北京林业大学学报
JOURNAL OF BEIJING FORESTRY UNIVERSITY

ViT-Count：面向冠层遮挡的Vision Transformer树木计数定位方法

张乔一 张瑞 霍光煜

ViT-Count: a Vision Transformer-based tree counting and localization method for canopy-occluded scenarios

Zhang Qiaoyi, Zhang Rui, Huo Guangyu

引用本文:
张乔一, 张瑞, 霍光煜. ViT-Count：面向冠层遮挡的Vision Transformer树木计数定位方法[J]. 北京林业大学学报, 2025, 47(10):128-138. doi: 10.12171/j.1000-1522.20250035
Zhang Qiaoyi, Zhang Rui, Huo Guangyu. ViT-Count: a Vision Transformer-based tree counting and localization method for canopy-occluded scenarios[J]. Journal of Beijing Forestry University, 2025, 47(10):128-138. doi: 10.12171/j.1000-1522.20250035

在线阅读 View online: https://doi.org/10.12171/j.1000-1522.20250035

您可能感兴趣的其他文章
Articles you may be interested in

SSFYOLO：一种面向复杂森林场景的树干检测算法
SSFYOLO: a trunk detection algorithm for complex forest scenarios
北京林业大学学报, 2025, 47(2): 132-142. https://doi.org/10.12171/j.1000-1522.20240145

基于无人机影像和MDIEA-YOLO苗木识别模型的造林验收智能系统
An intelligent system for afforestation acceptance based on UAV images and MDIEA-YOLO seedling recognition model
北京林业大学学报, 2025, 47(5): 14-25. https://doi.org/10.12171/j.1000-1522.20240354

基于深度学习的树木根系探地雷达多目标参数反演识别
Deep learning-based inverse identification of multi-target parameters for tree rooting ground-penetrating radar
北京林业大学学报, 2024, 46(4): 103-114. https://doi.org/10.12171/j.1000-1522.20230259

基于YOLOv7的红外相机野生动物图像筛选
Wildlife image screening for infrared cameras based on YOLOv7
北京林业大学学报, 2024, 46(2): 143-154. https://doi.org/10.12171/j.1000-1522.20230112

激光雷达辅助的样地调查单木定位
LiDAR based individual tree localization for sample plot inventory
北京林业大学学报, 2023, 45(1): 99-108. https://doi.org/10.12171/j.1000-1522.20220106

面向城市绿化精细管理的低密度树木数字孪生模型
A low-density tree digital twin model for refined urban greening management: a case study of tree wind disaster risk management
北京林业大学学报, 2025, 47(3): 139-150. https://doi.org/10.12171/j.1000-1522.20240400

DOI:10.12171/j.1000−1522.20250035

ViT-Count: 面向冠层遮挡的 Vision Transformer 树木计数定位方法

张乔一[1,2]　张　瑞[1]　霍光煜[1]
（1. 北京林业大学信息学院（人工智能学院），北京 100083；2. 首都师范大学数学科学院，北京 100048）

摘要：【目的】针对复杂场景中树木检测的挑战，如遮挡、背景干扰及密集分布等，本研究提出一种基于 Vision Transformer（ViT）的树木检测方法（ViT-Count），提升模型对复杂场景中树木的检测精度与鲁棒性。**【方法】**采用 ViT 作为基础模型，其在捕捉图像中全局上下文信息方面具有天然优势，尤其适用于形态多变的复杂环境。设计针对树木的视觉提示调优 VPT 机制，其通过在特征中注入可学习提示（prompts），优化模型在林地高密度树冠、光照变化及不同树种结构下的特征提取能力，提高对不同林分类型的适应性。设计卷积模块的注意力机制模块，利用其在局部感知基础上的长距离依赖建模能力，有效强化模型对树木遮挡、重叠及形态相似目标的辨别能力，提高整体检测的鲁棒性与准确性。设计一个树木检测解码器，通过多层卷积、归一化、GELU 激活与上采样操作逐步还原空间分辨率，以生成的目标密度图实现树木计数与定位。**【结果】**该方法在提升森林、城市场景下的树木检测鲁棒性的同时，增强了模型在多尺度树木目标上的泛化能力。在 Larch Casebearer 数据集和 Urban Tree 数据集上进行的实验显示，与其他主流模型相比，该方法的 MAE 和 RMSE 最多分别降低了 2.53、3.99，表明其泛化能力更强，具有最优的树木检测性能。可视化实验结果表明，在密集森林场景和复杂城市场景中，所提模型均具有较高的树木检测准确率。消融实验的结果证明了模型主要模块的有效性。**【结论】**基于 Vision Transformer 的面向复杂场景的树木计数与定位方法能够充分发挥 ViT 的全局建模能力及视觉提示调优机制任务适应性，结合卷积模块的注意力机制，有效提升复杂场景树木计数与定位的精度与鲁棒性。

关键词：目标识别；树木计数；树木定位；复杂场景；Vision Transformer (ViT)；视觉提示调优 (VPT)；注意力机制
中图分类号：S718.5；TP181　　**文献标志码：**A　　**文章编号：**1000−1522(2025)10−0128−11
引文格式：张乔一，张瑞，霍光煜. ViT-Count：面向冠层遮挡的 Vision Transformer 树木计数定位方法［J］. 北京林业大学学报，2025，47(10)：128−138. Zhang Qiaoyi, Zhang Rui, Huo Guangyu. ViT-Count: a Vision Transformer-based tree counting and localization method for canopy-occluded scenarios［J］. Journal of Beijing Forestry University, 2025, 47(10): 128−138.

ViT-Count: a Vision Transformer-based tree counting and localization method for canopy-occluded scenarios

Zhang Qiaoyi[1,2]　Zhang Rui[1]　Huo Guangyu[1]
(1. School of Information Science (School of Artificial Intelligence), Beijing Forestry University, Beijing 100083, China;
2. School of Mathematical, Capital Normal University, Beijing 100048, China)

Abstract: ［**Objective**］Against the backdrop of deep integration between remote sensing and computer vision, tree detection in complex scenes has long been hindered by occlusion, background noise, and high-density canopies. Traditional methods struggle to balance global semantics and local details, resulting in insufficient detection accuracy and robustness. To overcome this bottleneck, this study proposes an end-to-

收稿日期：2025−02−05　**修回日期：**2025−09−28
基金项目：国家自然科学基金项目（62402046），北京林业大学科技创新计划项目（BLX202358）。
第一作者：张乔一。主要研究方向：树木计数检测。Email：1410346674@qq.com　地址：100048 北京市海淀区西三环北路 105 号首都师范大学数学科学学院。
责任作者：霍光煜，讲师。主要研究方向：人工智能。Email：gyhuo@bjfu.edu.cn　地址：100083 北京市海淀区清华东路 35 号北京林业大学信息学院（人工智能学院）。
本刊网址：http://j.bjfu.edu.cn；http://journal.bjfu.edu.cn

end tree detection framework (ViT-Count) centered on the Vision Transformer (ViT), aiming to leverage its global context modeling capability to significantly enhance the accuracy of tree counting and localization in dense forest and urban environments. [**Method**] Firstly, the Vision Transformer (ViT) was employed as the backbone model due to its inherent advantage in capturing global contextual information, making it particularly well-suited for complex environments with diverse visual patterns. Secondly, a Visual Prompt Tuning (VPT) strategy tailored for tree detection was introduced. By injecting learnable prompts into the feature space, this method enhanced the model's capability to extract discriminative features under conditions such as high canopy density, illumination variation, and structural differences across tree species, thereby improving its adaptability to various forest stand types. Thirdly, the convolutional block attention mechanisms were designed, leveraging its ability to model long-range dependencies based on local perception. This effectively strengthened the model's capacity to distinguish between occluded, overlapping, and morphologically similar tree targets, thereby enhancing the overall robustness and accuracy of detection. Finally, a tree detection decoder was constructed, which progressively restored spatial resolution through stacked convolutional layers, normalization, GELU activation, and upsampling operations, ultimately generating a tree density map for tree counting and localization. [**Result**] Comparative experiments on the public Larch Casebearer and Urban Tree datasets showed that, compared with mainstream models, the proposed method achieves reductions up to 2.53 in MAE and 3.99 in RMSE, demonstrating stronger generalization capability. Visualization results indicate that the model maintains high accuracy in both dense forests and complex urban scenes. Ablation studies further confirm the effective contributions of VPT, the attention module, and the decoder components to the overall performance. [**Conclusion**] This Vision Transformer-based method for tree counting and localization in complex scenes fully exploits the global modeling capability of ViT and the task adaptability of the visual prompt tuning mechanism. By integrating the attention mechanism of convolutional modules, it effectively enhances the accuracy and robustness of tree counting and localization in complex environments.

Key words: object detection; tree counting; tree localization; complex scenes; Vision Transformer (ViT); visual prompt tuning (VPT); attention mechanism

针对复杂场景下的树木计数与定位技术对实现森林资源的高效监测与智能管理具有重要意义。尤其在自然林地、丘陵山地、城市等复杂环境中，树木的自动识别和精确检测面临诸多挑战，如树种多样性、枝叶遮挡、密集分布以及复杂的光照与地形条件，均可能导致检测性能下降。树木计数与定位技术不仅为林业调查提供了关键数据支持，还能辅助实现生物多样性评估、碳储量估算和生态环境保护等多项任务，有效提升森林资源监测的智能化水平。因此，如何在复杂背景下实现快速、精准的树木计数与定位，已成为当前森林遥感与计算机视觉领域的研究热点之一。

传统的树木计数与定位方法主要依赖人工调查，如步行或乘车记录树木数量，过程繁琐且耗时，易受人为因素影响，导致精度低、成本高[1]。精准林业兴起后，树木数量与空间坐标被纳入生长监测、产量预估和健康评估中，为精细化管理提供基础数据[2]。目前主流数据源自 RGB、RGB-D 或多/高光谱影像，不同传感器需匹配专用算法。基于纹理、形状、颜色等图像特征的目标检测与计数方案，常与 K 均值、随机森林和支持向量机等传统机器学习模型结合[3]。但这些方法在跨场景应用中泛化性能不足，难以直接迁移至新数据集[4]。

随着深度学习技术的发展，利用深度神经网络进行密度估计以实现树木的计数与定位，已成为林业视觉感知中的一种主流方法[5]。典型的网络模型包括 CNN(convolutional neural network)[6]和其扩展模型 MCNN(multi-column convolutional neural network)[7]，这类方法通过对图像进行特征提取，生成密度图，并据此完成数量估计任务。在传统 CNN 结构已展现出良好特征提取能力的基础上，研究人员进一步关注如何提升深层网络的训练性能。He 等[8]提出的 ResNet-50 网络在传统 CNN 基础上引入了残差结构，有效缓解了深层网络训练中梯度消失与梯度爆炸的问题，从而提升了模型的训练效率和稳定性。Li 等[9]设计了 CSRNet(congested scene recognition network)，在结构中融合空间稀疏卷积操作，降低了计算复杂度的同时增强了对高密度区域

的处理能力。Zhang 等[7]构建的 MCNN 模型通过多列网络结构实现多尺度特征提取，不同卷积配置的列独立提取图像特征后再进行融合，使网络在复杂环境下的适应性和表达能力显著增强，提升了树木计数与定位的准确性。金远航等[10]提出了一种改进的 YOLOv4-tiny 算法，通过引入空间金字塔池化、ELU 激活函数和 ECANet 机制，显著提升了枯死树木计数与定位的精度。上述研究表明，树木计数与定位模型的性能提升需围绕“场景痛点–技术适配”展开，这为本文模型的结构设计提供了清晰指导：需兼顾计算效率与密集目标处理能力、强化多尺度特征适配性、引入针对性的特征优化机制，以有效应对遮挡和密集树冠场景的挑战。

近年来，Transformer 结构[11]因其在自然语言处理领域中的卓越表现而受到广泛关注，并逐渐拓展至计算机视觉领域。其核心优势在于能够通过自注意力机制实现远距离依赖建模和全局特征捕捉，突破了传统卷积结构在感受野和信息传递方面的局限。Chen 等[12]提出了新型高效深度学习模型 Dent，在设计中结合了多感受野的卷积结构，用以提取局部与全局上下文特征，同时引入 Transformer 编码器，利用注意力机制建模视觉特征之间的关系，实现更优的上下文信息传递与融合。Jiang 等[13]提出了一种面向非林地树木计数与定位的 Transformer 框架，高效完成了大尺度非林地树木计数与定位任务，展现出优越的检测精度和良好的跨区域泛化能力。Gibril 等[14]提出了一种基于 Transformer 的遥感方法，结合改进的多尺度视觉 Transformer、特征金字塔网络与改进的切片辅助推理策略，实现了对无人机影像中椰枣树的识别、计数。王凯等[15]提出改进的 YOLOv7-ACGDmix 模型，通过引入注意力机制、可变形卷积和优化损失函数，有效提升了杨梅树（*Myrica rubra*）单木检测的精度与泛化能力。综上，当前基于 Transformer 及改进模型的树木计数与定位研究，在针对性适配、小目标与密集目标的特征处理优化以及树木遮挡特性的专用处理模块方面均存在不足，制约了其在复杂林业实际场景中的应用效能。

尽管现有的基于深度学习的树木计数与定位方法在常规环境中表现良好，但在遮挡和密集树冠场景下的树木计数与定位任务仍面临诸多挑战，例如目标被遮挡导致特征不完整，背景干扰引发误检，树木密集分布造成边界难以分割等问题，这些因素显著影响模型的鲁棒性与准确性。为解决上述问题，本文提出一种基于 ViT 的复杂场景树木计数与定位方法。一方面，设计一个针对树木计数与定位场景的视觉调优技术，通过在 ViT 的特征空间中注入可学习提示（prompts），实现模型对树木计数与定位任务的快速适应。该技术可有效增强模型对不同树木类型、不同生长阶段及季节变化下树木形态的识别能力。另一方面，为进一步提升模型对个体树木的边界感知能力，设计一种卷积注意力模块，将 CNN 的局部细节感知优势与 Transformer 的全局依赖建模能力有机结合。该模块利用卷积结构提取树冠边界等细粒度特征，并通过注意力机制实现全图范围内的特征依赖建模，使模型在处理密集遮挡、纹理变化显著的区域时具备更强的表达能力。特别是在城市环境中，建筑遮挡和树木尺度差异显著的情况下，该模块能够引导模型关注局部显著特征，提高部分遮挡树木的计数与定位精度，并增强模型在不同航拍分辨率下的适应能力。本研究旨在通过人工智能技术实现复杂场景的树木计数与定位，为森林资源监测和管理提供高效可靠的工具支持。

1 基于 Vision Transformer 的树木计数与定位方法

本文提出的基于 Vision Transformer（ViT）的树木计数与定位模型框架图包括多个关键模块。首先，ViT 模块用于提取图像的全局特征，借助自注意力机制有效捕捉树木的空间信息。接下来，视觉调优模块（VPT）进一步优化视觉特征，提升模型对复杂环境的适应性。卷积模块的注意力机制模块（convolutional block attention mechanisms，CBAM）则通过增强重要区域的特征响应，改善了模型对细节的关注。最后，树木检测解码器模块负责将提取的特征进行解码，精确检测树木的数量、位置（图 1）。整体架构结合了 Transformer 和卷积神经网络的优势，能够在不同场景中实现高效、准确的树木计数与定位。

1.1 ViT 模型

ViT 是一种基于 Transformer 架构的图像处理模型。相比于 CNN，ViT 更擅长捕捉远距离区域之间的上下文关系，适用于处理枝干交叉、冠层重叠等复杂结构。与传统的卷积神经网络（CNN）不同，ViT 将森林图像分割成若干小块（patch），然后将这些图像块视为一系列序列数据，输入到 Transformer 模型中进行处理。具体地，首先将输入图像 $X \in \mathbb{R}^{H\times W\times C}$（其中 H 是图像的高度，W 是宽度，C 是通道数）划分为大小为 $P\times P$ 的小块（patch），每个小块被展平，并通过线性映射转换为固定长度的向量。

假设图像被划分为 N 个大小为 $P\times P$ 的小块（其

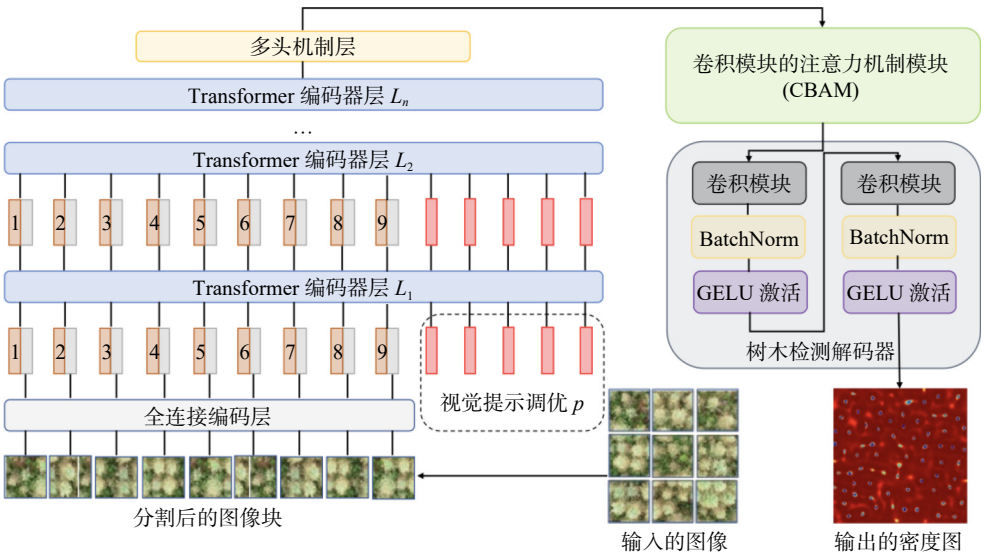


图 1　基于 Vision Transformer 的树木计数与定位模型（ViT-Count）总体框架

Fig. 1　Tree counting and localization model framework based on Vision Transformer(ViT-Count)

中 $N=\frac{HW}{P^2}$），则每个小块被映射到一个嵌入空间，得到一个向量 $\mathbf{z}_i=\boldsymbol{W}_{\mathrm{e}}x_i+b_{\mathrm{e}}$，其中 $\boldsymbol{W}_{\mathrm{e}}$ 是线性映射矩阵，b_{e} 是偏置项。接下来，为了保留位置信息，ViT 将位置编码 $\boldsymbol{P}\in\boldsymbol{R}^{N\times D}$ 加入到每个图像块的嵌入向量中，得到最终的输入序列。

$$\mathbf{z}_i^{\mathrm{input}}=\mathbf{z}_i+\boldsymbol{P}_i \tag{1}$$

然后，所有的嵌入向量（包括图像块的嵌入和一个额外的分类提示标记）被输入到标准的 Transformer 架构中，通过自注意力机制来建模图像块之间的依赖关系。具体的自注意力计算公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别表示查询、键和值，d_k 是查询和键的维度。通过多层的自注意力和前馈网络，ViT 模型对图像的全局上下文进行建模，并最终完成特征提取。

1.2　视觉提示调优

视觉提示调优是一种高效的微调办法[16]，主要用于训练参数规模较大的预训练模型。在本研究中，针对森林与城市环境中树木分布密集、遮挡严重、形态异质等特点，重新设计并定制了 VPT 模块，以更好地适配 ViT 在树木计数与定位任务中的表现。具体而言，ViT 模型将输入图像划分为固定大小的 patch，并通过嵌入层映射为 token 序列。在此基础上引入少量可学习的提示 token（prompt tokens），这些 token 可插入到输入端或 Transformer 中间层，与图像 token 共同参与特征建模。该设计不仅保留了 ViT 强大的全局建模能力，还通过提示引导模型聚焦于树木区域的显著特征，从而在复杂环境下显著提升检测精度和模型泛化能力。

假设输入图像的 patch 序列为 $X=[x_1,x_2,\cdots,x_N]$，其中 x_i 表示图像中第 i 个 patch，N 是 patch 的总数。VPT 通过在输入端引入可训练的提示 tokens，得到一个新的输入序列。

$$X'=[p_1,p_2,\cdots,p_M,x_1,x_2,\cdots,x_N] \tag{3}$$

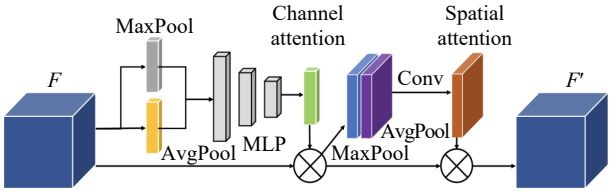
式中：$p_1,p_2,\cdots,p_M$ 表示提示 token，M 为提示 token 的数量。输入序列 X' 通过 Transformer 模型进行处理，并通过注意力机制计算图像中各个 tokens 之间的关系。通过这种方式，提示 tokens 与图像 tokens 共同参与计算，从而在不改变原有 ViT 结构的情况下，优化模型对树木计数与定位任务的适应性。在航拍树木计数与定位任务中，树木目标通常因高分辨率而细节丰富，且背景复杂多变，同时还面临树冠密集、相互遮挡以及光照条件变化等多重挑战。为了应对这些挑战，VPT 技术能够通过可训练的提示 tokens（即 $p_1,p_2,\cdots,p_M$），帮助 ViT模型更好地聚焦于树木的关键特征，从而提升树木计数与定位的精度和鲁棒性。

1.3　卷积模块的注意力机制模块

为了更有效地应对森林与城市等场景中树木分布密集、尺度差异大、纹理复杂等挑战，本研究设计了针对树木计数与定位的卷积模块的注意力机制（convolutional block attention module, CBAM），并进行了针对性优化，以增强模型的多尺度感知能力与泛化能力。CBAM 融合了通道注意力与空间注意力机制，能够在不显著增加模型计算量的前提下，自适应地强化关键区域和特征维度的表达能力。

与仅关注通道信息的传统注意力机制相比，CBAM 通过空间维度建模显著区域（如树干、树冠、阴影区域）的位置分布，提升模型对局部小目标与大

范围结构的同时感知能力；同时在通道维度上，进一步突出与树木语义相关的特征响应，从而提升对结构相似、遮挡重叠目标的分辨能力。通过这种局部—全局信息融合机制，CBAM 有效增强了模型在面对尺度变化显著、结构复杂多样的树木目标时的适应能力与鲁棒性。如图 2 所示，CBAM 由通道注意力模块与空间注意力模块串联组成，具备轻量、高效且可插拔的特性，非常适用于本研究所面向的多场景树木计数与定位任务。



F 为 ViT 模型输出的特征图，也是卷积模块的注意力机制模块（CBAM）模块的输入特征图。*F*′为 CBAM 模块输出的特征图。MLP 为全连接层。

图 2 卷积模块的注意力机制模块（CBAM）结构图

Fig. 2 Structure diagram of convolutional block attention module (CBAM)

1.3.1 通道注意力模块

CBAM 中的通道注意力模块通过生成一个注意力图来突出输入特征图中重要的通道。其方法是通过全局平均池化和全局最大池化操作，捕捉特征图的全局上下文信息。接下来，这些池化后的特征会通过一个共享的全连接层，并通过 Sigmoid 激活函数生成通道注意力图。通道注意力图与原始特征图进行逐元素相乘，从而强调重要的通道并抑制不重要的通道。通道注意力图$\boldsymbol{M}_{\mathrm{c}}$可以表示为

$$\boldsymbol{M}_{\mathrm{c}}=\sigma(\mathrm{FC}([\mathrm{GAP}(\boldsymbol{F}),\mathrm{GMP}(\boldsymbol{F})])) \tag{4}$$

式中：σ表示 Sigmoid 激活函数，$\mathrm{GAP}(\boldsymbol{F})$、$\mathrm{GMP}(\boldsymbol{F})$分别表示对特征图$\boldsymbol{F}$执行全局平均池化和全局最大池化操作，FC表示全连接层。

1.3.2 空间注意力

CBAM 中的空间注意力模块通过聚焦于特征图中重要的空间区域来实现。在空间维度上，空间注意力旨在捕捉位置相关的信息。它通过在通道维度上应用平均池化和最大池化操作来生成一个空间注意力图。随后，这些池化后的特征被连接起来，并通过一个大小为 7 × 7 的卷积层，再经过 Sigmoid 激活函数，生成空间注意力图。最后，空间注意力图通过逐元素相乘的方式应用于原始特征图。空间注意力图可以被表示为

$$\boldsymbol{M}_{\mathrm{s}}=\sigma(\mathrm{Conv}([\mathrm{GAP}(\boldsymbol{F}),\mathrm{GMP}(\boldsymbol{F})])) \tag{5}$$

式中：Conv表示卷积层，$\mathrm{GAP}(\boldsymbol{F}),\mathrm{GMP}(\boldsymbol{F})$是在通道维度上对特征图$\boldsymbol{F}$执行的全局平均池化和全局最大池化操作。

1.3.3 组合注意力机制

CBAM 的注意力机制是按顺序操作的，首先应用通道注意力，然后应用空间注意力。每个注意力机制后，特征图都会被更新，从而逐步精细化其表示。这种组合使得模型能够自适应地强调重要的通道和关键的空间区域，从而有效提高模型在关注重要特征，抑制无关噪声方面的能力。CBAM 的最终输出为

$$\boldsymbol{F}'=\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{M}_{\mathrm{c}}(\boldsymbol{F})) \tag{6}$$

式中：$\boldsymbol{M}_{\mathrm{c}}$是经过通道注意力处理后的特征图，$\boldsymbol{M}_{\mathrm{s}}$是应用于$\boldsymbol{M}_{\mathrm{c}}$的空间注意力图。

1.4 树木检测解码器模块

树木检测解码器模块设计为一个密度图回归神经网络，用于将经过 CBAM 处理的特征图$\boldsymbol{F}'$转换为目标密度图$\boldsymbol{M}$。具体而言，该模块通过一系列卷积层、归一化、激活函数及上采样操作实现特征的逐步解码与空间分辨率的还原。

首先，输入特征图$\boldsymbol{F}'$通过多个 2D 卷积层处理。每一卷积层紧接归一化操作（BatchNorm）与 GELU 激活函数，其输出特征的更新公式为

$$\boldsymbol{F}_{i+1}=\mathrm{GELU}(\mathrm{Norm}(\mathrm{Conv}(\boldsymbol{F}_i,\boldsymbol{W}_i,\boldsymbol{b}_i))) \tag{7}$$

式中：$\boldsymbol{W}_i$，$\boldsymbol{b}_i$分别表示第i层的卷积权重和偏置参数。随着层级的推进，特征图的通道数量逐步减少，以提取与密度图预测最相关的特征。在每层卷积操作后，特征图通过上采样函数进行空间分辨率的放大。本模块的上采样操作将特征图的分辨率放大为原来的两倍，直至特征图达到预设的目标高度和宽度。

解码器的最后一层卷积操作将特征图的通道数缩减至 1，用以生成密度图$\boldsymbol{M}$。随后，通过 Sigmoid 函数将输出值限制在 [0,1] 范围内，确保密度信息的数值范围合理，其计算公式为

$$\boldsymbol{M}=\sigma(\mathrm{Conv}(\boldsymbol{F}_{\mathrm{final}},\boldsymbol{W}_{\mathrm{final}},\boldsymbol{b}_{\mathrm{final}})) \tag{8}$$

式中：σ表示 Sigmoid 函数，$\boldsymbol{F}_{\mathrm{final}}$表示解码器的最终特征图。该解码器模块通过卷积层逐步减少特征通道数，并通过上采样逐步恢复空间分辨率，最终生成表示目标区域密度分布的密度图。

2 研究方法

2.1 数据集

本文使用了一个航拍树木数据集 Larch Casebearer 和一个遥感树木数据集 Urban Tree。Larch Casebearer 数据集为瑞典森林局在微软人工智能地球计划的支持下启动的一个项目的成果，它包含 1 543 张无人机在瑞典韦斯特哥特兰 5 个地区上空飞行时拍摄的图像，图像中主要包含了各种树木（表 1）。Urban

表 1　Larch Casebearer 数据集详情

Tab. 1　Details of Larch Casebearer dataset

数据子集	图像数量	落叶松数量	其他树木数量
Bebehojd_20190527	172	8 749	2 570
Ekbacka_20190527	215	6 423	2 702
Jallsavag_20190527	50	2 580	475
Kampe_20190527	179	15 508	3 219
Nordkap_20190527	224	11 262	2 987
Bebehojd_20190819	149	7 014	2 949
Ekbacka_20190819	181	6 768	3 868
Jallsavag_20190819	39	1 533	1 486
Kampe_20190819	113	8 134	1 537
Nordkap_20190819	221	8 656	3 458
Total	1 543	76 627	25 251

Tree 数据集包含来自城市、公园、河流、乡村的高分辨率航拍图像，涵盖了不同季节和天气条件下的树木分布情况。该数据集共包含 3 012 张图像，每张图像的分辨率为 256 × 256 像素，并标注了树木的位置信息(表 2)。Larch Casebearer 数据集的图像样本集中于森林场景，树木茂密并且树冠相互遮挡较多；Urban Tree 数据集的图像样本集中于城市场景，背景复杂且树木较小(图 3)。

2.2　模型有效性验证

为了证明本研究提出的 ViT-Count 方法的有效性和可行性，选择 13 种深度学习模型进行树木计数与定位与本文的方法进行比较。对两个数据集的数据进行了预处理，使其符合 13 个对比模型的要求，进而对每个模型进行相应的测试。实验的设计思路：首先利用模型对做好标记的图像进行训练，之后利用已经训练完成的模型对图像进行测试，通过生成图像对应的密度图得到检测的树木个数与位置，进而与真实密度图进行对比。

VGG[17]采用小尺寸 3 × 3 卷积核堆叠的方式加深网络，提高特征提取能力。ALEXNET[18]通过增加网络深度并使用 ReLU 激活函数，在 2012 年 ImageNet 竞赛中获得了巨大成功。该网络通过数据增强和 Dropout 技术，显著提高了图像分类的准确率。Inception-v3[19]通过采用不同大小的卷积核和精巧的网络设计，能够在较少的计算资源下实现高效的特征提取。其在图像分类任务中表现卓越，并在多个计算机视觉任务中取得了优异成绩。RESNET50[8]通过残差连接缓解梯度消失问题，并采用瓶颈结构减少计算量。该模型由 50 层组成。RESNET101[8]是 ResNet-50 的更深版本，包含 101 层残差单元，进一步提升了特征学习能力，但计算开销和存储需求也相应增加。MCNN[7]采用了多列结构，每列网络分别处理不同的输入区域，进而汇聚信息进行计数预测。该方法能够有效解决密集场景下的计数问题多列卷积神经网络。UNET[20]是一种典型的编码–解码结构神经网络，通过跳跃连接融合低层与高层特征。ViT[21]是一种将图像切分为小块并通过自注意力机制处理的模型，采用 Transformer 架构在图像分类等视觉任务中表现出色。METAMSNET[22]通过引入多源时空信息来进行更加精准的计数预测，尤其在动态场景中表现优异。该模型在处理具有时序特征和多模态数据的计数任务中，能有效提升预测的稳定性与精度。BCNET[23]是一种基于双层卷积结构的网络，通过深层和浅层特征的结合来提高计数精度。该模型采用双层网络架构，不仅能捕捉局部细节特征，还能有效捕捉全局信息，优化计数性能。MCNNT[24]结合多尺度卷积神经网络和 Transformer 架构，用于提升工业过程不平衡数据中特征提取与少数类故障识别能力。MSTNet[25]基于多尺度

表 2　Urban Tree 数据集详情

Tab. 2　Details of Urban Tree dataset

城市	气候带	树木种类	不同年度标注树木的数量		
			2016	2018	2020
比晓	美国内陆西部	10			682
奇科	内陆谷地	99		8 187	8 164
克莱蒙特	内陆帝国	92	4 858	4 880	4 678
尤里卡	加利福尼亚北部海岸	21			2 134
长滩	加利福尼亚南部海岸	100	6 470	6 403	5 845
棕榈泉	美国西南沙漠	100	4 433	4 707	4 109
河滨市	内陆帝国	90	5 015	4 400	4 087
圣塔莫尼卡	加利福尼亚南部海岸	92	5 824	5 830	5 841

注：该数据集涵盖了加利福尼亚州的8个城市和6个气候区，时间跨度为3年。

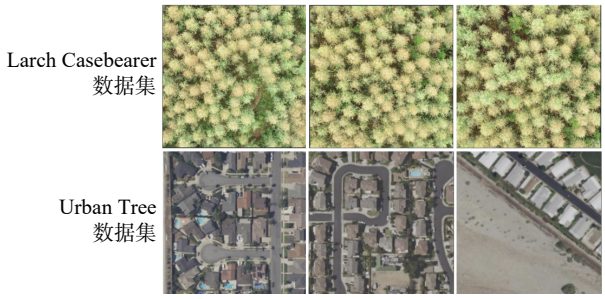


图 3 数据集样本展示

Fig. 3 Dataset sample display

Transformer 结构设计，更适应复杂场景下的目标分布。Gramformer[26]融合局部图结构与 Transformer 的全局建模能力，具备较强的上下文理解能力。

2.3 模块有效性验证

为进一步验证本文所提模型的核心设计价值，明确各关键模块对密度估计性能的实际贡献，本研究基于“控制变量法”设计并执行了消融实验：以 ViT 作为基础图像编码器提取输入图像的多层次特征信息，在编码器阶段引入视觉提示优化机制以动态调整特征注意力分布、增强关键区域的特征表征能力，最终由解码器将优化后的特征映射为与输入图像尺寸匹配的密度图，并以真实密度图为基准，通过定量指标评估模型性能；实验重点分析视觉提示模块（VPT）与通道注意力−空间注意力模块（CBAM），采用“√（启用）”“×（移除）”的二元控制策略构建多组不同模块组合的对照模型，通过选择启用和关闭每个模块，可以证明每个模块的有效性。

2.4 评价指标

使用平均绝对误差（mean absolute error，MAE）和均方根误差（root mean squared error，RMSE）来评估模型的计数性能。

$$\mathrm{MAE}=\frac{1}{t}\sum_{i=1}^{t}\left|N_{\mathrm{pred}}^{i}-N_{\mathrm{gt}}^{i}\right|$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{t}\sum_{i=1}^{t}\left(N_{\mathrm{pred}}^{i}-N_{\mathrm{gt}}^{i}\right)^{2}} \tag{9}$$

式中：t表示测试集中图像的总数目，N_{pred}^{i}表示检测到的图像中树木的数目，N_{gt}^{i}表示图像中树木的真实数目。

采用准确率（P）、召回率（R）和 F_1 分数来评估模型在定位任务中的性能。具体来说，当一个预测点与一个真实点之间的欧氏距离小于设定的距离阈值时，认为该预测是一次成功匹配，即该预测点正确地定位到了一个真实目标。评估时，系统会遍历所有预测点与真实点的距离，若该距离小于阈值则判定为匹配。每个真实点和预测点最多只能被匹配一次，匹配对确定后：成功匹配者（true positive）的数量记为 A_{TP}，没有成功匹配的预测点（false positive）的数量记为 A_{FP}，没有被任何预测点匹配到的真实点（false negative）的数量记为 A_{FN}。根据上述定义，3 个指标的计算公式如下。

$$P=\frac{A_{\mathrm{TP}}}{A_{\mathrm{TP}}+A_{\mathrm{FP}}}$$

$$R=\frac{A_{\mathrm{TP}}}{A_{\mathrm{TP}}+A_{\mathrm{FN}}}$$

$$F_1=\frac{2PR}{P+R} \tag{10}$$

通过参数量（parameters）和计算量（浮点运算次数，floating point operations，FLOPs）[27]两个核心指标来衡量算法复杂程度。参数量表示模型在训练过程中各个网络层中所有可学习参数的总和，反映了模型的存储规模和容量。计算量则指模型在进行一次前向推理时所需执行的 FLOPs，体现了模型的计算开销。这两个指标能够从存储和计算两个角度全面衡量算法的复杂程度。

2.5 实验细节和参数设置

采用 PyTorch 1.10.0 和 CUDA 版本 11.3，实验在 RTX 4090（24 GB）上进行，使用 Adam 优化器，设置权重下降率为 0.05，学习率为 0.000 1，VPT 宽度设置为 20，VPT 深度设置为 10，解码器深度设置为 2，解码器头数量设置为 8，线程数设置为 12，训练批量设置为 32。

3 结果与分析

3.1 模型有效性

本文提出的 ViT-Count 模型在所有模型中表现最为优越，取得了最低的 MAE（7.05）和 RMSE（9.25）值，显著优于现有主流方法，充分验证了其在复杂林业场景下的计数精度与鲁棒性。EFFICIENTNET 模型以 7.16 的 MAE 和 9.78 的 RMSE 紧随其后，为次优结果，体现出较好的性能平衡。整体来看，传统卷积模型如 VGG、RESNET 系列在精度上优于早期架构（如 Inception-v3 和 UNET），而 MCNN 系列与 MSTNet 通过融合多尺度特征也获得了较好效果。特别值得注意的是，ViT 及其变体 ViT-Count 显示出 Transformer 架构在捕捉全局上下文信息方面的强大能力，尤其在目标密集、尺度多变的场景中优势明显（表 3）。

从表 3 中 Urban Tree 数据集的实验结果可以看出，本文提出的 ViT-Count 模型在所有模型中表现最为优越，其在 MAE 和 RMSE 两个指标上分别达到了 10.58 和 13.75，显著优于其他现有方法。传统的卷积神经网络如 ALEXNET、RESNET50 和 RESNET101 虽然取得了一定的性能，但在误差控制

表 3　树木计数实验结果

Tab. 3　Tree counting experiment results

模型	Larch Casebearer		Urban Tree	
	MAE	RMSE	MAE	RMSE
Inception-v3	22.01	28.32	25.87	31.35
UNET	20.17	24.61	21.67	23.78
VGG	13.67	18.79	25.87	31.35
MCNN	12.47	16.72	15.79	20.54
BCNET	10.08	13.74	25.87	31.35
ViT	9.95	12.97	21.22	24.38
MCNNT	8.79	11.55	12.97	18.04
ALEXNET	8.42	11.16	12.79	17.58
Gramformer	8.33	10.30	13.20	17.68
RESNET50	8.22	11.09	12.11	16.80
MSTNet	7.38	<u>9.52</u>	12.19	16.34
RESNET101	7.25	10.39	<u>11.66</u>	<u>16.09</u>
EFFICIENTNET	<u>7.16</u>	9.78	12.07	16.68
ViT-Count	**7.05**	**9.25**	**10.58**	**13.75**

注：最优结果用加粗表示，次优的结果用下划线表示。

方面仍存在不足。相比之下，Inception-v3、VGG 等传统模型误差较大，MCNN、MSTNet 等多尺度模型表现中等，而 ViT-Count 在准确性与稳定性方面均表现最优，进一步验证了其在不同场景下的强泛化能力与鲁棒性。

本文提出的 ViT-Count 模型在 F_1 分数上取得最高分（65.76），同时 Recall 达到 83.07，显著优于其他模型，表明其在树木位置识别中的全面性能最优。尽管 MSTNet 在准确率上略高（88.41），但其召回率较低（49.28），导致整体 F_1 分数不及 ViT-Count。相比之下，其他模型如 RESNET101、RESNET50 等在精度或召回率方面存在偏差，难以兼顾检测完整性与准确性（表 4）。ViT-Count 在定位指标之间实现了更好的平衡，展现出在复杂背景下稳定可靠的树木定位能力。

本文提出的 ViT-Count 模型在 F_1 分数上取得最高值（47.14），同时在准确率（55.17）与召回率（41.15）之间实现了良好平衡，表现出优越的整体定位能力。相比之下，尽管 UNET 在召回率上达到 88.07，但其准确率极低（15.23），导致 F_1 分数仅为 25.96，定位结果不够稳定。其他模型如 EFFICIENTNET 和 RESNET50，虽在 F_1 分数上接近，但仍略逊于 ViT-Count（表 4）。总体来看，ViT-Count 在城市环境中复杂背景下展现出更强的泛化能力与定位精度，进一步验证了其在树木定位任务中的有效性与实用价值。

表 4 显示，ViT-Count 模型在参数量（11.65×10^6）和计算复杂度（1.82×10^9）上远低于多数高性能模型（如 UNET 和 Gramformer），同时显著优于传统大型网络（如 RESNET101 和 VGG）。相比轻量级模型虽计算更高，但性能更优。综上，ViT-Count 实现了精准计数和定位的同时，兼顾了较低的计算资源消耗，展现出较强的实用价值和部署潜力。

表 4　定位实验结果

Tab. 4　Localization experiment results

模型	Larch Casebearer			Urban Tree			参数量	计算量
	P	R	F_1	P	R	F_1		
Inception-v3	71.21	9.82	17.26	58.38	8.94	15.50	4.83×10^6	3.05×10^9
UNET	66.17	38.1	48.40	15.23	88.07	25.96	31.04×10^6	41.90×10^9
VGG	23.76	10.56	14.62	75.41	1.28	2.52	7.70×10^6	14.01×10^9
MCNN	14.77	49.60	22.76	41.66	9.48	15.44	7.03×10^6	0.13×10^9
BCNET	47.37	40.80	43.84	31.62	48.60	38.31	7.88×10^6	14.43×10^9
ViT	42.66	50.73	46.34	11.33	45.02	18.10	8.54×10^6	15.53×10^9
MCNNT	18.73	42.25	26.01	69.29	27.91	39.79	25.34×10^6	11.06×10^9
ALEXNET	42.4	27.03	33.04	17.67	41.25	24.74	2.50×10^6	0.75×10^9
Gramformer	64.51	32.24	42.99	46.34	37.28	41.32	38.43×10^6	20.73×10^9
RESNET50	50.71	77.73	60.54	49.17	39.34	45.51	8.67×10^6	7.53×10^9
MSTNet	88.41	49.28	63.29	45.22	34.55	40.08	15.26×10^6	1.33×10^9
RESNET101	51.73	72.91	61.65	47.84	33.67	39.52	27.68×10^6	22.46×10^9
EFFICIENTNET	56.14	55.54	55.84	52.98	41.64	43.37	4.56×10^6	1.05×10^9
ViT-Count	54.15	83.07	65.76	55.17	41.15	47.14	11.65×10^6	1.82×10^9

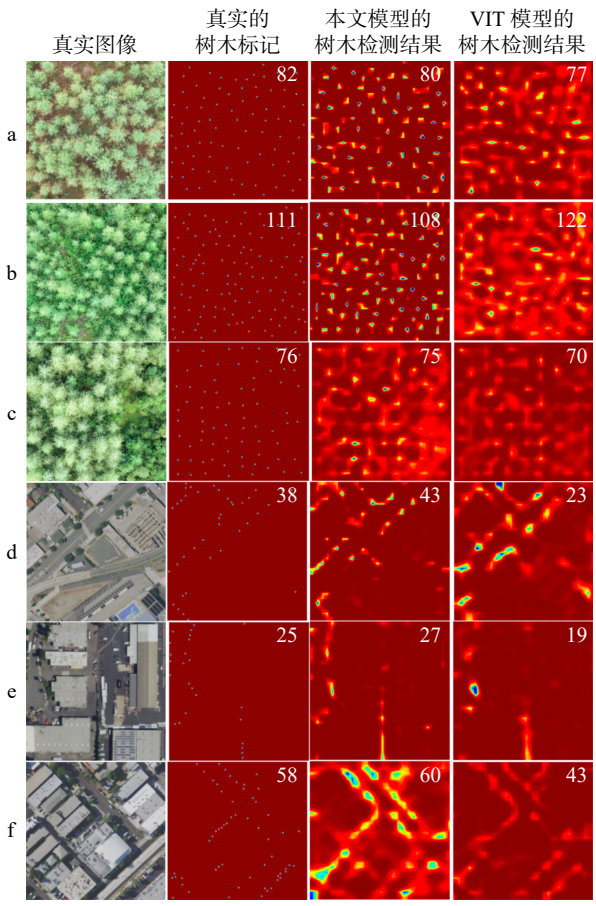
注：P为准确率，R为召回率，F_1分数，计算量用浮点运算次数FLOPs表示。

复杂场景下的目标特性对检测精度构成显著挑战，其中图 4a、b、c 图像代表典型森林复杂场景，该场景具有树木密集、树冠重叠严重、不同树种交错的特征。从图像对比可见，本文提出的 ViT-Count 模型相较于原始 ViT，在处理此类场景时展现出更强的树木目标区分能力与定位精度，能够有效应对树种交错带来的特征混淆问题，精准分离重叠树冠区域的独立目标。而图 4d、e、f 图像来自 Urban Tree 数据集，代表城市遥感图像场景，该场景存在背景干扰强(树木与建筑、道路等元素混杂)、树木目标尺寸小、尺度不一致的复杂性。在此场景中，ViT-Count 模型虽凭借核心模块协同作用保持了较高检测准确率，但受小尺度目标特征表达不充分、尺度差异导致的特征适配性不足等因素影响，仍存在一定程度的误检与漏检情况。

3.2 模块有效性

在 Larch Casebearer 数据集上，方法 1(VPT + CBAM)为最完整的设置，表现最佳，验证集和测试集上的 MAE 分别为 7.29 和 7.05，RMSE 为 9.49 和 9.69，说明 VPT 与 CBAM 的结合能够显著提升模型精度和稳定性。去除 CBAM 后的方法 2 性能下降，MAE 和 RMSE 均有所上升，表明 CBAM 在误差抑制方面具有积极作用。方法 3 在完全去除 VPT 和 CBAM 后，误差进一步扩大，验证了两个模块在提升模型性能方面的关键作用。相比之下，方法 4 虽然在一定程度上优于方法 3，但整体效果仍不及方法 1，说明 CBAM 在没有 VPT 配合的情况下作用有限(表 5)。整体结果表明，VPT 与 CBAM 协同使用对模型性能提升具有显著贡献。

在 Urban Tree 数据集的结果中，方法 1(VPT + CBAM)表现最佳，验证集和测试集上的 MAE 分别为 10.12 和 10.58，RMSE 为 12.86 和 13.75，表明两个模块在该数据集上同样具有显著效果。去除 CBAM 后的方法 2，误差略有上升(验证集和测试集上的 MAE 分别为 10.76 和 11.60，RMSE 为 13.95 和 15.88)，显示 CBAM 对提升精度具有积极作用。完全去除两个模块的方法 3 性能进一步下降，验证集和测试集上的误差均为最高，说明 VPT 与 CBAM 缺一不可。方法 4 仅保留了 CBAM，性能与方法 2



a、b、c 这 3 行图像来自 Larch Casebearer 数据集，代表典型的森林复杂场景，具有树木密集、树冠重叠、不同树种交错等特征；d、e、f 这 3 行图像来自 Urban Tree 数据集，代表城市遥感图像场景，具有背景干扰强、树木目标小、尺度不一致等复杂性。

图 4 树木计数与定位结果可视化对比

Fig. 4 Visual comparison of tree counting and localization results

表 5 消融实验结果

Tab. 5 Ablation experiment results

数据集	方法序号	VPT	CBAM	验证集		测试集	
				MAE	RMSE	MAE	RMSE
Larch Casebearer	1	√	√	7.29	9.49	7.05	9.69
	2	√	×	8.75	11.38	8.96	11.69
	3	×	×	10.26	13.35	10.59	13.89
	4	×	√	9.42	12.38	9.69	12.87
Urban Tree	1	√	√	10.12	12.86	10.58	13.75
	2	√	×	10.76	13.95	11.60	15.88
	3	×	×	11.29	14.45	11.62	15.43
	4	×	√	10.76	13.95	11.60	15.88

相近，但相比方法 1 仍有明显差距，表明 VPT 在模型性能提升中发挥了关键作用(表 5)。

从这两个数据集的消融实验结果来看，VPT 和 CBAM 的联合使用在提升模型性能方面发挥了关键作用。VPT 增强了模型的特征学习能力，而 CBAM 则帮助模型更加有效地关注到重要的特征区域。

3.3　鲁棒性

从图 5 中可以看出，本文提出的模型在未参与训练的不同类型数据上仍能保持良好的预测效果，误差维持在个位数范围内；相比之下，传统 ViT 模型在该数据集上存在较高的误检率，泛化性能相对较弱，进一步验证了本方法在实际应用中的有效性与鲁棒性。

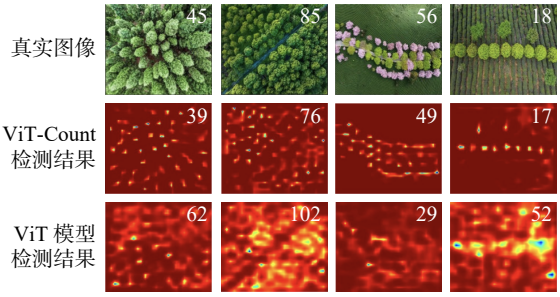


图 5　在未参与训练的不同类型数据上树木计数与定位结果可视化对比

Fig. 5　Visual comparison of tree counting and localization results on different types of untrained data

4　结　　论

本文面向复杂场景下树木计数与定位任务中存在的遮挡严重、背景干扰强及目标密集分布等难题，提出了一种基于 Vision Transformer(ViT)的高鲁棒性树木计数与定位方法。该方法充分发挥了 ViT 在建模全局上下文信息方面的优势，并结合本研究针对林地与城市场景定制设计的视觉提示调优机制，有效增强了模型对不同树种结构、光照条件及密度变化的适应能力。同时，设计卷积模块的注意力机制进一步强化了模型对局部结构和长距离依赖的感知能力，从而提升了对重叠、遮挡及形态相近树木目标的辨别性能，有效提升了树木计数与定位的准确性。尽管本文所提出的基于 Vision Transformer 的树木计数与定位方法在多种复杂场景中表现出良好的鲁棒性与泛化能力，但仍存在进一步研究的空间。当前林业数据的获取渠道增多，未来可结合多源遥感数据(如 LiDAR、高光谱等)与时序影像，提升模型对空间结构与树种变化的感知能力。未来的研究将侧重于多模态数据的协同建模与跨域适应能力的增强，旨在为大范围森林资源监测与城市生态管理提供更加高效且稳健的智能化解决方案。

参　考　文　献

[1] Maheswari P, Raja P, Apol O E, et al. Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques: a review[J]. Frontiers in Plant Science, 2021, 12: 684328.

[2] Ruiz-Garcia L, Lunadei L, Barreiro P, et al. A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends[J]. Sensors, 2009, 9(6): 4728−4750.

[3] Bao W, Lin Z, Hu G, et al. Method for wheat ear counting based on frequency domain decomposition of MSVF-ISCT[J]. Information Processing in Agriculture, 2023, 10(2): 240−255.

[4] Farjon G, Krikeb O, Hillel A B, et al. Detection and counting of flowers on apple trees for better chemical thinning decisions[J]. Precision Agriculture, 2020, 21: 503−521.

[5] Wan J, Wang Q, Chan A B. Kernel-based density map generation for dense object counting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(3): 1357−1370.

[6] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern Recognition, 2018, 77: 354−377.

[7] Zhang Y, Zhou D, Chen S, et al. Single-image crowd counting via multi-column convolutional neural network[C]//Lourdes A. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 589−597.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Lourdes A. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 770−778.

[9] Li Y, Zhang X, Chen D. Csrnet: dilated convolutional neural networks for understanding the highly congested scenes[C]// Michael S B. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Computer Society, 2018: 1091−1100.

[10] 金远航, 徐茂林, 郑佳媛. 基于改进 YOLOv4-tiny 的无人机影像枯死树木检测算法 [J]. 自然资源遥感, 2023, 35(1): 90−98.
Jin Y H, Xu M L, Zheng J Y. A dead tree detection algorithm based on improved YOLOv4-tiny for UAV images[J]. Remote Sensing for Natural Resources, 2023, 35(1): 90−98.

[11] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 1−15.

[12] Chen G, Shang Y. Transformer for tree counting in aerial images[J]. Remote Sensing, 2022, 14(3): 476.

[13] Jiang T, Freudenberg M, Kleinn C, et al. Detection transformer-based approach for mapping trees outside forests on high resolution satellite imagery[J]. Ecological Informatics, 2025, 87: 103114.

[14] Gibril M B A, Shafri H Z M, Shanableh A, et al. Large-scale assessment of date palm plantations based on UAV remote

sensing and multiscale vision transformer[J]. Remote Sensing Applications: Society and Environment, 2024, 34: 101195.

[15] 王凯, 陈优良, 胥寒莉. 基于改进 YOLOv7 的杨梅树单木检测 [J]. 农业工程学报, 2024, 40(11): 159−167.
Wang K, Chen Y L, Xu H L. Detection of individual trees of bayberry using improved YOLOv7[J]. Transactions of the Chinese Society of Agricultural Engineering, 2024, 40(11): 159−167.

[16] Jia M, Tang L, Chen B C, et al. Visual prompt tuning[C]//Shai A. European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 709−727.

[17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. arXiv: 1409.1556. [2025−09−28]. https://arxiv.org/abs/1409.1556.

[18] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25: 1−9.

[19] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Lourdes A. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 2818−2826.

[20] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation[C]//Nassir N. Medical image computing and computer-assisted intervention–MICCAI 2015. Munich: Springer International Publishing, 2015: 234−241.

[21] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[C]// Alexander R. Proceedings of the International Conference on Learning Representations. Seattle: OpenReview, 2020: 1−22.

[22] Fang S, Pan X, Xiang S, et al. Meta-MSNet: meta-learning based multi-source data fusion for traffic flow prediction[J]. IEEE Signal Processing Letters, 2020, 28: 6−10.

[23] Ke L, Tai Y W, Tang C K. Deep occlusion-aware instance segmentation with overlapping bilayers[C]//Nicole F. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville: IEEE Computer Society, 2021: 4019−4028.

[24] Lu R. A novel fault diagnosis method for imbalanced datasets based on MCNN-Transformer model in industrial processes[J]. International Journal of Adaptive Control and Signal Processing, 2024, 38(6): 1−17.

[25] 邵凯, 王明政, 王光宇. 基于 Transformer 的多尺度遥感语义分割网络 [J]. 智能系统学报, 2024, 19(4): 920−929.
Shao K, Wang M Z, Wang G Y. Transformer-based multiscale remote sensing semantic segmentation network[J]. CAAI Transactions on Intelligent Systems, 2024, 19(4): 920−929.

[26] Lin H, Ma Z, Hong X, et al. Gramformer: learning crowd counting via graph-modulated transformer[C]//Ida C. Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2024: 3395−3403.

[27] 齐建东, 马鐘添, 郑尚姿. 基于 YOLOv7 的红外相机野生动物图像筛选 [J]. 北京林业大学学报, 2024, 46(2): 143−154.
Qi J D, Ma Z T, Zheng S Z. Wildlife image screening for infrared cameras based on YOLOv7[J]. Journal of Beijing Forestry University, 2024, 46(2): 143−154.

（责任编辑 吴 娟
责任编委 陈志泊）