸至对应高度，构建出覆盖研究区域的城市级 LoD1（Level of Detail 1）三维白模。根据 CityGML 标准[119]，LoD1 属于规则块体模型，通常以建筑物的二维轮廓为基础，按照统一或属性给定的高度进行拉伸，不包含屋顶结构细节、立面纹理及复杂构件，仅表达建筑物的整体体量和空间位置关系。其中，CityGML 是一种由国际标准化组织（OGC，开放地理空间联盟）制定的三维城市模型数据标准。该层级模型在计算效率与空间表达之间取得了良好平衡，适用于大尺度城市形态分析、城市规划辅助决策、日照模拟、风环境评估等多种应用场景。

5.3 城市级三维白模建立结果分析

5.3.1 研究区域与数据准备

本为验证上述方法在实际城市场景中的适用性与鲁棒性，本文选取长沙市中心城区开展大范围三维建模实验。选取的建模区域如图 5.2 所示，该区域涵盖了多种典型城市功能分区与地形条件，能够全面检验方法在不同场景下的适应能力与重建质量。

针对该研究区域，下载最新 Google 光学遥感影像，空间分辨率为 0.3m。将原始影像裁剪为512×512像素的标准图像块，共获得 11583 张样本，覆盖面积约 240km^2，涉及建筑物约 40223 栋，共耗时 3 小时。基于 SPR-Net 和 Segformer-H 网络分别完成建筑轮廓提取与建筑高度估计任务后，将各图像块的轮廓预测图与高度预测图进行空间拼接，生成区域尺度的连续预测结果，如图 5.3 所示。其中，图 5.3 a）为拼接后的建筑轮廓预测图，图 5.3 b）为对应的建筑高度预测图，二者共同构成后续三维建模的数据输入。

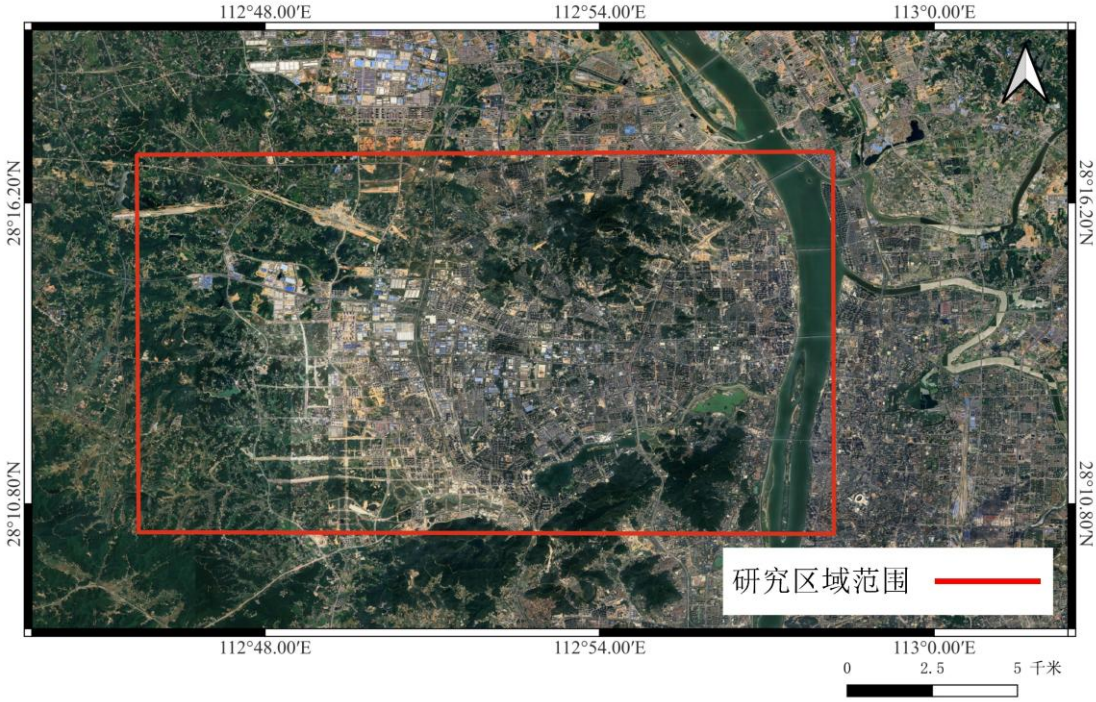
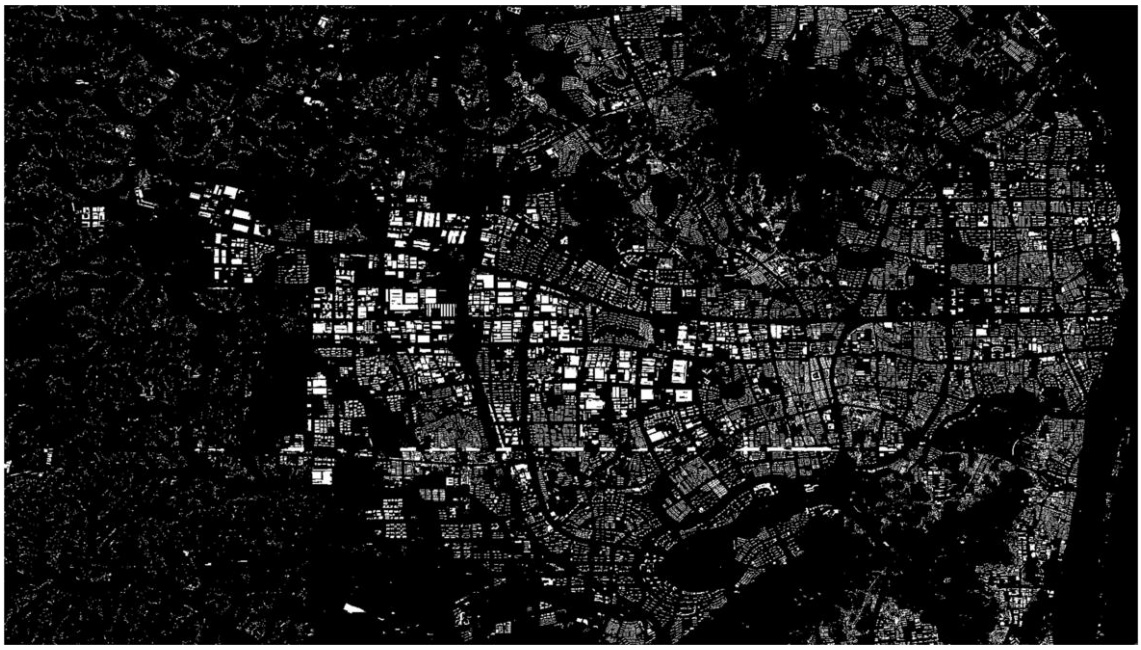


图 5.2 城市级建筑三维重建区域



a）建筑轮廓图

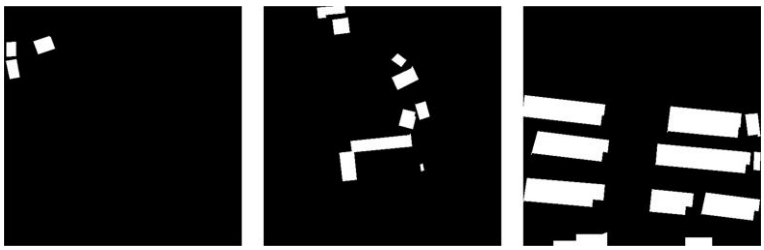
b）建筑高度图

图 5.3 应用区域内预测结果展示

5.3.2 后处理过程与中间结果

对拼接后的建筑轮廓图依次执行连通域提取、RDP 简化及正交化校正操作，正则化前后的对比结果如图 5.4 所示。可以观察到，原始预测轮廓（图 5.4 a）存在明显的边缘锯齿、弯折与不规则扰动，经正则化处理后（图 5.4 b），建筑边界恢复为棱角分明的规则多边形，直角特征得到有效还原，整体轮廓的几何表达质量显著提升。这一改善对于后续三维拉伸建模至关重要。

a）原始预测轮廓



b）正则化处理后建筑轮廓

图 5.4 建筑正则化校准前后对比图

在正则化轮廓的基础上，通过空间叠加分析提取各建筑轮廓范围内的像素高度值，并以中位数统计确定单体建筑的统一高度，生成的建筑中位数高度图如图5.5 所示。与原始像素级高度图相比，中位数高度图中每个建筑内部的高度值呈现一致性，有效消除了因边缘混合像素或预测不确定性造成的高度波动，为后续矢量化拉伸提供了稳定可靠的高度属性。

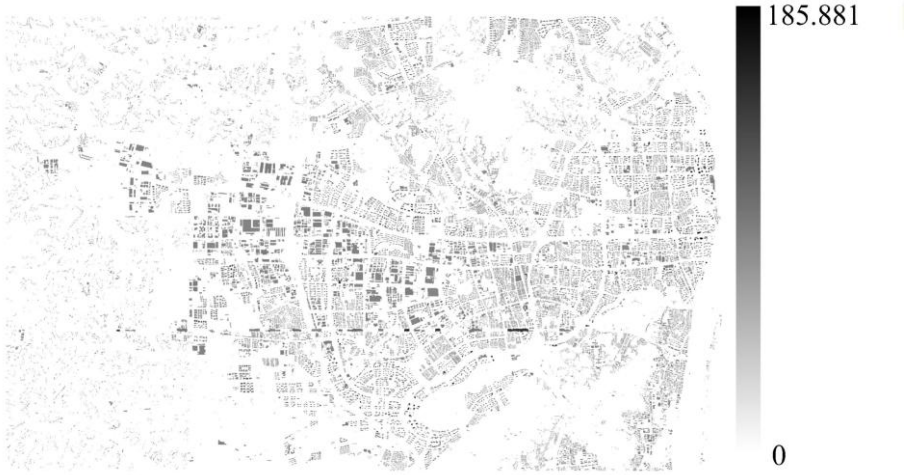
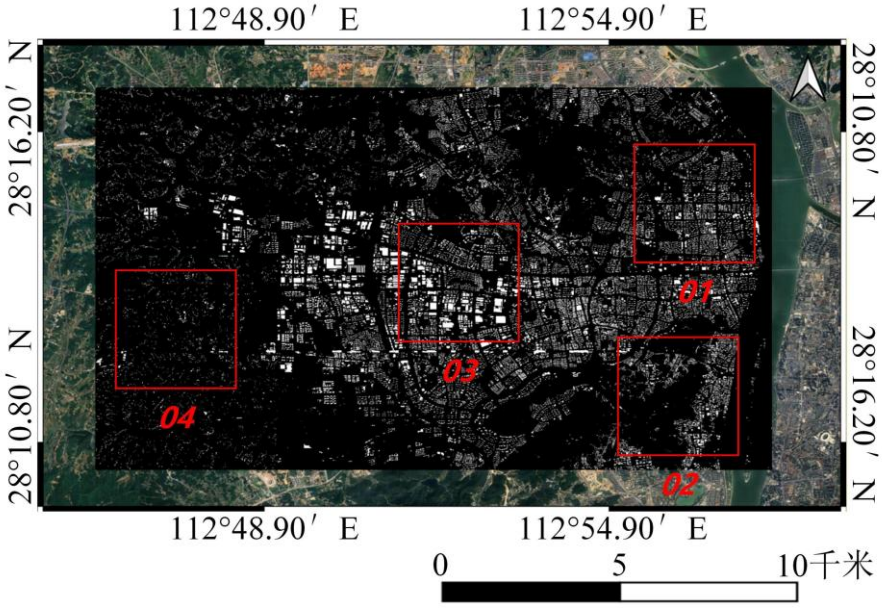


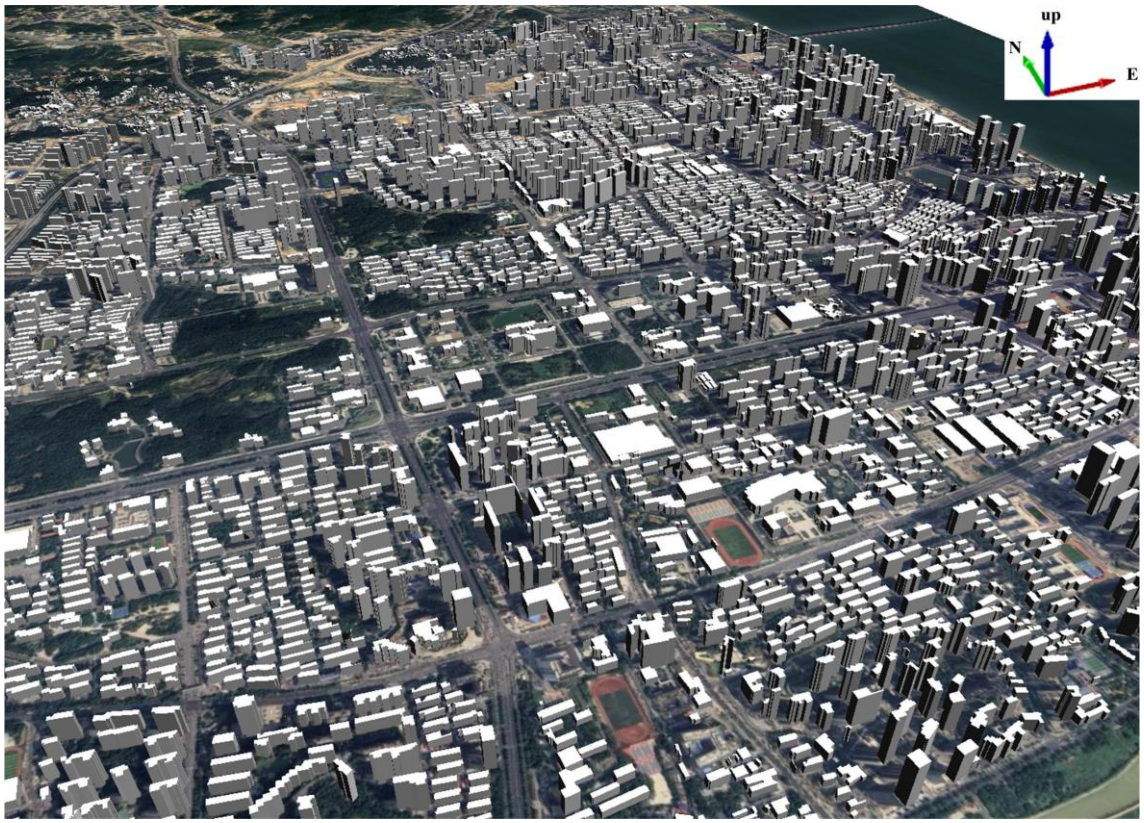
图 5.5 建筑中位数高度图（单位:m）

5.3.3 三维重建结果可视化与分析

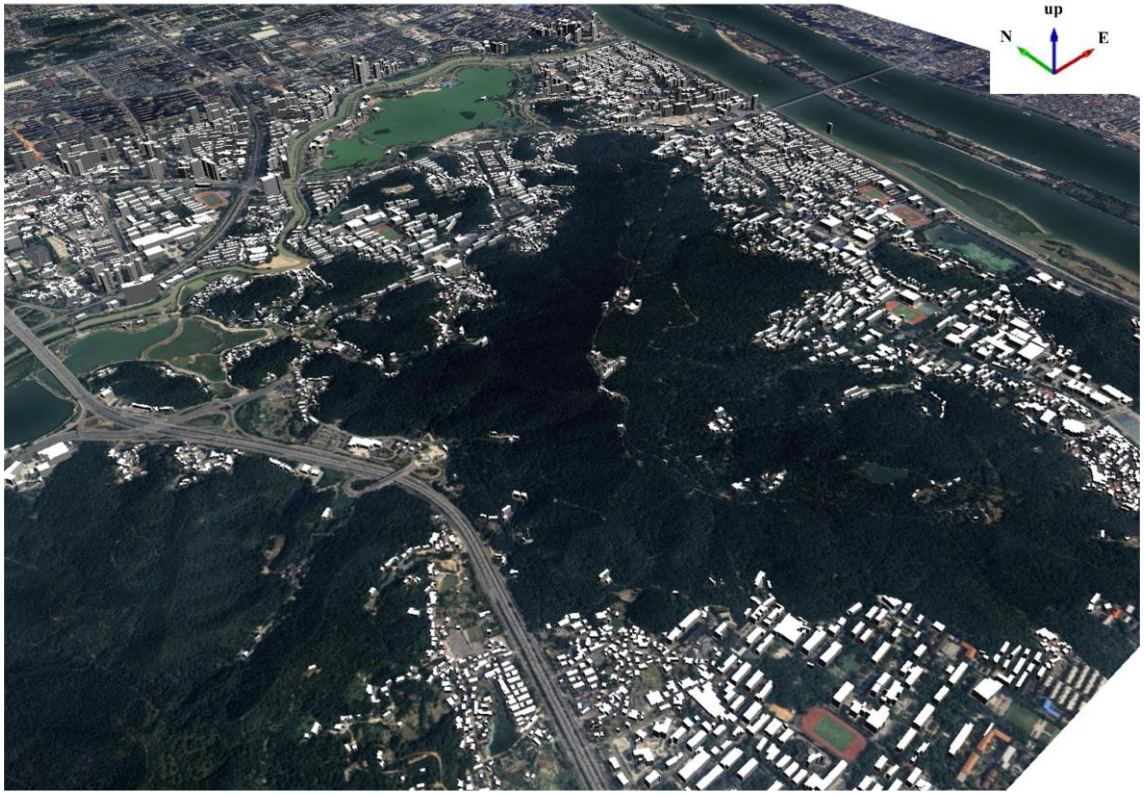
经过轮廓正则化、高度赋值、栅格转矢量及垂直拉伸等完整流程后，成功构建了覆盖研究区域的城市级 LoD1 三维白模。重建结果总览如图 5.6 a）所示，从宏观视角呈现了研究区域内建筑群的整体空间分布与高度格局。



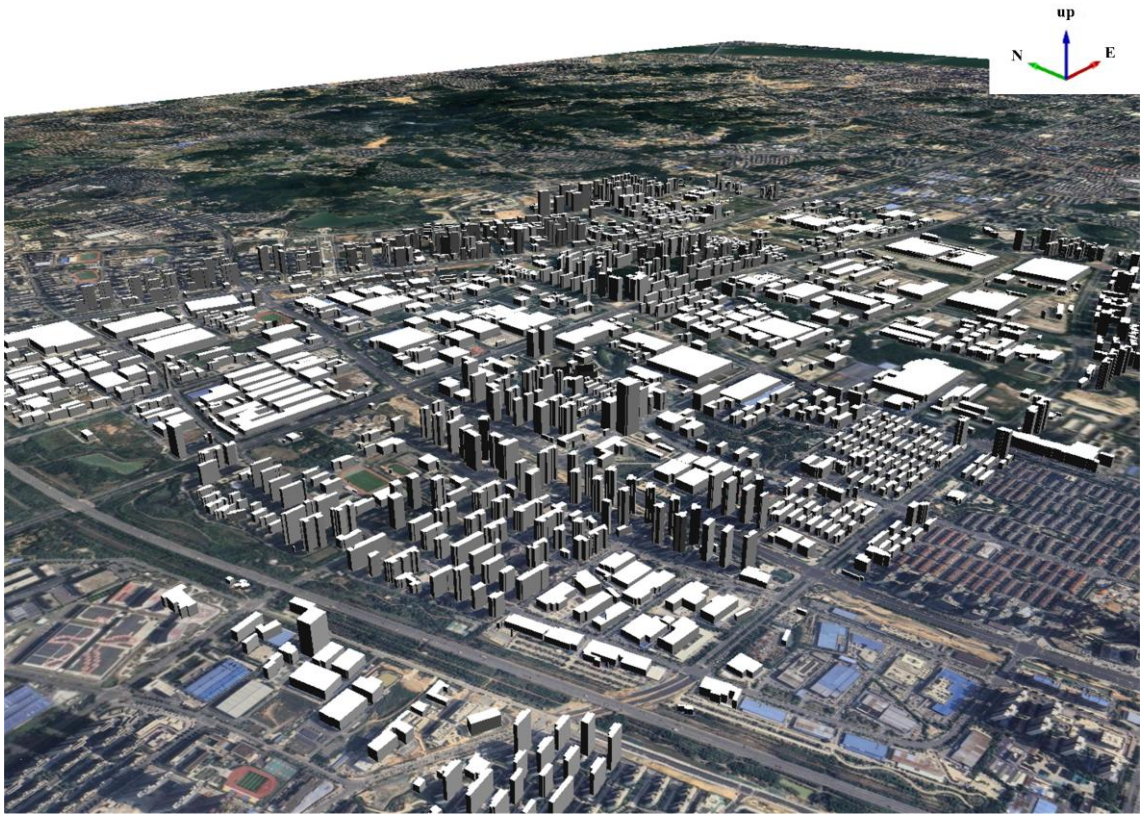
a) 研究区域总览及四个感兴趣区域分布



b) 区域 01：高密度高层住宅区



c) 区域 02：复杂地形混合区



d) 区域 03：大尺度工商业建筑区

e) 区域 04：稀疏农村低矮建筑区

图 5.6 不同城市功能区的典型三维重建结果

为全面评估模型在不同城市形态下的适用性与重建质量，本文选取了四类具有代表性的典型区域（分别是高密度高层住宅区、复杂地形混合区、大尺度工商业建筑区和稀疏农村低矮建筑区）进行可视化分析与详细讨论，具体结果如图 5.6 b）-5.6 e）所示。

不同城市功能区的典型三维重建结果：

区域 01：高密度高层住宅区 5.6 b）。该区域建筑排列紧密，楼栋之间间距较小，建筑高度普遍较高。重建结果显示，模型能够准确区分相邻建筑个体，高度分异明显，建筑轮廓经正则化后呈现出规整的几何形态，与实际城市景观高度吻合。

区域 02：复杂地形混合区 5.6 c）。该区域地处山地地形，建筑分布不均，高度差异显著，且地形起伏对建筑高度估算构成额外挑战。重建结果表明，本文方法在复杂地形条件下仍能有效提取建筑物并赋予合理高度，体现了良好的鲁棒性和环境适应能力。

区域 03：大尺度工商业建筑区 5.6 d）。该区域以大跨度厂房、仓库等工商业建筑为主，建筑单体面积大但高度相对较低。正则化算法能够较好地恢复此类建筑的长条状或矩形轮廓，拉伸后的三维模型准确反映了工业建筑的空间特征。

区域 04：稀疏农村低矮建筑区 5.6 e）。该区域建筑分布稀疏，单体面积小、高度低，是建筑提取的难点区域之一。重建结果显示，模型仍能有效识别并重建小型建筑，验证了方法对小目标建筑的检测与建模能力。

综合四类典型区域的分析结果可以看出，本文提出的城市级 LoD1 三维白模构建方法能够在多样化的城市功能分区与地形条件下稳定生成结构合理、几何规整的三维建筑模型，具备良好的场景泛化能力与工程实用价值，可为城市规划管理、空间分析以及智慧城市建设提供有效的三维数据支撑。

5.4 本章小结

本章针对现有城市级三维白模构建流程缺乏系统性、深度学习预测结果难以直接应用于工程建模等问题，提出了一套从深度学习预测结果到城市级 LoD1 三维白模的全流程自动化构建方法，并在长沙市大范围城区开展了实验验证。主要研究工作与结论如下：

（1）提出了从图像预测结果到城市级三维模型的完整自动化生成流程，解决了深度学习预测结果难以直接应用于工程建模的问题。其中，采用 RDP 算法进行轮廓简化，通过主方向判定与几何正交校正恢复建筑的直角特征与规则形态，有效消除了预测轮廓的边缘锯齿化问题；采用中位数统计策略进行高度赋值，显著提升了建筑高度属性的稳定性与可靠性。

（2）在长沙市约 40 223 栋建筑的大范围实验中验证了方法的场景泛化能力，实现了在不同城市形态、地形条件与建筑类型下的模型稳定预测。实验覆盖高密度高层住宅区、复杂地形混合区、大尺度工商业建筑区以及稀疏农村低矮建筑区等四类典型城市功能区。结果表明，本文方法在不同城市形态、地形条件和建筑类型下均能稳定生成结构合理、几何规整的三维建筑模型，展现了良好的环境适应性与工程实用性。

第6章 城市级三维白模应用研究

6.1 引言

近年来，随着遥感技术与深度学习算法的快速发展，利用高分辨率遥感影像自动提取建筑轮廓与估计建筑高度已成为可能，为大范围、低成本的三维城市建模提供了新的技术路径。然而，现有三维白模的应用研究多处于初级阶段，与土木工程实际需求存在一定脱节。现阶段，城市级三维白模大多局限于场景可视化展示与基本的空间浏览，未能充分挖掘其蕴含的丰富建筑几何信息。在城市建筑安全管理与结构健康监测领域，诸如形变点与建筑单体的空间归属判定、违法加层建筑的自动化识别等关键问题，均迫切需要以高精度的三维建筑模型作为空间参考底座。目前，鲜有研究将三维白模与建筑变化检测、雷达干涉测量等土木工程核心技术进行深度融合。这种应用层面的局限，导致三维白模的工程价值未能得到有效释放，也限制了其作为“数字底座”在城市精细化管理中的支撑作用。针对上述痛点，本章紧扣城市建筑安全监测需求，探索三维白模在 PS-InSAR 形变监测点配准与漏检识别、建筑违法加层智能识别等典型工程场景中的深层次应用，旨在推动三维白模从“可看”向“可用”的功能跃迁，为构建空天地一体化的智慧城市安全监测体系提供切实可行的技术支撑。

6.2 三维白模与 PS 点的配准与应用

6.2.1 问题背景与研究动机

永久散射体合成孔径雷达干涉测量技术（Persistent Scatterer Interferometric Synthetic Aperture Radar，PS-InSAR）是一种重要的地表形变监测技术，能够在长时间序列条件下获取毫米级精度的地表及建筑物形变信息，已广泛应用于城市地面沉降监测、基础设施安全评估以及建筑物健康诊断等领域[120]。将 PS-InSAR 形变监测点（简称 PS 点）与 LOD1 级建筑白模进行精确配准，可以实现 PS 点与具体建筑物之间的空间对应关系，从而快速判定形变信息所属的建筑单体，并进一步用于计算城市尺度 PS-InSAR 监测结果的漏检率和覆盖率，对于提升城市安全监测精度和建筑物形变分析的可靠性具有重要意义。

然而，由于合成孔径雷达（SAR）成像机制与光学成像机制在成像原理、观测几何及空间分辨特性等方面存在显著差异，同时受不同时相光学影像配准误差等因素影响，通过深度学习算法预测生成的建筑白模往往会产生整体或局部位置偏移，导致生成的 LOD1 建筑模型与 PS-InSAR 形变点云在空间上难以准确匹配。

此外，现有的 PS-InSAR 漏检率评估方法大多基于建筑物二维平面缓冲区思想，即在建筑物平面范围基础上向四周扩展一定距离，通过统计缓冲区内 PS 点数量来评估建筑物的检测完整性[121]。然而，该类方法未充分考虑雷达成像的空间位置及视线方向特性，简单采用各向同性的平面缓冲方式缺乏物理意义，容易引入不属于建筑物本体的 PS 形变点，从而导致 PS 点覆盖率被高估，评估结果准确性不足。

针对上述问题，本文提出了一套基于分步空间配准与参数化三维置信空间判定的解决方案，形成完整的“配准—归属—漏检评估”技术链路。首先实现 PS 点与三维白模的精确空间配准，继而基于 SAR 侧视成像物理机制构建参数化三维置信空间，完成 PS 点的建筑归属判定与漏检识别。

6.2.2 三维白模与 PS 点的配准方法

配准的核心目标是消除三维白模与 PS 点云之间因坐标系差异、成像几何差异等因素导致的空间偏移，使 PS 点能够精确附着在建筑表面。本文采用分步配准策略，依次在垂直方向、水平方向和卫星视线方向进行三次校正。优先进行垂直方向配准，旨在消除 SAR 干涉高程与光学预测高度之间因高程基准面不同产生的系统性偏差，防止高程误差耦合至后续水平配准过程；随后进行水平方向配准，在统一高程基准的基础上，高效校正由坐标系转换及深度学习模型预测带来的整体平面位置偏移，实现宏观空间对齐；最后进行基于卫星视线方向的精细配准，由于前两步仅能实现 PS 点与建筑整体轮廓的对齐，而 SAR 侧视成像机制决定了 PS 点实际散射位置位于建筑可见外立面而非建筑中心，因此需结合雷达成像几何，将 PS 点沿视线方向精确校正至真实的散射表面。该分步策略有效解耦了多维空间误差，兼顾了配准效率与物理机制的严谨性。

1. 垂直方向配准

由于 PS 点的高程信息来源于 SAR 干涉处理，而建筑白模的高度来源于光学遥感影像的深度学习预测，二者在垂直方向上通常存在系统性偏差。为消除该偏差，本文引入网格空间，将建筑白模数据和 PS 点数据分别转换为高程分布曲线，通过相关系数最大化原则确定最优垂直移动量。

具体而言，首先将目标区域进行网格化处理。对于建筑白模，计算每个网格中心点所处建筑多边形的最大高度作为该网格的高度值，形成建筑高度直方图；对于 PS 点，提取每个网格内所有 PS 点的最大高程作为网格高度，形成 PS 点高度直方图。为提高高层建筑在配准中的引导作用，对两个直方图分别进行考虑高度权重的加权归一化处理。以建筑高程分布曲线为例，其加权归一化过程表示为：

$$C_b\left(Z_p\right)=\frac{\alpha\left(Z_p\right)\times H_b\left(Z_p\right)}{\sum_q^n \alpha\left(Z_p\right)\times H_b\left(Z_p\right)} \tag{6.1}$$

式中，$C_b(Z_p)$表示第p个网格的归一化高程；$H_b(Z_p)$表示第p个网格的高度；$\alpha(Z_p)$表示第p个网格的高度权重，其计算方式为：

$$\alpha\left(Z_p\right)=\left(\frac{Z_p-Z_{\min}}{Z_{\max}-Z_{\min}}+\varepsilon\right)^{\gamma} \tag{6.2}$$

式中，Z_{min}表示所有网格的最低高度；Z_{max}表示所有网格的最高高度；ε为极小值，防止低层为 0；γ表示高度权重调节指数，用于调节高度对最终权重$\alpha(Z_p)$的非线性贡献程度。

设定 PS 点高程分布曲线在垂直方向上的平移量ΔZ，通过搜索使两条曲线的 Pearson 相关系数达到最大的最优垂直偏移量ΔZ_{opt}来完成垂直配准：

$$\Delta Z_{\text{opt}}=\arg\max\left[\rho\left(C_b\left(Z\right),C_{ps}\left(Z+\Delta Z\right)\right)\right] \tag{6.3}$$

2. 水平方向配准

垂直配准完成后，需要进一步消除水平面内的位置偏移。本文定义$M\times N$网格化的水平配准网络，分别基于建筑白模数据和 PS 点数据构建考虑高度权重的平面分布特征$F_b(x_i,y_i)$和$F_{\text{ps}}(x_i,y_i)$，并进行归一化处理。通过定义 PS 点平面分布特征在水平配准网络内的移动量Δx、Δy，计算每个移动量下两个平面分布特征的相关系数$\rho(\Delta x,\Delta y)$，以相关系数最大为目标进行移动量的寻优：

$$\left(\Delta x_{\text{opt}},\Delta y_{\text{opt}}\right)=\arg\max\rho\left(\Delta x,\Delta y\right) \tag{6.4}$$

3. 基于 SAR 视线方向的精细配准

完成垂直和水平配准后，PS 点在平面位置上已基本与建筑白模对齐，但由于 SAR 侧视成像的特殊性，PS 点实际散射位置位于建筑的可见外立面上，而非建筑正上方。因此，需要根据 SAR 卫星的成像几何进行最终的视线方向精细校正。

首先，根据 SAR 卫星的入射角θ和航向角α计算雷达视线方向单位矢量：

$$v_{\text{LOS}}=\left[-\sin\theta\cdot\cos\theta,-\sin\theta\sin\alpha,\cos\theta\right] \tag{6.5}$$

然后，基于 LOD1 建筑白模数据计算每个建筑外立面的法向量n。对于每个外立面，通过其法向量与雷达视线方向单位矢量的夹角φ判断其可见性：

$$\varphi=\arccos\left(n\cdot v_{\text{LOS}}\right) \tag{6.6}$$

当$\varphi<90°$时，即$n\cdot v_{LOS}>0$，表明该外立面朝向卫星视线方向，判定为可见外立面；反之则为不可见外立面。

最终，以建筑的至少两个可见外立面均存在与之对应平行的 PS 点序列为条件，筛选出代表性建筑，计算其附近 PS 点序列到对应可见外立面的距离在卫星视线方向上投影的平均值作为全局移动距离r，将所有 PS 点沿卫星视线方向移动

该距离，使 PS 点精确定位到建筑的可见外立面上。

6.2.3 基于三维置信空间的漏检识别方法

完成 PS 点与建筑白模的配准后，需要进一步判定每栋建筑是否被 PS-InSAR 成功监测，识别出漏检建筑。本文提出了一种双域置信判别法，用于建筑漏检识别。

1. 轮廓内强判定

将 PS 点和建筑正投影到同一二维平面，判断建筑的平面投影轮廓区域内是否存在 PS 点。若建筑轮廓内存在 PS 点，则表明该建筑被雷达正常“照亮”，判定为被成功监测，并将轮廓内的 PS 点标记为已归属，不参与后续的识别过程。

2. 三维置信空间判定

对于轮廓内不存在 PS 点的建筑，由于雷达视线方向、入射角、建筑高度、遮挡效应及测量误差等多重因素的影响，PS 点可能位于建筑轮廓外但仍属于该建筑。为此，本文引入三维置信空间的概念，以建筑可见外立面为参考面，沿雷达视线方向向外延展出一个参数化的三维梯形空间，见图 6.1。

三维置信空间的边界通过投影深度约束和平面内横向扩展约束来限定：

投影深度约束为：

$$0 \le d_i \le D_{\max} \tag{6.7}$$

其中，$d_i=(P-P_{\Pi_i})\cdot v_{LOS}$为 PS 点$P$沿雷达视线方向向参考面的投影深度；最大置信深度$D_{\max}$表示为：

$$D_{\max} = \exp\left(-\kappa\cdot\Gamma\right)\cdot\sqrt{\left(\frac{\delta_r}{\sin\theta}\right)^2+\left(\sigma_h\cdot\cot\theta\right)^2} \tag{6.8}$$

式中，δ_r为雷达分辨率，σ_{h}为建筑模型的高程误差，κ为经验修正系数，Γ为建筑复杂度（取值为[0, 1]），用于通过指数衰减项修正复杂城市背景下的信号多路径效应。

平面内横向扩展约束为：

$$\mathrm{dist}_{\Pi_i}\left(P\right) \le D_{\max}\cdot\tan\left(\theta+\frac{r}{2}\right) \tag{6.9}$$

式中，$\mathrm{dist}_{\Pi_i}(P)$表示 PS 点到其在外立面上正交投影点的平面内距离；r为雷达方位向分辨角。

当 PS 点同时满足投影深度约束与平面内横向扩展约束时，判定该 PS 点位于三维置信空间内部。当同一 PS 点同时落入多栋建筑的三维置信空间内时，引入基于建筑高度与视线距离的加权响应判别机制，将该 PS 点归属于高度加权响应值最大的建筑单体。

$$W_k = \alpha \cdot \frac{h_k}{H_{\max}} + \beta \cdot \left(1 - \frac{D_k}{D_{\max}}\right) \tag{6.10}$$

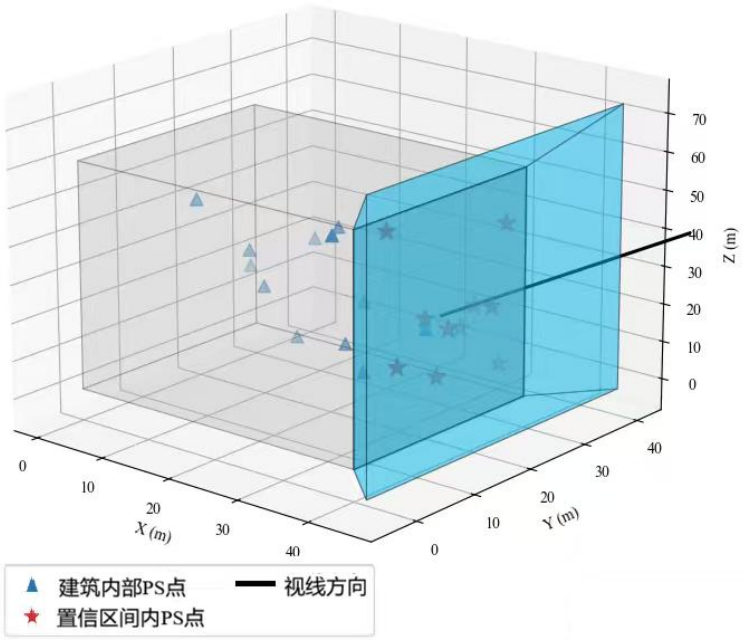


图 6.1 三维置信空间示意图

式中，h_k为第k栋建筑的高度，H_{max}为目标区域内所有建筑的最大高度；D_k为 PS 点到第k栋建筑参考面的投影距离，α，β为非负权重系数，权重系数α和β分别用于平衡建筑高度因素与雷达视线方向几何约束在 PS 点归属判别中的相对影响程度。

综合上述判定过程：当某建筑单体轮廓内存在 PS 点时，判定该建筑被成功监测；当轮廓内不存在 PS 点但三维置信空间内存在 PS 点时，判定该建筑被成功监测；当轮廓内和三维置信空间内均不存在 PS 点时，判定该建筑发生漏检。

6.2.4 实例验证

为验证上述方法的有效性，选取长沙市某区域进行实验。该区域包含建筑 1783 栋，获取区域内 2022 年 9 月至 2024 年 5 月期间共 20 景分辨率为 3m×3m（升轨）的 CSK 条带状图像，处理 SAR 影像提取 PS 点数量为 52853 个，卫星入射角为 24.31°，航向角为 11.24°。

将建筑白模与 PS 点结果统一转换为 2000 国家大地坐标系（CGCS2000），投影方式为高斯-克吕格 3 度分带投影。配准前建筑白模与 PS 点的位置关系如图 6.2 所示，可以观察到二者在空间上存在明显的偏移。

经过垂直方向配准，最佳垂直偏移量为ΔZ_{opt}=-15.18m，最佳水平偏移量为Δx=60m，Δy=-20m；经过视线方向精细配准，全局移动距离r=56.764m。配准后 PS 点成功附着在建筑可见外立面上，如图 6.3 所示。

在漏检识别阶段，根据计算得到雷达视线方向单位矢量 v_{LOS}=[-0.4041, -0.0803, 0.9112]。轮廓内强归属判定结果显示 16286 个 PS 点位于建筑内部；对剩余 36570 个轮廓外 PS 点进行三维置信空间判定，置信空间归属 PS 点 29505 个，仍未归属 PS 点 7065 个。综合两级判定结果，成功监测建筑 1494 栋，漏检建筑 289 栋，城市监测漏检率为 16.21%，最终结果见图 6.4。该结果可为后续建筑健康监测和风险评估提供可靠的数据支撑。

图 6.2 配准前建筑白模与 PS 点的位置关系

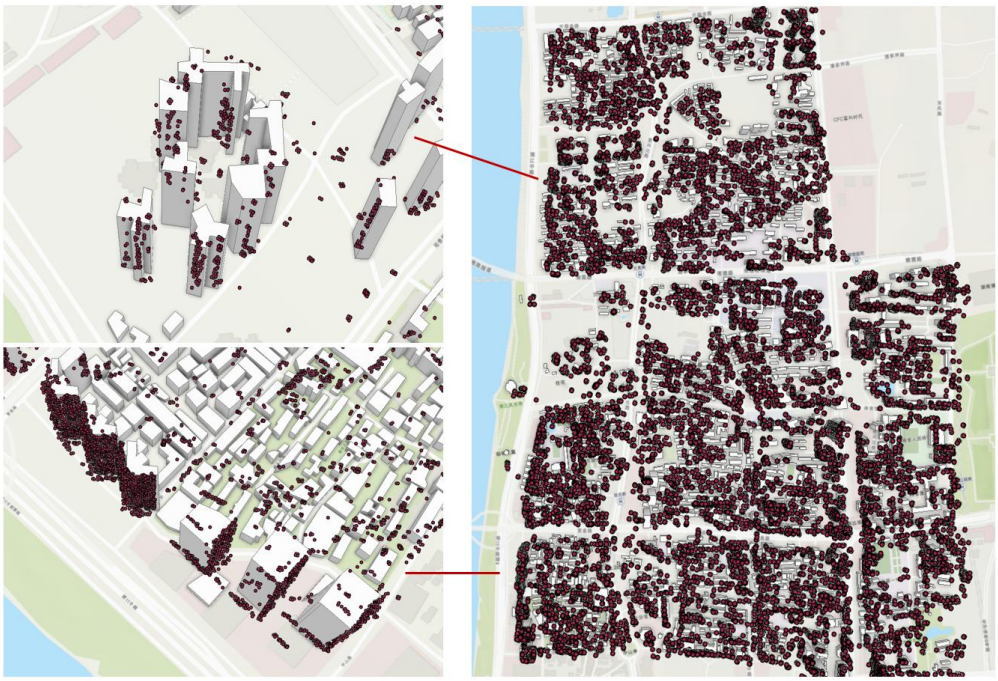


图 6.3 配准后建筑白模与 PS 点的位置关系

图 6.4 城市建筑监测结果及漏检空间分布图

为了进一步验证本文方法的有效性与优越性，将其与传统的基于建筑缓冲区的归属判定方法进行对比。在传统方法中（见图 6.5），将建筑物二维轮廓向外扩展 8m 作为缓冲区，其识别结果显示漏检建筑为 272 栋，城市监测漏检率为 15.26%，与本文提出的三维置信空间判定结果相比，传统方法识别出的漏检建筑数量明显偏少。

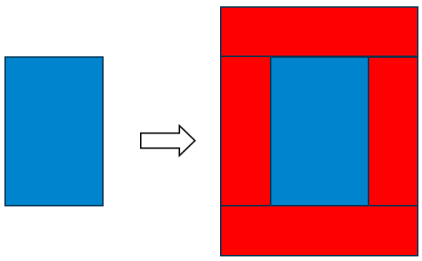


图 6.5 传统增加建筑缓冲区的方法

导致这一现象的主要原因在于：传统的二维缓冲区方法在平面上向四周均匀扩展，忽略了雷达侧视成像的几何特性。这种方法无法正确模拟真实的三维空间关系，极易将实际上并不属于该建筑的周边 PS 点（如相邻地物）错误地囊括进来，放宽了判定标准，最终导致建筑漏检率被严重低估。

为了更直观地揭示两种方法的差异，选取局部典型区域进行对比分析，见图 6.6 与图 6.7。研究发现，许多被传统方法判定为已成功监测的房屋，实际上发生了漏检。结合图 6.8 的具体案例可以看出，由于雷达视线存在倾角，位于目标房屋后方的其他散射体产生的 PS 点，在二维投影面上极易落入前方房屋的扩展缓冲区内。传统方法将其错误地归属给前方房屋，从而掩盖了前方房屋无有效 PS 点，即真实漏检的事实。

图 6.6 传统方法漏区域检建筑空间分布图

图 6.7 本方法区域漏检建筑空间分布图

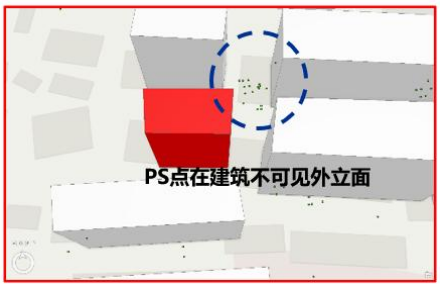


图 6.8 房屋前后向空间关系错位误判实例

6.3 建筑加层与新建建筑快速识别

6.3.1 问题背景与研究动机

随着城镇化进程加快，城市建筑密度不断上升，部分地区出现大量未经审批的自建房。这些建筑普遍缺乏专业设计和施工，部分居民为扩大使用空间，擅自加层，带来严重的结构安全隐患。因此，需要定期对建筑加层及违规新建行为进行检测，排除潜在的隐患。

传统的建筑加层检测主要依赖人工巡查，存在人力成本高、效率低、周期长等问题，难以实现对大范围城市区域的持续监测和及时干预。近年来，虽有部分技术采用无人机航拍图像来进行建筑加层的识别，但航拍图像受限于飞行路径、天气和操作条件等因素，且难以保证多次拍摄时的影像位置和角度完全一致，影响后续变化检测的准确性；航拍作业覆盖范围相对有限，难以满足大范围城市区域的高效、连续监测需求。

本文提出一种基于多模态遥感数据与三维白模时序分析的建筑加层智能识别方法，同时本节内容主要基于作者已发表的专利“一种基于多模态遥感数据的建筑加层智能识别方法及系统”[122]进行扩展和深化。该方法利用本文构建的城市级三维白模作为建筑高度的基准参考，结合多时相遥感影像的持续监测能力，通过时序差分分析与空间一致性检验实现对违法加层建筑的自动化识别。需要说明的是，新建建筑在本方法中可视为原始高度为 0 的建筑发生“加层”的特殊情形，二者在高度时序差分上均表现为高度的显著增加，因此可在统一的方法框架下实现协同识别。

6.3.2 方法框架

本方法的整体技术流程包括四个核心步骤：多模态遥感数据采集与预处理、基于深度学习的建筑高度时序预测、自适应阈值初步筛选、以及基于空间一致性的最终判定。

1. 多模态遥感数据采集与建筑高度预测

本方法持续获取包括光学遥感影像和雷达遥感影像在内的多模态、多视角遥感影像。

将不同时间节点采集的遥感影像输入前述章节所训练的深度学习预测模型中，实现城市建筑群高度的精确预测。针对任意一次建筑加层识别任务，选定包含多个连续时间节点的时间序列（至少包括 3 个时间节点，以消除季节性因素及临时性结构变化的干扰），在每个时间节点输出目标建筑的高度预测值。

2. 时序差分分析

计算当前时间节点目标建筑高度预测值与前一时间节点高度预测值的一阶差分作为目标建筑的高度变化值ΔH，遍历所有时间节点，形成目标建筑高度变化序列。相邻两个时间节点的高度差值反映了目标建筑在短期时间内的高度变化情况，包含了是否发生高度变化、变化的速率和幅度。

3. 基于自适应阈值的初步筛选

建筑存在加层或新建的情况下，其高度会发生明显的变化。本方法设定第一阈值T为建筑能够允许的高度变化值，一旦目标建筑的高度变化值超过第一阈值，则初步判定该建筑存在加层或属于新建建筑。

为了使阈值能够自适应地适配不同区域的建筑特征，需要首先确定合理的目标区域。本文将遥感影像进行网格化处理，计算每个网格单元内的建筑密度$x=A_b/A_g$（A_b为网格单元内的建筑总面积，A_g为网格单元的总面积），然后在每个网格单元内采用滑动窗口法确定目标区域，滑动窗口尺寸根据建筑密度动态调整：

$$W = W_0\left(1+p\left(1-x\right)\right) \tag{6.11}$$

式中，W_0为基准窗口尺寸，p为密度调节系数。该设计确保了建筑密度较低的区域采用较大的窗口以包含足够的建筑样本，建筑密度较高的区域采用较小的窗口以减少建筑间的相互干扰。

第一阈值T并非固定值，而是根据目标区域内所有建筑的高度变化情况自适应地进行调整。若目标区域所有建筑的高度变化值服从正态分布，则第一阈值设定为：

$$T = \mu + k\sigma\left(1+\lambda CV\right)\left(1-e^{-\frac{N}{N_0}}\right) \tag{6.12}$$

式中，μ为目标区域内所有建筑高度变化值的平均值；k为控制灵敏度的系数因子（取值范围 2-3）；σ为标准差；λ为区域异质性调整系数（取值范围 0.1-0.5）；CV=σ/μ为变异系数；N为目标区域内建筑的样本总量；N_0为基准样本数（取值 50-100）。

若目标区域所有建筑的高度变化值偏态严重，则采用基于中位数和绝对中位差的稳健统计量替代均值和标准差，或直接采用第 95 百分位数P_{95}作为阈值。

4. 基于空间一致性的最终判定

建筑违法加层一般表现为局部高度突变，而周围建筑高度不变。为了量化这种空间异质性，排除因模型预测误差或季节性因素导致的误判，本方法进一步采用加权局部空间异常度量方法计算目标建筑与目标区域内其他建筑的高度变化一致性指标。

空间异常度量值的计算过程表示为：

$$\mathrm{SA}\left(i\right) = \frac{\sum_{j\in\Omega_i}\omega_{ij}\left|\Delta H_i - \Delta H_j\right|}{\sum_{j\in\Omega_i}\omega_{ij}\sigma_{\mathrm{local}}} \tag{6.13}$$

式中，SA(i)表示目标建筑i的空间异常度量值；ΔH_i和ΔH_j分别表示目标建筑i和区域内建筑j的高度变化值；ω_{ij}表示空间权重矩阵，与目标建筑i以及邻近建筑j的距离成反比；Ω_i表示包含目标建筑i的目标区域；σ_{local}表示目标区域内所有建筑高度变化的标准差。

设定第二阈值（取值范围 2.5-3.5），若空间异常度量值超过第二阈值，则表明目标建筑的高度变化和邻近建筑的高度变化不一致，最终判定该建筑存在违法

加层或违规新建；反之，则表明目标建筑的高度变化和邻近建筑的高度变化存在一致性，判定为正常建筑。

6.3.3 实例验证

选取长沙某区域作为实验对象，采集 2019 年 9 月 13 日、2021 年 5 月 9 日、2026 年 3 月 3 日，3 个时间节点谷歌高分辨率光学遥感影像，构成多时相影像序列。三期影像均覆盖相同的空间范围，空间分辨率一致，满足时序差分分析的基本要求。研究区域多时相遥感影像如图 6.9 所示。

将三期遥感影像依次输入深度学习预测模型，获取各时间节点的建筑轮廓和高度预测结果并建立三维白模，随后按照本文所述方法流程依次进行时序差分分析、自适应阈值初步筛选及空间一致性判定。最终识别结果如图 6.10 所示，其中红色框标注为正确识别的加层或新建建筑，黄色框标注为误识别建筑，橙色框标注为漏识别建筑。

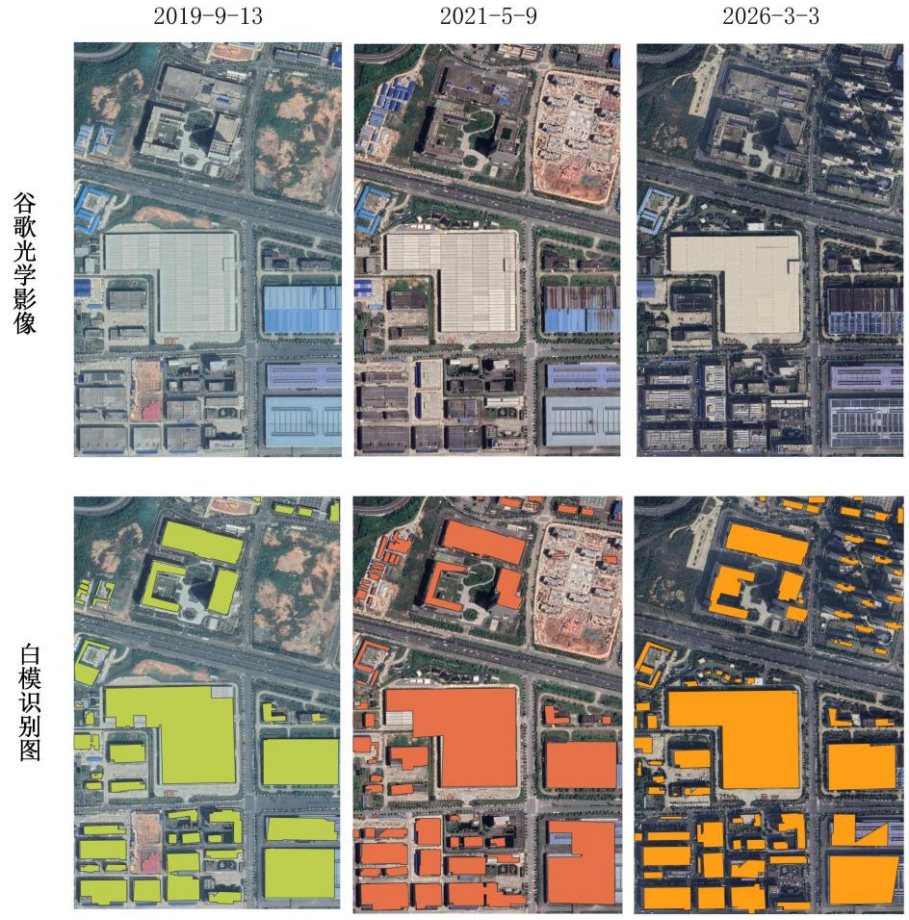
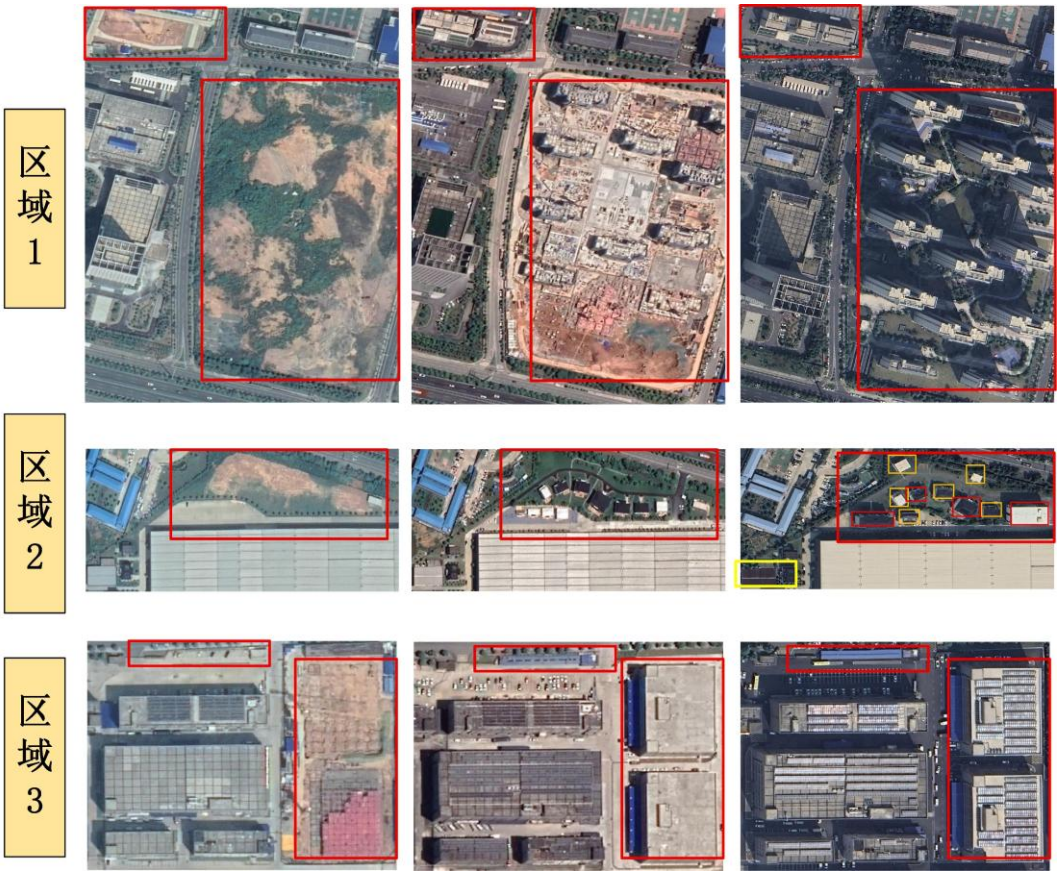


图 6.9 研究区域光学影像与识别结果

a）检测结果总览



b）检测区域对比图

图 6.10 建筑物加层与新建检测结果

经统计，本方法共检出 25 栋疑似加层建筑，其中 24 栋经人工核验确认为新建建筑，1 栋为误识别；同时存在 6 栋漏识别建筑。受限于实验区域和监测时段，

本次实验未发现既有建筑发生加层的情况，但由于新建建筑与加层建筑在本方法中遵循相同的高度变化判别逻辑，识别结果可有效验证该方法对建筑加层行为的适用性。由此可得，该方法的查准率为 96.0%（24/25），查全率为 80.0%（24/30）。

对误识别和漏识别案例进行逐一分析，其误差来源主要包括以下两方面：

（1）建筑屋顶特征变化导致的高度预测偏差。深度学习模型对建筑高度的预测在很大程度上依赖屋顶的光谱与纹理特征。当同一建筑在不同时期的屋顶颜色或材质发生变化时（如由浅色变为深色、翻新或老化等），模型可能对其高度产生不一致的预测结果。本实验中唯一的 1 栋误识别建筑即属于此类情况，该建筑屋顶在前期影像中呈白色，而在后期影像中变为黑色，显著的光谱差异导致模型在两期中给出了不同的高度估计值，其差值超过自适应阈值，从而被错误判定为加层建筑；

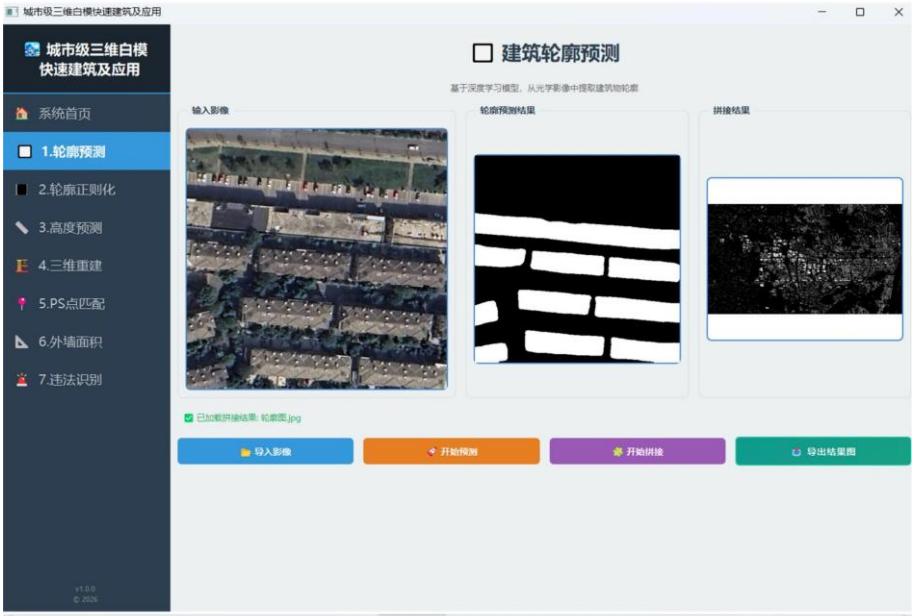
（2）建筑轮廓提取与后处理引起的目标丢失。不同时期影像中建筑轮廓的提取完整性和几何精度存在差异。一方面，部分小体量或密集排列的建筑在某一时期可能未被完整检出或发生几何偏移，导致在空间一致性判定环节无法建立有效的跨期对应关系；另一方面，轮廓提取后的正则化处理在优化建筑多边形几何形状的同时，会将面积过小的建筑轮廓作为噪声过滤，致使部分小面积真实建筑在后处理阶段即被消除，无法进入后续的变化检测流程。上述两类问题共同构成了本实验中漏识别的主要原因。

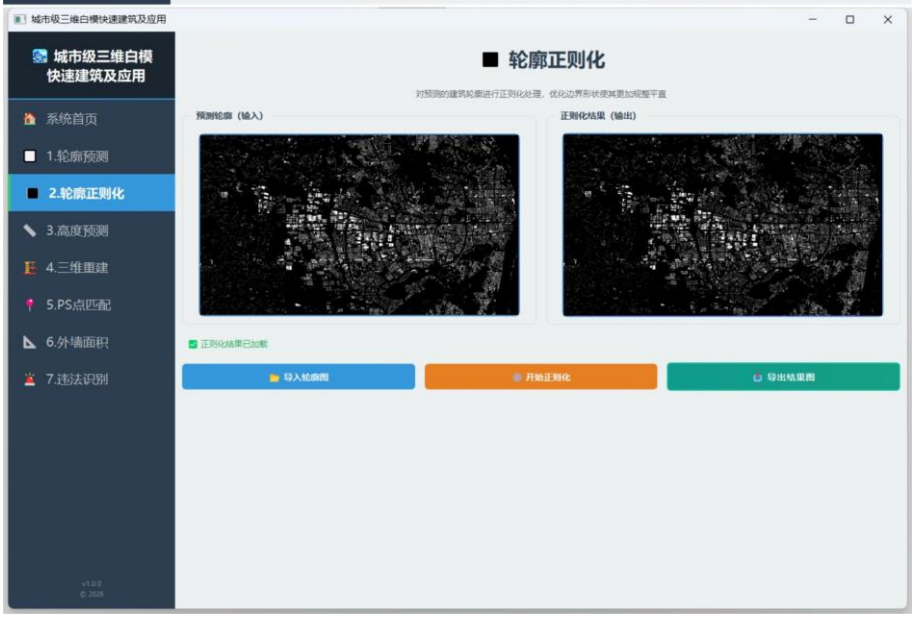
本方法建立在城市级三维白模的持续更新基础之上，只需设定合理的数据采集频率，即可实现违法加层的持续性监测，对违法加层及违规新建行为进行早期发现和干预。随着监测的持续进行，深度学习预测模型的参数还会不断更新和修正，进一步完善预测的准确性。判定结束后，系统可自动生成违法加层建筑的精确坐标、加层高度、遥感影像截图等关键信息，形成违法加层房屋清单及空间分布图，为城市管理和执法部门提供科学依据和技术支持。

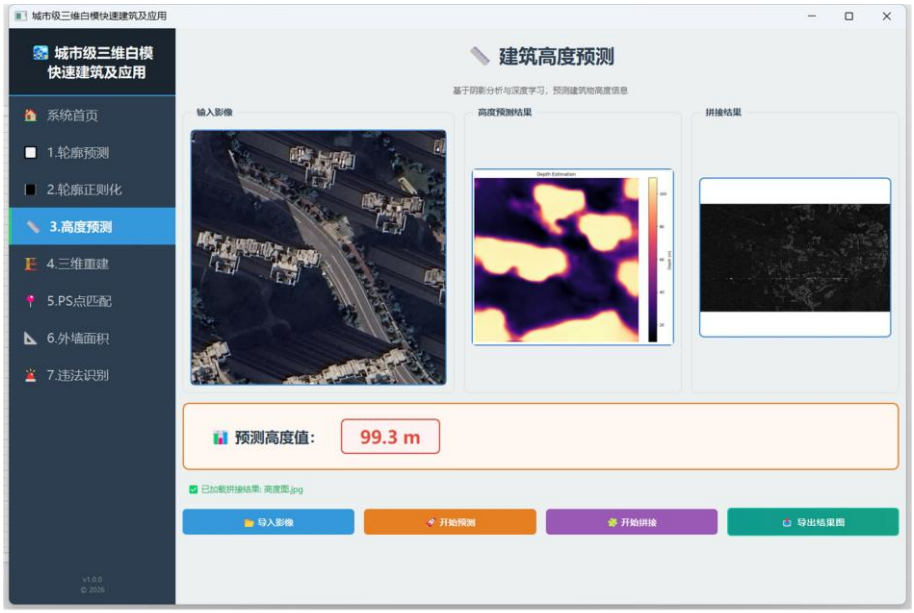
6.4 可视化平台建立

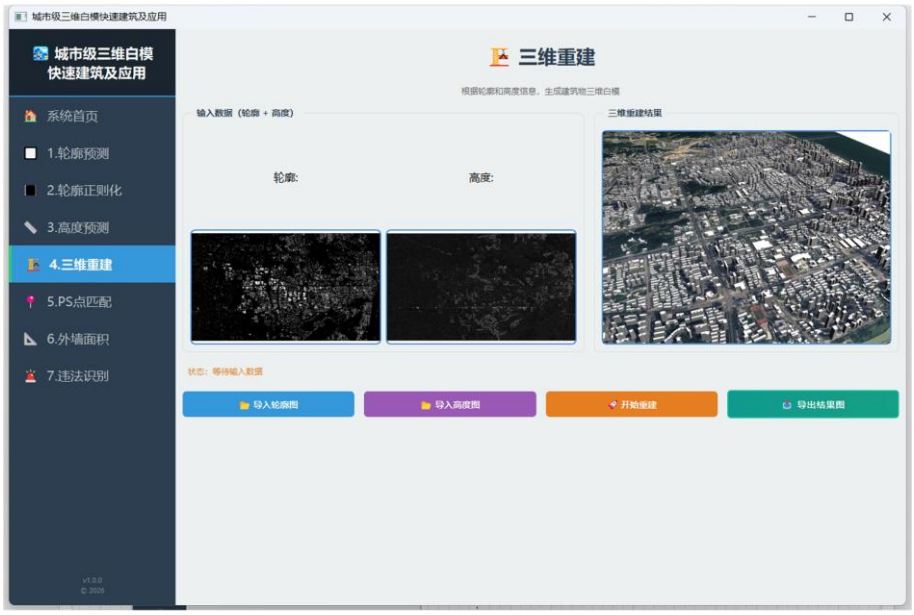
为将上述三维白模快速构建流程及其在城市建筑安全监测中的各项应用成果进行集成展示与交互操作，本文设计并开发了基于 Python 与 PyQt5 框架的城市建筑安全监测可视化平台。该平台以“城市级三维白模快速建筑及应用”为主题，以城市级三维白模为空间底座，集成了建筑轮廓预测、轮廓正则化、建筑高度预测、三维白模重建、PS-InSAR 形变监测点与白模的空间配准、外墙面积自动提取以及违法加层/修建识别等七大核心功能模块，实现了多源异构数据的统一管理与可视化呈现，平台界面见图 6.11。


城市级三维白模快速建筑及应用
系统首页
1.轮廓预测
2.轮廓正则化
3.高度预测
4.三维重建
5.PS点匹配
6.外墙面积
7.违法识别
城市级三维白模快速建筑及应用
Urban-Scale 3D Building White Model Rapid Construction & Application
建筑轮廓预测
轮廓正则化
建筑高度预测
三维重建（白模生成）
PS点与白模匹配校准
外墙面积自动提取
违法加层/修建识别
请点击左侧导航栏选择功能模块


城市级三维白模快速建筑及应用
建筑轮廓预测
输入影像
轮廓预测结果
拼接结果
导入影像
开始预测
开始拼接
导出结果图


城市级三维白模快速建筑及应用
轮廓正则化
预测轮廓（输入）
正则化结果（输出）
导入轮廓图
开始正则化
导出结果图


城市级三维白模快速建筑及应用
城市级三维白模
快速建筑及应用
系统首页
1.轮廓预测
2.轮廓正则化
3.高度预测
4.三维重建
5.PS点匹配
6.外墙面积
7.违法识别
建筑高度预测
基于阴影分析与深度学习，预测建筑物高度信息
输入影像
高度预测结果
拼接结果
预测高度值:
99.3 m
已加载拼接结果: 高度图.jpg
导入影像
开始预测
开始拼接
导出结果图


城市级三维白模快速建筑及应用
城市级三维白模
快速建筑及应用
系统首页
1.轮廓预测
2.轮廓正则化
3.高度预测
4.三维重建
5.PS点匹配
6.外墙面积
7.违法识别
三维重建
根据轮廓和高度信息，生成建筑物三维白模
输入数据（轮廓 + 高度）
轮廓:
高度:
三维重建结果
状态：等待输入数据
导入轮廓图
导入高度图
开始重建
导出结果图


城市级三维白模快速建筑及应用
城市级三维白模
快速建筑及应用
系统首页
1.轮廓预测
2.轮廓正则化
3.高度预测
4.三维重建
5.PS点匹配
6.外墙面积
7.违法识别
PS点与白模匹配校准
导入PS点数据与三维白模，进行位置校准与模型匹配
PS点数据
三维白模
匹配校准结果
PS点可视化
导入PS点数据
白模截图
导入白模数据
匹配结果
将在此显示
状态：等待导入PS点与白模数据
导入PS点
导入白模
开始匹配
导出结果图

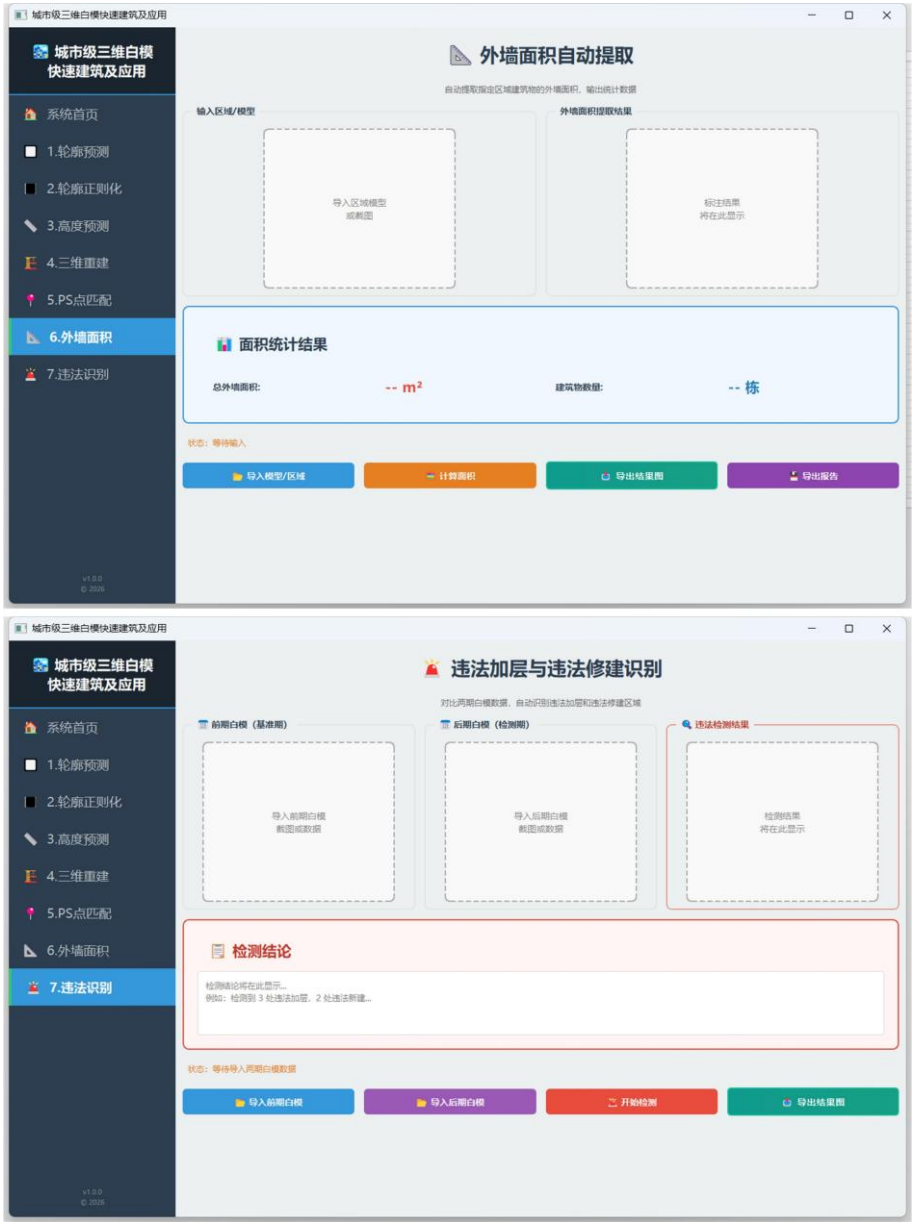


图 6.11 可视化平台操作界面

平台采用 Python 语言进行开发，以 PyQt5 作为图形用户界面（GUI）框架，兼顾了开发效率与界面美观性。整体界面采用左侧导航栏与右侧工作区相结合的经典布局模式：左侧导航栏以深色主题呈现，设有系统首页及七个功能模块的入口按钮，用户可通过点击相应按钮快速切换至目标功能页面；右侧工作区则根据当前所选模块动态加载对应的操作界面，提供数据导入、参数设置、算法执行、结果展示与导出等完整的交互流程。系统首页集中展示了平台名称、版本信息以及全部功能模块的概览列表，为用户提供了清晰的功能导引。

该可视化平台将三维白模的各项应用成果有机整合，为城市建筑安全的精细化管理提供了直观、高效的技术工具，展示了城市级三维白模作为“数字底座”在智慧城市建设中的综合应用价值。

6.5 本章小结

本章针对现有城市级三维白模应用与土木工程实际需求脱节的问题，紧扣城市建筑安全监测的核心目标，深入探索了三维白模在 PS-InSAR 形变监测配准、建筑漏检识别以及违法加层智能识别等典型工程场景中的应用，实现了三维白模从“可视化展示”向“工程化应用”的深度拓展。主要研究工作与结论如下：

（1）构建了一套基于分步空间配准与参数化三维置信空间判定方法，实现了完整的“配准—归属—漏检评估”技术链路。通过垂直方向、水平方向与 SAR 视线方向的分步配准策略消除空间偏移，并提出基于 SAR 侧视成像物理机制的参数化三维置信空间概念，结合“轮廓强判定+三维置信空间弱判定”的两级归属判定机制，实现了 PS 点的精确建筑归属与漏检识别。在长沙市某区域 1783 栋建筑、52853 个 PS 点的实验中，成功识别漏检建筑 289 栋，城市监测漏检率为 16.21%，为建筑健康监测与风险评估提供了可靠的数据支撑。

（2）设计了一种基于多时相遥感影像与三维白模时序分析的建筑违法加层智能识别方法，实现了大范围建筑违法加层的智能监测。利用高度时序差分，结合基于建筑密度的自适应动态阈值与空间一致性检验，有效排除了预测误差与季节性干扰。长沙市时序影像实验中，该方法查准率达 96.0%、查全率达 80.0%，验证了其在大范围自动化安全监测中的有效性。

（3）设计并开发了城市建筑安全监测可视化平台，实现了数据的统一管理与交互式可视化全流程闭环。基于 Python 与 PyQt5 框架，该平台以三维白模为空间底座，集成了建筑轮廓预测、建筑高度估计、白模建立、PS-InSAR 形变点配准、违法加层识别等七大核心功能模块。平台将本研究各项理论与算法成果有机整合，为智慧城市的安全精细化管理提供了直观、高效的工程化技术工具。

结论与展望

结论：

本文面向智慧城市建设与城市安全精细化管理对大范围、高精度三维建筑信息的迫切需求，针对传统城市三维建模方法成本高昂、算力依赖严重、更新周期漫长等突出瓶颈，提出了一套基于深度学习的遥感影像建筑物轮廓提取、高度估计及城市级三维白模快速构建与工程化应用的完整技术体系。从建筑轮廓的精细化语义分割到建筑高度的回归估计，从深度学习预测结果到城市级 LoD1 三维白模的全流程自动化生成，再到三维白模在 PS-InSAR 形变监测配准与建筑违法加层智能识别等典型工程场景中的深层次应用，实现了城市三维建筑模型从数据获取到工程落地的全链路贯通。本文的主要结论如下：

（1）提出了基于伪多模态结构信息的建筑物轮廓精细提取网络 SPR-Net，实现了复杂背景与多尺度条件下建筑物轮廓的高精度、精细化提取。针对遥感影像建筑物轮廓提取中边界模糊、尺度差异显著及复杂背景干扰等问题，以 SegFormer 为主干架构，创新性地设计了编码器端 PMM_S 与解码器端 PMM_C 两类伪多模态结构增强模块，在仅依赖单一光学影像输入的前提下实现了对建筑轮廓边界的有效强化；同时构建了覆盖长沙市约 115km²的高分辨率建筑数据集，并设计了复合损失函数以提升分割边界锐度。在 WHU 航空影像数据集上的对比实验中，SPR-Net-b2 以 89.51%的 IoU、93.34%的 Recall、95.6%的 Precision 和 94.47%的 F1 值全面领先 SegNext-b、Deeplabv3plus-r50 等主流分割网络；在长沙自建数据集上，SPR-Net-b2 以 76.93%的 IoU 较基线 SegFormer-b2 提升 4.19 个百分点，在复杂背景与多尺度条件下均实现了稳定的轮廓提取效果。同时，SPR-Net-b5 在 WHU 数据集上以 91.90%的 IoU、96.26%的 Precision 和 95.78%的 F1 值取得全部最优，超越 MAP-Net 等 SOTA 模型。消融实验系统验证了 PMM 模块等各组件的有效性与协同增益。此外，复杂度实验表明模型参数量与计算量均保持在较低水平，进一步证明了所提网络在建筑轮廓提取任务中兼具高精度与高效性。

（2）基于级联动态上采样的建筑物高度估计网络 Segformer-H，实现了复杂城市场景下建筑物高度的精度预测与稳定泛化。针对遥感影像建筑物高度估计任务中特征融合阶段空间信息丢失、边界模糊以及模型跨城市形态泛化困难等问题，在保留 SegFormer 层次化 Transformer 编码器全局语义建模能力的基础上，将解码器中的双线性插值上采样替换为基于 DySample 的级联动态上采样模块，通过可学习的偏移量生成器自适应地计算采样位置，有效克服了固定插值核在建筑物边界与高度突变区域引入的模糊效应。同时，自建了涵盖 9m 至 250m 大跨度高

度范围的长沙建筑高度数据集，弥补了国内相关场景数据的空白。在 ISPRS Potsdam、ISPRS Vaihingen 及长沙建筑高度三个数据集上的系统对比实验表明，Segformer-H 在所有评价指标上均取得最优性能，在 Potsdam 数据集上 RMSE 为 2.586m，在 Vaihingen 数据集上 RMSE 为 2.354m，在长沙数据集上 RMSE 为 11.004m，全面优于 Segnext-b 等对比方法。消融实验与可视化分析进一步验证了动态上采样模块的有效性，表明 Segformer-H 在建筑边缘刻画锐利度与不同复杂度场景下的预测稳定性方面均具有优势。

（3）构建了城市级 LoD1 三维白模全流程自动化生成方法，并探索了三维白模在城市建筑安全监测中的深层次工程应用。针对现有城市级三维白模构建流程缺乏系统性以及白模应用与土木工程实际需求结合不足的问题，通过预测结果空间拼接融合、建筑轮廓正则化、中位数高度赋值及矢量化拉伸建模等关键技术，构建了完整的城市级 LoD1 三维白模自动化生成流程。在长沙市约 40223 栋建筑的大范围实验中，该方法在高密度高层住宅区、复杂地形混合区、大尺度工商业建筑区以及稀疏农村低矮建筑区等四类典型城市功能区均能稳定生成结构合理、几何规整的三维建筑模型。在工程应用方面，构建了一套“配准—归属—漏检评估”的 PS-InSAR 形变监测点与三维白模空间匹配技术链路，在长沙市某区域 1783 栋建筑、52853 个 PS 点的实验中，成功识别漏检建筑 289 栋，城市监测漏检率为 16.21%；设计了基于多时相遥感影像与三维白模时序分析的建筑违法加层智能识别方法，在长沙市某区域三期遥感影像的实验验证中，查准率达 96.0%，查全率达 80.0%。此外，基于 Python 与 PyQt5 开发了集成七大功能模块的可视化平台，实现了三维白模构建与应用全流程的统一管理与交互操作。

展望：

尽管本文取得了一定成果，但仍存在诸多值得深入探索的方向，未来可从以下几个方面开展进一步研究：

（1）引入更先进的网络架构与多源数据融合策略。未来可探索将视觉基础大模型引入建筑物轮廓提取与高度估计任务，通过提示学习或参数高效微调等策略，释放大模型在复杂遥感场景下的表征能力。同时，鉴于单一数据源在信息表达上的局限性，未来研究可进一步拓宽数据获取渠道，引入 SAR 影像、LiDAR 点云、倾斜摄影、街景影像及众源地理数据等多源异构信息，提升城市三维建模的准确性与完整性。

（2）解决城中村等高密度区域建筑粘连问题。在城中村等场景中，建筑之间紧密相连、边界极其模糊，现有分割方法难以实现逐栋精准分离，严重制约了三维白模在城市精细化治理中的应用。未来可引入边界感知等技术手段，突破密集

粘连建筑的分离瓶颈，实现复杂场景下建筑物的逐栋高精度提取。

（3）融合土木工程专业知识，实现建筑风险智能研判。未来可将结构力学分析、承载力计算、抗震性能评估等专业知识引入识别后的决策环节，建立“检测—评估—预警”一体化的智能分析框架，对识别出的加层建筑结合其原始设计层数、结构类型等属性信息，自动进行承载力余量估算与风险等级判定，实现从发现问题到量化风险的完整技术闭环。

参考文献

[1] 凌敏. 基于数字孪生技术的智慧城市建筑群协同规划模式 [J]. 新型城镇化, 2026, (02): 47-9.

[2] 自然资源部. 实景三维中国建设总体实施方案(2022-2025年)[EB/OL].(2022-2-24)[2022-2-24]. https://www.gov.cn/xinwen/2022-03/01/content_5676226.htm

[3] 国家发展改革委, 国家数据局, 财政部, 等. 关于深化智慧城市发展 推进城市全域数字化转型的指导意见[EB/OL].(2024-5-14)[2024-5-14].https://www.gov.cn/zhengce/zhengceku/202405/content/6952353.htm

[4] 中共中央办公厅, 国务院办公厅. 关于推进新型城市基础设施建设打造韧性城市的意见[EB/OL].(2024-11-26)[2024-11-26].https://www.gov.cn/zhengce/202412/content_6991171.htm

[5] 石晓冬, 杨明, 王吉力. 城市体检：空间治理机制、方法、技术的新响应 [J]. 地理科学, 2021, 41(10): 1697-705.

[6] He Y, Niu L, Zhang G, et al. Detailed health monitoring of large-scale urban infrastructure by combining optical and SAR images [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025.

[7] Xu Y, Tuttas S, Hoegner L, et al. Geometric primitive extraction from point clouds of construction sites using VGS [J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(3): 424-8.

[8] He Y, Xu G, Kaufmann H, et al. Integration of InSAR and LiDAR technologies for a detailed urban subsidence and hazard assessment in Shenzhen, China [J]. Remote Sensing, 2021, 13(12): 2366.

[9] Varol B, Yılmaz E Ö, Maktav D, et al. Detection of illegal constructions in urban cities: Comparing LIDAR data and stereo KOMPSAT-3 images with development plans [J]. European Journal of Remote Sensing, 2019, 52(1): 335-44.

[10] Liu W, Zhou L, Zhang S, et al. A new high-precision and lightweight detection model for illegal construction objects based on deep learning [J]. Tsinghua Science and Technology, 2024, 29(4): 1002-22.

[11] Levine N M, Spencer Jr B F. Post-earthquake building evaluation using UAVs: A BIM-based digital twin framework [J]. Sensors, 2022, 22(3): 873.

[12] Khanmohammadi S, Arashpour M, Bai Y. Applications of building information modeling (BIM) in disaster resilience: Present status and future trends[C]. ISARC Proceedings of the International Symposium on Automation and Robotics in Construction. IAARC Publications, 2020: 1380-7.

[13] Akhavi Zadegan A, Vivet D, Hadachi A. Challenges and advancements in image-based 3D reconstruction of large-scale urban environments: a review of deep learning and classical methods [J]. Frontiers in Computer Science, 2025, 7: 1467103.

[14] Zeng X, Luo B, Zhang S, et al. RMTDepth: Retentive Vision Transformer for Enhanced Self-Supervised Monocular Depth Estimation from Oblique UAV Videos [J]. Remote Sensing, 2025, 17(19): 3372.

[15] Mahmoud M, Zhao Z, Chen W, et al. Automated Scan-to-BIM: A deep learning-based framework for indoor environments with complex furniture elements [J]. Journal of Building Engineering, 2025, 106: 112596.

[16] 冯楚乔, 刘宇飞, 岳岩, 等. 基于三维视觉的房屋智能巡检方法研究[J]. 建筑结构学报, 2024, 45(06): 133-42.

[17] Li W, Meng L, Wang J, et al. 3d building reconstruction from monocular remote sensing images[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 12548-57.

[18] Li Q, Mou L, Sun Y, et al. A review of building extraction from remote sensing imagery: Geometrical structures and semantic attributes [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.

[19] Ji S, Wei S, Lu M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set [J]. IEEE Transactions on geoscience and remote sensing, 2018, 57(1): 574-86.

[20] Wu J, Yuan M, Wang T, et al. HeightFormer: Single-imagery height estimation transformer with bilateral feature pyramid fusion [J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.

[21] Jabbar S, Taj M. Stereollax net: Stereo parallax-based deep learning network for building height estimation [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[22] Luo H, Zhang J, Liu X, et al. Large-scale 3D reconstruction from multi-view imagery: A comprehensive review [J]. Remote Sensing, 2024, 16(5): 773.

[23] Nex F, Zhang N, Remondino F, et al. Benchmarking the extraction of 3D geometry from UAV images with deep learning methods [J]. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives, 2023, 48(1/W3-2023): 123-30.

[24] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-40.

[25] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical image computing and computer-assisted intervention. Springer, 2015: 234-41.

[26] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137-49.

[27] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-95.

[28] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-90.

[29] Chen L-C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-48.

[30] Chen L-C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation [J]. arXiv preprint arXiv:170605587, 2017.

[31] Chen L-C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 801-18.

[32] He K, Gkioxari G, Dollár P. Girshick ross. mask r-cnn[C]. Proceedings of the IEEE international conference on computer vision. 2017: 2961-9.

[33] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]. International workshop on deep learning in medical image analysis. Springer, 2018: 3-11.

[34] Yu C, Wang J, Peng C, et al. Bisenet: Bilateral segmentation network for real-time semantic segmentation[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 325-41.

[35] Poudel R P, Liwicki S, Cipolla R. Fast-scnn: Fast semantic segmentation network [J]. arXiv preprint arXiv:190204502, 2019.

[36] Wu H, Zhang J, Huang K, et al. Fastfcn: Rethinking dilated convolution in the backbone for semantic segmentation [J]. arXiv preprint arXiv:190311816, 2019.

[37] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-54.

[38] Zhang H, Wu C, Zhang Z, et al. Resnest: Split-attention networks[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2736-46.

[39] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:201011929, 2020.

[40] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-22.

[41] Strudel R, Garcia R, Laptev I, et al. Segmenter: Transformer for semantic segmentation[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 7262-72.

[42] Chu X, Tian Z, Wang Y, et al. Twins: Revisiting the design of spatial attention in vision transformers [J]. Advances in neural information processing systems, 2021, 34: 9355-66.

[43] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers [J]. Advances in neural information processing systems, 2021, 34: 12077-90.

[44] Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1290-9.

[45] Guo M-H, Lu C-Z, Hou Q, et al. Segnext: Rethinking convolutional attention design for semantic segmentation [J]. Advances in neural information processing systems, 2022, 35: 1140-56.

[46] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces[C]. First conference on language modeling. 2024.

[47] Yuan J. Learning building extraction in aerial scenes with convolutional networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(11): 2793-8.

[48] Zhu Q, Liao C, Hu H, et al. MAP-Net: Multiple attending path neural network for building footprint extraction from remote sensed imagery [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(7): 6169-81.

[49] Zhao Y, Sun G, Zhang L, et al. MSRF-Net: Multiscale receptive field network for building detection from remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-14.

[50] Huang X, Zhang Z, Li J. China's first sub-meter building footprints derived by deep learning [J]. Remote Sensing of Environment, 2024, 311: 114274.

[51] Chen S, Ogawa Y, Zhao C, et al. Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 195: 129-52.

[52] Huang H, Liu J, Wang R. Easy-net: A lightweight building extraction network based on building features [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 62: 1-15.

[53] Li Q, Shi Y, Huang X, et al. Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF) [J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 58(11): 7502-19.

[54] Li Y, Hong D, Li C, et al. HD-Net: High-resolution decoupled network for building footprint extraction via deeply supervised body and boundary decomposition [J]. ISPRS journal of photogrammetry and remote sensing, 2024, 209: 51-65.

[55] Hu A, Wu L, Xu Y, et al. SANET: A shape-aware building footprints extraction method in remote sensing images by integrating Fourier shape descriptors [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.

[56] Wu Y, Xu L, Chen Y, et al. TAL: Topography-aware multi-resolution fusion learning for enhanced building footprint extraction [J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

[57] He L, Shan J, Aliaga D. Generative building feature estimation from satellite images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61: 1-13.

[58] Li Q, Zorzi S, Shi Y, et al. RegGAN: An end-to-end network for building footprint generation with boundary regularization [J]. Remote Sensing, 2022, 14(8): 1835.

[59] Wei S, Zhang T, Yu D, et al. From lines to polygons: Polygonal building contour extraction from high-resolution remote sensing imagery [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2024, 209: 213-32.

[60] Zhang T, Wei S, Zhou Y, et al. P2PFormer: A primitive-to-polygon method for regular building contour extraction from remote sensing images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[61] Xu B, Xu J, Xue N, et al. Accurate polygonal mapping of buildings in satellite imagery [J]. arXiv preprint arXiv:220800609, 2022.

[62] Li W, Zhao W, Yu J, et al. Joint semantic–geometric learning for polygonal building segmentation from high-resolution remote sensing images [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 201: 26-37.

[63] Huang J, Zhang X, Xin Q, et al. Automatic building extraction from high-resolution aerial images and LiDAR data using gated residual refinement network [J]. ISPRS journal of photogrammetry and remote sensing, 2019, 151: 91-105.

[64] Hosseinpour H, Samadzadegan F, Javan F D. CMGFNet: A deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images [J]. ISPRS journal of photogrammetry and remote sensing, 2022, 184: 96-115.

[65] Sun Y, Hua Y, Mou L, et al. Conditional GIS-aware network for individual building segmentation in a VHR SAR image[C]. 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, 2021: 4532-5.

[66] Jing H, Sun X, Wang Z, et al. Fine building segmentation in high-resolution SAR images via selective pyramid dilated network [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 6608-23.

[67] Li X, Zhang G, Cui H, et al. Progressive fusion learning: A multimodal joint segmentation framework for building extraction from optical and SAR images [J]. ISPRS Journal of photogrammetry and remote sensing, 2023, 195: 178-91.

[68] Xie Y, Tu J, Zhao Y, et al. Single-Image Building Height Estimation Using Spatial Distribution-Aware Optimization in Complex Urban Areas [J]. Remote Sensing, 2026, 18(5): 801.

[69] Zhu B, Wang Y, Yu H. An algorithm measuring urban building heights by combining the PS-InSAR technique and two-stage programming approach framework [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 16: 7624-35.

[70] Cao Y, Weng Q. A deep learning-based super-resolution method for building height estimation at 2.5 m spatial resolution in the Northern Hemisphere [J]. Remote Sensing of Environment, 2024, 310: 114241.

[71] Li X, Wang M, Fang Y. Height estimation from single aerial images using a deep ordinal regression network [J]. IEEE Geoscience and Remote Sensing Letters, 2020, 19: 1-5.

[72] Sun Y, Mou L, Wang Y, et al. Large-scale building height retrieval from single SAR imagery based on bounding box regression networks [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184: 79-95.

[73] Nascetti A, Yadav R, Ban Y. A CNN regression model to estimate buildings height maps using Sentinel-1 SAR and Sentinel-2 MSI time series[C]. IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2023: 2831-4.

[74] Gong X, Hou Z, Wan Y, et al. Multispectral and SAR image fusion for multiscale decomposition based on least squares optimization rolling guidance

filtering [J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20.

[75] Lyu S, Ji C, Liu Z, et al. Four seasonal composite Sentinel-2 images for the large-scale estimation of the number of stories in each individual building [J]. Remote Sensing of Environment, 2024, 303: 114017.

[76] Kakooei M, Baleghi Y. Mapping Building Heights at Large Scales Using Sentinel-1 Radar Imagery and Nighttime Light Data [J]. Remote Sensing, 2024, 16(18): 3371.

[77] Chen Y, Yan Q, Huang W. MFTSC: A semantically constrained method for urban building height estimation using multiple source images [J]. Remote Sensing, 2023, 15(23): 5552.

[78] Sun X, Huang X, Mao Y, et al. GABLE: A first fine-grained 3D building model of China on a national scale from very high resolution satellite imagery [J]. Remote Sensing of Environment, 2024, 305: 114057.

[79] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521(7553): 436-44.

[80] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives [J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-828.

[81] Bengio Y, Goodfellow I, Courville A. Deep learning [M]. MIT press Cambridge, MA, USA, 2017.

[82] Wang Q, Ma Y, Zhao K, et al. A comprehensive survey of loss functions in machine learning [J]. Annals of Data Science, 2022, 9(2): 187-212.

[83] Bottou L. Large-scale machine learning with stochastic gradient descent[C]. Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers. Springer, 2010: 177-86.

[84] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010: 249-56.

[85] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-8.

[86] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. International conference on machine learning. pmlr, 2015: 448-56.

[87] Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks[C]. International conference on machine learning. Pmlr, 2013: 1310-8.

[88] Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts [J]. arXiv preprint arXiv:160803983, 2016.

[89] Goyal P, Dollár P, Girshick R, et al. Accurate, large minibatch sgd: Training imagenet in 1 hour [J]. arXiv preprint arXiv:170602677, 2017.

[90] Cybenko G. Approximation by superpositions of a sigmoidal function [J]. Mathematics of control, signals and systems, 1989, 2(4): 303-14.

[91] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]. Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-23.

[92] Polyak B T. Some methods of speeding up the convergence of iteration methods [J]. Ussr computational mathematics and mathematical physics, 1964, 4(5): 1-17.

[93] Tieleman T. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude [J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26.

[94] Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv preprint arXiv:14126980, 2014.

[95] Loshchilov I, Hutter F. Decoupled weight decay regularization [J]. arXiv preprint arXiv:171105101, 2017.

[96] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 2002, 86(11): 2278-324.

[97] 香超. 基于计算机视觉与深度学习的裂缝检测算法研究 [D]. 湖南: 湖南大学, 2023.

[98] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[99] Zhang A, Lipton Z C, Li M, et al. Dive into Deep Learning [M]. Cambridge University Press, 2023.

[100]Malambo L, Popescu S. Image to image deep learning for enhanced vegetation height modeling in Texas [J]. Remote Sensing, 2023, 15(22): 5391.

[101]Wang Y, Ding W, Zhang R, et al. Boundary-aware multitask learning for remote sensing imagery [J]. IEEE Journal of selected topics in applied earth observations and remote sensing, 2020, 14: 951-63.

[102]Bakici S, Erkek B, Ayyildiz E, et al. The use of 3d city models form oblique images on land administration [J]. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2017, 4: 117-21.

[103]Bui N-T, Hoang D-H, Nguyen Q-T, et al. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation[C]. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024: 7985-94.

[104]Zhang X, Liu C, Yang D, et al. RFAConv: Innovating spatial attention and standard convolutional operation [J]. arXiv preprint arXiv:230403198, 2023.

[105]Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[106]Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]. 2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-71.

[107]Kervadec H, Bouchtiba J, Desrosiers C, et al. Boundary loss for highly unbalanced segmentation[C]. International conference on medical imaging with deep learning. PMLR, 2019: 285-96.

[108]Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-24.

[109]Fu Y, Lou M, Yu Y. SegMAN: Omni-scale context modeling with state space models and local attention for semantic segmentation[C]. Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 19077-87.

[110]Zhu Q, Liao C, Hu H, et al. MAP-net: Multiple attending path neural network for building footprint extraction from remote sensed imagery [J]. IEEE Trans Geosci Remote Sensing, 2021, 59(7): 6169-81.

[111]Chang J, He X, Li P, et al. Multi-Scale Attention Network for Building Extraction from High-Resolution Remote Sensing Images [J]. Sensors, 2024, 24(3): 1010.

[112]Huang H, Liu J, Wang R. Easy-net: A lightweight building extraction network based on building features [J]. IEEE Trans Geosci Remote Sensing, 2024, 62: 1-15.

[113]Liu W, Lu H, Fu H, et al. Learning to upsample by learning to sample[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2023: 6027-37.

[114]Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network [J]. Advances in neural information processing systems, 2014, 27.

[115]Rottensteiner F, Sohn G, Gerke M, et al. Results of the ISPRS benchmark on urban object detection and 3D building reconstruction [J]. ISPRS journal of photogrammetry and remote sensing, 2014, 93: 256-71.

[116]Cramer M. The DGPF-test on digital airborne camera evaluation overview and test design [J]. Photogrammetrie-Fernerkundung-Geoinformation, 2010: 73-82.

[117]Wei S, Ji S, Lu M. Toward automatic building footprint delineation from aerial images using CNN and regularization [J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 58(3): 2178-89.

[118]Teng J, Wang F, Liu Y. An efficient algorithm for raster-to-vector data conversion [J]. Geographic Information Sciences, 2008, 14(1): 54-62.

[119]Gröger G, Plümer L. CityGML–Interoperable semantic 3D city models [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2012, 71: 12-33.

[120]Crosetto M, Monserrat O, Cuevas-González M, et al. Persistent scatterer interferometry: A review [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 115: 78-89.

[121]Hu M, Xu B, Wei J, et al. Correcting the location error of persistent scatterers in an urban area based on adaptive building contours matching: A case study of Changsha [J]. Remote Sensing, 2024, 16(9): 1543.

[122]陈祎林, 周云. 一种基于多模态遥感数据的建筑加层智能识别方法及系统, CN120747747B [P]. 2026-01-23.

攻读学位期间的学术或实践成果

1．学术论文：

[1] 周云, **陈祎林**, 李子卿等. 基于光学遥感的城市建筑群三维建模[J]. 建筑结构学报.（已录用）

[2] 周云, 张文杰, **陈祎林**等. 基于无人机单目视频的道路裂缝定位与量化方法[J/OL]. 工程力学, 1-14[2025-04-13].

[3] 周云, 邹少豪, 曹爽, **陈祎林**, 罗星. 基于星载 InSAR 对地观测的城市建筑群变形风险模糊评估理论与应用研究[J/OL]. 建筑结构学报, 1-19[2026-06-10].

2．发明专利：

[1] **陈祎林**, 周云. 一种基于多模态遥感数据的建筑加层智能识别方法及系统: 202510861212.7[P]. 2026-01-23.

[2] 周云, **陈祎林**. 一种基于时序 InSAR 的高层建筑群健康监测方法与系统: 202510767574.X[P]. 2025-12-30.

[3] 周云, **陈祎林**, 陈建炜等. 一种基于 InSAR 位移测量的多颗 SAR 卫星角反射器跟踪控制方法、跟踪控制系统以及跟踪式角反射器: 202510020888.3[P]. 2025-04-25.

[4] 周云, 罗先明, 谈忠坤, **陈祎林**等. 一种计算 RC 构件轴向时变变形参数的智能方法及装置: 202510218137.2[P]. 2025-09-02.

[5] 谈忠坤, 罗先明, 周云, **陈祎林**等. 一种考虑连廊刚度的 RC 构件时变变形智能计算方法: 202510218126.4[P]. 2025-08-15.

[6] 罗星, 黄一钊, **陈祎林**, 周云. 一种建筑白模与时序 InSAR 形变点的匹配方法及系统: 2026101685352[P].（实质审查）

[7] 罗星, 李子卿, 邹少豪, 周云, **陈祎林**. 一种城市建筑形变监测漏检点识别方法及系统: 2026101685348[P].（实质审查）

致谢

不知不觉，我的研究生生涯即将画上句号。这段人生中重要的旅程始于2023年的6月，终于2026年的6月。回顾三年研究生时光，有欢喜，也有失落，有收获，也有遗憾。曾奋力拼搏，力争最好成绩，也曾迷茫无助，看不见前行的方向。但是我从未放弃，这源于一路上大家对我的帮助与支持。

由衷感谢我的导师周云教授，感谢您在科研道路上对我的指导与帮助。您为我提供了许多横向项目机会，让我提升了自身的能力。感谢您不辞辛苦地帮助我获取数据，使实验得以顺利开展并成功落地。

感谢我的父母，是你们的安慰与鼓励，陪我度过无数灰心失望的日子，也让我能够坚强地面对科研中的困难与挫折。

感谢我的挚友邹浩，在无数平凡而略显枯燥的日子里，我们一起畅谈理想与未来，彼此鼓励、相互支撑。还记得那段时光，我们上谈论国际局势，下细品人间百味，一起聊天，玩游戏，让我度过了研究生生涯中最快乐的一段时光。同时也感激你在我伤心低落时给予的鼓励，陪我走过迷茫与困苦的日子。你是一位值得信赖、可以交心的朋友。人生得一知己，此生无憾，能与你同行一程，已是幸事。愿未来各自奔赴山海之时，仍不失今日之赤诚与坦荡，最后愿我们的友谊长存。

感谢甘道平老师和罗先明师兄对我的知遇之恩，让我得以顺利迈向人生的新阶段。感谢张文杰师兄，你是我科研生活中最重要的引路人，让我少走了许多弯路，让我的科研之路更加顺畅。

感谢我的室友马龙博，让我的研究生生活增添了更多欢乐。

感谢师弟李子卿，在我的科研过程中，不遗余力地帮我制作数据集、开展实验、修改论文格式等，有了师弟的帮助，我的科研才能更加顺利，真的是一位非常可靠的师弟。同时也感谢师弟黄一钊帮我制作数据集。感谢课题组各位同门对我的帮助。

感谢自己，当然要感谢自己，在研究生三年时光里，我没有辜负自己，一直在努力奋斗，结局也许并非完美，但感谢自己在一个又一个艰难的日子里没有放弃，始终坚持了下来。

行文至此，落笔为终。

陈祎林

2026年4月2日于湖南大学

答辩委员会名单

答辩委员会	姓名	职称	工作单位
主席	黄远	教授	湖南大学土木工程学院
委员	易伟建	教授	湖南大学土木工程学院
委员	郭杰标	高工	湖南大兴加固改造工程有限公司
秘书	王征	中级其他	湖南大学土木工程学院