面向抓斗轨迹稳定提取的空间与时序协同优化方法

陈晓玉[①] 张凤卓[①] 陈 杨[②] 刘文远[③] 孔德明*[④⑤]

[①](燕山大学信息科学与工程学院 秦皇岛 066004)

[②](大连华锐智能化科技有限公司 大连 116013)

[③](燕山大学人工智能学院 秦皇岛 066004)

[④](燕山大学电气工程学院 秦皇岛 066004)

[⑤](河北省测试计量技术及仪器重点实验室燕山大学 河北秦皇岛 066000)

摘 要：在港口散货装卸作业中，抓斗作为门座式起重机的关键执行部件，其连续轨迹提取易受尺度变化、局部遮挡、边界退化和帧间形态波动等因素影响。为此，该文提出一种面向抓斗轨迹稳定提取的空间与时序协同优化方法。该方法从空间表征增强与时序一致性约束两个层面展开：前者通过多尺度特征交互和边界细节增强，提高复杂背景与遮挡条件下抓斗区域的分割完整性；后者在分割表征层引入时序一致性约束，减弱局部误分、边界波动和短时漏检对轨迹连续性的影响，并通过联合优化目标实现端到端训练。实验结果表明，该方法在DAVIS2016、SegTrackV2和实际门座式起重机抓斗数据集上均表现出较优综合性能，其中在DAVIS2016上区域相似度J和轮廓精度F分别达到74.79%和74.70%，在实际抓斗数据集上平均绝对误差MAE和均方根误差RMSE较YOLOv8-seg分别下降约55.3%和52.6%，漏检率Miss rate由2.95%降至0.56%。遮挡实验进一步表明，该方法能够减弱轨迹抖动与短时中断，在复杂港口作业场景下具有较强的连续目标提取能力。同时，改进后模型推理速度保持在33fps以上，具备实时应用潜力。

关键词：视频目标分割；目标轨迹提取；抓斗轨迹稳定提取；空间表征增强；时序一致性约束

中图分类号：TN911.73; TP391.41 **文献标识码**：A **文章编号**：1009-5896(2026)00-0001-11

DOI: 10.11999/JEIT260512 **CSTR**: 32379.14.JEIT260512

1 引言

随着港口装卸作业自动化与无人化水平不断提升，作业目标感知与运动分析已成为智能监测和安全控制的重要基础。在散货装卸作业中，抓斗是门座式起重机抓取和搬运物料的关键部件，其运动轨迹反映了作业状态，并为设备协同、作业调度和碰撞预警提供依据。受复杂环境、背景干扰和运动状态变化影响，抓斗在连续作业中易出现尺度变化、局部遮挡、边界模糊和短时缺失等问题，增加了轨迹提取误差与连续中断的风险。因此，面向复杂港口视频场景，开展抓斗轨迹稳定提取研究对于提升无人化装卸作业中的连续感知与安全决策能力具有重要工程意义。

近年来，深度学习推动了视频目标分割与连续目标提取技术的发展。以掩码区域卷积神经网络(Mask Region-based Convolutional Neural Network, Mask R-CNN)[1]为代表的实例分割方法为目标区域提取奠定了基础，单样本视频目标分割(One-Shot Video Object Segmentation, OSVOS)[2]、时空记忆网络(Space-Time Memory, STM)[3]和XMem[4]等方法进一步增强了跨帧目标表达能力。相关综述研究表明，视频目标分割已由早期逐帧目标提取逐步发展为兼顾区域表征、跨帧关联和连续感知的时空协同建模框架[5,6]。在此基础上，围绕多尺度特征融合、动态感知更新和局部匹配等方向，相关研究持续改善了复杂场景下的目标分割与连续表达效果[7,8]；针对轻量化建模、动态记忆增强和跨场景泛化等问题，也有工作进一步拓展了方法的效率与适用范围[9–11]。此外，相关研究还通过时序重校准、时空相似性建模和跨场景泛化提升了连续目标提取的鲁棒性[12–14]。

在工业现场视频监控任务中，目标提取方法不仅需要具备较高的分割精度，还需满足无人工初始化和实时推理要求。以YOLO系列为代表的实时实例分割方法能够在单帧图像中同时完成目标检测与像素级掩膜预测，具有推理速度快、部署便利和场景适应性强等特点，已成为复杂工业场景中目标自动提取的重要基线。然而，现有实时实例分割方法多以单帧区域预测为主要目标，对连续帧间局部误分、边界波动和短时漏检在轨迹层面的累积影响考

收稿日期：2026-04-24；改回日期：2026-07-01；网络出版：2026-07-23

*通信作者： 孔德明 demingkong@ysu.edu.cn

基金项目：河北省自然科学基金资助项目(No.F2025203055)

Foundation Item: The Natural Science Foundation of Hebei Province (No.F2025203055)

虑不足。对于港口抓斗作业视频，抓斗在升降、摆动和开合过程中常伴随尺度变化、局部遮挡与边界模糊，背景中还存在门机结构、钢丝绳和料斗等强干扰，即使单帧分割总体正确，局部边界扰动和短时漏分仍可能导致质心偏移，并在时间维上累积为轨迹抖动、漏检增加乃至连续中断。因此，有必要在实时实例分割框架基础上进一步引入空间表征增强与时序一致性约束，以同时提升抓斗区域提取质量和轨迹连续性。针对上述问题，本文提出一种面向抓斗轨迹稳定提取的空间与时序协同优化方法，从空间表征增强与时序一致性约束两个层面提升轨迹提取的准确性与连续性。主要工作概括如下：

(1)针对抓斗运动中的尺度变化、遮挡干扰与边界退化问题，设计空间表征增强网络，提高目标区域提取质量，为轨迹提取提供可靠的空间基础。

(2)针对抓斗轨迹在连续帧中的抖动、漏检与中断现象，构建时序一致性约束机制，在时间维度上减弱局部误分和短时缺失引起的轨迹波动与连续中断。

(3)在DAVIS2016、SegTrackV2和实际门座式起重机抓斗数据集上开展实验验证，围绕分割性能、轨迹稳定性及遮挡条件下的连续目标提取能力进行分析，评估所提方法在港口工业视频场景中的有效性。

2 相关工作

2.1 面向目标区域稳定提取的空间表征增强相关工作

为提升复杂场景下的视频目标分割质量，现有研究主要从注意力增强、多尺度融合和跨层聚合等方面改善目标区域表征。侯志强等人[15]提出基于多尺度特征增强与全局-局部特征聚合的视频目标分割算法，通过融合不同尺度特征并结合全局与局部信息，提高目标区域表达能力和边缘细化效果。陈雷等人[16]提出一种基于Transformer特征金字塔的自蒸馏目标分割方法，在不增加网络参数规模的情况下提升目标分割性能，为多尺度特征表达与轻量化分割优化提供了参考。Woo等人[17]提出卷积块注意力模块，通过通道与空间重标定突出目标响应、抑制冗余背景干扰。Lin等人[18]提出特征金字塔网络，增强不同尺度特征的语义表达能力；Liu等人[19]进一步提出路径聚合网络，通过双向信息传递改善浅层细节与深层语义融合效果。上述工作为复杂场景中的目标区域稳定提取提供了重要基础，但大多以分割精度为目标，对局部边界波动和短时误分引起的质心偏移关注不足。受此启发，本文构建空间表征增强策略，以提高抓斗区域完整性和边界稳定性，为后续轨迹提取提供可靠的空间基础。

2.2 面向连续轨迹提取的时序一致性建模相关工作

跨帧连续表达能力直接影响视频目标提取的可靠性。现有研究通过时序建模提升连续分割效果：丁建睿等人[20]融合邻域注意力和状态空间模型增强时空特征建模；Li等人[21]利用历史帧信息辅助当前帧预测；Wang等人[22]通过多尺度时空记忆网络增强跨帧关联；Miao等人[23]采用混合记忆约束时序一致性，减弱跨帧预测波动。上述方法主要优化掩膜传播质量或分割连续性，对局部掩膜波动、边界破碎和短时漏检在质心轨迹上的累积影响关注不足，因而难以直接约束轨迹层面的抖动和中断问题。需要指出的是，以OSVOS、STM、XMem为代表的半监督视频目标分割方法通常依赖首帧人工掩膜初始化，其核心目标是基于已知目标进行跨帧掩膜传播与更新，难以直接满足无人工交互的港口监控场景需求[3,4]。本文面向抓斗轨迹提取任务，要求模型在无首帧人工掩膜、无人工交互条件下自动提取抓斗区域，并保持轨迹输出的连续性。因此，本文未采用首帧掩膜驱动的显式记忆传播机制，而是在自动实例分割框架的分割表征层引入时序一致性约束，抑制局部误差在时间维上的累积传播，从而提升抓斗轨迹提取的连续性与可靠性。

3 方法

3.1 方法总体框架

针对港口作业视频中抓斗目标受复杂背景、尺度变化、局部遮挡和帧间波动影响而难以稳定提取的问题，本文提出一种面向抓斗轨迹稳定提取的空间与时序协同优化方法，整体框架如图1所示。以YOLOv8-seg为基础分割网络，方法从空间表征增强、时序一致性约束和联合优化训练三个层面展开。借鉴局部匹配与特征融合有助于提升视频目标细节保持与连续表达能力的思想[24]，模型先提取多层特征并生成初始掩膜，再通过主干、颈部和原型分支增强目标区域、边界和局部结构表达；随后在分割表征层引入时序一致性约束，抑制局部误差在时间维上的累积传播，并通过统一损失实现端到端训练，从而协同提升单帧分割质量与轨迹提取能力。

3.2 面向稳定提取的抓斗空间表征增强

复杂背景下，抓斗与背景的可分性不足。其瓶颈在于，标准卷积对通道和空间维度的特征响应缺乏区分机制：抓斗边缘易与煤堆、船舱结构等背景纹理混淆，导致目标响应被冗余信息削弱。为此，本文在主干中间层特征上引入CBAM[17]。设输入特

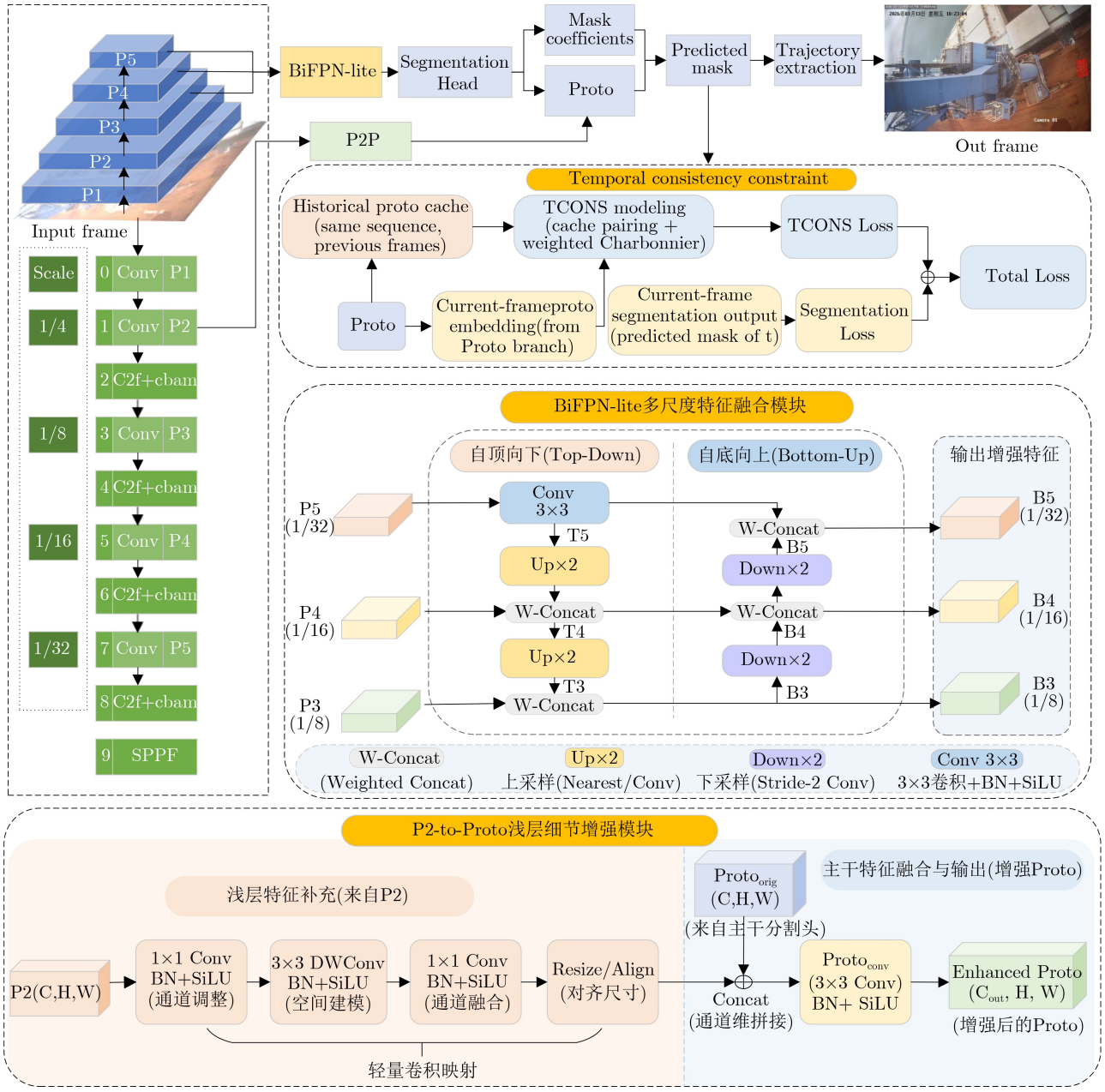

图 1 方法整体框架图

征为F，则经过通道注意力和空间注意力重标定后的输出分别为

$$F_{\mathrm{C}} = M_{\mathrm{C}}(F) \odot F \tag{1}$$

$$F' = M_{\mathrm{S}}(F) \odot F_{\mathrm{C}} \tag{2}$$

其中，$M_{\mathrm{C}}(\cdot)$和$M_{\mathrm{S}}(\cdot)$分别表示通道注意力映射和空间注意力映射，$\odot$表示逐元素乘法。该过程通过增强抓斗相关通道和关键空间位置的响应，减弱复杂背景中冗余干扰的影响，从而缓解了逆光、阴影等工况下因纹理相似导致的可分性不足问题，为后续特征融合提供更具判别性的区域表示。

仅增强区域响应仍不足以应对抓斗在升降和摆动过程中的剧烈尺度变化。传统FPN结构中，不同尺度特征间的信息交互以单向传递为主且连接方式固定；当抓斗尺度变化时，高层特征易丢失细节信息，浅层特征语义表达不足，进而导致层间响应失衡。本文将原有Concat融合节点替换为加权拼接节点W-Concat，构建BiFPN-lite多尺度特征融合结构。设完成尺度对齐后的输入特征为$\{\tilde{P}_k\}$，则第i个融合节点输出为

$$F_i = \mathrm{Concat}(\tilde{w}_1\tilde{P}_1, \tilde{w}_2\tilde{P}_2, \cdots, \tilde{w}_n\tilde{P}_n) \tag{3}$$

其中，归一化权重$\tilde{w}_k$定义为

$$\tilde{w}_k = \frac{\mathrm{ReLU}(w_k)}{\sum_{j=1}^{n}\mathrm{ReLU}(w_j) + \mathcal{E}} \tag{4}$$

式中，w_k为可学习融合权重，$\mathcal{E}$为防止分母为零的微小常数。各分支特征在完成尺度对齐后先依据归一化权重进行加权，再按通道维进行拼接。该设计通过双向跨尺度连接和可学习加权，增强了不同尺度信息之间的交互与互补，有效缓解了臂架变幅时层间响应失衡导致的尺度敏感问题，提高了抓斗在尺度变化条件下的区域表达和边界保持能力。

尽管多尺度融合改善了区域层面的表达稳定性，但原型分支生成的掩码在边界处仍趋于模糊，导致局部结构被弱化。这是由于深层特征在反复下采样中不可避免会丢失细粒度空间信息：抓斗作业时，钢索悬挂和抓瓣开合会产生微小但频繁的局部抖动，边界模糊会直接放大这种抖动，导致提取的质心轨迹出现非真实的锯齿状偏移。为补充浅层细节，本文将P_2特征引入原型生成分支。设原始原型特征为$\mathrm{Proto}_{\mathrm{orig}}$，由$P_2$生成的补充特征为$\mathrm{Proto}_{\mathrm{p2}}$。$P_2$先经过$1\times1$卷积完成通道映射，再经过深度可分离$3\times3$卷积进行细化，并在尺寸不一致时通过双线性插值完成空间对齐。随后，将二者融合得到增强后的原型特征：

$$\mathrm{Proto}_{\mathrm{new}}=\Phi(\mathrm{Concat}(\mathrm{Proto}_{\mathrm{orig}},\mathrm{Proto}_{\mathrm{p2}})) \tag{5}$$

其中，$\mathrm{Concat}(\cdot)$表示按通道维拼接，$\Phi(\cdot)$表示卷积映射操作。该设计在保持原型分支语义表达能力的同时补充了边界与局部结构信息，从而减弱由边界退化引起的非真实轨迹波动。

经空间表征增强(Spatial Representation Enhancement, SRE)后，模型获得的目标特征对相邻帧位置变化更加稳定，对局部扰动更不敏感，从而为后续时序一致性约束提供了更加可靠的输入基础。

3.3 面向稳定提取的时序一致性约束

空间表征增强后，轨迹的稳定提取仍取决于相邻帧间的表征一致性。抓斗这类连续运动目标易因局部误分、边界退化和短时遮挡导致相邻帧分割结果的非真实波动。现有方法在轨迹点层面进行平滑，虽可减弱位置抖动，但难以抑制分割表征误差在时间维上的累积传播。其本质瓶颈在于，分割网络逐帧独立推理，相邻帧的原型特征缺乏显式一致性引导：某帧因遮挡或边界退化产生的表征偏移会直接传递至后续帧的轨迹提取，而轨迹点层面的后滤波仅能在输出端被动平滑，无法修正表征空间的分布漂移。为此，本文在分割头输出的原型特征上构建时序一致性约束(Temporal Consistency Constraint, TCONS)，以约束抓斗目标在短时间范围内的表征变化。TCONS仅作用于训练阶段，推理时不参与计算。

设第t帧分割头输出的原型特征为Proto_t。为将空间特征转化为可比较的帧级表示，本文对空间维度进行全局平均，得到第t帧的嵌入向量：

$$\boldsymbol{z}_t=\left[{z_t}^{(1)},{z_t}^{(2)},\cdots,{z_t}^{(C)}\right]^{\mathrm{T}}\in\mathbb{R}^C,$$
$${z_t}^{(d)}=\frac{1}{HW}\sum_{u=1}^{H}\sum_{\nu=1}^{W}{\mathrm{Proto}_t}^{(d,u,v)},d=1,2,\cdots,C \tag{6}$$

由此可形成同一视频序列的帧级表示集合$\{\boldsymbol{z}_t\}$。训练过程中，本文按视频序列维护一个特征缓存，并仅保留该序列最近一次参与约束计算的历史表示。设当前帧索引为t，缓存中对应的最近历史帧索引为t'，其特征表示为$\boldsymbol{z}_{t'}$。当两帧满足$0<t-t'\le G$时，认为二者构成有效跨帧配对，其中G表示允许的最大帧间隔。按本文视频帧率(20 fps)，$G=50$对应约2.5 s，可覆盖抓斗一个典型短时动作片段，能维持语义一致性；进一步增大G时，预实验中时序约束损失趋于不收敛，故取50。

每个当前帧仅与同序列中满足间隔约束的最近历史帧建立一次配对关系。对于每个有效配对，采用Charbonnier损失度量当前帧与缓存帧之间的特征差异：

$$\ell_t=\frac{1}{C}\sum_{d=1}^{C}\sqrt{\left(z_t^{(d)}-z_{t'}^{(d)}\right)^2+\mathcal{E}^2} \tag{7}$$

其中，$\mathcal{E}$为数值稳定项，用于避免特征差异接近零时根号运算导致的数值不稳定取$\mathcal{E}=10^{-6}$。

考虑到相邻更近的帧应具有更高约束强度，进一步引入间隔衰减权重$\omega_t=\gamma^{(t-t'-1)}$，其中，$\gamma$为衰减系数，其使时间上更接近的帧对且具有更高约束权重，同时避免远距离帧对对一致性约束产生过强干扰，结合预实验结果取$\gamma=0.7$。由此，时序一致性约束项定义为

$$\mathcal{L}_{\mathrm{tcons}}=\frac{\sum_t\omega_t\ell_t}{\sum_t\omega_t} \tag{8}$$

该约束直接作用于连续帧的原型特征表达，强制同一抓斗目标在短时间内保持相对稳定的表征分布。与轨迹点层面的后验平滑相比，本文将一致性约束前移至分割表征层，从训练阶段就显式抑制了局部误分和边界扰动引起的时间维表征漂移，从而在源头上阻断误差累积传播，为稳定轨迹提取提供更稳定的跨帧表征基础。由此得到的$\mathcal{L}_{\mathrm{tcons}}$用于后续联合优化。

3.4 联合优化目标与训练策略

时序一致性约束以相对稳定的分割表征为前

提。若训练初期直接引入跨帧约束，原型特征波动较大，不仅难以形成有效对齐，反而会将未收敛的噪声梯度反向传播至分割分支，导致区域提取与时序约束同时退化。基于此，本文采用分阶段训练策略：前E_{w}个epoch仅优化分割损失，待区域表达基本稳定后，再引入时序约束进行联合训练。经预实验取$E_{\mathrm{w}}=5$。

在此基础上，总损失定义为

$$\mathcal{L}=\mathcal{L}_{\mathrm{seg}}+\lambda_{\mathrm{tcons}}\mathcal{L}_{\mathrm{tcons}} \tag{9}$$

其中，$\mathcal{L}_{\mathrm{seg}}$为分割损失，$\mathcal{L}_{\mathrm{tcons}}$定义见式(8)；$\lambda_{\mathrm{tcons}}$为时序一致性约束损失权重，用于控制TCONS在总优化目标中的整体占比。当其取值过大时，易削弱模型对单帧区域判别的关注，取值过小时则难以形成有效时序约束，综合消融实验结果取$\lambda_{\mathrm{tcons}}=0.3$。

联合训练时，每个epoch开始清空序列缓存，并沿视频帧顺序依次处理。仅当当前帧与缓存中的历史帧成功配对后，方计算一致性损失并利用当前帧特征更新缓存。该策略保证跨帧约束始终基于当前轮次的表征分布，避免历史缓存对优化造成干扰。由此，空间表征增强与时序一致性约束被统一至同一目标函数下，模型在保持单帧目标区域提取能力的同时，有效抑制了局部误差在时间维上的累积传播，从而提升连续轨迹提取的稳定性。

4　实验与结果分析

4.1　数据集与实验设置

为验证所提方法在抓斗目标分割与轨迹稳定提取任务中的有效性，本文在DAVIS2016[25]、SegTrackV2[26]和实际门座式起重机抓斗数据集上开展实验。其中，DAVIS2016和SegTrackV2为公开视频目标分割数据集，DAVIS2016用于评估模型在通用视频场景中的目标区域提取能力，SegTrackV2用于考察模型在目标形变、快速运动、局部遮挡和运动模糊等复杂条件下的适应能力。实际门座式起重机抓斗数据集来源于港口真实散货装卸连续作业视频，原始分辨率为1920×1 080，帧率为20fps。数据集由连续作业视频抽帧并进行像素级人工标注，标注时以抓斗可见主体区域为前景，门机结构、钢丝绳、料斗、煤堆和船舱背景等均作为背景处理，样本覆盖抓斗升降、摆动、开合、局部遮挡、尺度变化和复杂背景干扰等典型工况，共3000张图像，按2100/400/500划分为训练集、验证集和测试集，后文简称抓斗数据集。

所有对比模型均采用统一训练配置，输入图像尺寸为640×640，批大小为4，优化器为AdamW，初始学习率为0.002。考虑到本文任务要求模型在无首帧人工掩膜、无人工交互条件下自动提取抓斗区域，并兼顾实时推理能力，因此实验选取与该任务设定一致的自动实例分割模型作为对比方法。对比模型包括YOLOv8-seg、YOLO11-seg、YOLO12-seg和YOLO26-seg，均基于Ultralytics实例分割框架中的对应模型配置实现，并在相同数据划分、输入尺寸和训练配置下重新训练与测试。实验在Windows环境下完成，硬件平台为RTX5060显卡，软件环境为Python 3.12、Ultralytics 8.4.14和Torch 2.11.0+cu128。

轨迹评价时，预测质心和真实质心分别由预测掩膜和人工标注掩膜的前景像素几何中心计算得到；若某帧未预测到目标前景掩膜，则判定为漏检，否则按时间顺序连接预测质心形成目标轨迹。MAE和RMSE分别表示预测质心与真实质心之间欧氏距离误差序列的平均值与均方根；Mean jitter定义为相邻帧预测位移与真实位移之差的标准差；Miss rate表示漏检帧数占总帧数的比例；Max consecutive miss表示连续漏检帧段的最大长度。

4.2　分割性能对比实验

由表1可见，本文方法在边界质量和主要综合指标上均取得最优结果，表明面对真实港口连续作业场景中的复杂背景、局部遮挡和边界退化问题，所提方法能够在稳定分割抓斗主体的同时有效改善区域边界表达。表2进一步显示，该方法在公开数据集上同样具备较强的适应能力。其中，在DAVIS2016上的提升最为突出，本文方法三项指标均排名第一，验证了空间与时序协同优化策略不仅适用于港口抓斗场景，在通用单目标视频场景下同样有效。SegTrackV2包含目标形变、快速运动、局部遮挡和运动模糊等更为复杂的情形，更能考察模型的跨场景泛化能力。在该数据集上，因部分序列相邻帧稳定对应关系较弱，TCONS模块对单帧区域重叠和轮廓匹配指标的增益受到限制，但本文方法仍保持了与基线相近并略有提升的分割性能，表明即使场景动态变化剧烈，所提方法整体上仍能维持较稳定的区域表达。从图2可以进一步看出，相较于YOLOv8-seg，本文方法在抓斗场景中获得了更

表 1　抓斗数据集分割性能对比实验结果(%)

方法	J	F	Mask mAP50-95
YOLOv8-seg	89.45	97.83	88.24
YOLO11-seg	84.63	93.61	85.85
YOLO12-seg	84.84	94.14	85.73
YOLO26-seg	78.58	87.22	86.01
本文方法	**90.05**	**98.56**	**88.54**

表 2 公开数据集分割性能对比实验结果(%)

方法	DAVIS2016数据集			SegTrackV2数据集		
	J	F	Mask mAP50-95	J	F	Mask mAP50-95
YOLOv8-seg	69.38	69.05	47.54	47.53	53.44	**82.23**
YOLO11-seg	64.40	62.51	28.39	45.58	51.30	76.14
YOLO12-seg	69.85	68.07	31.16	45.16	50.11	76.37
YOLO26-seg	64.80	63.84	27.74	41.51	46.82	73.86
本文方法	**74.79**	**74.70**	**51.21**	**47.65**	**53.48**	82.10

(a) 原始图像 (b) 真值 (c) YOLOv8-seg (d) 本文方法

图 2 分割定性对比结果图

完整的目标区域，在DAVIS2016场景中减轻了局部缺损；在SegTrackV2部分序列中，本文方法减少了快速运动条件下的轮廓破碎和孤立误分，说明时序一致性约束对局部分割完整性仍具有一定改善作用，但由于该改善主要体现在局部区域，在全局平均J和F指标上的体现相对有限。结合表3可知，模型参数量较基线仅略有增加，推理速度仍保持在33fps以上，说明所提方法以较小的计算代价取得了更优的分割表现。

4.3 轨迹稳定性分析

为评估不同方法的轨迹稳定提取能力，本文从各帧预测掩膜中提取目标质心并构建轨迹序列，从轨迹偏差、帧间抖动和漏检情况三个方面进行比较。表4和表5的结果表明，在两个数据集上，本文方法在全部4项指标上均取得最优结果，说明分割质量的提升已有效传递至轨迹层面，进而同步改善了定位精度与连续性。在DAVIS2016数据集上，本文方法的Miss rate较YOLOv8-seg下降48.5%，表明所提方法能更有效地抑制轨迹提取中的短时漏检。在抓斗数据集上，MAE和RMSE较YOLOv8-seg分别下降约55.3%和52.6%，Miss rate由2.95%降至0.56%，体现出该方法在连续作业场景下更强的轨迹保持能力，也印证了其对误差传播和漏检累积的抑制作用。以抓斗数据集为例，从图3可以看出，

表 3 模型复杂度比较

方法	Params(M)	FPS	Inference Time(ms)
YOLOv8-seg	11.7905	35.80	27.93
YOLO11-seg	2.8428	42.78	23.38
YOLO12-seg	2.8210	37.09	26.96
YOLO26-seg	3.0531	35.91	27.84
本文方法	11.8378	33.24	30.08

表 4 DAVIS2016数据集轨迹稳定性对比实验结果

方法	MAE (pixel)	RMSE (pixel)	Mean jitter (pixel)	Miss rate (%)
YOLOv8-seg	38.43	56.58	37.59	7.87
YOLO11-seg	61.75	82.17	46.96	6.54
YOLO12-seg	37.38	55.03	31.93	7.73
YOLO26-seg	50.49	77.23	57.97	9.58
本文方法	**32.11**	**49.26**	**31.04**	**4.05**

表 5 抓斗数据集轨迹稳定性对比实验结果

方法	MAE (pixel)	RMSE (pixel)	Mean jitter (pixel)	Miss rate (%)
YOLOv8-seg	6.74	7.13	1.66	2.95
YOLO11-seg	7.20	7.88	3.82	13.74
YOLO12-seg	14.81	16.04	3.02	7.73
YOLO26-seg	9.50	11.42	8.99	18.19
本文方法	**3.01**	**3.38**	**1.52**	**0.56**

相较于YOLOv8-seg，本文方法的轨迹误差分布更加集中，整体波动范围更小，高误差离散点明显减少，表明所提方法不仅降低了轨迹定位误差，还有效抑制了由局部误分和边界扰动引起的非真实波动。总体而言，空间与时序协同优化策略能够将区域提取质量的提升进一步传递到轨迹层面，从而实现更稳定的抓斗轨迹提取。

4.4 遮挡鲁棒性实验

遮挡是影响抓斗连续作业场景下目标稳定提取的关键因素，容易导致目标区域破碎、轨迹中断及漏检累积。实际港口环境中的遮挡通常由钢丝绳、门机结构、料斗边缘及复杂背景干扰等因素引起，遮挡形态具有随机性和不可控性。为在统一扰动条件下比较不同方法对目标区域缺失和短时漏检的鲁棒性，本文在测试样本上分别施加10%、20%和30%的人工遮挡，构建轻度、中度和较重遮挡条件。该实验并非完全复现实港口遮挡形态，而是作为真实场景测试基础上的补充鲁棒性验证，用于分析遮挡强度增加时各方法性能退化趋势，结果如表6所示。整体来看，随着遮挡比例增大，各方法的区域提取质量与轨迹稳定性均出现不同程度的退化，J持续下降，Miss rate和RMSE同步上升。相比之下，本文方法表现出更强的综合鲁棒性：RMSE和Miss rate较YOLOv8-seg分别下降约34.5%和23.6%，最大连续漏检帧数由52帧降至47帧，表明该方法在遮挡条件下仍能维持较稳定的目标响应，在一定程度上抑制了短时失检引发的时序累积效应。从图4的定性结果来看，不同遮挡比例下本文方法提取的目标区域更为完整，局部破碎和位置偏移明显更少。图5进一步揭示，随着遮挡加剧，各对比方法的J均单调下降，Miss rate和RMSE持续上升，而本文方法的变化幅度始终较小，在较高遮挡条件下仍保持了良好的区域完整性与连续提取能力。综合而言，空间表征增强与时序一致性约束的协同设计有效提升了模型对遮挡干扰的适应能力，在连续提取的稳定性方面表现出更好的综合均衡性。

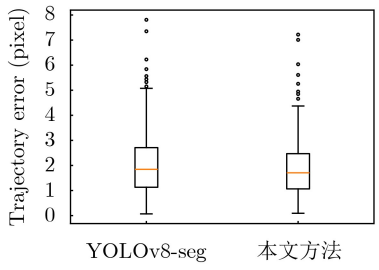

图 3 轨迹误差分布箱线图

4.5 消融实验

为评估各模块的独立贡献与协同效果，本文以YOLOv8-seg为基线，依次引入SRE和TCONS进行消融实验，结果如表7所示。单独引入SRE后，模型部分指标有提升，说明SRE能够提升目标区域表达与质心定位精度，但Miss rate由2.95%升至3.67%，表明仅依靠SRE尚不足以保证连续帧中的目标保持能力。单独引入TCONS后，MAE、RMSE和Miss rate均下降，说明TCONS能够改善相邻帧间响应连续

表 6　抓斗数据集遮挡鲁棒性总体对比实验结果

方法	J(%)	F(%)	MAE(pixel)	RMSE(pixel)	Miss rate(%)	Max consecutive miss (frames)
YOLOv8-seg	61.67	70.22	36.38	156.41	22.95	52
YOLO11-seg	43.80	51.34	48.26	133.65	37.00	85
YOLO12-seg	48.40	56.59	74.97	185.75	26.10	56
YOLO26-seg	57.10	64.87	**32.19**	102.42	28.02	87
本文方法	**63.51**	**72.42**	33.28	**102.38**	**17.54**	**47**

(a) 原始图像　(b) 遮挡图像　(c) 真值　(d) YOLOv8-seg　(e) 本文方法

图 4 遮挡实验对比结果图

性，但其单独作用下分割区域质量提升有限。当两者协同作用时，模型取得最佳综合表现，MAE和RMSE较基线分别下降约55.3%和52.6%，Miss rate下降约81.0%，说明空间表征增强与时序一致性约束之间形成了有效互补，共同提升了连续轨迹提取的稳定性。

表8给出了TCONS关键参数敏感性实验结果。实验采用单因素变量控制方式，每次仅改变一个参数，其余参数保持默认设置不变，用于分析时序一致性损失权重λ_{tcons}、时间衰减系数γ、warmup轮数E_{w}和最大帧间隔G对模型性能的影响。

由表8可知，当$\lambda_{\text{tcons}}=0.3$时，模型在分割质量、轨迹误差和漏检率之间取得较好平衡；当$\gamma=0.7$时，模型综合性能较优，表明适中的时间衰减能够平衡近邻帧约束和远距离帧间干扰；当$E_{\text{w}}=5$时，模型获得较低的轨迹误差和漏检率，说明在分割特征初步稳定后再引入TCONS更有利于连续目标提取；当$G=50$时，模型在轮廓精度和轨迹误差方面表现较优，说明该时间窗口能够较好平衡跨帧信息利用和远距离帧间变化干扰。总体来看，所选参数兼顾分割精度、轨迹误差和漏检率，具有较好的综合稳定性。

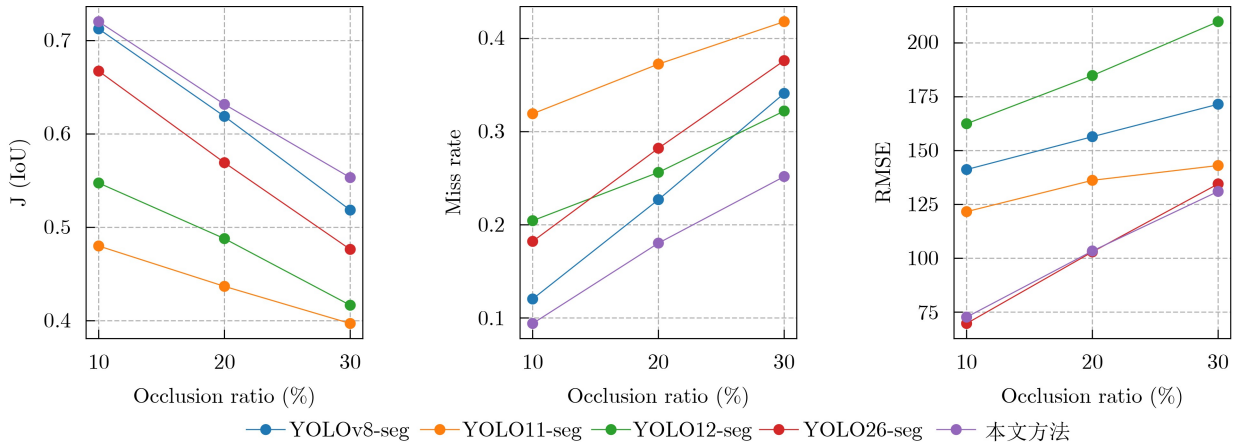

图 5 遮挡实验折线图

表 7 抓斗数据集消融实验结果

Baseline	SRE	TCONS	J(%)	F(%)	MAE(pixel)	RMSE(pixel)	Miss rate(%)	Mask mAP50-95(%)
√			89.45	97.83	6.74	7.13	2.95	88.24
√	√		89.49	97.81	5.33	5.43	3.67	88.40
√		√	89.39	97.78	5.33	5.76	2.89	88.43
√	√	√	**90.05**	**98.56**	**3.01**	**3.38**	**0.56**	**88.54**

表 8 TCONS关键参数敏感性实验结果

参数	取值	J(%)	F(%)	MAE(pixel)	RMSE(pixel)	Miss rate(%)
	0.2	89.53	97.80	3.02	**3.10**	1.11
λ_{tcons}	0.3	**90.05**	**98.56**	**3.01**	3.38	**0.56**
	0.4	89.91	98.07	4.17	4.35	0.67
	0.6	89.90	98.22	3.16	3.53	0.89
γ	0.7	**90.05**	**98.56**	**3.01**	**3.38**	**0.56**
	0.8	89.70	97.95	3.06	3.43	1.00
	0	89.39	97.79	3.02	3.39	1.34
E_{w}	5	**90.05**	**98.56**	**3.01**	**3.38**	**0.56**
	10	89.00	97.59	3.82	4.19	1.22
	30	**90.08**	98.41	3.75	4.12	**0.56**
G	50	90.05	**98.56**	**3.01**	**3.38**	**0.56**
	70	90.10	98.44	3.13	3.50	**0.56**

4.6 适用性与局限性分析

本文方法主要面向固定监控视角下的门座式起重机抓斗作业场景。在该类场景中，摄像机位置相对稳定，抓斗目标外观和运动过程具有一定连续性，因此空间表征增强和时序一致性约束能够较好地发挥作用。对于不同港口环境，若光照条件、背景结构、作业物料和抓斗外观差异较大，模型需要结合目标场景样本进行再训练或微调，以提高跨场景适应能力。对于桥式抓斗卸船机、岸桥等其他设备类型，若目标结构、运动轨迹和遮挡关系发生明显变化，也需要重新构建相应数据样本并进行适配。另一方面，本文方法未专门针对动态摄像视角变化进行设计。当摄像机存在明显移动、变焦或视角快速变化时，相邻帧间背景和目标尺度变化会增强，可能削弱时序一致性约束的有效性，从而影响轨迹提取稳定性。后续可进一步结合视频稳像、运动补偿、多视角融合和跨场景自适应方法，提高模型在动态视角及复杂港口环境下的泛化能力。

5 结束语

港口作业视频中，抓斗目标常因复杂背景、尺度变化、遮挡和帧间形态波动而导致轨迹难以稳定提取。为此，本文提出了一种面向抓斗轨迹稳定提取的空间与时序协同优化方法。通过空间表征增强改善区域及边界表达，通过时序一致性约束抑制帧间误差累积，有效提升了目标分割完整性与轨迹提取连续性。实验结果表明，该方法在通用单目标视频数据集DAVIS2016和实际抓斗数据集上取得了较好的分割与轨迹提取效果，在复杂公开视频目标分割数据集SegTrackV2上保持了与基线相近并略有提升的分割性能，验证了空间表征增强与时序一致性约束的协同设计能够在保证单帧分割精度的同时提升连续轨迹提取的稳定性，为港口工业视频场景下连续运动目标的稳定感知提供了一种可行方案。未来将进一步研究复杂工况下的跨场景泛化与长期连续感知问题，提升模型在实际作业环境中的鲁棒性与适用性。

参 考 文 献

[1] HE Kaiming, GKIOXARI G, DOLLÁR P, *et al*. Mask R-CNN[C]. Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 2980–2988. doi: 10.1109/ICCV.2017.322.

[2] CAELLES S, MANINIS K K, PONT-TUSET J, *et al*. One-shot video object segmentation[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 5320–5329. doi: 10.1109/CVPR.2017.565.

[3] OH S W, LEE J Y, XU Ning, *et al*. Video object segmentation using space-time memory networks[C]. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019: 9225–9234. doi: 10.1109/ICCV.2019.00932.

[4] CHENG H K and SCHWING A G. XMem: Long-term video object segmentation with an Atkinson-Shiffrin memory model[C]. 17th European Conference on Computer Vision – ECCV 2022, Tel Aviv, Israel, 2022: 640–658. doi: 10.1007/978-3-031-19815-1_37.

[5] MONNIER Q, POULI T, and KPALMA K. Survey on fast dense video segmentation techniques[J]. *Computer Vision and Image Understanding*, 2024, 241: 103959. doi: 10.1016/j.cviu.2024.103959.

[6] XU Guoping, UDUPA J K, YU Yajun, *et al*. Segment anything for video: A comprehensive review of video object segmentation and tracking from past to future[J]. *Neurocomputing*, 2026, 682: 133439. doi: 10.1016/j.neucom.2026.133439.

[7] HOU Zhiqiang, LI Fucheng, DONG Jiale, *et al*. Video object segmentation based on dynamic perception update and feature fusion[J]. *Image and Vision Computing*, 2024, 150: 105156. doi: 10.1016/j.imavis.2024.105156.

[8] WANG Jingxin, ZHANG Yunfeng, BAO Fangxun, *et al*. Video object segmentation by multi-scale attention using bidirectional strategy[J]. *Image and Vision Computing*, 2024, 148: 105136. doi: 10.1016/j.imavis.2024.105136.

[9] KIM J, KIM J, and HONG S. G-TRACE: Grouped temporal recalibration for video object segmentation[J]. *Image and Vision Computing*, 2024, 147: 105050. doi: 10.1016/j.imavis.2024.105050.

[10] HOU Zhiqiang, WANG Chenxu, MA Sugang, *et al*. Lightweight video object segmentation: Integrating online knowledge distillation for fast segmentation[J]. *Knowledge-Based Systems*, 2025, 308: 112759. doi: 10.1016/j.knosys.2024.112759.

[11] LIU Yong, YU Ran, YIN Fei, *et al*. Learning high-quality dynamic memory for video object segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, 47(5): 3452–3468. doi: 10.1109/TPAMI.2025.3532306.

[12] SUN Maojin and SUN Minghui. STSim-Mamb: A spatiotemporal similarity learning framework for unsupervised video object segmentation[J]. *Image and Vision Computing*, 2026, 169: 105945. doi: 10.1016/j.imavis.2026.105945.

[13] KAZEMI E S, TOUBAL I E, RAHMON G, *et al*. Domain generalization for multiple video object segmentation and tracking using transformers and smart memory[J]. *International Journal of Computer Vision*, 2026, 134(5): 206. doi: 10.1007/s11263-026-02742-1.

[14] LU Hannan, TIAN Zhi, WEI Pengxu, *et al*. Integrating instance-level knowledge to see the unseen: A two-stream network for video object segmentation[J]. *Neurocomputing*, 2024, 602: 127878. doi: 10.1016/j.neucom.2024.127878.

[15] 侯志强, 董佳乐, 马素刚, 等. 基于多尺度特征增强与全局-局部特征聚合的视频目标分割算法[J]. 电子与信息学报, 2024, 46(11): 4198–4207. doi: 10.11999/JEIT231394.
HOU Zhiqiang, DONG Jiale, MA Sugang, *et al*. Video object segmentation algorithm based on multi-scale feature enhancement and global-local feature aggregation[J]. *Journal of Electronics & Information Technology*, 2024, 46(11): 4198–4207. doi: 10.11999/JEIT231394.

[16] 陈雷, 杨吉斌, 曹铁勇, 等. 一种基于Transformer特征金字塔的自蒸馏目标分割方法[J]. 电子与信息学报, 2025, 47(2): 551–560. doi: 10.11999/JEIT240735.
CHEN Lei, YANG Jibin, CAO Tieyong, *et al*. A self-distillation object segmentation method based on Transformer feature pyramid[J]. *Journal of Electronics & Information Technology*, 2025, 47(2): 551–560. doi: 10.11999/JEIT240735.

[17] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional block attention module[C]. 15th European Conference on Computer Vision – ECCV 2018, Munich, Germany, 2018: 3–19. doi: 10.1007/978-3-030-01234-2_1.

[18] LIN T Y, DOLLÁR P, GIRSHICK R, *et al*. Feature pyramid networks for object detection[C]. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 936–944. doi: 10.1109/CVPR.2017.106.

[19] LIU Shu, QI Lu, QIN Haifang, *et al*. Path aggregation network for instance segmentation[C]. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8759–8768. doi: 10.1109/CVPR.2018.00913.

[20] 丁建睿, 张听, 刘家栋, 等. 融合邻域注意力和状态空间模型的医学视频分割算法[J]. 电子与信息学报, 2025, 47(5): 1582–1595. doi: 10.11999/JEIT240755.
DING Jianrui, ZHANG Ting, LIU Jiadong, *et al*. Medical video segmentation algorithm integrating neighborhood attention and state space model[J]. *Journal of Electronics & Information Technology*, 2025, 47(5): 1582–1595. doi: 10.11999/JEIT240755.

[21] LI Jun, SUN Lijuan, REN Hengyi, *et al*. Learning effective feature representation for video object segmentation via memory[J]. *Knowledge-Based Systems*, 2024, 299: 112020. doi: 10.1016/j.knosys.2024.112020.

[22] WANG Hui, ZHAO Yuqian, ZHANG Fan, *et al*. Multi-scale spatio-temporal memory network for semi-supervised video object segmentation[J]. *Neurocomputing*, 2025, 642: 130487. doi: 10.1016/j.neucom.2025.130487.

[23] MIAO Bo, BENNAMOUN M, GAO Yongsheng, *et al*. Temporally consistent referring video object segmentation with hybrid memory[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(11): 11373–11385. doi: 10.1109/TCSVT.2024.3419119.

[24] HOU Zhiqiang, CUI Hao, WANG Chenxu, *et al*. Frequency-aware fusion for improved video object segmentation[J]. *Neurocomputing*, 2025, 656: 131585. doi: 10.1016/j.neucom.2025.131585.

[25] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, *et al*. A benchmark dataset and evaluation methodology for video object segmentation[C]. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 724–732. doi: 10.1109/CVPR.2016.85.

[26] LI Fuxin, KIM T, HUMAYUN A, *et al*. Video segmentation by tracking many figure-ground segments[C]. Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 2013: 2192–2199. doi: 10.1109/ICCV.2013.273.

陈晓玉：女，副教授，研究方向为信号设计与计算机视觉.
张凤卓：女，硕士生，研究方向为计算机视觉、视频目标分割.
陈 杨：男，高级工程师，研究方向为港口机械与智能化装备技术研发及工程管理.
刘文远：男，教授，研究方向为计算机图形学、人工智能.
孔德明：男，教授，研究方向为计算机视觉、目标跟踪.

责任编辑：余 蓉

A Spatial-Temporal Collaborative Optimization Method for Stable Grab Trajectory Extraction

CHEN Xiaoyu[①] ZHANG Fengzhuo[①] CHEN Yang[②]
LIU Wenyuan[③] KONG Deming[④⑤]

[①](*School of Information Science and Engineering, Yanshan University, Qinhuangdao* 066004, *China*)

②(*Dalian Huarui Intelligent Technology Co., Ltd.*, *Dalian* 116013, *China*)
③(*School of Artificial Intelligence*, *Yanshan University*, *Qinhuangdao* 066004, *China*)
④(*School of Electrical Engineering*, *Yanshan University*, *Qinhuangdao* 066004, *China*)
⑤(*Hebei Key Laboratory of Measurement Technology and Instrumentation*, *Yanshan University*, *Qinhuangdao* 066000, *Hebei Province*, *China*)

Abstract:

Objective In port operation videos, the grab is a continuously moving target, and accurate trajectory extraction is of great importance for operation monitoring, equipment coordination, and collision warning. However, complex backgrounds, scale variations, partial occlusion, and boundary degradation often make target region extraction unstable, leading to centroid deviation, trajectory jitter, missed detections, and trajectory interruption. To address these issues, this paper proposes a spatial-temporal collaborative optimization method for stable and continuous grab trajectory extraction. While maintaining a relatively high inference speed, the proposed method effectively improves both the accuracy and continuity of trajectory extraction, providing a practical solution for stable perception of continuously moving targets in port industrial video scenarios.

Methods Built on YOLOv8-seg, the proposed framework integrates spatial representation enhancement and temporal consistency constraint. First, CBAM, BiFPN-lite, and shallow feature aggregation are introduced to improve target-background separability, strengthen multi-scale representation, and preserve boundary details. Then, a temporal consistency constraint is imposed on prototype features through global average pooling, a cache-based pairing mechanism, and a weighted Charbonnier loss, thereby suppressing the temporal accumulation of local errors. In addition, a stage-wise training strategy with warmup epochs and a unified loss function is adopted to ensure stable convergence.

Results and Discussions Experiments are conducted on DAVIS2016, SegTrackV2, and a real portal crane grab dataset to evaluate the proposed method in terms of segmentation performance, trajectory stability, and occlusion robustness. The results show that the proposed method achieves the best segmentation performance on the grab dataset, with J and F scores of 90.05% and 98.56%, respectively. It also shows clear improvement on DAVIS2016 and maintains comparable performance with slight gains on SegTrackV2 (Tables 1 and 2, Fig. 2). In terms of trajectory stability, compared with YOLOv8-seg, the proposed method reduces MAE and RMSE by approximately 55.3% and 52.6%, respectively, while lowering the miss rate to 0.56% (Table 5). It also produces a more concentrated trajectory error distribution and a smaller fluctuation range (Fig. 3). Occlusion robustness experiments further show that, under different occlusion ratios, the proposed method maintains good region integrity and continuous extraction capability, reducing the maximum number of consecutive missed frames from 52 to 47 (Table 6, Figs. 4 and 5). Ablation studies verify the complementarity of spatial representation enhancement and the temporal consistency constraint, while parameter analysis shows that setting the TCONS weight to 0.3 provides the best balance between segmentation quality and trajectory stability (Tables 7 and 8).

Conclusions This paper proposes a spatial-temporal collaborative optimization method to address the challenge of stable grab trajectory extraction in port operation videos. Experimental results demonstrate that the proposed method achieves favorable segmentation accuracy and trajectory stability on DAVIS2016 and the real grab dataset, maintains comparable segmentation performance on SegTrackV2, shows strong continuous extraction capability under occlusion, and incurs no significant loss in inference speed. Since the current study is limited to fixed crane viewpoints, future work will focus on cross-scene generalization and long-term continuous perception under more complex operating conditions to further enhance the robustness and applicability of the proposed method in real-world environments.

Key words: video object segmentation; target trajectory extraction; stable grab trajectory extraction; spatial representation enhancement; temporal consistency constraint