面向动态环境的巡检机器人轻量级语义视觉SLAM框架

余浩扬[②] 李艳生*[①] 肖凌励[②] 周继源[②]

[①](重庆邮电大学人工智能学院(重邮科大讯飞人工智能学院) 重庆 400065)

[②](重庆邮电大学集成电路学院(重庆国际半导体学院) 重庆 400065)

摘 要：为提升巡检机器人在城市动态环境中的定位精度与鲁棒性，该文提出一种基于第3代定向快速与旋转简要同步定位与建图系统 (ORB-SLAM3)的轻量级语义视觉同步定位与建图(SLAM)框架。该框架通过紧耦合所提出的轻量级语义分割模型(DHSR-YOLOSeg)输出的语义信息，实现动态特征点的精准剔除与稳健跟踪，从而有效缓解动态目标干扰带来的特征漂移与建图误差累积问题。DHSR-YOLOSeg基于YOLO第11代轻量级分割模型(YOLOv11n-seg)架构，融合动态卷积模块(C3k2_DynamicConv)、轻量特征融合模块(DyCANet)与复用共享卷积分割(RSCS)头，在分割精度小幅提升的同时，有效降低了计算资源开销，整体展现出良好的轻量化与高效性。在COCO数据集上，相较于基础模型，DHSR-YOLOSeg实现参数量减少13.8%、10^9次浮点运算(GFLOPs)降低23.1%、平均精度指标(mAP50)提升约2%；在KITTI数据集上，DHSR-YOLOSeg相比其他主流分割模型及YOLO系列不同版本，在保持较高分割精度的同时，进一步压缩了模型参数量与计算开销，系统整体帧率得到有效提升。同时，所提语义SLAM系统通过动态特征点剔除有效提升了定位精度，平均轨迹误差相比ORB-SLAM3降低8.78%；在此基础上，系统平均每帧处理时间较主流方法如DS-SLAM和DynaSLAM分别降低约18.55%与41.83%。研究结果表明，该语义视觉SLAM框架兼具实时性与部署效率，显著提升了动态环境下的定位稳定性与感知能力。

关键词：巡检机器人；语义分割；视觉同步定位与建图；动态特征点剔除

中图分类号：TN919.85 **文献标识码**：A **文章编号**：1009-5896(2025)10-3979-14

DOI: 10.11999/JEIT250301 **CSTR**: 32379.14.JEIT250301

1 引言

近年来，巡检机器人作为智慧城市与智能运维的重要组成，已逐步部署于高校校园、产业园区等各类城市环境，承担着设施巡检[1]、环境监测[2]与安全巡逻等任务。作为支撑机器人实现自主导航与环境建图的关键技术，同步定位与地图构建(Simultaneous Localization And Mapping, SLAM)已成为移动机器人领域的研究热点。其中，视觉SLAM凭借硬件依赖低、信息密度高，广泛应用于各类中小型巡检平台。然而，城市环境普遍具有开放性强、动态物体频繁[3]、结构布局复杂[4]等特点，大量非结构化动态元素(如行人、车辆)频繁穿行于机器人路径，严重干扰视觉SLAM的特征提取与跟踪，导致定位误差累积、回环识别困难，最终影响地图的完整性与一致性。此外，受限于巡检机器人平台在功耗、体积与成本等方面的部署限制[5]，当前任务对感知算法的实时性与计算效率提出了实际需求。传统深度网络因模型复杂、推理成本高，难以直接部署在边缘设备上，进一步推动了轻量化视觉算法的研究。因此，如何设计一套兼具轻量性、动态感知与鲁棒定位能力的语义视觉SLAM框架，已成为城市巡检机器人亟需解决的关键问题。

多数视觉SLAM仍基于静态世界假设，难以有效应对动态场景中的频繁干扰，易引发特征漂移、建图误差积累甚至系统失效。其中，定向快速与旋转简要同步定位与建图系统 (Oriented FAST And Rotated BRIEF Simultaneous Localization and Mapping, ORB-SLAM)系列(ORB-SLAM[6], ORB-SLAM2[7], ORB-SLAM3[8])作为代表性视觉里程计方法，主要依赖静态背景特征点建图，在静态或弱动态环境下具备良好定位与建图性能，但未显式建模动态干扰，鲁棒性受限于环境稳定性。单目视觉惯性导航系统 (Visual-Inertial Navigation System-Monocular, VINS-Mono)[9]通过紧耦合融合视觉与惯性测量单元(Inertial Measurement Unit, IMU)数据，在高速运动下定位稳定性较好，但前端仍依赖静态特征，缺乏动态物体识别与剔除。部分研究引入目标检测辅助，提取潜在动态区域并剔除其中特征点，以增强动态环境下的鲁棒性，但主流方法多

收稿日期：2025-04-25；改回日期：2025-07-28；网络出版：2025-08-04

*通信作者： 李艳生 liyansheng@cqupt.edu.cn

基金项目：国家自然科学基金(52575102)，重庆市城市管理局项目(城管科2022-34)，重庆市自然科学基金(CSTB2022NSCQ-MSX0340)

Foundation Items: The National Natural Science Foundation of China (52575102), Chongqing Urban Administration Bureau Project (Urban Management Science Section 2022-34), Chongqing Natural Science Foundation (CSTB2022NSCQ-MSX0340)

基于检测框表达，易误剔静态特征，影响建图完整性。例如，Zang等人[10]提出改进型ORB-SLAM3，结合YOLOv5检测动态目标框，并引入卢卡斯–卡纳德光流法 (Lucas-Kanade optical flow method, LK)光流分析帧间运动一致性，缓解检测框边界粗糙导致的误剔问题，提升动态性判别精度，但也增加了系统复杂度；Wu等人[11]提出的语义-激光-全球导航卫星系统融合同步定位与建图系统(Semantic-Laser-GNSS Simultaneous Localization And Mapping, SLG-SLAM)在跟踪线程中引入单次多框检测器 (Single Shot multibox Detector, SSD)检测器，直接屏蔽检测框内特征点，虽实现简洁，但区域粗略，仍依赖额外传感器补偿误差。为提升动态特征剔除精度，部分研究采用语义分割掩码替代检测框，提供更精细的动态区域剔除。典型如动态同步定位与建图系统 (Dynamic Simultaneous Localization And Mapping, DynaSLAM)[12]采用掩码区域卷积神经网络 (Mask Region-based Convolutional Neural Network, Mask-RCNN)提取像素级动态区域，并结合ORB-SLAM2实现了在动态场景中的鲁棒建图。Vincent等人[13]进一步融合扩展卡尔曼滤波器与深度语义分割网络，实现了动态目标的跟踪与遮蔽，系统时可达到14 帧/s的处理速度。但此类方法通常依赖高复杂度深度分割网络，计算资源开销大，难以满足实时处理要求，限制了其在边缘设备和资源受限平台上的部署。

相较而言，YOLO(You Only Look Once)系列模型凭借端到端架构、快速推理与良好的检测分割一体化能力，在轻量化语义感知中具备更高部署适应性。其扩展的分割模型在兼顾精度与实时性的同时，降低了模型复杂度与计算消耗，为动态感知与实时SLAM融合提供了可行路径。同时，YOLOv11[14]作为该系列中近期具有代表性的版本，在结构上引入C3k2模块、快速空间金字塔池化(Spatial Pyramid Pooling-Fast, SPPF)与并行空间注意力(Convolutional block with Parallel Spatial Attention, C2PSA)机制，在提升特征提取能力的同时进一步压缩模型复杂度，相较YOLOv8等前代模型，在特征提取效率、多尺度信息融合与关键区域感知能力方面实现了结构性优化，展现出更强的多任务适应性与部署效率。为此，将YOLO框架扩展的分割模型集成至视觉SLAM系统，既能够在保持语义分割效率的同时实现轻量级的像素级动态区域感知，也能够为后续的语义建图与智能巡检任务提供丰富的场景语义信息。尽管此类方法具有较高的工程部署潜力，但现有研究多基于室内红绿蓝与深度(Red Green Blue and Depth, RGB-D)[15–17]图像数据集进行训练与验证，场景动态性有限，且大多仍采用YOLOv5[18]与YOLOv8[19]等早期模型，模型轻量化设计与实时性能优化不足，难以满足复杂户外环境下双目视觉输入、多目标动态感知与实时地图构建等综合应用需求。目前，在结构轻量、感知精准、计算高效且可与SLAM系统深度融合的端到端语义分割框架方面，仍存在明显空缺。

综上所述，针对双目视觉输入、动态目标感知与实时地图构建等多重约束条件下语义视觉SLAM系统缺乏结构轻量、性能可靠的端到端解决方案的问题，本文提出一种适用于动态环境的轻量级语义分割模型(DHSR-YOLOSeg)，并将其紧耦合至ORB-SLAM3双目模式中，构建了一个具备语义感知能力的视觉SLAM系统。主要工作如下：

(1)结合超轻量动态上采样算子(an ultra-lightweight and effective Dynamic upSampler, DySample)与通道注意力-分层特征金字塔网络(Channel Attention-Hierarchical Spatial Feature Pyramid Network, ChannelAttention_HSFPN)，本文设计了一种轻量特征融合模块(DyCANet)，并用于重构YOLOv11n-seg的Neck结构，从而提升动态区域的语义表达能力与边界分割精度。

(2)在YOLOv11n-seg基础上提出结构优化策略，集成融合动态卷积模块(C3k2 Dynamic Convolution module, C3k2_DynamicConv), DyCANet和复用共享卷积分割(Reused and Shared Convolutional Segmentation, RSCS)头，构建DHSR-YOLOSeg模型，在保持网络轻量化的同时，实现了对动态区域的高效精确分割。

(3)将DHSR-YOLOSeg模型输出的语义信息集成至ORB-SLAM3的跟踪线程，在双目输入下实现动态区域识别与特征点剔除，提升系统在城市动态环境下的精度与效率。与面向动态环境的语义视觉同步定位与建图系统(a semantic visual SLAM towards Dynamic environments, DS-SLAM), DynaSLAM等主流语义SLAM系统对比，结果表明本文方法在误差抑制、运行速度与部署适应性方面具备更优综合性能。

2 方法概述

2.1 系统整体架构

本文基于ORB-SLAM3框架构建了一个语义增强的视觉SLAM系统，其整体结构如图1所示。系统保留了ORB-SLAM3原有的3线程结构，包括：跟踪线程(Tracking)、局部建图线程(Local Mapping)与回环检测线程(Loop Closing)。在此基础

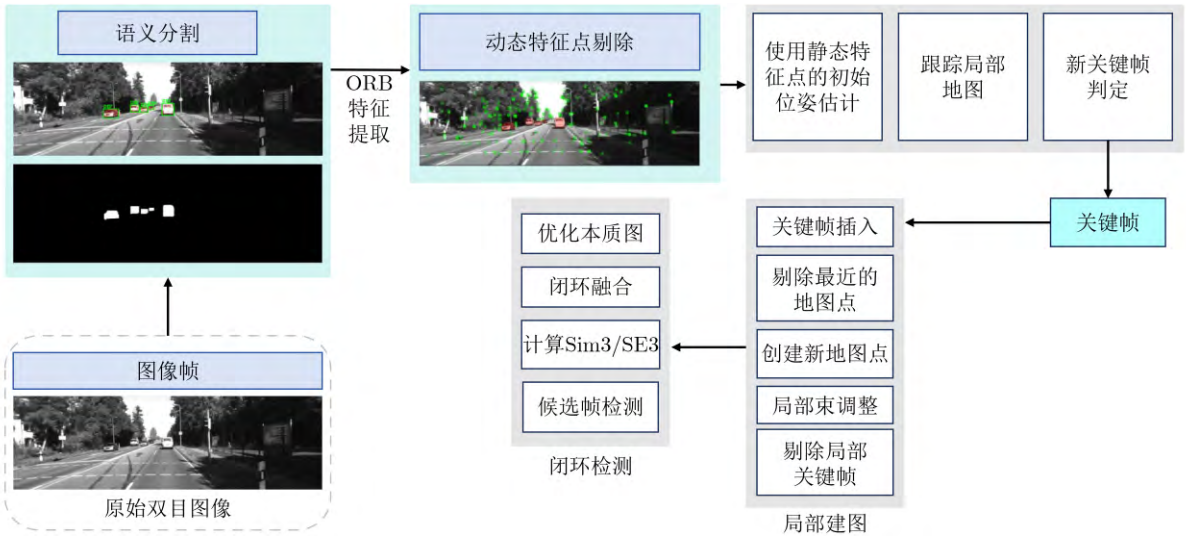


图 1 语义分割与ORB-SLAM3融合的系统框架图

上，本文在跟踪线程引入轻量级语义分割模块DH-SR-YOLOSeg，实现对图像中动态目标的像素级识别。所提语义信息用于引导跟踪线程识别并剔除动态特征点，有效增强系统在动态环境下的定位鲁棒性与建图一致性。相关特征剔除机制将在2.7节中详细介绍。其余两个线程维持ORB-SLAM3原有流程，分别负责关键帧优化与全局一致性维护，在保证定位精度的同时兼顾系统实时性。

2.2 YOLOv11n-seg模型架构

YOLOv11n-seg整体采用主干网络(Backbone)、特征融合颈部(Neck)和分割头部(Segment Head)3段式架构。主干通过多层卷积和C3k2结构逐步提取特征，高层引入快速空间金字塔池化模块增强上下文信息，结合并行空间注意力机制模块提升全局感知能力。颈部通过多次上采样与特征拼接(Concat)构建特征金字塔，融合多尺度语义信息。分割头部基于主干输出的特征图**P3**, **P4**和**P5**，利用Segment层在各尺度分别输出分割结果，实现多尺度动态目标的像素级识别与分割。该模型支持通过深度、宽度和最大通道数的复合缩放，灵活适配不同算力平台。整体结构如图2所示。

2.3 基于动态卷积的计算优化方法

近年来，随着计算硬件与数据工程的突破，大规模视觉预训练已成为计算机视觉发展的核心驱动力。这类模型通过自监督学习构建通用视觉表征，在图像识别、目标检测以及语义分割等任务中展现出强大的跨任务迁移能力。然而，主流模型常面临数据规模、参数量与计算成本(Giga FLoating-point Operations Per second, GFLOPs)的“三重困境”：性能提升往往伴随计算资源的指数级增长。在这一背景下，动态卷积(DynamicConv)[20]技术应运而生，针对上述问题提出了创新性方案。该方法通过动态权重机制，在不显著增加计算开销的前提下，实现了参数量的指数级扩展。具体而言，DynamicConv是一种基于条件计算(Conditional Computation)策略的动态卷积方法，其核心思想是根据输入特征动态调整卷积核参数。这种机制能够在几乎不增加浮点运算量的前提下，显著提升模型的表征能力和预测准确率。

DynamicConv借鉴了混合专家(Mixture-of-Experts, MoE)条件计算框架，其机制如图3所示，MoE通过并行部署多个专家网络增强模型表达能力，利用门控机制动态选择部分专家参与计算，在不显著增加计算量的情况下扩展模型容量。DynamicConv创新性地将MoE[21]中的专家网络替换为多个卷积核，结合输入特征自适应的动态权重调整，使不同输入激活不同卷积核组合。该设计既保留了卷积操作的低计算开销优势，又显著提升了特征提取的灵活性，提升视觉任务中的卷积效率与适应性。本文基于DynamicConv构建的C3k2模块通过动态参数调节机制有效提升了目标检测性能。该模块创新性地将条件计算策略融入网络结构，可根据输入特征实时动态调整卷积核参数。相比传统固定结构，动态自适应机制使网络更灵活地应对复杂检测场景中的尺度变化与遮挡问题。DynamicConv的工作原理可以通过式(1)和式(2)表示

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{W}^{'} \tag{1}$$

$$\boldsymbol{W}^{'} = \sum_{i=1}^{M} \alpha_i \boldsymbol{W}_i \tag{2}$$

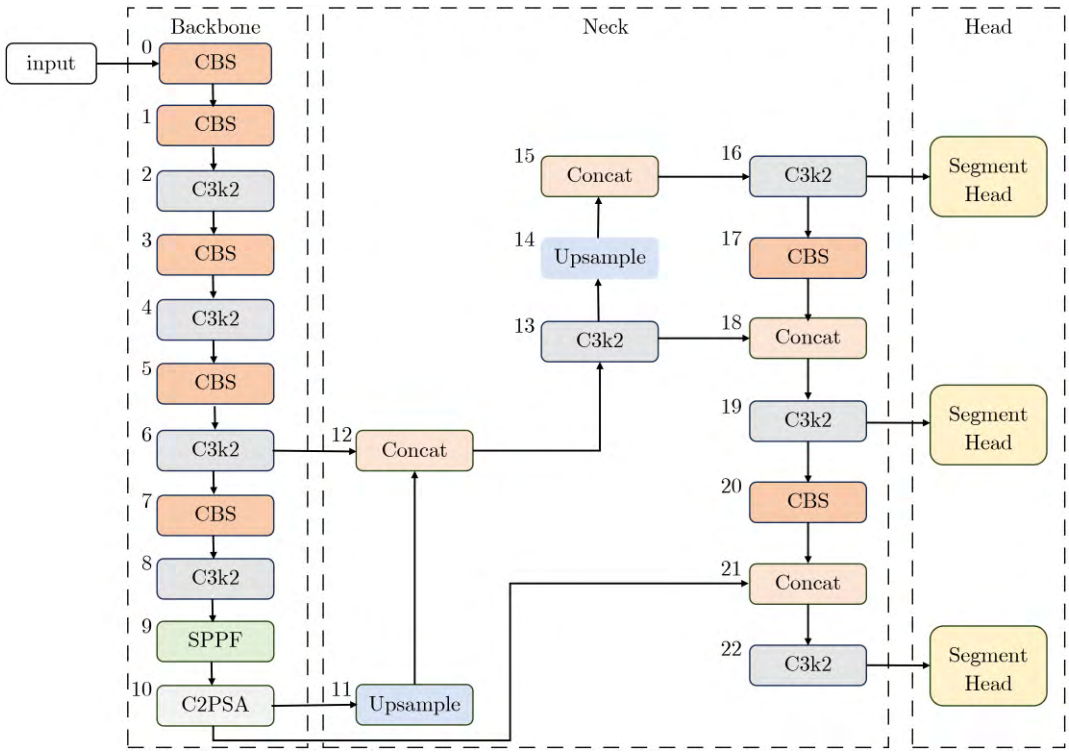


图 2 YOLOv11n-seg模型结构图

(a) 普通层 (b) MoE层，动态选择专家计算

图 3 MoE机制示意图

其中，$\boldsymbol{W}'$是动态生成的卷积核，$\boldsymbol{W}_i$表示第i个专家权重，α_i是通过输入特征动态计算的权重系数，具体生成方式为

$$\alpha = \mathrm{softmax}\left(\mathrm{MLP}\left(\mathrm{Pool}\left(\boldsymbol{X}\right)\right)\right) \tag{3}$$

这一设计通过全局平均池化(Pool)提取全局特征，再通过两层多层感知机(MultiLayer Perceptron, MLP)生成动态权重系数α_i。与传统卷积相比，DynamicConv不仅提升了模型对复杂语义特征的适应能力，同时相较于传统卷积结构显著降低了GFLOPs，从而在分割精度与推理效率之间取得了更优平衡。

2.4 基于DyCANet的特征融合结构

为提升动态场景下目标区域的语义感知能力与空间重建精度，本文在YOLOv11n-seg的Neck模块基础上进行结构改造，借鉴DySample方法[22]与ChannelAttention_HSFPN模块[23]的设计思想，提出一种融合动态点采样与通道注意力机制的轻量化特征融合结构，命名为DyCANet。该结构专为动态目标识别任务优化设计，在保持推理效率的同时，显著提升多尺度特征在动态区域内的判别能力与边界表达能力。

首先，本文引入了一种基于点重采样机制的轻量动态上采样算子——DySample，其核心思想是跳过传统动态卷积核的生成与计算，直接对输入特征图进行内容感知的空间采样。不同于内容感知特征重组方法和解码器编码器融合方法等依赖动态卷积的上采样机制，DySample回归“上采样的本质”，将整个过程建模为一个基于偏移点的网格采样任务，如图4所示。

在该框架中，DySample利用轻量的线性映射预测每个位置的采样偏移量，并通过PyTorch提供的内置函数grid_sample对输入特征$\boldsymbol{X}$进行重采样

$$\boldsymbol{X}' = \mathrm{grid\_sample}(\boldsymbol{X}, \boldsymbol{S}) \tag{4}$$

$$\boldsymbol{O} = \lambda \cdot \mathrm{Linear}\left(\boldsymbol{X}\right) \tag{5}$$

$$\boldsymbol{S}=\boldsymbol{G}+\boldsymbol{O} \tag{6}$$

其中，$\boldsymbol{G}$为标准采样网格，Linear(·)表示通过一层线性变换生成的偏移映射，$\boldsymbol{O}$为每个位置对应的偏移量，$\boldsymbol{S}$表示加权后的动态采样点集，最终通过grid_sample实现内容感知的上采样，得到上采样结果$\boldsymbol{X}'$。这里，s表示上采样倍率(决定特征图空间分辨率的放大倍数)，g表示分组数量(决定通道被划分的组数，每组生成二维偏移)。DySample带来的高质量语义重建能力为后续的通道注意力引导融合奠定了良好基础，以下将进一步探讨融合结构中的注意力机制设计。为进一步提升多尺度语义信息的融合与表达能力，本文引入了ChannelAttention_HSFPN模块，该模块融合了通道注意力机制与分层特征金字塔结构(Hierarchical Spatial Feature Pyramid Network, HS-FPN)，具备通道选择能力强、多尺度特征整合效果好的优势，尤其适用于复杂背景与小目标遮挡等场景，在保持轻量化的同时，有效增强了模型的特征表达能力。ChannelAttention_HSFPN的整体结构由两个阶段组成，分别为特征选择阶段与特征融合阶段，如图5所示。

在特征选择阶段，输入特征图首先通过最大池化(Max Pooling)和平均池化(Average Pooling)2条分支提取通道层面的统计特征，随后各分支分别通过2层1×1卷积进行通道映射。2条分支输出在通道维度相加，并经Sigmoid激活生成通道注意力图，最终用于对输入特征图进行通道加权，强化重要语义信息，抑制无关干扰特征。在特征融合阶段，模块采用自上而下的引导策略，利用转置卷积对高层特征图上采样，结合双线性插值调整尺寸，使其与低层特征图对齐。上采样特征图输入通道注意力模块，对低层特征图进行加权调制。最后，高层引导信息与低层细节特征相加，完成语义引导下的细节增强。通过上述两阶段协同，ChannelAttention_HSFPN模块在保持特征完整性的同时，突出关键通道，引导空间融合，最终实现语义与细节的高效协同表达，显著提升模型在复杂环境下的目标检测与语义分割性能。综上，DyCANet结合DySample的动态点采样与ChannelAttention_HSFPN的通道引导机制，在基础模型原有Neck架构上完成了重构优化。该结构在不显著增加计算成本的前提下，增强了多尺度语义建模与细节还原能力，尤其适用于复杂背景、小目标遮挡与动态干扰显著的场景，为分割模块提供了更具判别性的语义特征支撑。

2.5　基于共享卷积分割头的结构设计与优化方法

在YOLOv11n-seg中，如图6所示，来自特征

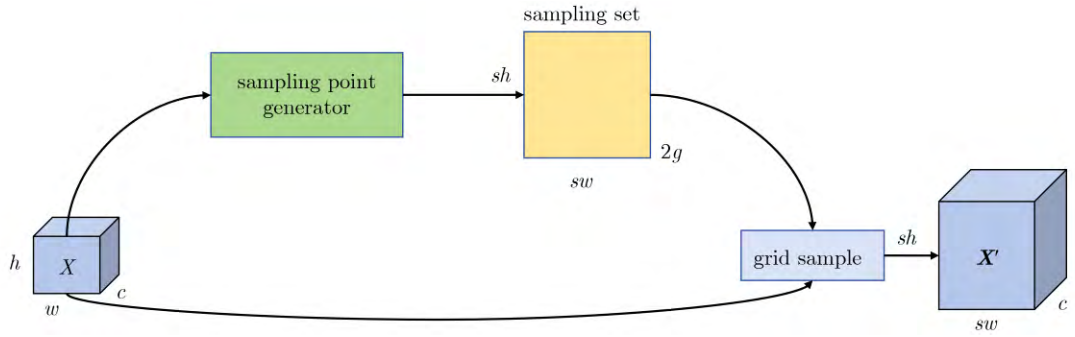


图 4 DySample的整体上采样流程

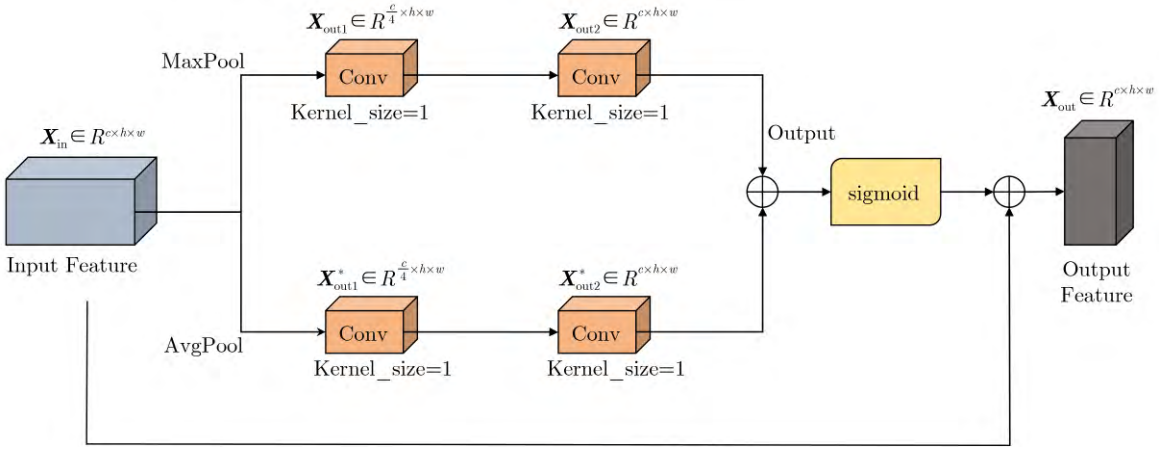


图 5 ChannelAttention_HSFPN结构图

金字塔不同尺度的特征图(**P3**, **P4**, **P5**)分别送入独立的分割头。每个分割头内部采用三分支解耦结构，语义掩码分支为主，边界框回归与类别预测分支为辅，以引导更精确的掩码生成。各分支在输出前均堆叠3次3×3卷积，并依次进行归一化(BN)与激活函数处理，提升特征表达与预测能力。然而由于各尺度分割头之间卷积模块未共享，带来大量参数与计算冗余，显著增加了模型体积与推理开销，制约了其在边缘平台与低功耗设备上的部署效率。

为进一步优化分割头结构效率，本文借鉴复用共享卷积检测头(Reused and Shared Convolutional Detection, RSCD)设计理念[24]，对其结构进行适配改造并集成，作为新的分割解码头模块，如图7所示。改进结构中，不同尺度特征图首先通过1×1卷积进行通道压缩，统一送入共享的3×3卷积模块，结合BN层与激活函数完成基础特征建模。随后，网络分为3路输出类别预测(cls)、边界回归(reg)与掩码生成(mask)，各分支均包含独立的后处理卷积层，

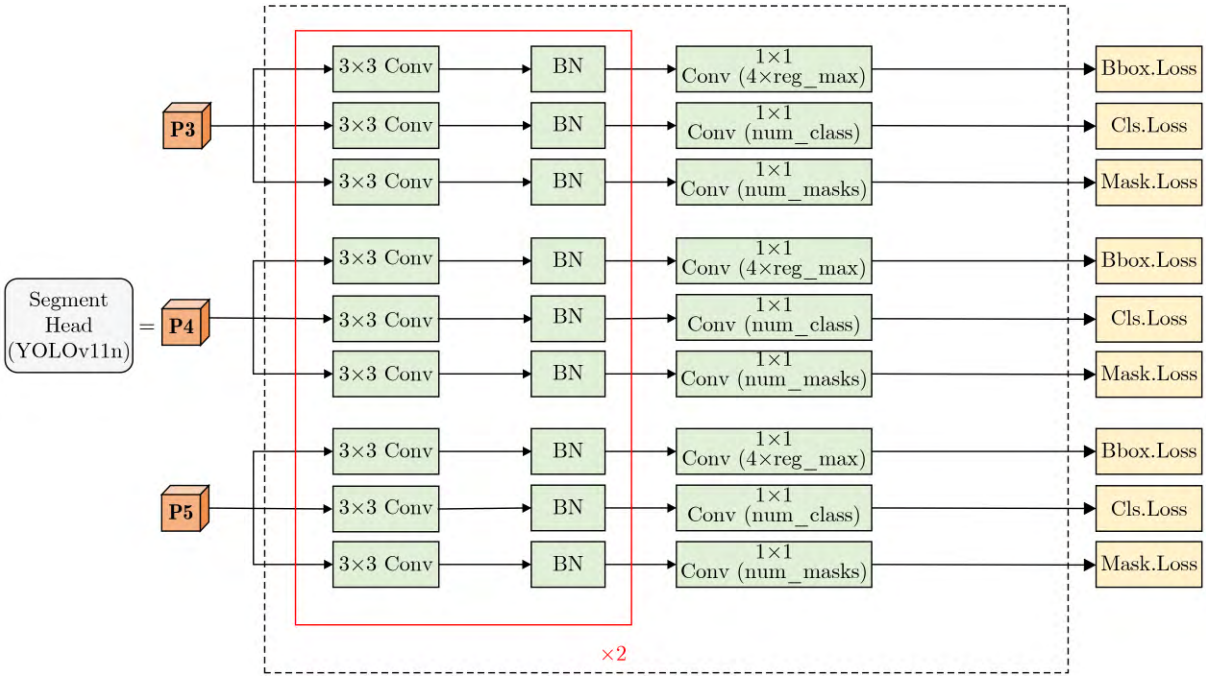


图 6 YOLOv11n分割头结构图

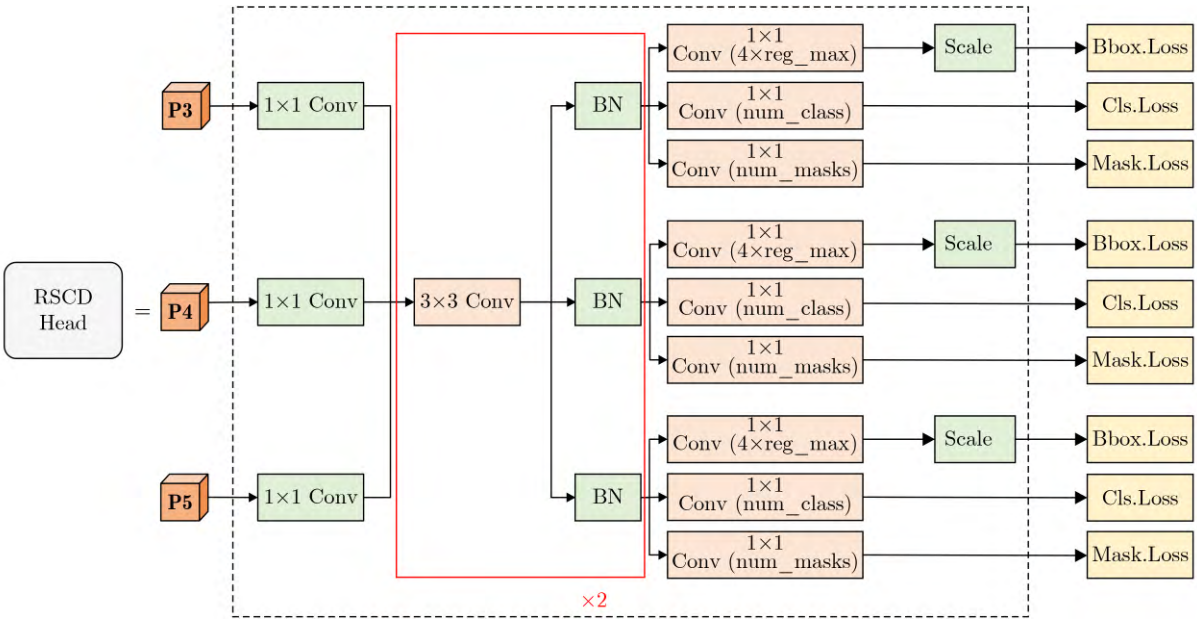


图 7 RSCD头结构图

前置共享模块有效减少了冗余计算。针对分割任务的尺度敏感性，边界回归分支后进一步引入尺度调整层(scale-layer)，提升目标边缘解析能力。

本文基于改进后的RSCD模块，重新设计了复用共享卷积分割头(Reused and Shared Convolutional Segmentation, RSCS)，并集成至YOLOv11n-seg，替代原有分割头结构。该结构通过参数共享与路径复用，有效降低了多尺度分支的冗余与计算开销，在大幅减小模型复杂度的同时，保持了分割精度与边界感知能力，提供了高效可部署的结构优化方案。

2.6 DHSR-YOLOSeg模型结构

基于前述模块设计，本文构建了一种面向动态环境视觉感知的轻量级语义分割模型DHSR-YOLOSeg，其网络结构图如图8所示，旨在实现高效动态区域掩码提取，满足语义SLAM系统对实时性与精度的双重需求。本模型以YOLOv11n-seg为基础，在原C3k2模块中引入动态卷积(DynamicConv)与条件计算策略，构建了轻量高效的C3k2_DynamicConv模块，在不显著增加计算量的前提下提升特征提取能力；同时，结合DySample上采样算子与ChannelAttention_HSFPN形成DyCANet模块，有效增强多尺度特征的语义表达与边界分割精度；此外，分割头部分引入RSCS结构，替代YOLOv11n-seg原分割头，实现多尺度卷积的共享与复用，显著减少参数冗余与计算开销。

通过上述多模块协同设计，DHSR-YOLOSeg实现了结构紧凑、语义表达充分、计算代价可控的目标掩码生成能力。该模型在保证运行效率的同时，有效提升了动态目标区域的分割质量，以精细语义分割结果引导特征点剔除，为视觉SLAM系统提供了鲁棒、可信的语义信息输入。

2.7 基于语义信息的ORB-SLAM3紧耦合优化

为缓解视觉SLAM在动态场景下因运动物体引起的特征漂移与地图污染问题，本文提出一种基于语义分割引导的特征点剔除方法。本方法利用DHSR-YOLOSeg模型提供的语义信息，在ORB-SLAM3的跟踪线程中实时过滤动态区域内的特征点，提升系统在高动态环境下的定位稳定性与建图精度。具体流程为：系统接收双目图像并同步输入至DHSR-YOLOSeg，生成当前帧语义掩码；通过图像坐标映射构建动态区域掩码；在特征提取与匹配阶段，依据掩码剔除动态区域内的ORB特征点，避免动态物体对位姿估计的干扰。本方法仅对Tracking线程进行轻量集成，保持原有系统结构与计算流程不变，具备非侵入式、低冗余的工程优势。同时，该

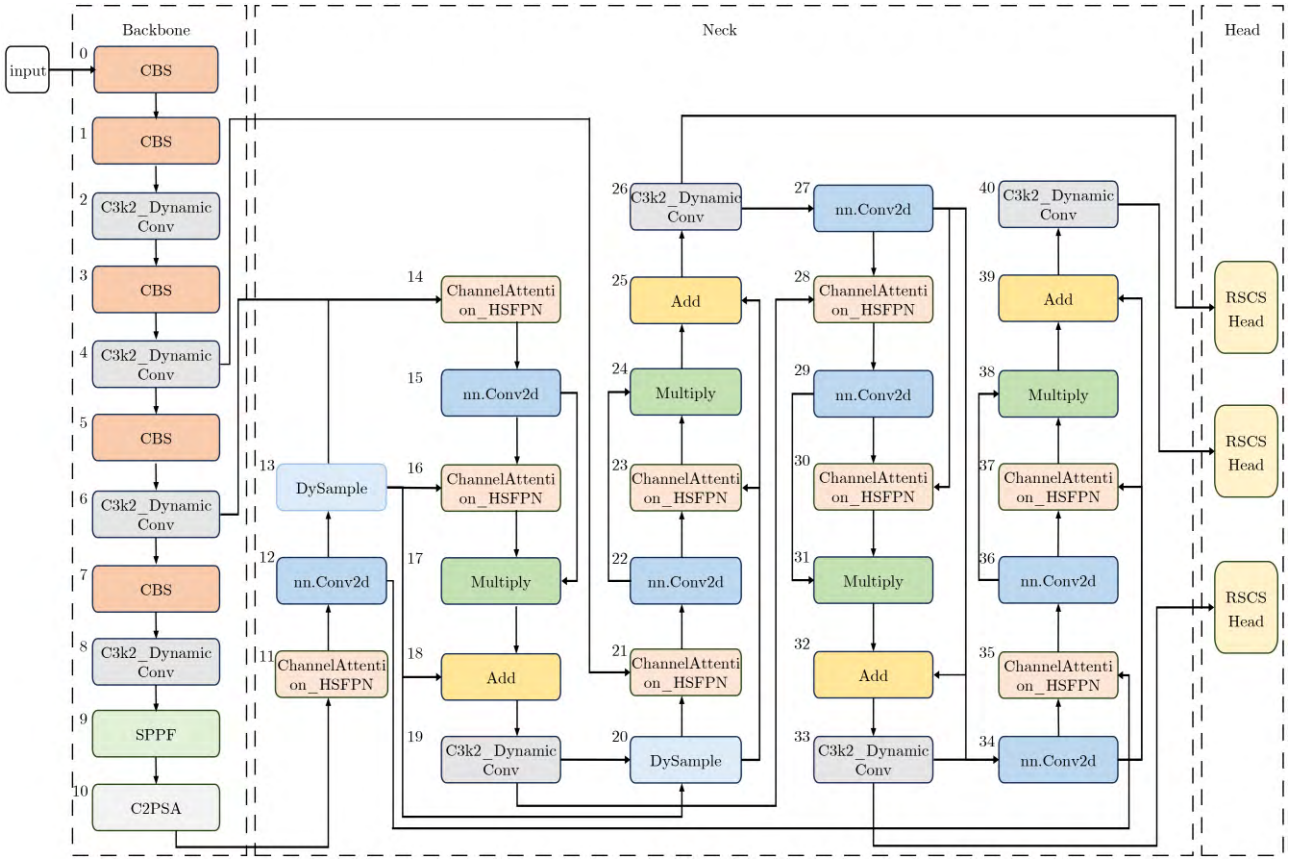


图 8 DHSR-YOLOSeg网络结构图

策略无需依赖光流或几何一致性等高计算成本模块，兼顾了实时性与特征筛选的鲁棒性。实验证明，基于语义感知的动态区域剔除显著增强了系统在多种高动态户外环境下的定位稳定性与精度一致性，为后续多源融合、语义建图及高层任务理解提供了可靠的前端感知基础。

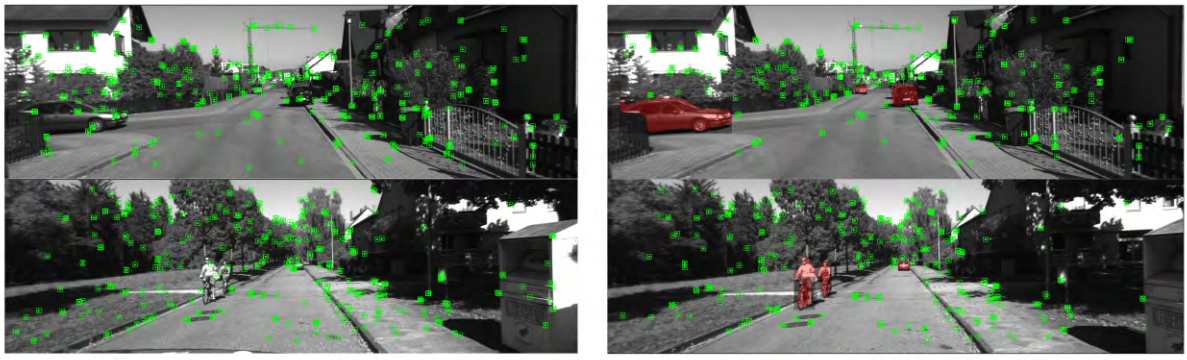
为直观展示融合语义信息后的动态特征点剔除效果，本文在典型城市动态街景中选取部分帧图像，对比ORB-SLAM3系统原始特征提取结果与引入语义剔除后的特征分布，如图9所示。图9(a)为传统ORB-SLAM3系统的特征提取结果，动态目标(如行人、自行车、车辆)区域内存在大量特征点，易受目标运动影响，导致特征漂移与匹配误差。图9(b)为引入DHSR-YOLOSeg模型后，系统基于语义分割结果对动态特征点进行筛除，仅保留静态背景特征用于位姿估计与地图优化。该剔除机制在城市道路场景中对典型动态干扰具有良好的适应性与抑制效果，为系统在复杂动态环境下的稳定定位与鲁棒建图提供了关键支撑。

3 实验结果与分析

3.1 消融实验

为验证DHSR-YOLOSeg模型中各结构模块对语义分割性能与计算效率的影响，本文在COCO数据集上设计并开展了多组消融实验。实验聚焦于“person”,“car”,“bicycle”和“motorcycle”等城市环境下的典型动态目标，覆盖常见动态干扰因素。为增强模型的泛化能力，训练过程中引入了随机缩放、平移、翻转和颜色扰动等多种数据增强策略。消融实验围绕3个关键结构模块依次展开，分别为C3k2_DynamicConv(模块1)、DyCANet(模块2)和RSCS头(模块3)，通过逐步剔除或组合模块以分析各自性能贡献。所有实验均在搭载NVIDIA GeForce RTX 3070 Ti(8 GB显存)的笔记本平台上完成，训练与推理均基于PyTorch框架，输入图像尺寸统一为640×640。评估指标包括mAP(IoU阈值设为0.5)、参数量、模型大小、GFLOPs与帧率(fps)，用于综合衡量模型检测精度与计算效率。

从表1可见，DHSR-YOLOSeg模型在不同结构模块引入后的性能表现呈现出典型的精度-复杂度权衡特性。引入DynamicConv模块(实验1)后，参数量增加31.1%、模型体积扩大30%，但mAP50与基准模型一样，说明其在提升特征表达方面具有正向贡献；DyCANet模块(实验2)则在参数量和模型大小分别下降26.5%和25%的同时，仅造成约2%的



(a) ORB-SLAM3 特征点图　　(b) 剔除后的特征点图

图 9 特征点剔除前后对比

表 1 消融实验结果

实验序号	模块1	模块2	模块3	参数量(M)	GFLOPs	模型大小	准确率	mAP50
基准模型	×	×	×	2.83	10.40	6.00	0.76	0.51
1	√	×	×	3.71	10.00	7.80	076	0.52
2	×	√	×	2.08	9.00	4.50	0.75	0.50
3	×	×	√	2.58	9.10	7.00	0.76	0.51
4	√	√	×	2.63	8.90	5.60	0.76	0.51
5	√	×	√	3.44	8.90	8.70	0.75	0.52
6	×	√	√	1.89	8.10	4.60	0.73	0.49
7	√	√	√	2.44	8.00	5.70	0.77	0.51

mAP50下降，表现出良好的轻量化与保精度能力；RSCS模块(实验3)在GFLOPs下降12.5%的条件下维持与基准模型相同的分割精度，验证其结构复用设计的高效性。在多模块组合配置中，完整集成3模块的DHSR-YOLOSeg(实验7)展现出最优综合性能：相比基准模型，参数量减少13.8%、GFLOPs下降23.1%、模型大小缩小5%，准确率提升1.3%，mAP50与基准模型一样，体现了各模块之间良好的协同增强效应。相比之下，实验6(DyCANet+RSCS)在精度下降约4%的代价下达到了最低GFLOPs与模型体积。综上，DHSR-YOLOSeg通过3模块有机集成，在保持计算代价低的前提下有效提升了分割精度与系统部署效率，验证了其在动态语义感知任务中的实用性与优越性。

3.2 对比实验

为进一步验证DHSR-YOLOSeg在动态视觉场景下的语义分割效率与实际部署性能，本文基于KITTI数据集中典型城市道路场景，选取多种具有代表性的语义分割模型，开展推理效率与计算复杂度评估。所选模型涵盖不同技术路线与模型类别，其中，Mask R-CNN 为广泛应用于语义感知任务的经典深层结构，YOLOv5n-seg, YOLOv8n-seg与YOLOv11n-seg分别体现了YOLO系列分割模型在不同阶段的轻量化演进，形成了具有针对性与层次性的模型对比体系。KITTI数据集包含城市街区、乡村道路及高速公路等多种典型户外场景，涵盖动态目标频繁、光照变化与遮挡复杂等视觉干扰因素，为分割模型在真实动态环境下的计算性能评估提供了丰富测试样本。本节综合采用mAP(IoU阈值设为0.5)、计算复杂度(GFLOPs)与推理帧率(帧/s)作为评估指标，系统性分析各模型在动态环境下的检测精度与计算效率表现，以全面衡量其实际部署性能与实时适应性。

从表2结果可以看出，DHSR-YOLOSeg在多个关键性能指标上均展现出显著优势。在模型参数方面，DHSR-YOLOSeg仅为1.45 M，较YOLOv11n-seg减少约48.9%，相较于Mask R-CNN更是下降超过96%，大幅减轻了模型在边缘计算设备上的部署负担。在计算复杂度方面，其GFLOPs仅为4.49，相比于Mask R-CNN的67.16下降超过93%，计算开销显著降低，有利于实时推理与低功耗运行。在帧率方面，DHSR-YOLOSeg实现了60.19 帧/s，为所有模型中最高，明显优于YOLOv8n-seg(56.51 帧/s)与YOLOv11n-seg(58.93 帧/s)，满足动态视觉SLAM系统对高实时性的要求。尽管模型极度轻量，DHSR-YOLOSeg仍取得了0.51的mAP50，在所有模型中排名第1，精度与效率兼具，表现出极强的语义感知能力。综上，DHSR-YOLOSeg在语义分割精度、模型轻量性与推理速度之间实现了最优平衡，验证了其在动态街景语义SLAM中的部署潜力与综合性能优势，为复杂场景下实现高效语义引导SLAM提供了有力的感知支撑。进一步地，如图10所示，DHSR-YOLOSeg能够在KITTI序列中稳定识别并分割典型动态目标(如“person”,“car”,“bicycle”)，生成的掩码边界清晰、轮廓完整，为特征点剔除与动态区域建图提供了高质量的语义支持。

表 2　对比实验结果

模型	参数量(M)	GFLOPs	mAP50	推理帧率(帧/s)
Mask R-CNN	35.92	67.16	0.47	16.05
YOLOv5n-seg	2.05	4.31	0.48	57.89
YOLOv8n-seg	3.26	6.78	0.50	56.51
YOLOv11n-seg	2.84	5.18	0.50	58.93
DHSR-YOLOSeg	1.45	4.49	0.51	60.19

3.3 语义信息紧耦合下的SLAM轨迹精度提升验证

为评估DHSR-YOLOSeg模型在SLAM系统中结合语义信息实现动态区域特征点剔除的实际效果，本文在KITTI Odometry数据集的00～10号序列上开展轨迹精度对比实验。通过分别运行ORB-SLAM3, DynaSLAM[12], DS-SLAM[25]以及引入DHSR-YOLOSeg模型的语义增强版本，本文在城市道路、乡村街区与高速场景下对比各系统的位姿估计性能，验证语义信息与前端跟踪的紧耦合设计对系统定位稳定性与轨迹一致性的提升效果。为全面衡量轨迹误差，本文采用主流评估指标：均方根误差(Root Mean Square Error, RMSE)。其中，RMSE对较大误差更敏感，可有效反映局部跟踪失效对整体精度的影响，其计算公式为

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2} \tag{7}$$

其中，n是观测值的总数，y_i是第i个观测值的真实值，$\widehat{y}_i$是第i个观测值的预测值。基于上述指标，本文对各序列的轨迹误差进行了定量评估，实验结果如表3所示。

本文方法、DynaSLAM与DS-SLAM的性能提升幅度，均通过与ORB-SLAM3的RMSE结果对比计算得出，而非直接比较各方法之间的优劣。从轨迹精度来看，本文方法在10个测试序列中有8个实现了RMSE降低，平均降幅为8.78%，在覆盖范围与适应性上优于DynaSLAM(7个序列)和DS-SLAM

图 10 DHSR-YOLOSeg模型在不同KITTI序列上的动态目标语义分割结果

表 3 不同语义SLAM方法在KITTI序列上的轨迹RMSE误差对比(m)

序列	ORB-SLAM3	DynaSLAM	DS-SLAM	本文
00	1.490	1.009	1.249	1.296
01	13.319	17.471	15.395	12.873
02	4.241	4.492	4.166	3.531
04	0.273	0.254	0.261	0.244
05	0.994	0.833	0.924	0.947
06	1.121	0.716	1.219	1.162
07	0.630	0.658	0.644	0.534
08	3.517	2.475	3.973	3.466
09	1.697	1.261	1.779	1.715
10	1.046	0.752	1.026	0.993

(5个序列)。其中，序列00, 02与07的降幅分别为13.02%, 16.74%和15.24%，展现出稳定的误差抑制效果。DynaSLAM虽在部分序列(如06与00)达到36.13%和32.28%的显著降幅，但整体仅在7个序列优于ORB-SLAM3，表现出一定的不均衡性。DS-SLAM则整体改进幅度有限，仅在5个序列上实现下降(平均6.26%)，且在多个序列上误差反而上升，稳定性较差。从计算效率来看，本文方法的平均处理时间仅约48.9 ms，远低于DynaSLAM的84.0 ms与DS-SLAM的60.0 ms，在引入语义分割进行动态目标抑制的同时，仍保持了较高的实时性。综上，本文方法在保证运行效率的前提下，实现了对ORB-SLAM3的整体精度提升，在适应性上优于DynaSLAM，在稳定性与实时性上明显优于DS-SLAM，体现出较强的鲁棒性与实用性。

3.4 语义SLAM系统的轻量化性能对比

为验证DHSR-YOLOSeg在语义SLAM系统中的轻量化特性与帧级实时处理能力，本文在KITTIOdometry数据集00～10号序列上，分别运行ORB-SLAM3, DynaSLAM, DS-SLAM及本文提出的语义增强系统，统计各方法在跟踪线程中的平均每帧处理时间(Mean Tracking Time)。上述序列涵盖城市道路、交叉路口与高速场景等典型城市动态环境，便于全面对比系统在实际干扰下的运行效率与响应能力。实验结果用于评估不同语义SLAM方法在动态场景中的帧级效率与实时性表现，进一步验证本文方法在语义增强的同时，具备更优的前端执行性能与平台部署适应性。

由表4可见，本文所提出的语义SLAM系统在各序列中的每帧平均处理时间稳定保持在41～55 ms之间，整体运行效率显著优于DynaSLAM与DS-SLAM等典型语义增强方法。统计结果显示，本文方法的平均处理时间为48.86 ms，较DynaSLAM降低约41.83%，较DS-SLAM降低约18.55%。尽管通

表 4 不同语义SLAM方法平均每帧跟踪处理时间对比(ms)

序列	ORB-SLAM3	DynaSLAM	DS-SLAM	本文
00	36.25	84.91	64.75	44.27
01	33.47	78.78	57.08	41.90
02	37.07	86.71	61.86	50.63
04	36.01	84.38	57.38	52.27
05	36.71	85.92	59.37	50.11
06	36.74	85.99	61.09	48.45
07	35.23	82.66	59.89	48.95
08	36.90	86.34	63.92	54.76
09	35.37	82.97	58.18	50.62
10	34.61	81.29	56.37	46.69

过语义分割的方式将语义信息引入SLAM系统前端，用于动态目标区域的识别与特征点剔除，系统仍保持良好的帧级处理效率与稳定性。该结果表明，所提出的DHSR-YOLOSeg模型在实现动态感知的同时具备较强的计算适应性，为城市动态环境下语义SLAM系统的实时部署提供了可行路径与性能保障。

从轨迹误差对比实验与平均每帧跟踪处理时间结果可以看出，本文所提出的语义增强SLAM系统在动态城市环境中兼顾了精度提升与计算效率，展现出良好的整体鲁棒性与系统实用性。系统所集成的核心模块DHSR-YOLOSeg提供了高效的像素级语义分割能力，为后续动态区域的检测与特征点剔除提供了有效支持，从而在保持较高帧率的同时实现了稳定的轨迹误差抑制。本方法在保障运行效率的基础上提升了系统在复杂场景下的稳定性。

4　结束语

本文针对城市动态环境下巡检机器人对轻量化感知与鲁棒定位的双重需求，提出了一种轻量级语义感知与视觉SLAM深度融合的动态感知方法。在模型设计方面，构建了融合DySample动态采样与ChannelAttention_HSFPN分层通道注意力机制的DyCANet模块，并在YOLOv11n-seg基础上引入C3k2_DynamicConv与RSCS头，形成DHSR-YOLOSeg网络，实现了分割精度与计算效率的协同优化。在系统集成方面，将DHSR-YOLOSeg输出的语义信息紧耦合至ORB-SLAM3的前端跟踪线程，用于动态区域识别与特征点剔除，有效提升了系统在城市动态场景下的定位精度与运行实时性。

在KITTI Odometry数据集上的实验结果表明，DHSR-YOLOSeg在精度、参数量和推理速度等维度实现了优于YOLOv5-seg与YOLOv8n-seg的综合表现。融合语义信息后的SLAM系统在全部8个典型动态序列中均获得了轨迹误差的显著降低，展现出更强的稳定性与运行效率。总体而言，所提方法在精度、效率与部署适应性之间实现了良好权衡，具备较高的实际应用潜力。

尽管所提方法在分割精度与系统效率方面展现出较好性能，仍存在以下局限性：一方面，轻量化设计中分割精度与计算复杂度之间仍需进一步平衡；另一方面，当前系统为纯视觉输入，尚未融合多模态数据，泛化能力在复杂场景下仍有待提升。后续工作将进一步在多类真实场景中开展泛化验证，并探索多源信息融合策略，以增强系统的环境适应性与鲁棒性。

参 考 文 献

[1] HALDER S and AFSARI K. Robots in inspection and monitoring of buildings and infrastructure: A systematic review[J]. *Applied Sciences*, 2023, 13(4): 2304. doi: 10.3390/app13042304.

[2] LI Yuhao, FU Chengguo, YANG Hui, *et al*. Design of a closed piggery environmental monitoring and control system based on a track inspection robot[J]. *Agriculture*, 2023, 13(8): 1501. doi: 10.3390/agriculture13081501.

[3] 罗朝阳, 张荣芬, 刘宇红, 等. 自动驾驶场景下的行人意图语义VSLAM[J]. 计算机工程与应用, 2024, 60(17): 107–116. doi: 10.3778/j.issn.1002-8331.2306-0159.
LUO Zhaoyang, ZHANG Rongfen, LIU Yuhong, *et al*. Pedestrian intent semantic VSLAM in automatic driving scenarios[J]. *Computer Engineering and Applications*, 2024, 60(17): 107–116. doi: 10.3778/j.issn.1002-8331.2306-0159.

[4] 李国逢, 谈嵘, 曹媛媛. 手持SLAM: 城市测量的新方法与实践[J]. 测绘通报, 2024(S2): 255–259. doi: 10.13474/j.cnki.11-2246.2024.S253.
LI Guofeng, TAN Rong, and CAO Yuanyuan. Handheld SLAM: Emerging techniques and practical implementations in urban surveying[J]. *Bulletin of Surveying and Mapping*, 2024(S2): 255–259. doi: 10.13474/j.cnki.11-2246.2024.S253.

[5] ZHANG Tianzhe and DAI Jun. Electric power intelligent inspection robot: A review[J]. *Journal of Physics: Conference Series*, 2021, 1750(1): 012023. doi: 10.1088/1742-6596/1750/1/012023.

[6] MUR-ARTAL R, MONTIEL J M M, and TARDÓS J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. *IEEE Transactions on Robotics*, 2015, 31(5): 1147–1163. doi: 10.1109/TRO.2015.2463671.

[7] MUR-ARTAL R and TARDÓS J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras[J]. *IEEE Transactions on Robotics*, 2017, 33(5): 1255–1262. doi: 10.1109/TRO.2017.2705103.

[8] CAMPOS C, ELVIRA R, RODRÍGUEZ J J G, *et al*. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM[J]. *IEEE Transactions on Robotics*, 2021, 37(6): 1874–1890. doi: 10.1109/TRO.2021.3075644.

[9] QIN Tong, LI Peiliang, and SHEN Shaojie. VINS-Mono: A robust and versatile monocular visual-inertial state estimator[J]. *IEEE Transactions on Robotics*, 2018, 34(4): 1004–1020. doi: 10.1109/TRO.2018.2853729.

[10] ZANG Qiuyu, ZHANG Kehua, WANG Ling, *et al*. An adaptive ORB-SLAM3 system for outdoor dynamic environments[J]. *Sensors*, 2023, 23(3): 1359. doi: 10.3390/s23031359.

[11] WU Hangbin, ZHAN Shihao, SHAO Xiaohang, *et al*. SLG-

SLAM: An integrated SLAM framework to improve accuracy using semantic information, laser and GNSS data[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2024, 133: 104110. doi: 10.1016/j.jag.2024.104110.

[12] BESCOS B, FÁCIL J M, CIVERA J, *et al*. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. *IEEE Robotics and Automation Letters*, 2018, 3(4): 4076–4083. doi: 10.1109/LRA.2018.2860039.

[13] VINCENT J, LABBÉ M, LAUZON J S, *et al*. Dynamic object tracking and masking for visual SLAM[C]. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, USA, 2020: 4974–4979. doi: 10.1109/IROS45743.2020.9340958.

[14] KHANAM R and HUSSAIN M. YOLOv11: An overview of the key architectural enhancements[EB/OL]. https://arxiv.org/abs/2410.17725, 2024.

[15] XU Ziheng, NIU Jianwei, LI Qingfeng, *et al*. NID-SLAM: Neural implicit representation-based RGB-D SLAM in dynamic environments[C]. 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, Canada, 2024: 1–6. doi: 10.1109/ICME57554.2024.10687512.

[16] GONG Can, SUN Ying, ZOU Chunlong, *et al*. Real-time visual SLAM based YOLO-fastest for dynamic scenes[J]. *Measurement Science and Technology*, 2024, 35(5): 056305. doi: 10.1088/1361-6501/ad2669.

[17] WU Peiyi, TONG Pengfei, ZHOU Xin, *et al*. Dyn-DarkSLAM: YOLO-based visual SLAM in low-light conditions[C]. 2024 IEEE 25th China Conference on System Simulation Technology and its Application (CCSSTA), Tianjin, China, 2024: 346–351. doi: 10.1109/CCSSTA62096.2024.10691775.

[18] ZHANG Ruidong and ZHANG Xinguang. Geometric constraint-based and improved YOLOv5 semantic SLAM for dynamic scenes[J]. *ISPRS International Journal of Geo-Information*, 2023, 12(6): 211. doi: 10.3390/ijgi12060211.

[19] YANG Tingting, JIA Shuwen, YU Ying, *et al*. Enhancing visual SLAM in dynamic environments with improved YOLOv8[C]. The Sixteenth International Conference on Digital Image Processing (ICDIP), Haikou, China, 2024: 132741Y. doi: 10.1117/12.3037734.

[20] HAN Kai, WANG Yunhe, GUO Jianyuan, *et al*. ParameterNet: Parameters are all you need for large-scale visual pretraining of mobile networks[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024: 15751–15761. doi: 10.1109/CVPR52733.2024.01491.

[21] YU Jiazuo, ZHUGE Yunzhi, ZHANG Lu, *et al*. Boosting continual learning of vision-language models via mixture-of-experts adapters[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2024: 23219–23230. doi: 10.1109/CVPR52733.2024.02191.

[22] LIU Wenze, LU Hao, FU Hongtao, *et al*. Learning to upsample by learning to sample[C]. The IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023: 6004–6014. doi: 10.1109/ICCV51070.2023.00554.

[23] HUANGFU Yi, HUANG Zhonghao, YANG Xiaogang, *et al*. HHS-RT-DETR: A method for the detection of citrus greening disease[J]. *Agronomy*, 2024, 14(12): 2900. doi: 10.3390/agronomy14122900.

[24] CAO Qi, CHEN Hang, WANG Shang, *et al*. LH-YOLO: A lightweight and high-precision SAR ship detection model based on the improved YOLOv8n[J]. *Remote Sensing*, 2024, 16(22): 4340. doi: 10.3390/rs16224340.

[25] YU Chao, LIU Zuxin, LIU Xinjun, *et al*. DS-SLAM: A semantic visual SLAM towards dynamic environments[C]. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018: 1168–1174. doi: 10.1109/IROS.2018.8593691.

余浩扬：男，硕士，研究方向为基于多传感器的室外巡检机器人视觉同步定位与建国.

李艳生：男，副教授，研究方向为机器人技术与智能制造技术.

肖凌励：女，硕士，研究方向为智能移动机器人.

周继源：男，硕士，研究方向为具身智能机器人.

责任编辑：余 蓉

A Lightweight Semantic Visual Simultaneous Localization and Mapping Framework for Inspection Robots in Dynamic Environments

YU Haoyang② LI Yansheng① XIAO Lingli② ZHOU Jiyuan②

①(*School of Artificial Intelligence* (*iFLYTEK Artificial Intelligence School*, *CQUPT*), *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

②(*School of Integrated Circuits* (*Chongqing International Semiconductor Institute*), *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)

Abstract:

Objective In complex dynamic environments such as industrial parks and urban roads, inspection robots depend heavily on visual Simultaneous Localization And Mapping (SLAM) systems. However, the presence of moving objects often causes feature drift and map degradation, reducing SLAM performance. Furthermore, conventional semantic segmentation models typically require extensive computational resources, rendering them unsuitable for embedded platforms with limited processing capabilities, thereby constraining SLAM deployment in autonomous inspection tasks. To address these challenges, this study proposes a lightweight semantic visual SLAM framework designed for inspection robots operating in dynamic environments. The framework incorporates a semantic segmentation-based dynamic feature rejection method to achieve real-time identification of dynamic regions at low computational cost, thereby improving SLAM robustness and mapping accuracy.

Methods Building upon the 11th generation lightweight YOLO segmentation model (YOLOv11n-seg), a systematic lightweight redesign is implemented. to enhance performance under constrained computational resources. First, the original neck is replaced with DyCANet, a lightweight multi-scale feature fusion module that integrates dynamic point sampling and channel attention to improve semantic representation and boundary segmentation. DyCANet combines DySample, a dynamic upsampling operator that performs content-aware spatial sampling with minimal overhead, and ChannelAttention_HSFPN, a hierarchical attention structure that strengthens multi-scale integration and highlights critical semantic cues, particularly for small or occluded objects in complex scenes. Second, a Dynamic Convolution module (DynamicConv) is embedded into all C3k2 modules to enhance the adaptability and efficiency of feature extraction. Inspired by the Mixture-of-Experts framework, DynamicConv applies a conditional computation mechanism that dynamically adjusts kernel weights based on the input feature characteristics. This design allows the network to extract features more effectively across varying object scales and motion patterns, improving robustness against dynamic disturbances with low computational cost. Third, the original segmentation head is replaced by the Reused and Shared Convolutional Segmentation Head (RSCS Head), which enables decoder structure sharing across multi-scale branches. RSCS reduces redundant computation by reusing convolutional layers and optimizing feature decoding paths, further improving overall model efficiency while maintaining segmentation accuracy. These architectural modifications result in DHSR-YOLOSeg, a lightweight semantic segmentation model that significantly reduces parameter count and computational cost while preserving performance. DHSR-YOLOSeg is integrated into the tracking thread of ORB-SLAM3. to provide real-time semantic information. This enables dynamic object detection and the removal of unstable feature points during localization, thereby enhancing the robustness and trajectory consistency of SLAM in complex dynamic environments.

Results and Discussions Ablation experiments on the COCO dataset demonstrate that, compared with the baseline YOLOv11n-seg, the proposed DHSR-YOLOSeg achieves a 13.8% reduction in parameter count, a 23.1% decrease in Giga Floating Point Operations (GFLOPs), and its Average Precision at IoU 0.5 (mAP50) is the sameas the baseline YOLOv11n-seg's (Table 1). On the KITTI dataset, DHSR-YOLOSeg reaches an inference speed of 60.19 frame/s, which is 2.14% faster than YOLOv11n-seg and 275% faster than the widely used Mask R-CNN (Table 2). For trajectory accuracy evaluation on KITTI sequences 00～10, DHSR-YOLOSeg outperforms ORB-SLAM3 in 8 out of 10 sequences, achieving a maximum Root Mean Square Error (RMSE) reduction of 16.76% and an average reduction of 8.78% (Table 3). Compared with DynaSLAM and DS-SLAM, the proposed framework exhibits more consistent error suppression across sequences, improving both trajectory accuracy and stability. In terms of runtime efficiency, DHSR-YOLOSeg achieves an average per-frame processing time of 48.86 ms on the KITTI dataset, 18.55% and 41.38% lower than DS-SLAM and DynaSLAM, respectively (Table 4). The per-sequence processing time ranges from 41 to 55 ms, which is comparable to the 35.64 ms of ORB-SLAM3, indicating that the integration of semantic segmentation introduces only a modest computational overhead.

Conclusions This study addresses the challenge of achieving robust localization for inspection robots operating

in dynamic environments, particularly in urban road settings characterized by frequent interference from pedestrians, vehicles, and other moving objects. To this end, a semantic-enhanced visual SLAM framework is proposed, in which a lightweight semantic segmentation model, DHSR-YOLOSeg, is integrated into the stereo-based ORB-SLAM3 pipeline. This integration enables real-time identification of dynamic objects and removal of their associated feature points, thereby improving localization robustness and trajectory consistency. The DHSR-YOLOSeg model incorporates three key architectural components—DyCANet for feature fusion, DynamicConv for adaptive convolution, and the RSCS Head for efficient multi-scale decoding. Together, these components reduce the model's size and computational cost while preserving segmentation performance, providing an efficient and deployable perception solution for resource-constrained platforms. Experimental findings show that: (1) ablation tests on the COCO dataset confirm substantial reductions in complexity with preserved accuracy, supporting embedded deployment; (2) frame rate comparisons on the KITTI dataset demonstrate superior performance over both lightweight and standard semantic segmentation methods, meeting real-time SLAM requirements; (3) trajectory evaluations indicate that the dynamic feature rejection strategy effectively mitigates localization errors in dynamic scenes; and (4) the overall system maintains high runtime efficiency, ensuring a practical balance between semantic segmentation and real-time localization performance. However, current experiments are conducted under ideal conditions using standardized datasets, without fully reflecting real-world challenges such as multi-sensor interference or unstructured environments. Moreover, the trade-offs between model complexity and accuracy for each lightweight module have not been systematically assessed. Future work will focus on multimodal sensor fusion and adaptive dynamic perception strategies to enhance the robustness and applicability of the proposed system in real-world autonomous inspection scenarios.

Key words: Inspection robot; Semantic segmentation; Visual Simultaneous Localization And Mapping (SLAM); Dynamic feature filtering