融合频域感知与对比学习的无人机影像森林全要素解译方法:FSC-Mask2Former

杨振[1,2,3]，姚宗琦[1,2,3]，张晓丽[1,2,3]

1. 北京林业大学 林木资源高效生产全国重点实验室, 北京 100083;
2. 北京林业大学 精准林业北京市重点实验室,北京 100083;
3. 北京林业大学 森林培育与保护教育部重点实验室,北京 100083

摘 要：单木尺度的树种分布与生境信息是森林生态系统科学管理的重要基础，无人机可见光（RGB）影像具备采集时间灵活、空间分辨率高、获取成本低等优势，为精细尺度的森林监测提供了数据支撑。高空间分辨率影像完整记录了林木的精细轮廓与森林的背景生境，利用全景分割技术对其进行统一解译，能够同步获取森林全要素的提取结果。然而，高空间分辨率影像在高郁闭度森林场景下的全景分割主要存在两方面难点：一是不同树种仅依靠的光谱信息区分度有限；二是传统方法多采用语义与实例分割任务分离的架构，导致前景与背景上下文利用不足，容易引发像素归属冲突。针对上述问题，本研究提出了一种端到端全景分割模型FSC-Mask2Former（Frequency and Supervised Contrastive Mask2Former）。该模型在统一的掩膜分类范式下进行了两项核心改进：（1）引入频域感知注意力模块，通过二维离散余弦变换在频域空间捕捉高频边缘信号，强化模型对树冠微观纹理的细粒度特征提取能力；（2）设计实例感知查询对比头，利用监督对比学习策略施加判别性约束，增加相似树种间的类间特征距离。在广西南宁高峰林场等多个研究区的验证结果表明，该模型的综合全景质量（All PQ）达到57.0%，其中反映目标边界拟合精度的分割质量（SQ）达到76.0%，反映类别区分能力的识别质量（RQ）提高至56.0%。研究结果表明，该模型能有效克服复杂森林场景下的树种混淆问题，实现单木个体与背景生境要素的高精度同步解译，为实现复杂生境下的森林全要素精细化制图提供了一套低成本、高效率的技术方案。

关键词：无人机遥感，全景分割，Mask2Former，频域分析，对比学习，单木识别

中图分类号：TP79;S758.1;TP391.41

引用格式：杨振,姚宗琦,张晓丽.XXXX.融合频域感知与对比学习的无人机影像森林全要素解译方法:FSC-Mask2Former.遥感学报,XX(XX):1-20

YANG Zhen, YAO Zongqi, ZHANG Xiaoli. XXXX. FSC-Mask2Former: A Forest Panoptic Segmentation Method Based on UAV Imagery. National Remote Sensing Bulletin, DOI:10.11834/jrs.20266167]

1 引言

森林生态系统的精准监测（Qiu et al. 2023）是森林生态系统科学管理与可持续经营的重要基础（Waser et al. 2017）。获取单木尺度的树种空间分布及其生长生境信息，对于生物量估算（Adhikari et al. 2021）、碳汇监测（Xu et al. 2023）以及生态系统评价（Kumar et al. 2022）具有重要意义。回顾遥感技术在林业领域的应用，观测手段经历了从航空摄影调查（Ulaby, Li, and Shanmugan 2007）到中高分辨率卫星影像大面积覆盖的历史演进（Chen 2021; Zhong, Dai, et al. 2024）。然而，在面对高郁闭度、结构复杂的天然林或混交林（Lu et al. 2016）时，卫星遥感由于空间分辨率限制（Pimentel et al. 2020）和混合像元效应（Keefe, Zimbelman, and Picchi 2022），难以支撑起单木级别的森林参数提取需求。近年来，小型无人机遥感平台的兴起（Wang, Zhu, and

收稿日期：2026-04-20；**预印本**：XXXX-XX-XX

基金项目：国家重点研发计划(编号:2023YFD2201700);国家自然科学基金(编号:32171779)

第一作者简介：杨振,研究方向为森林树种识别。E-mail：yangzhen2023@bjfu.edu.cn

通信作者简介：张晓丽,研究方向为资源环境遥感。E-mail：zhangxl@bjfu.edu.cn

Yun 2023），实现了由林分整体观测向单木精准解译的尺度转变（Ninomiya 2022），为森林资源的细粒度监测提供了高精度数据源（Nurhayati 2015）。

无人机平台的作用在很大程度上取决于所搭载传感器的物理特性。机载激光雷达能够获取森林的垂直三维结构（Kwong and Fung 2020），在单木定位与树高反演中具有优势（Liu et al. 2013），但在树种精细识别所需的光谱信息上明显不足（Cao 2020）。机载高光谱遥感能够通过数百个连续波段捕获树种间的细微反射差异（Gong， Pu，and Yu 1998），但其昂贵的设备成本与复杂的数据处理流程限制了其大范围推广应用（Zhong，Zhang， et al. 2024）。相比之下，可见光影像虽然仅包含红、绿、蓝三个波段，但其具备厘米级的空间分辨率（Pierdicca et al. 2023）。在森林全要素解译任务中，可见光影像提供的二维高频纹理（Söderkvist 2001）与几何轮廓信息（Gyawali，Aalto， and Ranta 2025），足以支持对单木个体及生境背景的识别需求，是目前森林生态系统科学管理中最具普适性的数据源（Elharrouss et al. 2021）。

随着计算机视觉技术的深入应用，利用深度学习进行遥感影像解译已成为当前的研究热点（Minaee et al. 2021）。然而，复杂的森林场景对现有的算法体系提出了严峻挑战。现有的研究多将解译任务解耦为语义分割与实例分割：语义分割（如DeepLab系列）通过全卷积神经网络提取像元级的特征映射（Long， Shelhamer， and Darrell 2015），侧重于林地、裸土等连片背景要素的整体判别，但在高郁闭度场景下，受限于感受野与池化操作造成的特征平滑，难以有效分离紧密相邻的单木个体；实例分割（如Mask R-CNN）则通过区域候选网络对目标进行定位与掩膜提取（Ren et al. 2015），致力于解决单木的独立性问题，但往往忽略了林木与其生长环境间的空间拓扑联系，且在多任务结果合并时常伴随像素归属冲突。森林是一个由可数单木与不可数生境构成的复合生态系统，单纯的前景提取或背景分类均无法完整表达其复杂的空间逻辑。全景分割技术的出现为解决语义分割、实例分割任务分离架构下遇到的问题提供了新路径（Kirillov et al. 2019； Bi et al. 2023），该技术在统一的特征空间内，同步实现对背景要素的语义分类与单木个体的实例分离，确保了全要素解译的结果一致性（Jia 2023）。

在全景分割算法的发展历程中，研究视角主要经历了从任务组合向统一掩膜分类的转变。早期的研究路径可分为自上而下与自下而上两个流派。以Panoptic FPN为代表的自上而下学派（Lin et al. 2017），其核心逻辑是在实例分割的基础上引入语义分支，通过检测框定位目标后再进行局部掩膜提取，最后利用启发式规则将前景与背景进行合并。然而，此类方法在处理重叠区域时常面临像素归属冲突，难以从底层特征层面实现空间一致性。相比之下，以Panoptic-DeepLab为代表的自下而上学派则侧重于像元级的特征聚类（Zust et al. 2025），通过预测语义标签结合像素偏移量或中心点信息进行实例关联。但在高郁闭度林区影像中，密集的单木分布导致像素偏移量的预测极易受到空间分辨率限制而失效，进而引发严重的实例掩膜粘连。近年来，以Mask2Former为代表的端到端掩膜分类范式脱颖而出（Hu et al. 2025）。该范式通过统一的对象查询（Object Query）机制，将前景目标与背景生境的提取整合在同一预测框架中，从物理逻辑上规避了繁琐的后处理逻辑与候选框依赖，为实现森林全要素的协同解译提供了更具一致性的架构支持（Patel et al. 2023）。

尽管端到端全景分割框架在通用场景下表现优异，但其在高度非结构化的森林影像中仍面临特征适配性问题。由于可见光波段信息有限，背景生境与单木个体在光谱空间内呈现出高度的同谱异物特征。常规模型在特征提取过程中，由连续下采样操作引起的高频信号流失，使得模型难以捕捉到区分不同树种的微观纹理。此外，由于缺乏针对林业特殊类别的判别性约束，模型在处理光谱特征极度近似的相邻树种时，特征表达往往不够清晰，容易产生类别判定偏差。如何在维持背景解译连贯性的同时，针对森林影像的频域特性提升模型的微观特征提取能力，并构建更具判别力的特征表达，是目前森林全要素解译研究中需要解决的关键问题。

针对上述问题，本研究构建了一套适配森林高密度前景实例场景的全景分割模型FSC-Mask2Former（Frequency and Supervised Contrastive Mask2Former），旨在实现复杂生境下背景与单木的高精度同步解译。本研究的核心改进逻辑如下：

（1）引入基于二维离散余弦变换的频域纹理

感知模块。针对森林解译中微观纹理易流失的问题，该模块通过在频域空间捕捉高频边缘信号，直接补偿空间域下采样过程中造成的细节损耗，强化模型对树冠微观纹理的细粒度特征提取能力（Li，Jiang，et al. 2023；Ge et al. 2025）；

（2）设计实例感知查询对比头。利用对比学习策略在特征空间中施加判别性约束，通过增加相邻相似树种间的类间特征距离，提升高密度混交林场景下的识别精度与鲁棒性。

本研究以广西南宁高峰林场作为算法开发与性能评估的核心研究区，开展了系统的对比实验与消融实验。同时，为进一步评估模型的应用潜力，将构建的模型迁移应用于内蒙古根河、安徽绩溪及广西横州等跨纬度、不同生境的林区影像解译中。研究结果旨在为复杂生境下的全要素精细化制图提供有效的技术支撑。

2 研究区与数据

2.1 研究区概况

本研究选取广西南宁高峰林场作为算法研发与性能评估的核心研究区，并选取内蒙古根河、广西横州及安徽绩溪作为模型的迁移应用研究区。各研究区在地理分布、树种组成及地物要素上呈现出明显的差异。

2.1.1 核心研究区：广西南宁高峰林场

高峰林场（东经107°45′—109°38′，北纬22°12′—23°32′）位于广西南宁市北部，地处南亚热带季风气候区，地形以丘陵为主。本研究将该区域作为模型训练、消融实验及性能评估的核心研究区。影像覆盖范围内优势树种包括桉树（Eucalyptus robusta）、杉木（Cunninghamia lanceolata）、红锥（Castanopsis hystrix）以及油茶（Camellia oleifera）。除乔木冠层外，影像还涵盖了低矮植被覆盖区、裸地以及不透水面等地物要素。该研究区地物分布密集，是检验模型在森林场景下全要素解译能力的理想场景。

2.1.2 迁移应用研究区

（1）内蒙古根河（东经120°12′—122°55′，北纬50°20′—52°30′）：代表寒温带针阔混交林生境。研究区内主要优势树种为落叶松（Larix gmelinii）与白桦（Betula platyphylla）。除单木个体外，解译对象包括林区周边的裸地、不透水面、低矮植被覆盖区以及水体。该研究区用于验证模型在典型高纬度寒温带森林场景下的迁移适应性与全要素制图潜力。

（2）广西横州（东经108°48′—109°37′，北纬22°34′—23°05′）：属于高密度南亚热带森林。该区域优势树种涵盖桉树、杉木、火力楠（Michelia macclurei）以及部分典型针叶与阔叶树种，共计5类核心树种前景类别。影像范围内地物结构复杂，除了常规的低矮植被、不透水面与裸土要素外，还包含分布广泛的坑塘与渠系水体。该研究区旨在通过对高郁闭度林业场景进行实验，评估模型在复杂生境下的解译鲁棒性与全场景稳定性。

（3）安徽绩溪（东经118°34′—118°53′，北纬30°00′—30°15′）：代表中亚热带天然次生林，具有极高的物种多样性。该区域共包含15个树种，其中优势树种包括马尾松（Pinus massoniana）、枫香（Liquidambar formosana）、山胡椒（Lindera glauca）、杉木（Cunninghamia lanceolata）以及榉树（Zelkova serrata）。影像解译范围不仅涉及复杂的乔木冠层，还考虑了河流、山塘等复杂水体背景以及林地周围的裸土、不透水面和低矮植被。由于涵盖了复杂的乔木冠层与多样化的生境背景，该研究区被选为验证模型全要素协同提取能力与多目标解译性能的验证区之一。

2.2 数据处理与数据集构建

2.2.1 数据获取

（1）影像获取与无人机航测平台本研究的外业调查与影像获取工作均在晴朗、无风的气象条件下开展，确保了样地调查数据与遥感影像的同步性。航飞任务以预设的调查样地为中心，通过覆盖样地及其外延生境的较大区域进行独立航摄，从而获取对应样地的航测区原始影像。样地布设遵循典型抽样原则，各研究区样地规格统一设置为25m×25m。各研究区的具体调查时间与样地数量如表1所示。

针对内蒙古根河、安徽绩溪及广西横州研究区，飞行平台选用大疆经纬M350 RTK多旋翼无人机，该设备内置RTK模块，能够提供高精度的位置信息。在载荷方面，无人机搭载大疆禅思P1（Zenmuse P1）全画幅相机进行影像采集。禅思P1

配备4500万像素的全画幅CMOS传感器，获取的可见光（RGB）影像能够达到厘米级地面分辨率（Ground Sample Distance，GSD），可清晰记录林冠形态及林隙边界特征。广西南宁高峰林场研究区的影像航测采用大疆Matrice 600 Pro六轴无人机平台，并搭载Nano-Hyperspec高光谱相机。为保证多研究区实验数据的一致性，本研究统一提取了红、绿、蓝（RGB）三个波段构建数据集，以进行后续的处理与解译。具体的无人机飞行及传感器参数如表2所示。

表1　各研究区外业调查情况表

Table 1　Details of field survey in each study area

研究区	外业调查时间	样地/条带数量
广西高峰林场	2022年7月	5个条带样区
内蒙古根河	2024年8月	5个样地
安徽绩溪	2025年8月	5个样地
广西横州	2025年6月	5个样地

表2　无人机平台与传感器参数表

Table 2　Parameters of UAV Platform and Sensors

设备类别	参数名称	核心研究区	迁移应用区
飞行平台	型号	大疆 Matrice 600 Pro	大疆 M350 RTK
	悬停精度(RTK开启)	垂直:±0.1　;±0.1 水平:±0.1　;±0.1	垂直:±0.1　;±0.1 水平:±0.1　;±0.1
	飞行高度	120m	150m
	航向/旁向重叠率	80%/70%	70%/30%
全画幅相机	型号	Nano-Hyperspec	大疆禅思P1
	传感器类型/尺寸	高光谱传感器	全画幅(35.9×24 mm)
	成像波段	270个通道 (仅提取RGB)	可见光(R,G,B)
	地面采样距离(GSD)	约7.5cm/pixel	2cm/pixel

（2）航测区外业核查与真值采集在无人机航飞的同时，在样地内同步开展了外业实地调查。利用手持RTK设备，调查人员对航测区内的主要乔木个体进行坐标定位，记录单木位置，并由专业人员现场核实、登记树种信息。此外，对林下低矮植被、裸土等背景覆盖物的分布范围进行了实地确认与记录。所有外业采集的单木及背景覆盖物实测数据，经内业人工校验后，统一汇总并导出为包含精确地理坐标与类别属性的矢量文件。

（3）倾斜影像空三解算与正射成图本研究的外业航飞采用了倾斜摄影测量（Oblique Photogrammetry）进行数据采集。相较于单一垂直视角的传统航飞，倾斜摄影通过同步获取垂直及多个倾斜视角的影像，能够更完整地记录林冠的三维几何结构，有效减少高郁闭度森林树冠边缘的投影遮挡。在影像预处理阶段，采用大疆智图（DJI Terra）软件对原始多视影像进行统一解算。处理流程主要包括：首先，导入包含高精度RTK定位信息的全视角原始照片，通过多视特征匹配进行空中三角测量（Aerial Triangulation）；其次，基于空三解算结果生成高密度点云与高精度数字表面模型（Digital Surface Model，DSM）；最后，利用DSM对垂直视角影像进行严格的投影差纠正与无缝镶嵌，最终生成各样区独立的高分辨率数字正射影像图（Digital Orthophoto Map，DOM），并统一导出为带有空间参考信息的TIFF格式文件。由于各样地在地理空间上相互独立，每个飞行架次获取的影像已涵盖样地及其外延生境，因此不同架次生成的DOM成果独立参与数据集构建，无需进行跨架次的全局拼接处理。

2.2.2　影像预处理

由于无人机单幅正射影像（DOM）涵盖的空间范围广且空间像素分辨率高，难以直接输入深度学习网络进行计算。本研究采用重叠滑窗技术对各研究区影像进行裁切，切片尺寸统一设定为512×512像素。考虑到森林高郁闭度场景下目标分布极其密集，为保证切片边缘处的树冠几何特征完整，滑窗步长设定为256像素，使相邻图像块间保持50%的空间重叠。该策略确保了每一个单木个体至少在一个完整的切片中呈现，从而提升了

模型在推理阶段的解译连续性。

2.2.3 基于SAM辅助的半自动标注

针对森林场景中单木目标基数大、边缘形态不规则的特点，本研究采用Segment Anything Model（SAM）（Ravi et al. 2024）结合Labelme进行高效的半自动标注工作。首先，通过种子点引导SAM生成单木冠层的初步掩膜；随后，由人工根据可见光影像提供的微观纹理与几何轮廓，对掩膜边界进行校验与修正。该流程重点处理高郁闭度区域的边缘遮挡问题，确保单木树冠掩膜的准确性，在保证标注质量的同时能够提升大规模全景数据集的构建效率。

2.2.4 全景分割分类原则与逻辑

本研究构建的数据集严格遵循全景分割任务的统一分类标准。在解译逻辑上，将所有具有明确独立轮廓的单木个体定义为前景实例（Thing），并为每一个体分配唯一的实例编号（Instance ID）以实现个体间的区分；将连片分布的生境要素定义为背景要素（Stuff）。

在语义类别的映射体系中，模型为不同的树种与地物要素分配了独立的类别标签（Category ID）。具体而言，地物类别从标签1开始顺序编号。前景实例所属的树种类别占据起始的标签序列，背景要素类别紧随其后。此外，考虑到滑动窗口裁剪在影像边缘产生的填充区域，本研究将这些非目标区域统一定义为忽略区域，并将其类别标签统一标记为255（ignore_index）。该策略确保了模型在损失计算阶段能够有效忽略无关区域的干扰，使网络聚焦于目标类别的特征提取。图1展示了全景分割标签示意图，通过不同的颜色与文字标注，清晰界定了前景单木实例与背景生境要素的属性归属与空间布局。

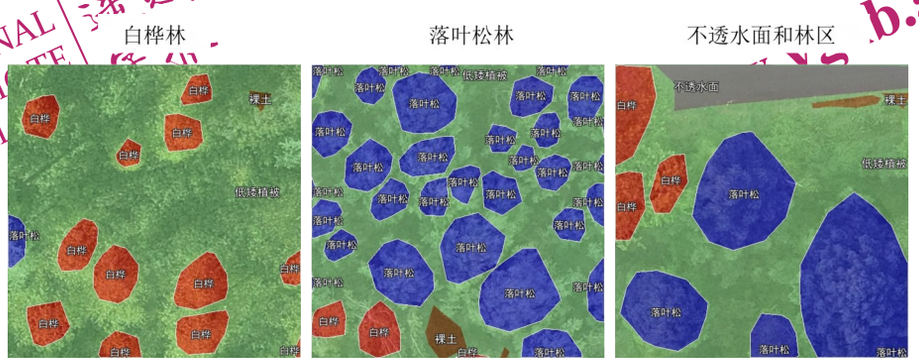

图1 全景分割数据集标签示意图

Fig.1 Illustration of labels in the panoptic segmentation dataset

2.2.5 数据集类别体系与样本统计

高质量的数据集是保障模型鲁棒性的基础。针对不同研究区的生境特征，本研究设计了差异化的类别映射体系。在标注策略上，本数据集采用了全要素密集标注（Dense Annotation），即除影像边缘的空白填充区外，所有有效像元均被赋予了精确的前景实例或背景要素标签。

在数据规模方面，为剔除无解译目标的冗余切片与低质量图像，本研究对原始滑窗切片进行了严格的人工筛选。最终，四个研究区共选取958张高质量切片纳入数据集，累计精细标注前景单木实例7636个。具体的各研究区入选数据量、类别划分准则及标签编码（ID）如表3所示。由于安徽绩溪研究区前景树种较多，该研究区的前景类别明细与样本量如表4所示。

结合表3的量化统计可以看出，在全要素密集标注体系下，各研究区地物的空间分布受到生境异质性的影响，具体呈现出以下结构特征：

（1）前景单木分布呈现长尾效应受自然群落演替规律的影响，各样区内不同树种的样本量呈现出典型的长尾分布（Long-tail Distribution）。特别是在物种多样性最丰富的绩溪研究区（详见表3），少数优势树种占据绝大多数实例，而大量次生树种的样本量较为稀少。

（2）背景生境要素的空间异质性在提取所有独立单木个体后，剩余的生境空间由背景要素完全覆盖。其中，低矮植被构成了各研究区基础的

背景要素，保持了极高的像素连通性；不透水面（主要为贯穿林区的道路网络及林缘建筑物）多呈现形态明确的连续带状或规则区块状分布；裸土要素则相对破碎，多以不规则斑块散布于林隙或人为活动干扰区；水体要素受限于影像的实际拍摄范围，仅在根河、横州与绩溪研究区中出现。

表3　各研究区全景分割数据集类别与样本统计明细表

Table 3　Category and Sample Statistics of the Panoptic Segmentation Datasets Across Study Areas

研究区	数据集数量/张	类别属性	类别明细与样本量	标签 ID
高峰林场	88	前景实例(4类)	红椎(679株)、油茶(1050株)、杉木(660株)、桉树(753株)	1–4
		背景类别(3类)	低矮植被、裸土、不透水面	5–7
内蒙古根河	216	前景实例(2类)	落叶松(555株)、白桦(376株)	1–2
		背景类别(4类)	低矮植被、裸土、不透水面、水体	3–6
安徽绩溪	370	前景实例(15类)	马尾松、枫香等15类（累计1560株，详见表4）	1–15
		背景类别(4类)	低矮植被、裸土、不透水面、水体	16–19
广西横州	284	前景实例(5类)	火力楠(105株)、马尾松(148株)、桉树(1076株)、针叶树(459株)、阔叶树(215株)	1–5
		背景类别(4类)	低矮植被、裸土、不透水面、水体	6–9

表4　安徽绩溪研究区前景单木树种实例数量统计表

Table 4　Statistics of Foreground Individual Tree Species Instances in the Anhui Jixi Study Area

树种名称	数量/株	树种名称	数量/株	树种名称	数量/株
马尾松	415	黄檀	36	漆树	9
杉木	289	山核桃	36	冬青	7
榉树	273	板栗	34	山合欢	4
枫香	248	白檀	22	麻栎	4
山胡椒	170	柿子	10	盐肤木	3

3　研究方法

3.1　网络总体架构设计

Mask2Former作为全景分割中端到端掩膜分类（Mask Classification）方法的经典网络，通过集合预测机制将全景分割简化为对象查询（Object Queries）与预测掩膜的对应匹配。该方法在技术逻辑上避免了传统架构对检测框（Box）及非极大值抑制（NMS）的依赖，同时也克服了密集目标场景下像素聚类易失效的问题。虽然Mask2Former在自然图像解译中表现出较好的场景解析与目标分割性能，但在处理物理结构复杂、光谱特征近似的林区遥感影像时，通用模型由于缺乏对微观纹理的针对性提取能力以及对相似类别的判别性约束，难以达到理想的提取精度。为此，本研究在Mask2Former的基础上，通过对特征提取路径与解码端特征空间的针对性重构，构建FSC-Mask2Former模型（如图2所示），以实现复杂生境下森林全要素的一体化精细提取。

FSC-Mask2Former的整体架构由五个核心组件整合而成：

3.1.1　主干网络与基础特征提取

主干网络（Backbone，如ResNet或Swin Transformer）是整个模型的基础特征提取器。输入的原始高分辨率图像经过骨干网络的逐层下采样计算，提取出包含不同空间分辨率与语义层级的多尺度特征金字塔（通常涵盖原图1/4至1/32比例的特征图）。这些多级特征为后续的像素级解码与目标查询提供了基础的空间与纹理信息。本研究在特征提取路径中引入频域纹理感知模块（FTA），以频域空间的特征增强，补偿下采样过程中造成的高频细节损耗，从而为后续解译过程提供细粒度的边缘特征输入。

3.1.2　多尺度像素解码器

骨干网络输出的多层级特征被送入多尺度像素解码器（Pixel Decoder），其具体网络结构如图3所示。该解码器通常采用多尺度可变形注意力模

块（Multi-Scale Deformable Attention，MSDeformAttn），仅在采样点附近进行局部注意力计算，从而在保持高分辨率特征的同时控制了计算复杂度。像素解码器的核心作用是逐步融合跨尺度信息，并最终输出用于后续掩膜生成的逐像素嵌入（Per-pixel Embeddings）。

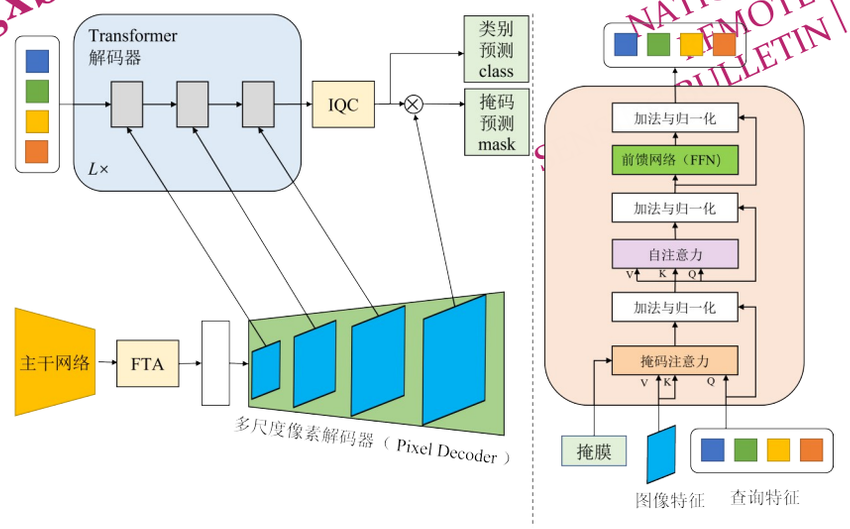

图2 FSC-Mask2Former总体网络结构图

Fig.2 Overall network architecture of FSC-Mask2Former

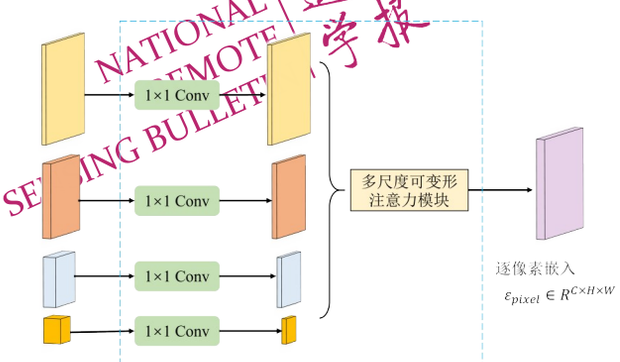

图3 多尺度像素解码器结构图

Fig.3 Architecture of the multi-scale pixel decoder

3.1.3 掩膜注意力机制

传统的Transformer交叉注意力机制需要全局计算所有空间位置的键（Key）与查询（Query）的相关性，计算量较大且易引入背景干扰。Mask2Former提出了掩膜注意力机制，通过引入前一层的掩膜预测结果作为注意力掩码矩阵$\mathcal{M}$，限制交叉注意力仅在目标区域内进行计算。其数学表达式如公式（1）所示：

$$MaskedAttn(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}} + \mathcal{M})V \quad (1)$$

式中，Q、K、V分别代表查询、键和值矩阵；d为特征维度。掩码矩阵M的取值规则为：当空间像素点（x，y）属于前一层预测的二值掩膜区域时，$\mathcal{M}(x,y)=0$；否则$\mathcal{M}(x,y)=-\infty$。这种机制迫使网络高度聚焦于目标物体自身的局部特征，有效加速了模型的收敛过程，如图4所示。

3.1.4 统一的查询机制

网络需要独立的前景区域提取（如RPN）不同，Mask2Former采用了一套统一的对象查询机制（Object Queries）来同时处理前景实例（Thing）与不可数背景（Stuff）。网络将图像中的潜在目标抽象为一组固定数量的N个可学习嵌入向量$Q_{obj} \in R^{C\times N}$（默认设定N=100）。

这组Query在Transformer解码器中与图像特征充分交互后，被转化为包含特定目标信息的实例嵌入qi。随后，分类头输出该目标的类别概率分布pi。在全景分割任务中，类别集合中不仅包含可数的前景类别，还包含连片的背景类别，以及一个特殊的无目标（∅）类别。这种统一的分类框架使得模型无需设计额外的前背景分离模块，掩膜头只需将实例嵌入qi与逐像素嵌入ε_{pixel}进行点乘，如公式（2）所示

$$m_i = Sigmoid q_i \cdot \varepsilon_{pixel} \quad (2)$$

即可直接输出对应前景或背景的二值掩膜预测m_i。

3.1.5 二分匹配与端到端损失

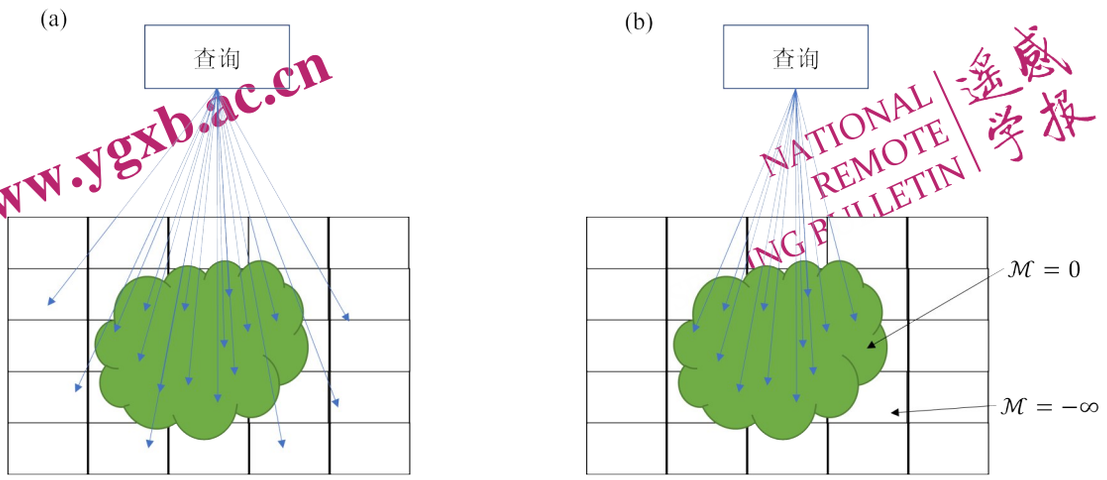

图4　传统交叉注意力和掩膜注意力对比图

Fig.4　Comparison diagram of traditional cross-attention and masked attention

(a)传统交叉注意力机制 (b) 掩膜注意力机制

(a)Traditional cross-attention mechanism (b)Masked attention mechanism

在训练阶段，网络输出的N个预测结果是无序的。Mask2Former采用匈牙利算法（Hungarian Algorithm）在预测集合与真实标注集合之间寻找全局最优的一一对应关系，如图5所示。假设匹配代价函数为$\mathcal{L}_{match}$，网络寻找使总代价最小化的排列组合。

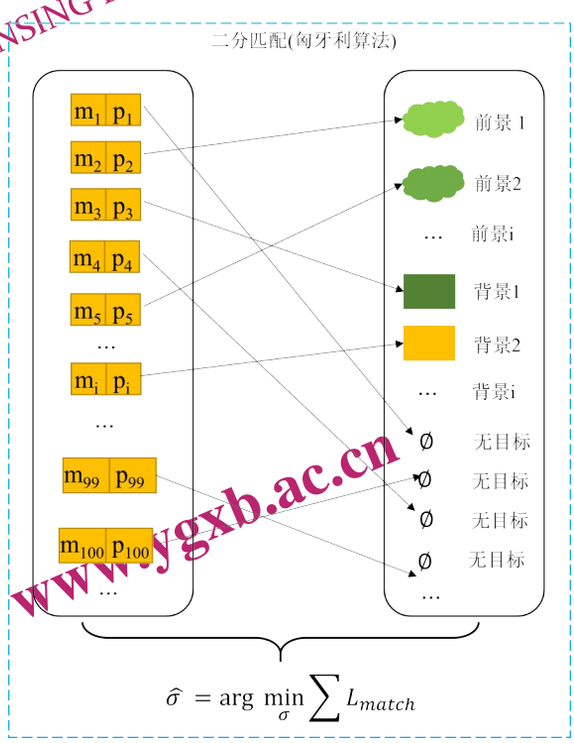

图5　匈牙利算法示意图

Fig.5　Illustration of the Hungarian algorithm

完成最优匹配后，网络对匹配对计算端到端的联合损失函数$\mathcal{L}_{total}$进行反向传播更新权重。该联合损失由分类交叉熵损失$\mathcal{L}_{cls}$、掩膜交叉熵损失$\mathcal{L}_{ce}$以及Dice损失$\mathcal{L}_{dice}$加权组成，如公式（3）所示：

$$\mathcal{L}_{total}=\lambda_{cls}\mathcal{L}_{cls}+\lambda_{ce}\mathcal{L}_{ce}+\lambda_{dice}\mathcal{L}_{dice} \tag{3}$$

式中，λ_{cls}、λ_{ce}和λ_{dice}为平衡各项损失权重的超参数。

完成标签分配后，考虑到森林生境中不同树种光谱特征极其近似，仅依靠常规的交叉熵损失难以在特征空间划定清晰的分类边界。为此，本研究在联合损失体系中针对性地引入了实例感知查询对比头（IQC-Head）。该模块利用对比学习策略对查询向量施加额外的判别约束，结合原有的分类与掩膜损失，增加相似类别在隐式特征空间中的类间特征距离，确保全要素解译在几何边界与逻辑归属上的一致性。

3.2　频域纹理感知注意力（FTA）模块

在林业遥感影像中，不同树种（如特定的针叶树与阔叶树）在宏观颜色与光谱分布上高度重合，区分它们往往依赖于树冠粗糙度、枝叶交错形态等高频微观纹理信息。然而，传统的通道注意力机制（如SE网络等）在获取全局感受野时，通常依赖空间域的全局平均池化（Global Average Pooling，GAP）操作来进行特征压缩。这种物理操作本质上是一种空间低通滤波器，会不可逆地抹平图像中细微的高频纹理差异。为保留并强化

这些关键纹理，本研究设计了频域纹理感知注意力（Frequency-domain Texture Awareness，FTA）模块（如图6所示）。该模块放弃了传统的空间域GAP操作，引入二维离散余弦变换（2D DCT）（Gao et al. 2026），将特征图转换至频域进行处理

$$F(u,v)=\alpha(u)\alpha(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\cos\left[\frac{(2x+1)u\pi}{2M}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{4}$$

其中，$F(u，v)$为频域系数，$\alpha(u)$与$\alpha(v)$为正交归一化系数。

进入频域空间后，FTA模块舍弃了代表背景均值的低频直流分量，专门提取代表树冠边缘与纹理突变的中高频波段系数。通过这种特定频段的增强，强化对细微物理结构差异的特征响应。这种前端特征的频域增强，能够提升模型在宏观特征相似条件下的细粒度特征感知能力。图6所示的频域增强机制，本质上是在特征空间构建了一个可学习的高通滤波器，从而在模型深层映射中保留了森林解译需要的边缘梯度信息。

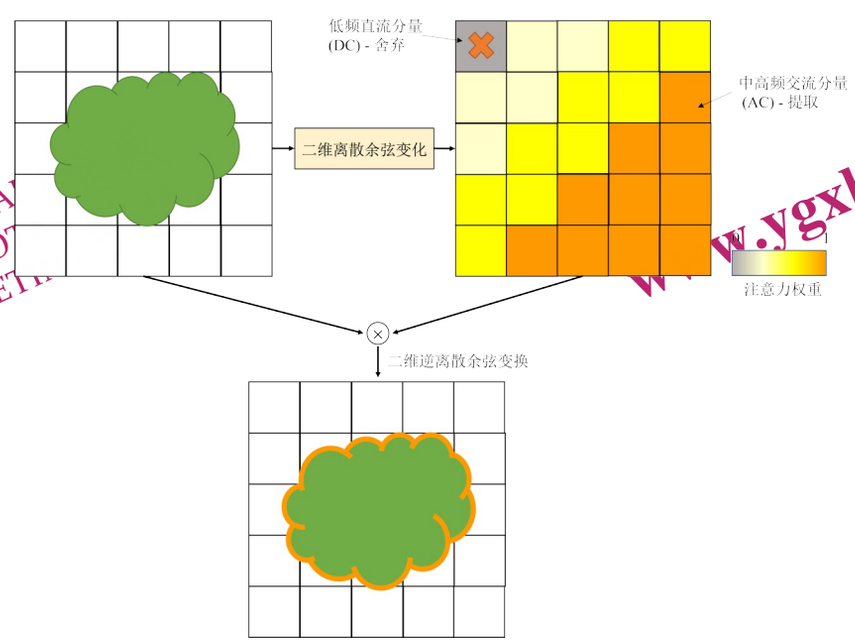

图6 FTA模块示意图

Fig.6 Schematic diagram of the FTA module

3.3 实例感知Query对比头（IQC-Head）模块

在全景分割质量（Panoptic Quality，PQ）的评估体系中，模型性能由分割质量（Segmentation Quality，SQ）与识别分类质量（Recognition Quality，RQ）共同决定。通过对Mask2Former的性能剖析中发现，原网络在划定树冠物理边界时表现尚可，但在区分具体树种类别时存在明显的性能瓶颈，表明网络未能有效拉开不同树种在特征空间中的类间距离。为解决这一特征混淆问题，本研究在Transformer解码器的输出端引入了实例感知Query对比头（Instance-aware Query Contrastive Head，IQC-Head），模块原理示意图如图7所示。该模块在模型训练阶段加入对比学习分支，利用匈牙利匹配的结果，提取与真实目标（Ground Truth）成功匹配的对象查询（Object Queries）高维特征向量（Li，Zhang，et al. 2023；Yuan et al. 2022）。IQC-Head模块采用InfoNCE对比损失函数（Khosla et al. 2020），以锚点Query qi为基准（Gwon et al. 2024），将同属一类的Query设定为正样本k^+，将属于其他树种或背景的Query设定为负样本集合$\{k_j^-\}$。其对比损失$\mathcal{L}_{IQC}$的数学表达如公式（5）所示：

$$\mathcal{L}_{IQC}=-\log\frac{\exp(q_i k^+/\tau)}{\exp q_i k^+/\tau+\sum_{j=1}^{k}\exp(q_i k_j^-/\tau)} \quad (5)$$

式中，τ为调节对比惩罚强度的温度系数（Temperature Parameter）；k为负样本数量。通过在Query层面施加该对比约束，网络在反向传播中被强制要求排斥不同种类的单木特征，并拉近同类单木的特征距离，从而在隐式特征空间中实现了类别的有效区分（Chen and He 2021）。图tu通过对比学习实现了特征分布的各向异性约束，让模型在隐式空间中拉大相似树种间的决策边界，从根本上抑制了识别过程中的类别归属冲突。

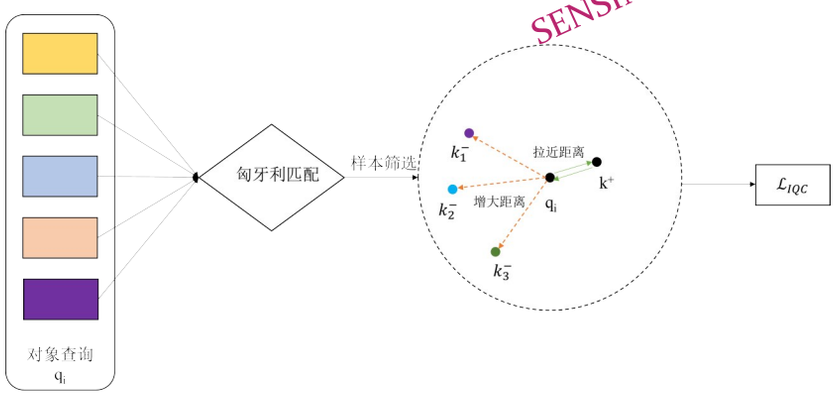

图7　IQC模块示意图

Fig.7　Schematic diagram of the IQC module

3.4　实验设置与评价指标

3.4.1　实验环境与参数设置

在软件环境方面，模型使用PyTorch（版本1.12.0）深度学习框架，根据全景分割任务的架构需求，引入了MMDetection3.3.0目标检测与分割代码库进行联合构建。为匹配网络输入要求，遥感影像切片尺寸统一设定为512×512像素。

在网络超参数与训练调度方面，全景分割模型选用以ResNet-50为主干网络（Backbone）的Mask2Former架构。针对林区单株树木密集分布的空间先验特征，本研究对底层参数进行了林业场景下的适配性调参，将网络的对象查询向量（Object Queries）数量由默认的100组显式扩容至300组。模型最大训练总轮次（Epoch）设定为200轮，为适配实验硬件的显存限制，批处理大小（Batch Size）设为2。为保证Transformer架构在训练过程中的稳定收敛，网络采用MultiStep策略进行学习率调度：在模型迭代至总轮次的80%（第160轮）与95%（第190轮）时，学习率分别按0.1的衰减系数（Gamma）进行步进式下降，以促使网络在训练后期平滑逼近全局最优解。

3.4.2　评价指标

全景分割任务要求在统一的框架下同时评估不可数背景的语义分割精度，以及可数前景的实例定位与分割精度。因此，传统的mIoU（Mean Intersection over Union）或AP指标难以全面衡量模型性能。本研究采用全景分割领域的标准评价体系——全景质量（Panoptic Quality，PQ）及其衍生指标。全景评价体系的核心在于实例匹配机制：给定预测掩膜集合与真实标注掩膜集合，只有当预测掩膜p与真实标注掩膜g的交并比满足IoU（p，g）>0.5时，该预测才被认定为匹配成功，记为真正例TP；未能匹配的预测掩膜记为假正例FP；未能被模型预测出的真实掩膜记为假负例FN。基于此，核心评价指标定义如下：

（1）分割质量（Segmentation Quality，SQ）分割质量主要衡量模型在已经成功匹配（即分类且大致定位正确）的目标中，其掩膜边界的精细程度。SQ值越高，说明模型对树冠轮廓和林地边界的刻画越贴合真实物理形态。其计算公式为成功匹配实例的平均IoU，如公式（6）所示：

$$SQ=\frac{\sum_{(p,g)\in TP}IoU(p,g)}{|TP|} \quad (6)$$

（2）识别质量（Recognition Quality，RQ）识别质量相当于传统评价体系中的F1-Score，主要衡量模型对目标的检出与分类能力，直接反映了模型在密集森林场景中是否存在漏检（FN）或误

检/类别混淆（FP）。其计算公式如公式（7）所示：

$$RQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{7}$$

（3）全景质量（Panoptic Quality，PQ）全景质量是全景分割的综合性评价指标，同时兼顾了像素级的分割精度（SQ）与实例级的识别准确率（RQ）。其在数学上严格定义为SQ与RQ的乘积，计算公式如公式（8）所示：

$$PQ = SQ \times RQ = \frac{\sum_{(p,g)\in TP} IoU(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \tag{8}$$

此外，为更细致地分析模型对不同类型地物的提取表现，将PQ进一步拆分为针对不可数背景类别的stuff PQ和针对可数前景个体的things PQ。同时，针对高郁闭度混交林中存在的树种误判问题，本研究还引入了各前景树种的识别质量（RQ）作为精细化评估指标，以深入解析模型对不同树种个体的分类判别精度。

3.4.3 对比实验模型选取

为了全面、客观地评估本文提出的FSC-Mask2Former模型在复杂高郁闭度森林全景要素提取中的综合性能，本研究选取了三种具有代表性的技术路线进行对比。这三种模型严密反映了全景分割任务从任务分离组合、自上而下、自下而上聚类到端到端统一的技术演进脉络，具体选取依据如下：

（1）独立分支后融合框架有别于端到端的统一网络，独立分支后融合是一种典型的两步式全景分割策略。该策略在技术路线上将全景任务物理拆解为两个平行的子任务：首先分别利用独立的语义分割网络提取不可数背景，利用实例分割网络提取可数前景个体；随后，依赖人工设定的规则对两个分支的输出进行合并，以处理交界处的分类冲突。其具体实现方案为：使用DeepLabV3+模型进行背景提取，同时使用Mask R-CNN模型进行单木实例分割，最后按照标准启发式规则实现掩膜融合。将此分步处理策略纳入对比实验，旨在客观验证端到端全景架构在消除像素归属冲突、提升复杂边界拓扑一致性方面，相比于传统任务解耦模式所具备的架构优势。

（2）Panoptic FPNPanoptic FPN是早期的自上而下双分支全景分割基准模型（Lin et al. 2017）。该网络分别利用Mask R-CNN进行实例分割，利用语义分支进行背景分割，最终通过启发式规则解决像素的分类冲突，选用该模型进行对比，用于验证本文端到端架构在处理前景单木与背景林隙时，无需依赖人工规则即可实现像素级精确匹配的结构优势。

（3）Panoptic-DeepLabPanoptic-DeepLab是基于自下而上（Bottom-up）策略的全景分割模型（Zust et al. 2025）。该网络放弃了提议框生成过程，通过预测像素语义标签以及指向实例中心的偏移量（Offset），利用聚类算法对单木进行实例分组。考虑到林业场景中单木树冠形态不规则且密集交叠，中心点回归往往容易产生位置偏移。引入该模型进行对比，可有效检验本文基于Transformer掩膜注意力机制在处理不规则边缘与较高密度实例分离时的鲁棒性。

4 实验结果

4.1 对比实验结果

各模型在相同测试集上的全景分割精度如表5所示。不同网络架构在高郁闭度森林下的分割表现存在较大差异。独立训练后融合框架的整体性能垫底，其综合全景质量（all PQ）仅为41.0%，且在前景分割边界的刻画上表现最差（SQ仅为60.0%）。自下而上聚类的Panoptic-DeepLab模型在背景的提取上取得了76.0%的高分（stuff PQ），但在前景单木的提取上表现极度乏力，其things PQ仅为20.0%，前景识别质量（RQ）更是低至30.0%，位列所有模型的最低。Panoptic FPN各项指标相对均衡，取得了43.0%的综合PQ，但整体精度上限不高。

采用集合预测机制的Mask2Former模型在前景边界的拟合上展现出优势，其SQ达到了75.0%；但该模型在背景上的表现（stuff PQ为61.0）甚至不及早期的Panoptic-DeepLab，且RQ为44.0%依然处于较低水平，导致其最终的综合all PQ仅停留在46.0%。

相比之下，本文构建的FSC-Mask2Former在所有核心指标上均取得了最优表现。在all PQ上，改进模型达到了57.0%，较基线模型提升了11.0个百分点。具体到各子任务：在背景提取方面，

stuff PQ由基线的61.0%跃升至79.0%；在前景提取方面，things PQ达到42.0%。更为关键的是，改进模型的RQ达到了56.0%，较基线网络提高了12.0个百分点，同时在SQ上维持了76.0%的高水准。

表5　主流全景分割模型提取精度对比（%）

Table 5　Comparison of Extraction Accuracy Among Mainstream Panoptic Segmentation Models (%)

	全景质量（all PQ）	背景质量（stuff PQ）	前景质量（things PQ）	前景识别（RQ）	前景分割（SQ）
独立训练后融合	41	69	35	43	60
Panoptic DeepLab	43	76	20	30	65
Panoptic FPN	43	62	30	43	70
Mask2Former	46	61	34	44	75
FSC-Mask2Former	57	79	42	56	76

表6　不同对比实验模型的前景树种分类质量（RQ）评价（%）

Table 6　Comparison of RQ for foreground individual tree species among different models (%)

	油茶	杉木	桉树	红锥
独立训练后融合	38.01	32.23	71.54	33.03
Panoptic DeepLab	25.77	22.12	36.92	28.4
Panoptic FPN	42.65	31.19	69.41	31.37
Mask2Former	42.67	34.1	69.57	32.11
FSC-Mask2Former	46.5	50.49	73.17	56.06

图8展示了不同全景分割模型在复杂林业场景下的可视化提取效果。由图8可知，各算法在处理交叠目标与非结构化背景时呈现出不同的视觉特征。Panoptic-DeepLab的预测掩膜中存在明显的实例粘连现象，大量相邻的单木实例个体被错误归并为单一的连片图斑。Panoptic FPN在林木类别分配上存在错分实例，且在道路等线性地物上发生了物理断裂。Mask2Former基线模型生成的掩膜边缘相对平滑，但在高郁闭区存在局部漏检，且常将空间邻近的不同树种赋予相同的实例标签。这种视觉层面的类别混淆在表6的定量评估结果中得到了进一步证实。受限于极度相似的光谱特征，对比实验模型在红锥和杉木上的分类精度普遍较低，其中Mask2Former在红锥上的全景质量仅为32.11%（见表6），这反映了该模型在复杂混交林中难以精准划定类别判定的特征边界。独立训练后融合框架的输出结果在前景与背景衔接区域存在几何重叠，掩膜边缘较为粗糙。相比之下，FSC-Mask2Former准确分离了树冠交叠的单木个体，保留了林隙间细小裸土与低矮植被斑块的独立轮廓，并确保了道路等大型背景地物的形态完整性，其生成的全景掩膜在几何形态与属性类别上均为最佳。

4.2　消融实验结果

为进一步探究本文提出的频域纹理感知模块（FTA）与实例感知对比头（IQC）对模型精度的具体贡献，本研究在Mask2Former基线模型的基础上，进行核心模块的分离与组合测试。

实验在保持相同超参数与训练策略的条件下进行，定量结果如表7所示。

由表7的精度指标数据可知，任何单一改进模块的引入，均能为基线模型带来直接的精度提升。在引入FTA模块后，精度指标最直观的改变体现在stuff PQ上，该指标由61.0%大幅提升至78.0%，涨幅高达17个百分点；同时，all PQ也提高至57.0%。在单独引入IQC模块后，RQ由基线的44.0%提升至53.0%，things PQ也上升至39.0%。

当在网络中同时加入FTA与IQC两大模块后，模型达到了性能的最高点。与单一模块的模型变体相比，FSC-Mask2Former在stuff PQ与things PQ上均实现了进一步的提高。特别是在RQ上，最终模型达到了56.0%的峰值。这表明各项优化机制在同一架构下不仅正常运转，且产生了正向的叠加效应。

消融实验的可视化推理结果（图9）直观展现了各模块加入后模型输出掩膜的形态变化。在包含道路等线性地物的研究区影像中，基线网络（Mask2Former）输出的道路掩膜边缘呈现明显的锯齿感，并伴随像素空洞；引入FTA模块后，道路图斑的边缘变得平滑且形态连续，同时，裸土、林隙等背景要素的边界由基线方案中的模糊连片状态转变为清晰定义的独立斑块。在不同树种交

错分布的混交林区域，基线方案在光谱相近的杉木与油茶个体间出现较多的类别标签误判；引入IQC模块后，单木个体的类别归属标签与真实林相特征保持一致，不同树种间的实例独立边界得到明确界定，这种视觉上的纠正效果直接反映在分树种分类质量的提升上（见表8）。通过引入实例感知查询对比机制，红锥的分类质量RQ从Mask2Former的32.11%提升至47.13%，杉木也从34.10%提高到47.06%。这进一步证实了该模块能有效拉大特征空间中的类间距离，从而从特征表征层面抑制了由同谱异物导致的误判现象。最终整合两大模块的FSC-Mask2Former在保证单木个体轮廓清晰准确的同时，较为完整还原了林地要素间的空间分布关系。

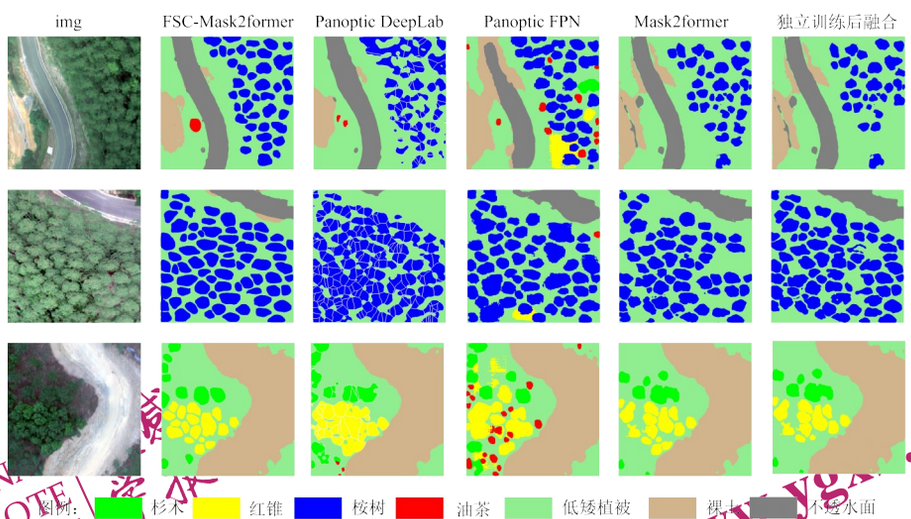

图8 各方法推理结果可视化对比图

Fig.8 Visual comparison of inference results among different methods

表7 核心优化模块消融实验精度对比（%）

Table 7 Accuracy Comparison of Ablation Studies on Core Optimization Modules (%)

	全景质量（all PQ）	背景质量（stuff PQ）	前景质量（things PQ）	前景识别（RQ）	前景分割（SQ）
Mask2Former	46	61	34	44	75
FTA-Mask2Former	57	78	39	52	73
IQC-Mask2Former	56	78	39	53	76
FSC-Mask2Former	57	79	42	56	76

4.3 全景效果展示

结合无人机航拍数据的空间分布特征，本研究选取高峰林场核心研究区内多个条带连续影像，对FSC-Mask2Former的大范围制图表现进行可视化验证。

由图10可见，在处理长条带空间尺度的影像时，模型输出的全景掩膜在语义属性与形态结构上表现出较好的连续性。针对复杂的乔木冠层，模型准确分割了不同的优势树种个体，各树种的单木实例掩膜在图上准确反映了实际各树种的群落聚集特征。同时，制图结果对贯穿林区的蜿蜒道路、边缘不规则的裸地以及大面积低矮植被等背景生境要素进行了完整的形态刻画。前景单木实例与背景要素图层之间形成了明确的物理边界，未出现大面积的像素错分或逻辑冲突。多个条带全景分割制图结果客观显示，模型在面对实际航拍样带中不断变化的林相结构时，依然能够稳定产出类别归属清晰、地物要素完整的全景分割结果。

4.4 多区域迁移实验结果

为验证FSC-Mask2Former在跨纬度、多生境条件下的迁移适应性，本节将模型应用于内蒙古

根河、安徽绩溪及广西横州三个独立验证区，并针对各区域局部的典型林相进行了重新训练与评估。各验证区全景分割的核心定量评价精度如表9所示。

由表9的精度指标可知，模型在应对异质性森林结构时呈现出符合生境物理特征的性能差异。在内蒙古根河验证区，模型取得了最优的全景分割综合精度，其all PQ达到60.58%。特别是在树冠形状评价上，该区域的things SQ高达77.37%，这主要是因为该研究区森林郁闭度较低，单木边缘受遮挡程度较低，模型能够精确地还原树冠物理形态。在广西南宁横州验证区，面对高郁闭度人工林，模型的树种分类识别精度things RQ达到48.66%，表明网络在密集重叠场景下依然保持了较强的单木实例分类能力；但受限于严重的冠层空间遮挡，单木轮廓的精细拟合难度增加，导致things SQ降低至58.78%，all PQ为55.09%。在安徽绩溪验证区，由于山区地形起伏剧烈且林木散布极不规则，背景生境与前景要素的提取难度同步增加，all PQ为51.39%，但模型仍保持了基础的场景全要素解析能力。

多区域全景分割的可视化推理结果（图11）展示了改进模型在不同地貌特征下的输出表现。在根河验证区，模型提取了贯穿林地的直线便道及两侧背景，并输出了边界独立的单木实例掩膜。在地形起伏明显的绩溪验证区，面对随山势不规则散布的林业场景，模型保持了蜿蜒道路等地物的形态连贯，并精确定位了不同尺度树冠的中心位置。在郁闭度极高的横州验证区，模型在密集的冠层交错区有效抑制了掩膜粘连现象，实现了单木轮廓的个体分离，同时对水体、裸土等背景给出了清晰的边界定义。

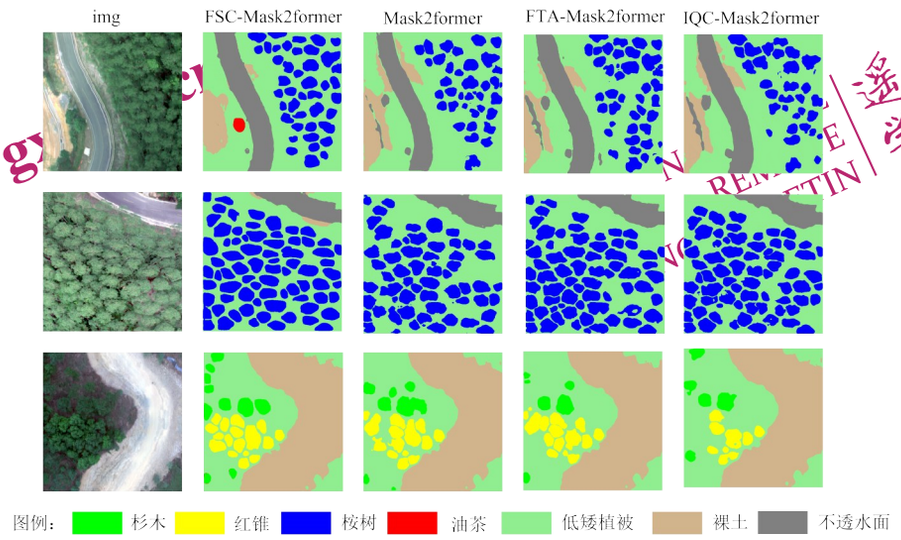

图9 消融实验推理结果可视化对比图

Fig.9 Visual comparison of inference results in the ablation study

表8 核心改进模块对各树种分类质量（RQ）提升的消融分析（%）

Table 8 Ablation study on the impact of core improvement modules on per-species RQ(%)

	油茶	杉木	桉树	红锥
独立训练后融合	42.67	34.1	69.57	32.11
Panoptic DeepLab	45.57	42.49	71.95	48.2
Panoptic FPN	46.15	47.06	72.73	47.13
Mask2Former	46.5	50.49	73.17	56.06

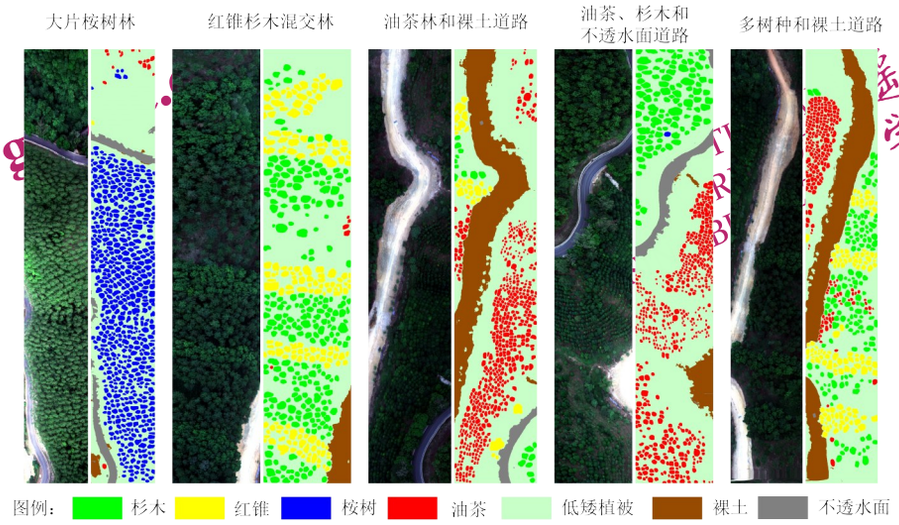

图10　FSC-Mask2Former大范围全景分割效果图

Fig.10　Large-scale panoptic segmentation results of FSC-Mask2Former

表9　FSC-Mask2Former在不同验证区的全景分割精度（%）

Table 9　Panoptic Segmentation Accuracy of FSC-Mask2Former in Different Validation Areas (%)

	全景质量（all PQ）	背景质量（stuff PQ）	前景质量（things PQ）	前景识别(RQ)	前景分割（SQ）
内蒙古根河	60.58	71.90	37.96	49.11	77.37
安徽绩溪	51.39	66.96	31.14	41.60	63.61
广西横州	55.09	69.23	33.99	48.66	58.78

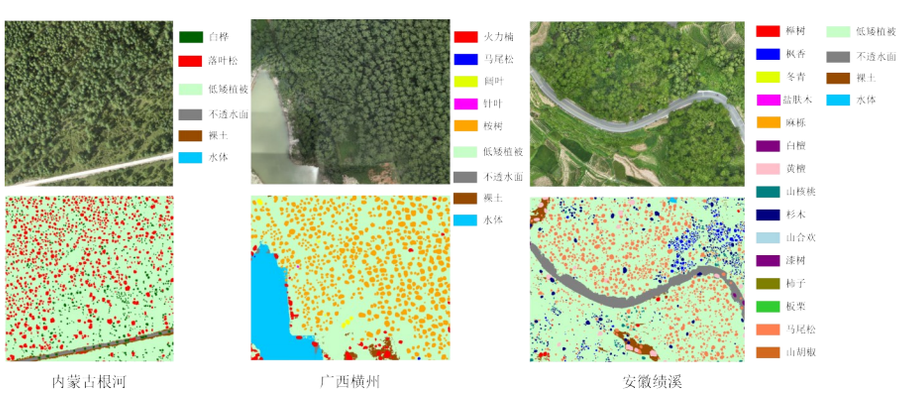

图11　改进模型在三大验证区的全景分割可视化结果

Fig.11　Visualization results of land cover classification by the improved model in the three validation areas

5　讨论

5.1　不同全景分割架构对森林场景的适配性分析

从表5的对比实验精度指标可以看出，不同架构在面对高郁闭度森林时表现出明显的精度差异。这种精度表现上的差异，本质上源于各算法内部的组织形式对森林生境复杂性的适应能力不同。现有的多数森林影像解译研究仍受限于单一维度的目标提取。传统的图像分类和语义分割方法侧重于获取林地、草地、裸土等连续生境地物的宏观分布，但无法有效区分单木个体，难以满足单木尺度的精细监测需求。而实例分割虽能实现单木的有效分离，却将林隙、林下植被、裸土等背景区域视为无效信息舍弃，造成关键生境特征的流失。这种任务分离的固有局限，使得传统方法难以实现森林全要素的完整表达。

独立训练后融合方法虽试图弥补这一缺陷，

但其与Panoptic FPN的all PQ较低，分别为41.0%和43.0%。这两类方法在处理树冠边缘与林隙交界等复杂区域时，由于背景语义与单木实例任务在特征层面是分离的，预测结果经常出现像素归属上的冲突，即便采用人为设定的逻辑规则进行物理合并，也难以准确还原真实的林地空间拓扑。而采用自下而上聚类机制的Panoptic-DeepLab的stuff PQ虽达到76.0%，但RQ仅为30.0%。这反映了偏移量回归机制在森林场景下的局限性。在高郁闭度林区，相邻树冠质心的物理距离较近，像素偏移量往往难以在受限的特征空间内精确锁定独立的单木树冠中心，从而引发严重的实例粘连，导致其在前景个体提取上的表现位列所有模型末端。Mask2Former通过掩膜分类机制，在前景边界刻画上展现出优势，其SQ达到了75.0%。然而，该模型在背景提取（stuff PQ 61.0%）和前景识别（RQ 44.0%）上仍存在短板。这说明在特征提取阶段，仅依靠空间域的下采样操作会平滑掉区分地物要素的关键纹理；同时，常规的分类约束也无法有效拉开相似树种间的特征距离，导致模型在面对光谱特征相近的个体时容易发生分类偏差。图8观测到的标签错位情况与表6中的各树种精度指标结果相符合。受限于相似的光谱特征，对比实验模型在红锥和杉木上的识别质量普遍较低。这说明光谱信息的局限性使得传统掩膜分类机制在复杂混交林中无法有效划定类间特征边界，进而引发严重的类别混淆。

相比之下，FSC-Mask2Former的all PQ指标达到57.0%，较基线模型提升了11.0个百分点。其中，反映实例判别能力的RQ指标从44.0%提升至56.0%，背景提取精度stuff PQ达到了79.0%。上述精度提升得益于频域特征补偿（FTA）与特征空间分离（IQC）的共同作用，该机制从特征映射的底层阶段，有效克服了通用架构固有的纹理平滑与类别混淆缺陷。结合基准区的多个条带制图与多个研究区（图10、图11）的解译结果，模型在应对跨纬度、高异质性的复杂林相结构时，均保持了稳定的泛化表现。这表明FSC-Mask2Former具有较强的特征提取鲁棒性，能够有效克服环境变化对树种分类的干扰。无论是条带形研究区的整体解译，还是复杂地形下的大尺度全要素分割，该模型均能确保全要素解译在几何边界与逻辑归属上的空间连贯性。

5.2 核心改进模块的机理分析

针对Mask2Former网络在复杂森林生境下的局限性，本研究引入的各优化模块的具体机理分析如下：

频域特征转换有效解决了背景要素在特征下采样过程中的细节流失问题。在传统的卷积架构中，连续的空间池化操作本质上具有低通滤波效应，导致图像中细碎裸土、林间小道等非结构化背景的高频信号发生不可逆的损耗，反映在解译结果上即为道路断裂或背景边缘模糊。FSC-Mask2Former引入的FTA模块通过二维离散余弦变换在频域空间捕获边缘梯度信息，从底层补偿了空间维度的特征损耗。这种结合空间与频域的特征表示机制，使得网络在进行深层映射时，既能够保留常规卷积提取的宏观语义信息，又能将反映树冠边缘与林隙的细粒度纹理有效融入其中，从而避免了小尺度生境要素在多尺度特征金字塔传递中的同化。广西高峰林场的实验结果显示，该模块使得模型在处理不规则背景地物时拥有更完整的结构信息，stuff PQ由Mask2Former的61.0%提升了17.0个百分点，确保了森林背景解译的连贯性。

实例对比约束机制则降低了相似树种间的分类误判率。针对可见光影像中不同树种光谱特征相似的难点，IQC模块在解码阶段引入了对比损失函数，通过有监督的对比学习，增大了不同树种在特征空间中的类间特征距离。该机制通过度量空间的重塑，使得相同树种的查询向量相互靠近，同时推斥光谱响应极其相似的异类特征，克服了传统方法单纯依赖局部像素交叉熵进行分类的局限性。该机制减轻了同谱异物现象带来的特征混淆，使得识别指标RQ提升了9.0个百分点（由44.0%提升到53.0%），在视觉效果上纠正了杉木与油茶等相似林种的标签错乱，提高了单木的分类准确率。单木类别属性划分的准确度提升在表8统计的各树种RQ指标中表现得十分明显。引入IQC模块后，原本极易发生标签混淆的红锥与杉木的分类质量均获得了大幅改善。这说明该模块通过优化特征表达，减少了由于同谱异物现象造成的识别错误。

这种前端细节补偿与后端特征空间重塑形成了有效的协同优化。FTA模块提供的清晰空间边

界为后端的IQC模块提供了高质量的形态基础，而IQC模块则在清晰的掩膜基础上精准界定了语义属性。在两者的共同作用下，FSC-Mask2Former在提高前景单木识别精度的同时，也保证了背景生境提取的完整性，提高了高郁闭度森林场景下全景分割的整体性能。上述底层特征感知与后端语义判别的协同优化，有效兼顾了前景单木的精细化提取与背景生境的连续性表达，突破了高空间分辨率林业影像全要素解译的精度瓶颈，为复杂森林的精细化监测提供了可靠的技术支撑。

5.3 局限性与未来展望

尽管FSC-Mask2Former在多个研究区的无人机影像中均表现出较好的全要素提取效果，但在复杂的森林全景解译工程中仍存在改进空间。

目前的模型主要依赖无人机高空间分辨率的可见光影像，在应对光谱特征极度相似且交叠严重的树种时，依然存在一定的误判概率。特别是在光照条件复杂的森林内部，冠层阴影会对掩膜的边缘确定产生干扰，导致部分受遮挡的个体在空间结构上出现欠分割现象。此外，全景分割任务对高质量标注数据的依赖程度极高，单木水平的像素级标注耗费大量的人力与时间成本，这在一定程度上限制了模型在更大范围、更多样化森林生境中的快速迁移与应用。

未来的研究将重点探索多源遥感数据的深度融合。一方面，可以尝试引入激光雷达（LiDAR）点云数据提取森林的垂直结构信息，借助三维高度信息修正可见光特征的误判，从而实现高郁闭度森林下相邻树冠的精准分离。另一方面，针对高质量标注样本稀缺的问题，可以进一步研究半监督学习或弱监督学习机制，利用少量的精确掩膜样本引导模型学习大规模未标注影像中的通用特征，降低对精细标注的依赖。此外，将全景分割技术与时间序列影像相结合，监测林木在生长周期中的空间演变过程，以及在后续研究中引入混淆矩阵开展具体的树种级定量精度评估，以进一步探明极端复杂环境下特定类别的误判机制，也将是未来实现森林精准监测的重要方向。

6 结 论

针对复杂森林全景分割任务，本文构建了全景分割模型FSC-Mask2Former，并在多个研究区完成了模型验证。实验结果表明，该模型在综合全景质量（all PQ）上达到57.0%，相较于主流全景分割架构具备更优的全要素解译能力。频域特征补偿模块通过在频域内捕获中高频空间分量，有效弥补了下采样过程导致的空间信息流失，背景提取精度（stuff PQ）达到79.0%，缓解了背景要素掩膜不连贯的情况。实例特征区分机制利用对比学习策略增加类间特征距离，将前景识别质量（RQ）提升至56.0%，降低了光谱相似树种的类别误判现象。在不同研究区的应用结果证明，FSC-Mask2Former具备较好的泛化能力，能够保持掩膜几何质量与语义属性的一致性，充分验证了依托常规可见光（RGB）影像实现高精度森林全要素精细化制图的可行性，为大规模森林生态系统的科学管理提供了兼具经济性与可靠性的技术支撑。

参考文献（References）

Adhikari A., Kumar M., Agrawal S., and S. R. 2021. "An integrated object and machine learning approach for tree canopy extraction from UAV datasets." Journal of the Indian Society of Remote Sensing 49 (3):471-8.

Bi Y., Zheng Y., Shi C., Zhang K., and Liu J. 2023. "Review of image panoptic segmentation based on deep learning." Journal of Frontiers of Computer Science & Technology 17 (11). (毕阳阳,郑远帆,史彩娟,等.基于深度学习的图像全景分割综述[J].计算机科学与探索,2023,17(11):2605-2619. [DOI: 10.3778/j.issn.1673-9418.2304063])

Cao K. 2020. "Individual tree species classification combining airborne CCD imagery and LiDAR point cloud data." Beijing Forestry University. (曹凯利.融合机载CCD影像和LiDAR点云数据的单木树种分类[D].北京林业大学,2020. [DOI:10.26949/d.cnki.gblyu.2020.000742.])

Chen L. 2021. "Application and prospect of remote sensing technology in forestry in China." China High-Tech (03):141-2. (陈立舟.遥感技术在我国林业中的应用与展望[J].中国高新科技,2021,(03):141-142.[DOI: 10.3969/j.issn.2096-4137.2021.03.063])

Chen X., and He K. 2021. Exploring simple siamese representation learning. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Elharrouss O., Al-Maadeed S., Subramanian N., Ottakath N., Almaadeed N., and Himeur Y. 2021. "Panoptic segmentation: A review." arXiv preprint arXiv:2111.10250.

Gao N., Du X., Xu P., Gao E., and Yang Y. 2026. "High-Resolution Mapping Coastal Wetland Vegetation Using Frequency-Augmented Deep Learning Method." Remote Sensing 18 (2):247.

Ge X., Qi L., Yan Q., Sun J., Zhu Y., and Zhang Y. 2025. "Enhancing Real-Time Aerial Image Object Detection with High-Frequency

Feature Learning and Context-Aware Fusion." Remote Sensing 17 (12):1994.

Gong P., Pu R., and Yu B. 1998. "Identification and analysis of coniferous species with hyperspectral data in different seasons." Journal of Remote Sensing 2 (3):211-7.(宫鹏,浦瑞良,郁彬 . 不同季相针叶树种高光谱数据识别分析[J]. 遥感学报,1998,(03):211-217.)[DOI：10.11834/jrs.19980310]

Gwon M.-G., Um G.-M., Cheong W.-S., and Kim W. 2024. Instance-aware contrastive learning for occluded human mesh reconstruction. Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Gyawali A., Aalto M., and Ranta T. 2025. "Tree Species Detection and Enhancing Semantic Segmentation Using Machine Learning Models with Integrated Multispectral Channels from PlanetScope and Digital Aerial Photogrammetry in Young Boreal Forest." Remote Sensing 17 (11):1811.

Hu J., Cao L., Jin X., Zhang S., and Ji R. 2025. "Universal Image Segmentation With Efficiency." IEEE transactions on pattern analysis and machine intelligence.

Jia B. 2023. "Research on panoptic segmentation network based on feature enhancement." 硕士, North University of China.(贾博慧 . 基于特征增强的全景分割网络研究[D]. 中北大学,2023.)[DOI: 10.27470/d.cnki.gbfgc.2023.001129]

Keefe R. F., Zimbelman E. G., and Picchi G. 2022. "Use of individual tree and product level data to improve operational forestry." Current Forestry Reports 8 (2):148-65.

Khosla P., Teterwak P., Wang C., Sarna A., Tian Y., Isola P., Maschinot A., Liu C., and Krishnan D. 2020. "Supervised contrastive learning." Advances in neural information processing systems 33: 18661-73.

Kirillov A., He K., Girshick R., Rother C., and Dollár P. 2019. Panoptic segmentation. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Kumar G., Kumar A., Saikia P., Roy P., and Khan M. 2022. "Ecological impacts of forest fire on composition and structure of tropical deciduous forests of central India." Physics and Chemistry of the Earth, Parts a/b/c 128:103240.

Kwong I. H., and Fung T. 2020. "Tree height mapping and crown delineation using LiDAR, large format aerial photographs, and unmanned aerial vehicle photogrammetry in subtropical urban forest." International Journal of Remote Sensing 41 (14):5228-56.

Li F., Zhang H., Xu H., Liu S., Zhang L., Ni L. M., and Shum H.-Y. 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. Paper presented at the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Li N., Jiang S., Xue J., Ye S., and Jia S. 2023. "Texture-aware self-attention model for hyperspectral tree species classification." IEEE Transactions on Geoscience and Remote Sensing 62:1-15.

Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., and Belongie S. 2017. Feature pyramid networks for object detection. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Liu L., Pang Y., Fan W., Li Z., Zhang D., and Li M. 2013. "Tree species classification of temperate natural forest using airborne LiDAR and hyperspectral data." Journal of Remote Sensing 17 (3): 679-95.(刘丽娟,庞勇,范文义,等 . 机载 LiDAR 和高光谱融合实现温带天然林树种识别[J]. 遥感学报,2013,17(03):679-695.)[DOI：10.11834/jrs.20131067]

Long J., Shelhamer E., and Darrell T. 2015. Fully convolutional networks for semantic segmentation. Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Lu J., Wang K., Li C., and Ma T. 2016. "Forest type classification based on wavelet transform and random forest." Journal of Northwest Forestry University 31 (6):264-7.(吕杰, 汪康宁, 李崇贵, 等 . 基于小波变换和随机森林的森林类型分类研究[J]. 西北林学院学报, 2016, 31(6): 264-267.)[DOI：10.3969/j. issn. 1001-7461.2016.06.45]

Minaee S., Boykov Y., Porikli F., Plaza A., Kehtarnavaz N., and Terzopoulos D. 2021. "Image segmentation using deep learning: A survey." IEEE transactions on pattern analysis and machine intelligence 44 (7):3523-42.

Ninomiya S. 2022. "High-throughput field crop phenotyping: current status and challenges." Breeding Science 72 (1):3-18.

Nurhayati R. 2015. "Individual Tree Crown Delineation in Tropical Forest Using Object Based Analysis of Orthoimage and Digital Surface Model." Wageningen University Wageningen, The Netherlands.

Patel Y., Xie Y., Zhu Y., Appalaraju S., and Manmatha R. 2023. "Simcon loss with multiple views for text supervised semantic segmentation." arXiv preprint arXiv:2302.03432.

Pierdicca R., Nepi L., Mancini A., Malinverni E., and Balestra M. 2023. "UAV4TREE: Deep learning-based system for automatic classification of tree species using RGB optical images obtained by an unmanned aerial vehicle." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 10: 1089-96.

Pimentel C. S., McKenney J., Firmino P. N., Calvão T., and Ayres M. P. 2020. "Sublethal infection of different pine species by the pinewood nematode." Plant Pathology 69 (8):1565-73.

Qiu S., Fang M., Yu Q., Niu T., Liu H., Wang F., Xu C., Ai M., and Zhang J. 2023. "Study of spatialtemporal changes in Chinese forest eco-space and optimization strategies for enhancing carbon sequestration capacity through ecological spatial network theory." Science of the Total Environment 859:160035.

Rajaei A., Abiri E., and Helfroush M. S. 2024. "Balanced spatio-spectral feature extraction for hyperspectral and multispectral image fusion." Computers and Electrical Engineering 118:109391.

Ravi N., Gabeur V., Hu Y.-T., Hu R., Ryali C., Ma T., Khedr H., Rädle R., Rolland C., and Gustafson L. 2024. "Sam 2: Segment anything in images and videos." arXiv preprint arXiv:2408.00714.

Ren S., He K., Girshick R., and Sun J. 2015. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advanc-*

es in neural information processing systems 28.

Söderkvist O. 2001. "Computer vision classification of leaves from swedish trees." In.

Ulaby F. T., Li R. Y., and Shanmugan K. 2007. "Crop classification using airborne radar and Landsat data." IEEE Transactions on Geoscience and Remote Sensing (1):42-51.

Wang L., Zhu Z., and Yun T. 2023. "Individual tree crown detection algorithm based on improved YOLOv3." Computer Simulation 40 (01):510-6. (王丽文,朱正礼,云挺.基于改进YOLOv3的单木树冠检测算法[J].计算机仿真,2023,40(01):510-516.[DOI: 10.3969/j.issn.1006-9348.2023.01.092])

Waser L. T., Boesch R., Wang Z., and Ginzler C. 2017. "Towards automated forest mapping." In Mapping Forest Landscape Patterns, 263-304. Springer.

Xu B., Wei H., Cai Z., Yang J., Zhang Z., Wang C., Li J., Zhao J., Qu Y., and Yin G. 2023. "Exploring the potential of Gaofen-1/6 for crop monitoring: Generating daily decametric-resolution leaf area index time series." IEEE Transactions on Geoscience and Remote Sensing 61:1-14.

Yuan H., Li X., Yang Y., Cheng G., Zhang J., Tong Y., Zhang L., and Tao D. 2022. Polyphonicformer: Unified query learning for depth-aware video panoptic segmentation. Paper presented at the European Conference on Computer Vision.

Zhong H., Zhang Z., Liu H., Wu J., and Lin W. 2024. "Individual tree species identification for complex coniferous and broad-leaved mixed forests based on deep learning combined with UAV LiDAR data and RGB images." Forests 15 (2):293.

Zhong L., Dai Z., Fang P., Cao Y., and Wang L. 2024. "A review: Tree species classification based on remote sensing data and classic deep learning-based methods." Forests 15 (5):852.

Zhou Y., Zhang M., and Wang Y. 2025. "Global-frequency-domain network: a semantic segmentation method for high-resolution remote sensing images based on fine-grained feature extraction and global context integration." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

Zust L., Cabon Y., Marrie J., Antsfeld L., Chidlovskii B., Revaud J., and Csurka G. 2025. Panst3r: Multi-view consistent panoptic segmentation. Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.

FSC-Mask2Former: A Forest Panoptic Segmentation Method Based on UAV Imagery

YANG Zhen[1,2,3], YAO Zongqi[1,2,3], ZHANG Xiaoli[1,2,3]

1.State Key Laboratory of Efficient Production of Forest Resources, College of Forestry, Beijing Forestry University, Beijing 100083, China;

2.Beijing Key Laboratory of Precision Forestry, College of Forestry, Beijing Forestry University, Beijing 100083, China;

3.Key Laboratory of Forest Cultivation and Protection, Ministry of Education, Beijing Forestry University, Beijing 100083, China

Abstract: Objective Accurate monitoring of forest resources at the individual tree level is fundamental for forest ecosystem management. Unmanned aerial vehicle (UAV) visible light (RGB) imagery provides a cost-effective and high-spatial-resolution data source for these wide-area monitoring tasks. High-spatial-resolution imagery comprehensively records the fine contours of trees and the background habitat of the forest. Utilizing panoptic segmentation technology for unified interpretation enables the synchronous extraction of all forest elements. Nevertheless, interpreting highly closed-canopy forest scenes remains a critical challenge. Traditional deep learning approaches often decouple semantic segmentation for background elements and instance segmentation for individual trees, leading to severe pixel-level classification conflicts and spatial topology inconsistencies. Furthermore, the limited spectral information in RGB imagery frequently causes severe spectral confusion among adjacent trees. To systematically address these challenges, this study proposes an end-to-end forest panoptic segmentation model named FSC-Mask2Former. Method The proposed FSC-Mask2Former builds upon the Mask2Former baseline by introducing two core architectural improvements tailored to the unstructured features of forests. First, a Frequency-domain Texture Awareness (FTA) module is incorporated into the feature extraction pathway to compensate for the loss of micro-texture details caused by spatial downsampling, essentially functioning as a learnable high-pass filter in the feature space to retain critical edge gradients. Second, an Instance-aware Query Contrastive (IQC) head is integrated at the output of the Transformer decoder to maximize the inter-class feature distance between spectrally similar tree species, imposing an anisotropic constraint on the feature distribution to enlarge decision boundaries and fundamentally suppress category assignment conflicts. To evaluate the model, a densely annotated dataset was constructed using UAV RGB imagery from Gaofeng Forest Farm in the Guangxi Zhuang Autonomous Region, supplemented by data from Genhe City in the Inner Mongolia Autonomous Region, Jixi County in Anhui Province, and Hengzhou City in the Guangxi Zhuang Autonomous Region to validate model transferability.Result Comprehensive experiments demonstrate that FSC-Mask2Former significantly outperforms existing mainstream networks. The model achieves an overall Panoptic Quality (PQ) of 57.0%, a substantial gain of 11.0 percentage points over the baseline. Most notably, the foreground Recognition Quality (RQ) reaches 56.0%, representing a 12.0 percentage point increase. Visualizations confirm that FSC-Mask2Former effectively separates touching instances in high-canopy-closure forest areas, precisely delineates boundaries