LoMTrack: 基于自适应稀疏的视觉目标跟踪方法

刘大千，张嘉豪，费博雯†

(辽宁工程技术大学 软件学院，辽宁 葫芦岛 125105)

摘　要: 在视觉目标跟踪领域, 基于 transformer 的跟踪方法通常使用全面微调对预训练模型进行调整, 但其会覆盖预训练模型学到的通用知识, 引发灾难性遗忘, 进而削弱模型的通用视觉表征能力. 此外, 单一网络结构在处理多样化跟踪挑战时易产生知识干扰, 限制模型的泛化性能. 针对上述问题, 提出一种自适应稀疏驱动的目标跟踪方法. 在编码器中, 设计低秩适配 (LoRA) 进行参数高效微调, 通过冻结预训练权重保留通用知识, 仅更新低秩适配器实现快速领域适应. 在解码器中, 设计三重正则化路由混合专家 (TR^2-MoE) 以增强模型的表达和场景适应力, 通过专业化专家网络应对不同类型的跟踪挑战, 并利用动态路由机制实现场景自适应的计算路径选择. 与传统稀疏路由策略相比, TR^2-MoE 通过路由阶段的随机正则化与专家内部的结构化约束, 能够增强复杂场景下的泛化稳健性, 并抑制了专家单元间的过度共适应. 此外, 负载均衡约束采用与序列实际长度相适配的归一化策略, 从而在变长推理中维持更稳定的专家利用率. LoMTrack 在多个基准数据集上进行实验验证, 与当前主流目标跟踪器相比, 均能够取得较为先进的性能. 其中, 在 LaSOText 数据集上的 AUC 达到 53.8%, 明显优于现有跟踪器. 实验结果表明, 在多种场景下展现出更优的动态响应能力与训练鲁棒性.

关键词: 目标跟踪；低秩适配器；混合专家；动态路由；负载均衡

中图分类号: TP391.4　　**文献标志码:** A

DOI: 10.13195/j.kzyjc.2026.0459

引用格式: 刘大千, 张嘉豪, 费博雯. LoMTrack: 基于自适应稀疏的视觉目标跟踪方法 [J]. 控制与决策, xxxx, x(x): xxxx-xxxx.

LoMTrack: visual object tracking via adaptive sparsity

LIU Da-qian，*ZHANG Jia-hao*，*FEI Bo-wen*†

(College of Software，Liaoning Technical University，Huludao 125105，China)

Abstract: In visual object tracking, Transformer-based methods typically fully fine-tune pre-trained models. Full fine-tuning overwrites the general knowledge learned during pre-training, causing catastrophic forgetting and weakening the model's general visual representations. Moreover, a single network architecture easily suffers from knowledge interference across diverse tracking challenges, limiting generalization. We propose an adaptive sparsity-driven tracker to address these issues. For the encoder, we design a parameter-efficient fine-tuning scheme based on Low-Rank Adaptation (LoRA). It freezes the pre-trained weights to preserve general knowledge and updates only the low-rank adapters for fast domain adaptation. For the decoder, we design a Tri-Regularized Router Mixture of Experts (TR^2-MoE) to strengthen expressiveness and scene adaptability. Multiple specialized expert networks handle different tracking challenges, and a dynamic routing mechanism selects scene-adaptive computation paths. Compared with conventional sparse routing, TR^2-MoE introduces stochastic regularization during routing and structural constraints inside experts. This design enhances generalization robustness under complex conditions and suppresses excessive co-adaptation among expert units. The load balancing constraint also adopts a normalization strategy aligned with the actual sequence length. This maintains more stable expert utilization during variable-length inference. Extensive experiments on multiple benchmark datasets show that LoMTrack achieves state-of-the-art performance. Notably, it reaches an AUC of 53.8% on the LaSOText dataset, clearly outperforming existing trackers. The experimental results demonstrate stronger dynamic response capability and training robustness across diverse scenarios.

Keywords: object tracking；low-rank adaptation；mixture of experts；dynamic routing；load balancing

收稿日期: 2026-05-18；**录用日期:** 2026-09-04.

基金项目: 国家自然科学基金项目 (62303477, 62302509)；辽宁省科技厅人工智能创新发展计划项目 (2023JH26/10300027)；辽宁省教育厅基本科研项目 (LJ212510147054, LJ212410147090).

责任编委: 夏元清.

†通信作者. E-mail: feibowen@lntu.edu.cn.

本文附带电子附录文件, 可登录本刊官网该文“资源附件”区自行下载阅览.

0 引 言

视觉目标跟踪是计算机视觉中的基本任务之一, 旨在给定目标初始帧边界框, 并识别定位后续帧中的目标, 广泛应用于无人机追踪、运动估计和车辆自动驾驶等多个领域[1]. 虽然近年来视觉目标跟踪有了很大进展, 但在追踪性能和计算效率方面仍存在巨大的挑战.

近年来, 得益于 Transformer 架构在长序依赖建模与特征交互方面展现的卓越能力, 基于 Transformer 的单流范式逐渐成为视觉目标跟踪领域的主流架构. 该范式将模板和搜索区域整合为单一序列输入编码器, 实现模板与搜索区域图像块间的全局密集交互. 该范式通常使用全量微调对预训练模型进行任务迁移, 然而全量微调面临双重瓶颈: 一方面, 其高昂的计算开销与存储需求严重制约了方法的实际训练效率[2-5]. 更为关键的是, 这种全局参数更新策略会引发显著的灾难性遗忘现象, 导致模型在适配新任务过程中逐步弱化其预训练阶段习得的基础视觉表征能力[6-9]. 此外, 现有跟踪器普遍采用单一稠密网络应对所有场景, 而实际跟踪环境具有高度异质性, 涵盖形变、遮挡、光照变化、运动模糊和相似物干扰等多种复杂情形. 当模型试图将所有场景知识压缩到同一组参数中时, 不同场景属性之间极易诱发参数内部的知识冲突, 从而在根本上限制了模型在多变场景下的泛化表现.

为解决上述问题, 本文提出一种高效跟踪框架 LoMTrack(LoRA with Tri-Regularized Router Mixture of Experts for Efficient Visual Object Tracking). 在编码器部分, 原始预训练权重被冻结以维持模型的通用视觉先验, 同时设计一种基于低秩适配 (Low-Rank Adaptation, LoRA)[10] 的参数高效微调方案, 在参数空间的低秩子空间中进行适应性更新. 这种设计使得编码器能够快速适应下游跟踪任务, 并将可训练参数量压缩数个数量级, 有效抑制灾难性遗忘, 进而实现参数高效性与模型稳定性的良好平衡. 在解码器部分, 构建一种基于三重正则化路由混合专家 (TR^2-MoE) 的动态计算范式. 通过构造多个专业化专家网络, 各专家可专注于捕获特定跟踪挑战中的判别性特征. 为了优化专家协作效率, 引入动态门控路由机制, 依据输入特征的语义表征自适应地选择最相关的专家组合. 在此基础上, 为增强路由决策的泛化鲁棒性, 在专家分配阶段注入适度随机扰动, 迫使模型避免对单一专家组合的过度依赖. 此外, 本文在各专家子网络内部还引入隐空间正则化约束, 以抑制训练过程中专家间的过度共适应. 同时, 通过引入负载均衡约束, 并采用与序列实际规模相适配的归一化策略, 确保各专家在网络训练期间中获得均衡的激活概率, 从而缓解专家利用不足或过度使用的问题. 在推理阶段, 该架构采用稀疏激活策略, 仅调用部分相关专家参与计算, 前述正则化设计能够进一步保障稀疏推理的稳定性, 在维持模型总体表征能力的同时, 降低计算开销. 这种 LoRA 与 MoE 的协同设计, 既继承预训练模型的通用视觉先验, 又通过专家分工机制增强对多样化跟踪场景的专门化处理能力.

本研究主要贡献总结如下:

• 提出一种高效跟踪框架 LoMTrack, 解决单一微调策略在复杂跟踪场景下适应性与鲁棒性不足的问题, 增强跟踪器对多样化挑战的动态响应能力.

• 设计一种基于 LoRA 的编码器参数高效微调方案, 解决全量微调导致的可训练参数冗余与灾难性遗忘问题, 仅更新低秩适配器压缩可训练参数规模, 实现对下游任务的高效适配.

• 提出 TR^2-MoE 动态计算范式, 通过动态门控与专家专业化实现场景自适应路由, 同时设计长度感知负载均衡策略, 维持变长推理下的专家利用率, 进而解决传统 MoE 路由泛化弱与专家共适应问题.

• 在 GOT-10k[11]、LaSOText[12] 和 OTB100[13] 等多个大规模公开数据集的各项挑战中与当前流行的跟踪器进行对比, 验证提出的 LoMTrack 方法的优异性.

1 相关工作

1.1 基于 Transformer 的跟踪器

早期, 基于孪生网络的跟踪方法通常采用双流范式[14-16]: 首先使用一个共享权重的骨干网络提取模板帧和搜索帧的特征, 再通过互相关操作进行特征匹配, 从而直接回归目标位置. 相比之下, 单流范式通过整合模板和搜索区域信息, 输入到 Transformer 框架中进行联合特征提取和匹配. 这种方式实现了更深入灵活的特征交互, 从而从根本上避免了传统双流方法因“先分离后融合”处理所造成的信息隔阂. Chen 等人[17] 通过空域抑制与选择性查询回忆策略优化了 Transformer 训练过程. Hou 等人[18] 通过多维度特征增强与自适应位置编码提升了跟踪性能与实时性.

总体而言, 基于 Transformer 的单流跟踪器凭借全局特征交互已取得显著进展, 但其成功高度依赖于大规模预训练模型的全量微调, 这不仅带来高昂

的计算与存储开销, 还容易引发灾难性遗忘、削弱预训练基础表征. 针对这一瓶颈, 本文在编码端引入 LoRA 参数高效微调策略, 通过冻结预训练权重并仅优化低秩适配器, 在保留通用视觉先验的同时大幅压缩可训练参数量, 有效抑制了灾难性遗忘.

1.2 参数高效微调

参数高效微调 (PEFT) 旨在解决全参数微调在资源消耗和计算效率上的局限性. 其一系列方法通过冻结预训练模型主体、仅优化少量额外参数以实现高效的任务适配. 其主要分为三种方法: Prompt-Tuning[19] 通过在模型输入层或中间隐层引入k个可训练的前缀标记, 仅优化这些新增的语义前缀参数. Adapter-Tuning[20] 则是在预训练模型的每个层级中嵌入轻量级的神经网络模块, 即适配器, 微调过程中仅更新这些植入的适配器参数. LoRA[10] 通过低秩分解技术, 采用可训练的低秩矩阵逼近权重更新量, 在保持原始模型参数冻结的前提下实现高效优化.

然而, Prompt-Tuning 的效果受限于前缀长度: 若长度不足, 则表征能力有限. 若长度过长, 则挤占上下文资源且增加计算开销. Adapter-Tuning 的串行结构会在推理过程中产生固有延迟, 影响实时性能. 适配器的植入位置需精确设计, 不当的配置可能中断原模型的信息传递路径, 进而降低知识迁移效率.

LoRA 作为该领域的代表性技术, 通过数学上的低秩假设近似参数更新过程. 该方法在微调过程中保持原始权重矩阵固定, 通过训练一组低秩矩阵表征参数变化的方向和幅度. 这种设计不仅实现了极高的参数效率, 还具备独特的工程优势: 训练完成后, 低秩矩阵可以与原权重直接合并, 既不会引入任何额外的推理延迟, 又能保持与全参数微调相当的模型性能. 而在许多任务中, LoRA 已经被证实可以达到全面微调相当甚至更好的性能[21]. Lin 等人[22] 通过解决位置编码冲突与卷积头归纳偏置, 将 LoRA 应用于视觉跟踪. Liu 等人[23] 引入稀疏门控动态调整秩, 结合近端梯度优化与训练后剪枝实现紧凑微调.

以 LoRA 为代表的参数高效微调方法虽然实现了轻量级适配, 但本质上仍基于单一的稠密网络, 将所有场景知识压缩至同一组可调参数中, 难以化解不同跟踪挑战之间的知识冲突, 从而限制了模型在复杂多变环境下的泛化能力. 为了弥补这一缺陷, 本文在解码端进一步引入混合专家机制, 利用多个专业化专家网络对不同跟踪场景进行差异化处理, 以缓解单一网络的知识冲突.

1.3 混合专家模型

混合专家模型 (Mixture of Experts, MoE)[24] 的核心在于其稀疏激活特性, 从而将模型的总参数量与单次计算成本解耦. 这使得该模型能够在参数规模显著增长的同时, 将计算开销控制在合理范围内. 近年来, Raposo 等人[25] 通过动态令牌路由与 top-k 选择机制, 在保持性能的同时提升了推理效率. Li 等人[26] 通过任务敏感提示驱动的混合专家框架, 实现了开放世界持续学习中知识稳定性与可塑性的平衡.

尽管混合专家模型在多个领域取得了成功, 但其在目标跟踪中的应用尚未得到充分的探索. Tang 等人[27] 采用多专家混合架构, 通过置信度动态选择最优结果, 自适应解决了多模态融合时机问题. Cai 等人[28] 提出面向跟踪的专家混合模块, 结合多专家能力处理多样化关系建模, 实现了多尺度跟踪.

混合专家模型通过稀疏激活实现了参数规模与计算成本的解耦, 为动态场景自适应提供了基础, 但现有工作大多未与参数高效微调深度协同, 且传统 MoE 普遍面临路由决策泛化性弱、专家间过度共适应及负载不均衡等问题, 直接迁移至跟踪器容易导致训练不稳定和专家利用低效. 鉴于此, 本文提出 TR²-MoE, 通过注入适度随机扰动增强路由泛化性、引入隐空间正则化抑制专家共适应, 并设计长度感知的负载均衡策略保障变长推理下的专家利用率, 从而在继承预训练通用视觉先验的前提下, 实现对多样化跟踪挑战的稳定、高效自适应.

2 LoMTrack 跟踪器

本节致力于详细解读 LoMTrack 跟踪器的技术方案. 首先, 整体呈现跟踪器的系统框架, 阐述其工作流程. 接着, 深入解析编码器与解码器内部结构. 最终, 详细介绍负责输出预测的头部网络以及指导模型训练的损失函数.

2.1 总体框架

LoMTrack 跟踪器框架如图 1 所示, 其核心由编码器与解码器构成. 编码器基于 LoRA 机制实现参数高效微调, 从中提取判别性目标特征. 解码器基于 TR²-MoE 结构实现上下文信息的动态维护与传递. 最终由预测头生成跟踪结果.

在单流跟踪器的设计框架下, 采用基于卷积的块嵌入方法分别对模板图像与搜索图像进行序列化建模. 假设模板图像为$T \in \mathbb{R}^{H_z \times W_z \times 3}$和搜索图像为$S \in \mathbb{R}^{H_x \times W_x \times 3}$, 首先将每幅图像划分为规则的图像块. 块尺寸设为$P_{size} = (P, P)$, 则模板图像被划分为$N_T = N_T^H \times N_T^W$个块, 其中$N_T^H = H_T/P$, $N_T^W =$

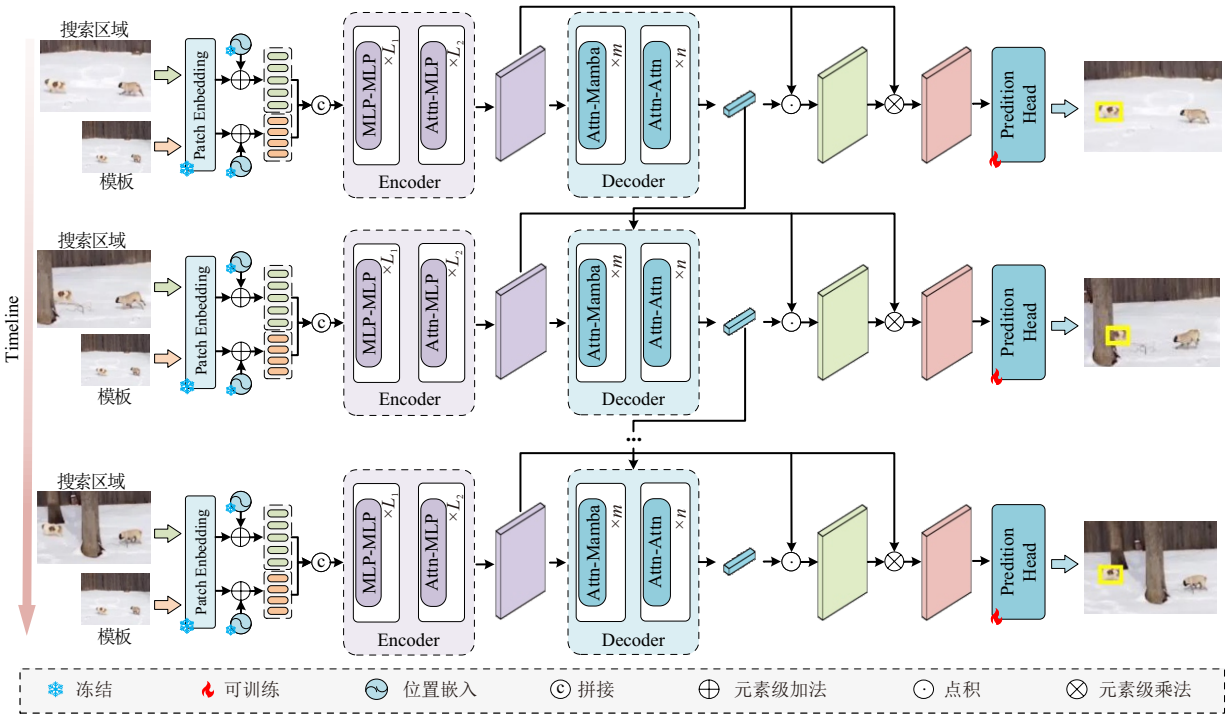


图 1 LoMTrack 跟踪框架

$W_T/P.H_T$和W_T分别为模板图像的高度和宽度. 搜索区域图像被分割为$N_S = N_S^H \times N_S^W$个块, 其中$N_S^H = H_S/P$, $N_S^W = W_S/P$. H_S和W_S分别为搜索图像的高度和宽度. 每个图像块进一步细分为$M \times M$个内部单元. 其中M为每个图像块在高度和宽度方向上的划分份数. 通过二维卷积操作将每个内部单元投影到嵌入空间, 卷积核尺寸为(K, K), 其中$K = P/M$, 即每个内部单元的大小恰好等于卷积核的感受野, 步长设置与卷积核尺寸保持一致. 该过程分别生成模板令牌嵌入$\mathbf{T} \in \mathbb{R}^{N_T \times d}$和搜索区域令牌嵌入$\mathbf{S} \in \mathbb{R}^{N_S \times d}$, 其中$d$为嵌入维度. 在此基础上, 再分别在模板令牌和搜索令牌添加位置嵌入, 进而得到模板序列嵌入$\mathbf{T}_e \in \mathbb{R}^{N_T \times d}$与搜索区域序列嵌入$\mathbf{S}_e \in \mathbb{R}^{N_S \times d}$, 其计算过程可表示为:

$$\mathbf{T}_e = \mathbf{T} + \mathbf{E}_T, \tag{1}$$

$$\mathbf{S}_e = \mathbf{S} + \mathbf{E}_S, \tag{2}$$

其中, $\mathbf{E}_T \in \mathbb{R}^{N_T \times d}$为模板图像的位置嵌入, $\mathbf{E}_S \in \mathbb{R}^{N_S \times d}$为搜索区域图像的位置嵌入. 随后, 将模板序列嵌入与搜索区域序列嵌入合并为联合序列f_0作为编码器的输入. 编码器基于 HiViT[29] 架构构建, 对输入序列进行多层次的时空特征整合. 编码器输出的时空增强特征随后送入解码器, 该解码器融合自注意力、双扫描选择性 Mamba(DualScan Selective Mamba, DSSMamba)、多头交叉注意力与 TR^2-MoE, 通过结合 DSSMamba 的状态空间建模能力与注意力的全局交互机制, 显著增强模型对长程上下文信息的捕获能力. 最终, 解码后的特征由预测头生成目标跟踪结果.

$$f_0 = \mathrm{Concat}(\mathbf{T}_e, \mathbf{S}_e), \tag{3}$$

$$f_n = \mathrm{Encoder}(f_{n-1}),\ n = 1 \ldots N, \tag{4}$$

$$f_m = \mathrm{Decoder}(f'_{m-1}),\ m = 1 \ldots M \tag{5}$$

2.2 LoRA

预训练权重通常在大规模数据集上经过充分训练, 蕴含了丰富的通用知识. 传统的全参数微调方法通过更新模型全部参数进而适配下游任务, 需要大量的计算资源且易引发灾难性遗忘. 基于模型在适配下游任务时, 其权重更新矩阵具备低秩性这一核心假设, LoRA 能够提供一种高效的参数微调策略. 该假设指出, 尽管模型参数规模庞大, 但其适应新任务所需的变化方向蕴藏在一个低维的本征子空间中. 具体实现方式是在原始冻结的预训练权重旁, 引入一组可训练的低秩矩阵对B和A. 前向传播过程被修正为:

$$W' = W_0 + \Delta W = W_0 + BA, \tag{6}$$

其中, $W_0 \in \mathbb{R}^{d \times k}$是被冻结的预训练权重, ΔW是权重更新矩阵, $A \in \mathbb{R}^{r \times k}$和$B \in \mathbb{R}^{d \times r}$是可训练的低秩矩阵. $r \ll \min(d, k)$即远小于模型维度的秩, 是 LoRA 的核心超参数. 将总参数量从$d \times k$降低到$(d + k) \times r$, 相较于全参数微调呈现数量级式的压缩.

在训练过程中, W_0被冻结且不参与梯度更新, 仅优化低秩矩阵A和B. 对于输入x, 模型的前向传

播过程 $h_0 = W_0x$被更新为:

$$h = W_0x + \Delta Wx = W_0x + BAx. \tag{7}$$

此外, 使用 LoRA 并不会增加推理延迟, 训练完成后, 其参数可以通过简单的矩阵加法与原始预训练权重W_0合并, 合并后的模型在结构上与原始预训练模型完全一致, 不引入任何额外的层或参数.

2.3 编码器

HiViT[29] 采用分层 Transformer 架构, 通过渐进式多尺度设计兼顾局部特征提取与全局语义建模. 其核心在于分阶段处理机制: (1) 在浅层阶段, 图像被划分为细粒度图像块并执行局部变换. (2) 在深层阶段, 通过块合并构建金字塔结构, 在压缩空间分辨率的同时扩展通道维度.

在架构的早期阶段, 多层感知机 (Multi-Layer Perceptron, MLP) 模块主要处理局部特征. 最后一个阶段则引入全局自注意力以实现全图上下文交互. 该设计在无需依赖复杂窗口机制的前提下, 能够达成多尺度表示能力与架构简洁性之间的有效平衡.

为了保持预训练模型强大泛化能力的同时, 快速适配特定的跟踪场景. 本文在 HiViT 架构中系统性地集成 LoRA 技术, 编码器的具体结构如图 2 所示.

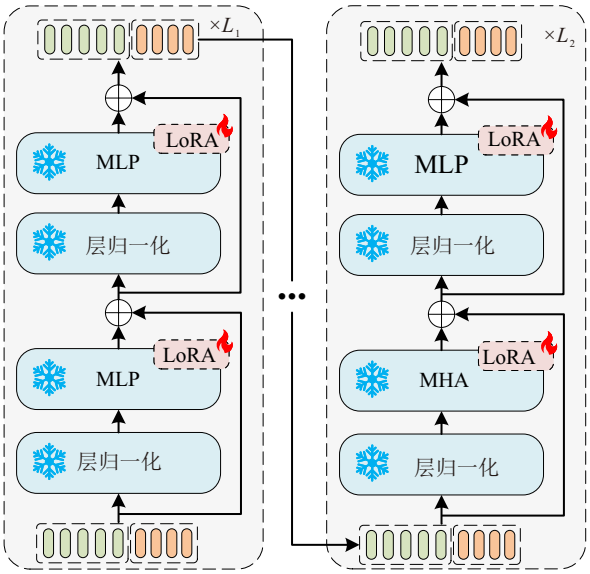


图 2 编码器结构

在注意力机制中, 为每个核心投影层都集成独立的 LoRA 适配器. 具体而言, 查询投影、键投影、值投影和输出投影四个线性变换层都分别嵌入专用的 LoRA 模块. 每个投影层可以独立调整其低秩适配强度, 从而规避合并投影方式中三个分量须共享相同秩维度的约束. 相较于合并投影策略, 本文所采用的全面适配方案赋予模型更丰富的表达灵活性, 使其能够更精确地建模模板与搜索区域之间的复杂对应关系. LoRA 的主要过程可表示为:

$$f^{(l)}_{\mathrm{mlp}_1} = \mathrm{LoRAMLP}_1(\mathrm{LN}(f^{(l-1)})) + f^{(l-1)} \quad l = 1 \ldots L_1, \tag{8}$$

$$f^{(l)}_{\mathrm{mlp}_2} = \mathrm{LoRAMLP}_2(\mathrm{LN}(f^{(l-1)}_{\mathrm{mlp}_1})) + f^{(l)}_{\mathrm{mlp}_1} \quad l = 1 \ldots L_1, \tag{9}$$

$$f^{(l)}_{\mathrm{attn}} = \mathrm{LoRAAttn}(\mathrm{LN}(f^{(l-1)}_{\mathrm{mlp}_2})) + f^{(l-1)}_{\mathrm{mlp}_2} \quad l = 1 \ldots L_2, \tag{10}$$

$$f^{(l)}_{\mathrm{mlp}} = \mathrm{LoRAMLP}(\mathrm{LN}(f^{(l-1)}_{\mathrm{attn}})) + f^{(l)}_{\mathrm{attn}} \quad l = 1 \ldots L_2. \tag{11}$$

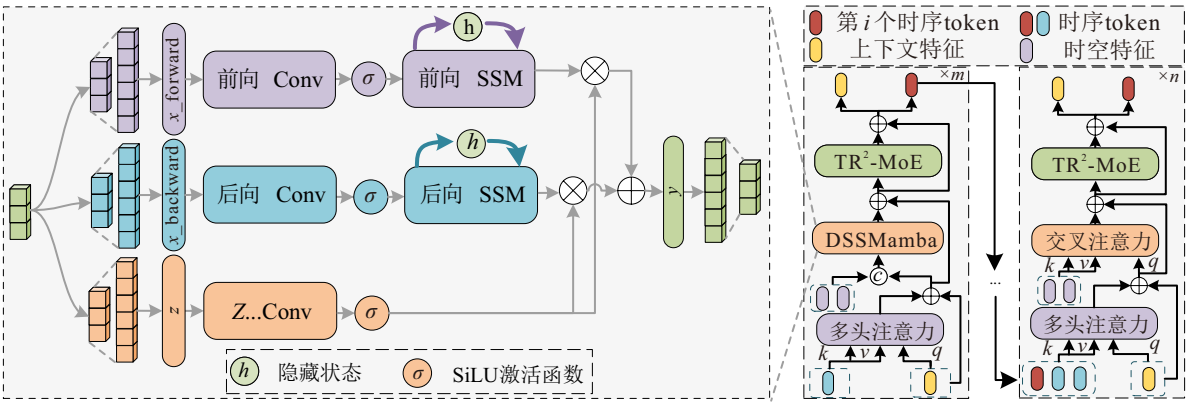
在训练过程中, 本文采用分阶段的参数优化策略. 首先冻结所有预训练的主干网络参数, 仅保留 LoRA 适配器为可训练状态. 由此将需优化的参数量压缩至原始模型的极小比例, 旨在降低计算与存储开销. 此外, 标准 LoRA 对低秩矩阵A和B赋予相同学习率, 会导致大宽度模型微调陷入低效, 为实现更高效的特征学习和更快的收敛速度, 本文对两个低秩矩阵分别设定不同的学习率.

2.4 解码器

在解码器中, 本文构建一个多阶段协同的混合计算框架. 每一阶段有机融合三种功能各异的处理单元: 多头注意力机制、DSSMamba 与交叉注意力交替使用的跨模态交互单元以及基于动态路由的 TR^2-MoE 网络. 解码器结构如图 3(a) 所示, 各模块以前馈方式顺序连接, 并辅以残差连接与层归一化确保训练稳定性. 具体而言, 自注意力单元聚焦于内部查询向量的关系建模. 跨模态交互单元在相邻层级间交替采用双向状态空间模型与交叉注意力——前者通过隐式状态传递捕获跨序列的长期依赖, 后者通过显式交互计算建立模板与搜索区域的细粒度对应关系. 而 TR^2-MoE 单元则根据输入特征自适应调度多个专用网络进行特征重构.

DSSMamba 旨在通过隐式状态捕获长距离依赖关系, 其结构如图 3(b) 所示. 该模块设计包含两个核心组件: 首先, 通过线性投影将输入特征映射到高维空间, 并利用一维因果卷积进行局部上下文特征的提取. 其次, 通过选择性扫描机制动态整合序列信息, 该机制根据当前输入自适应地调整状态转移参数, 从而实现对长程依赖关系的有效建模. 特别地, 通过独立参数化的前向与反向扫描路径分别捕捉两个方向的序列上下文, 实现支持双向处理模式. 最终将两者的输出进行融合, 以构建包含更完整上下文信息的序列表征.

TR^2-MoE 模块的目的是增强 LoMTrack 的特征



(b) DSSMamba (a) 解码器

图 3 解码器结构

处理能力, 其结构如图 4 所示, TR^2-MoE 模块由多个专家网络组成, 每个专家均为一个独立的前馈层, 具备特征变换和非线性激活能力. 通过可学习的路由机制, 模型能够针对不同输入特征动态选择最适配的专家组合进行响应.

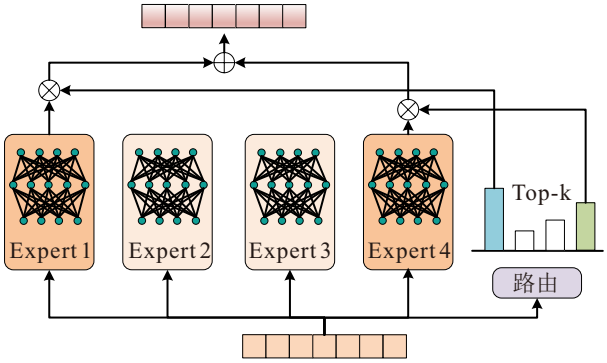


图 4 TR^2-MoE 结构

对于每个输入特征向量, 路由网络计算其与各个专家的匹配度:

$$\text{logits} = W_r \cdot x, \tag{12}$$

其中, W_r路由权重矩阵. 为增强路由决策的泛化鲁棒性, 本文在计算匹配度后随机将部分匹配度置零. 该乘性扰动迫使模型在训练过程中不依赖于固定的专家组合, 而是探索多样化的路由路径, 从而有效抑制路由决策对单一专家的过度依赖. 经过温度系数τ调节后, 通过 softmax 函数生成归一化的路由概率分注入适度随机扰动, 具体而言, 以概率$p = 0.1$ 对路由 logits 施加 Dropout 操作.

$$\mathbf{g} = \text{softmax}(\mathbf{logits}/\tau), \tag{13}$$

每个输入被分配给最相关的k个专家构成激活专家集S_k, 各专家的输出按路由概率加权融合:

$$y_{\text{route}} = \sum_{i \in S_k} \frac{g_i}{\sum_{j \in S_k} g_j} \cdot \text{Expert}_i(x), \tag{14}$$

这种动态选择机制使模型能够根据跟踪过程中不同目标的外观特性、运动模式和环境上下文, 进而自适应地调度最匹配的特征处理专家. 此外, 每个专家网络采用标准的双层前馈结构, 并在非线性激活后引入隐层正则化以抑制不同专家单元间的过度共适应.

$$\text{Expert}(x) = \text{fc2}(\text{GELU}(\text{fc1}(x))), \tag{15}$$

其中, fc1 和 fc2 为线性层.

2.5 预测头部和损失函数

预测头通常由多个子卷积网络组成, 集成分类头和边界框回归头的功能. 这些子网络输出三类核心信息: 用于目标类别判定的分类得分图、用于补偿分辨率降低所引起的离散化误差的偏移量, 以及归一化的边界框尺寸, 通过选取分类得分最高的位置, 再将该位置与偏移量及归一化边界框尺寸相结合, 进而计算得到最终的跟踪结果.

在损失函数方面, 本文采用多任务损失融合策略. 在分类任务中引入焦点损失L_{cls}以缓解类别不均衡问题. 在边界框回归任务中, 将L_1损失[29] 与广义 IoU(GIoU)[30] 损失相结合, 以提升边界框的定位精度.

此外, 为优化 TR^2-MoE 模块的专家负载分配, 本文设计基于负载分布的均衡损失$\mathcal{L}_{\text{balance}}$. 该损失通过促进专家重要性分布与负载分布的一致性优化路由决策. 具体而言, 定义专家重要性为所有输入位置上路由概率的累积和:

$$I_i = \sum_{b=1}^{M} p_i^{(b)}, \tag{16}$$

其中, M为一个批次内所有 token 的总数, $p_i^{(b)}$表示第b个 token 被路由至第i个专家的概率. 定义专家负载为所有输入位置上被选中的次数:

$$L_i = \sum_{b=1}^{M} \mathbb{I}[i \in S_k^{(b)}], \tag{17}$$

其中, $\mathbb{I}(\cdot)$为指示函数, $S_k^{(b)}$是第b个 token 经路由层 top-k 选择后确定的专家子集. 基于上述定义, 负载均衡损失通过促进重要性分布与负载分布的一致性优化路由决策:

$$\mathcal{L}_{\text{balance}} = \lambda \cdot \frac{N}{M^2} \sum_{i=1}^{N} I_i \cdot L_i, \tag{18}$$

其中, λ为平衡权重系数, N为专家总数. 平方归一化因子M^2能够消除批次规模变化对负载均衡损失幅值的影响, 使该正则项在变长序列训练场景下保持一致的惩罚力度, 从而维持更稳定的专家利用率分布.

最终, 总损失函数定义如下:

$$L = L_{cls} + \lambda_{giou} L_{giou} + \lambda_{L_1} L_1 + \mathcal{L}_{\text{balance}} \tag{19}$$

其中, $\lambda_{giou} = 2$和$\lambda_{L_1} = 5$分别是正则化参数.

3 实 验

在本节中, 首先阐述所提方法的实现细节, 随后在七个主流基准数据集上将 LoMTrack 与现有 SOTA 方法进行性能对比, 具体包括 UAV123[31]、LaSOT[32]、GOT-10k[11]、LaSOText[12]、TNL2K[33]、OTB100[13] 以及 TrackingNet[34]. 最后, 设计消融实验, 对所提出的 LoMTrack 各组件的有效性进行定性和定量分析, 同时补充目标跟踪结果与可视化分析内容, 为 LoMTrack 的核心工作机制提供直观解读.

3.1 实验细节

本节的实验环境为: 软件环境为 Python 3.8 和 PyTorch 1.17.0, 硬件为 4 块 NVIDIA 4090D GPU. 为适配多样化的实验需求, 设计两种参数配置不同的 LoMTrack 变体 (LoMTrack-256 和 LoMTrack-384). LoMTrack 主要由编码器和解码器两个组件构成: 编码器采用 HiViT-Base 架构, 配置 20 层. 解码器包含 4 层, 隐藏维度为 256, 注意力头数为 8, TR^2-MoE 的隐藏维度为 512. 训练遵循主流跟踪器的通用范式, 使用四个数据集:COCO[35]、LaSOT[32]、TrackingNet[34] 和 GOT-10k[11]. 并采用亮度扰动与水平翻转等标准数据增强策略. 使用 AdamW 优化器训练 LoMTrack, 矩阵B的学习率设为1×10^{-4}, 是矩阵A所分配学习率 2.5×10^{-5}的四倍, 其余参数的学习率为 4×10^{-4}, 权重衰减系数为10^{-4}. 上述配置与 OSTrack[1] 保持一致. LoMTrack 总计训练 150 个周期, 每周期使用 6 万对图像, 学习率在第 120 周期后执行衰减. 在 GOT-10k[11] 基准上, 训练总周期数降至 100, 学习率衰减于第 80 周期执行.

3.1.1 训练

在 COCO[35]、LaSOT[32]、TrackingNet[34] 和 GOT-10K[11] 数据集上进行训练. 对于 LoMTrack-256 变体, 搜索区域大小设为256×256, 模板大小设为128×128. 每个训练周期包含 15,000 个样本, 批次大小为 8, 共训练 150 个周期. 对于 LoMTrack-384, 搜索区域设为384×384, 模板大小设为192×192. 每个周期同样包含 15,000 个样本, 但批次大小降至 2, 训练周期数保持 150.

当仅在 GOT-10K[11] 数据集上训练时, 搜索区域和模板大小均设为256×256. 在此配置下, 每个周期包含 7,500 个样本, 批次大小为 4, 共训练 100 个周期. 对于 LoMTrack-384, 搜索区域设为384×384, 模板大小设为192×192. 每个周期仍为 7,500 个样本, 批次大小设为 2, 训练周期数同为 100.

3.1.2 推理

为了清晰展示 LoMTrack 的效率, 表 1 对不同方法性能指标进行比较. 与同样具备相同输入规模的基线模型 SeqTrack-B256[6] 相比, 所提出的 LoMTrack-256 在保持相当参数量的同时, 显著降低了计算复杂度. 此外, 随着输入规模的增大, LoMTrack-384 仍能保持 30.4FPS 的实时推理速度, 相较于 SeqTrack-B384 的 19FPS 展现出明显优势——这突显了本文方法在计算效率和实用性方面的提升.

表 1 不同跟踪器推理对比结果

跟踪器	速度	参数量	乘加运算次数	设备
TATrack-B[36]	14.1	112.8	45.1	V100
$ARTrack_{256}$[37]	13.5	181.0	172.1	V100
SeqTrack-B256[6]	48	89	66	4090D
SeqTrack-B384[6]	19	89	148	4090D
LoMTrack-256	31.9	91.9	32.7	4090D
LoMTrack-384	30.4	91.9	73.6	4090D

3.1.3 损失曲线

四种训练配置下的损失曲线如图 5 所示, 分别为: (a) LoMTrack-256 在 COCO、LaSOT、TrackingNet 和 GOT-10K 四个数据集上联合训练的损失曲线. (b) LoMTrack-256 仅在 GOT-10K 上训练的损失曲线. (c) LoMTrack-384 在四个数据集上联合训练的损失曲线. (d) LoMTrack-384 仅在 GOT-10K 上训练的损失曲线. 四组损失曲线均呈现清晰的收敛态势, 随着训练推进, 损失整体呈稳定下降趋势. 在各自放大的关键训练区间内, 曲线未出现明显反弹或振荡, 表明 LoMTrack-256 和 LoMTrack-384 在不同训练配置下均能有效收敛, 且收敛后仍具备可靠的泛化能力. 综上所述, 两个模型最终均达成良好的拟合状态, 充分验证训练过程的稳定性与有效性.

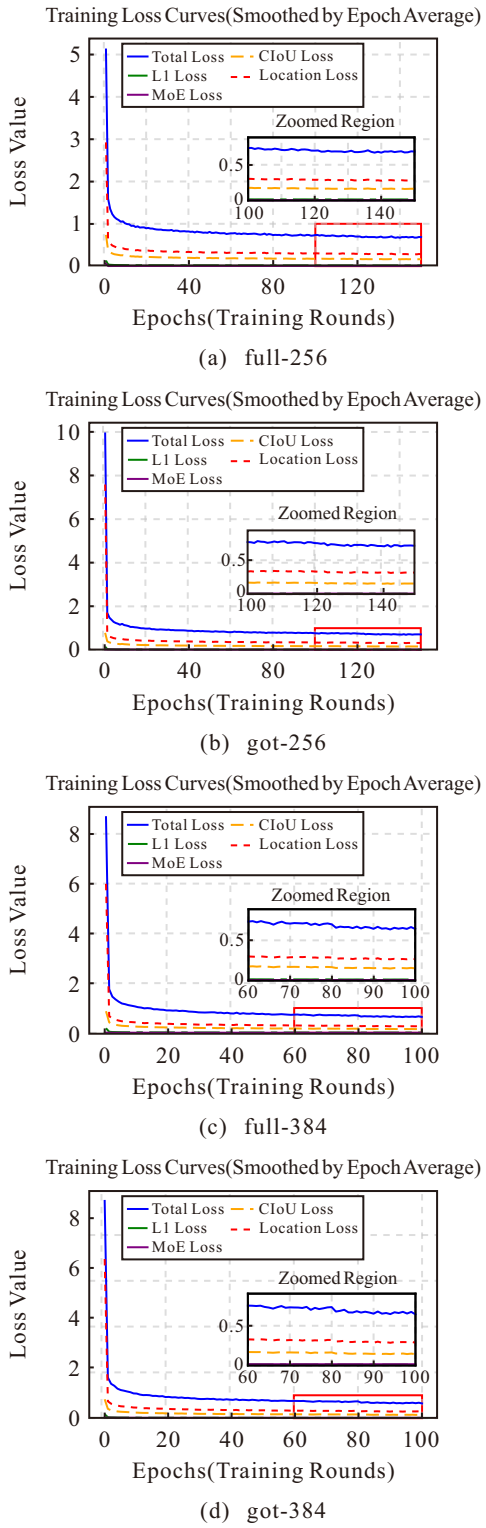


图 5 训练损失曲线

3.2 结果与对比

为进一步验证所提 LoMTrack 的优越性, 本节将其与七个主流基准数据集上的现有 SOTA 跟踪方法展开全面对比. 在训练阶段, 模型采用 COCO[35]、LaSOT[32]、TrackingNet[34] 和 GOT-10k[11] 四个数据集联合训练; 在测试阶段, 则覆盖 UAV123[31]、LaSOT、GOT-10k、LaSOText、TNL2K[31]、OTB100[13] 和 TrackingNet 共七个基准. 其中, UAV123、TNL2K、LaSOText、OTB100 等数据集完全未参与训练, 可严格检验模型的跨域迁移泛化能力. 在 GOT-10K 上添加符号*表示相应方法仅使用 GOT-10K 训练集进行训练. 最优结果以**粗体**突出显示, 次优结果以下划线标示.

3.2.1 LaSOT

LaSOT[32] 是专注于单目标长时跟踪任务的高质量基准数据集, 包含 1 120 个训练序列与 280 个测试序列两部分. 其测试序列平均长度逾 2 500 帧, 且广泛涉及目标消失、外观剧烈变化等挑战, 是评估长时间跟踪鲁棒性的核心基准. 不同跟踪器在此数据集上的对比结果见表 2, LoMTrack 在与当前先进的跟踪器比较中展现出具有竞争力的性能. LoMTrack-384 以 72.6% 的 AUC 分数领先于其他前沿跟踪器, 其优异性能得益于本文所采用的微调方案——在冻结主干网络原始权重的前提下, 仅利用 LoRA 模块进行高效优化. 该策略有效避免了基础知识的遗忘, 确保了跟踪器的稳健性能.

3.2.2 LaSOText

LaSOText[12] 作为 LaSOT 的扩展数据集, 涵盖 150 个新增视频序列. 这些序列纳入遮挡、快速移动小物体等挑战性跟踪场景, 进一步丰富长时跟踪任务的评估维度. 不同跟踪器在此数据集上的对比结果见表 2, LoMTrack-384 和 LoMTrack-256 在 AUC、P_{Norm} 和 P 指标上表现尤为突出, 均领先于其它先进跟踪器, 这充分证明了 LoMTrack 在面对复杂视觉干扰时仍能保持高度鲁棒性与跟踪准确性, 展现出优异的场景适应能力与泛化性能.

3.2.3 TrackingNet

TrackingNet[34] 是一款面向单目标跟踪任务的大规模基准数据集, 包含 511 个视频样本, 涵盖丰富的物体类别与多样化场景. 值得强调的是, 其测试序列在长度分布上具有显著的长时特性——部分序列持续时间超过 4 000 帧, 且频繁出现目标消失、运动模糊、光照剧变和尺度大幅变化等复合挑战, 对跟踪器的长期稳定性和重检测能力提出了严苛要求. 不同跟踪器在此数据集上的对比结果见表 2 所示, LoMTrack-384 与 LoMTrack-256 在该数据集上取得了优良的性能, 其 AUC 指标超越多种先进跟踪器. 这一领先优势再一次验证了 LoMTrack 所采用的 TR^2-MoE 架构的有效性, 不同的专用网络能够自适应地处理相应场景, 从而显著提升其在复杂环境下的判别能力与鲁棒性.

3.2.4 GOT-10K

GOT-10K[11] 作为大规模单目标跟踪数据集, 包

表 2 四种不同数据集的对比结果

方法	来源	LaSOT			LaSOText			TrackingNet			GOT-10K*		
		AUC	P_{Norm}	P	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.5}$	$SR_{0.75}$
MixFormer-22k[38]	CVPR22	69.2	78.7	74.7	–	–	–	83.1	88.1	81.6	70.7	80	67.8
OSTrack-256[1]	ECCV22	69.1	78.7	75.2	47.4	57.3	53.3	83.1	87.8	82	71	80.4	68.2
ROMTrack-256[39]	ICCV23	69.3	78.8	75.6	48.9	59.3	55	83.6	88.4	82.7	72.9	82.9	70.2
SeqTrack-B256[6]	CVPR23	69.9	79.7	76.3	49.5	60.8	56.3	83.3	88.3	82.2	74.7	84.7	71.8
EVPTrack-224[40]	AAAI24	70.4	80.9	77.2	48.7	59.5	55.1	83.5	88.3	–	73.3	83.6	70.7
DiffusionTrack-B256[41]	CVPR24	70.8	79.8	76.7	–	–	–	83.8	88.2	82.1	74.8	85.4	72.0
MPIT[42]	TCSVT25	69.4	78.4	–	–	–	–	–	–	–	73.3	82.7	68.2
SeqTrack-B256-MPT[43]	ICML25	70.1	80.4	76.7	50.8	62.4	58.0	83.6	88.8	82.6	–	–	–
MFDSTrack-OS[44]	TCSVT25	70.5	79.8	76.3	48.6	58.1	54.2	83.8	88.4	82.7	73.6	82.7	70.5
TBIMT-256[45]	KBS25	70.5	80.2	76.8	48.0	57.9	53.5	83.6	88.3	82.6	74.4	84.5	71.7
RIDTrack[46]	KBS25	69.7	79.4	75.0	49.5	60.3	56.1	**84.4**	**89.1**	83.5	74.8	85.0	72.1
LoMTrack-256	Ours	**71.2**	**81.5**	**78.3**	**51.8**	**63.1**	**59.3**	84.2	88.9	**83.6**	**75.7**	**85.8**	**73.3**
一些分辨率更高的追踪器													
OSTrack-384[1]	ECCV22	71.1	81.1	77.6	50.5	61.3	57.6	83.9	88.5	83.2	73.7	83.2	70.8
ROMTrack-384[39]	ICCV23	71.4	81.4	78.2	51.3	62.4	58.6	84.1	89	83.7	74.2	84.3	72.4
SeqTrack-B384[6]	CVPR23	71.5	81.1	77.8	50.5	61.6	57.5	83.9	88.8	83.6	74.5	84.3	71.4
OneTracker-384[47]	CVPR24	70.5	79.9	76.5	–	–	–	83.7	88.4	82.7	–	–	–
STCFormer[48]	AAAI24	71.5	81.5	78.0	52.0	63.0	59.6	–	–	–	74.3	84.2	72.6
TBTrack[49]	TCSVT24	72.5	81.9	77.9	52.6	63.9	60.3	84.0	88.3	82.9	75.9	85.8	74.4
MGTrack-Small[50]	TCSVT25	70.2	79.5	76.5	49.6	60.4	56.7	83.6	88.3	82.5	74.7	84.6	72.3
OSTrack-B384-MPT[43]	ICML25	70.7	80.5	79.2	52.8	64.0	60.7	84.6	89.1	83.8	–	–	–
MGTrack-Base[50]	TCSVT25	71.3	81.0	77.6	50.2	61.1	58.2	**84.7**	**89.7**	**84.6**	76.2	85.4	73.6
TBIMT-384[45]	KBS25	71.8	**83.1**	78.9	52.5	63.5	60.1	84.2	89.1	83.9	75.3	85.7	72.9
LoMTrack-384	Ours	**72.6**	82.9	**80.5**	**53.8**	**65.4**	**61.9**	**84.7**	89.4	84.4	**77.4**	**87.0**	**75.8**

表 3 在 UAV123[31] 和 TNL2K[33] 数据集的对比结果

数据集	SiamFC	ECO	SiamRPN++	TransT	OSTrack	SeqTrack	ARTrack	F-BDMTrack	EVPTrack	AQATrack	TemTrack	LoMTrack
UAV123	46.8	53.5	61	69.1	68.3	69.2	67.7	69	70.2	70.7	**70.8**	**70.8**
TNL2K	29.5	32.6	41.3	50.7	54.3	54.9	57.5	56.4	57.5	57.8	58.8	**59.7**

含超 10,000 个视频序列, 其基准测试制定了独特协议: 跟踪器仅允许使用训练集进行训练, 且训练集与测试集的目标类别不存在重叠. 本文严格遵循这一协议训练框架, 并将结果提交至官方评估服务器完成性能评估. 不同跟踪器在此数据集上的对比结果见表 2, LoMTrack-384 和 LoMTrack-256 在关键指标 AO 上均取得了优于当前流行算法的成绩, 展现出显著的性能优势. 这一成果有力验证了所提框架在面对未知类别物体时的强大跟踪能力.

3.2.5 UAV123

UAV123[31] 是面向无人机视角单目标跟踪任务的专用基准数据集, 涵盖 123 个视频样本. 该数据集聚焦低空航拍场景, 包含目标尺度多变、背景复杂等无人机跟踪特有的挑战. 不同跟踪器在此数据集上的对比结果见表 3, LoMTrack 在此基准上达到了当前最先进的性能水平, 充分验证了 LoMTrack 在应对航拍视角复杂条件与小目标跟踪任务时的高效性与强鲁棒性.

3.2.6 TNL2K

TNL2K[33] 是专门面向单目标跟踪任务的测试基准数据集, 包含 700 个视频样本. 不同跟踪器在此数据集上的对比结果见表 3, LoMTrack 在此基准上达到最优性能, 这主要归因于 TR^2-MoE 的稀疏激活与路由正则化设计, 使 LoMTrack 在未知类别与新颖场景下仍能维持稳定的泛化能力.

3.2.7 OTB100

OTB100[13] 是单目标跟踪领域极具影响力的经典基准数据集, 包含 100 个视频序列. 该数据集系统覆盖了光照突变、尺度缩放、运动模糊等多种典型跟踪挑战场景. 不同跟踪器在此数据集上的对比结果见表 4, LoMTrack 以 70.8% 的 AUC 得分位居第一, 同时其在 P_{Norm} 与 P 指标上亦实现双重领先. 该优异表现得益于 LoRA 以低秩适配保留主干通用特征, TR^2-MoE 由专家分工实现场景差异化适配. 二者配

合使得专家精准适配又避过拟合, 从机制层面支撑 LoMTrack 在 OTB100 指标领先.

表 4 OTB100 数据集算法对比

#	Tracker	AUC	P_{Norm}	P
1	DiMP	68.4	84.2	89.4
2	STARK	68.1	84.2	88.7
3	ToMP	70.0	85.9	90.8
4	OSTrack	69.4	84.2	91.0
5	SuperDiMP	70.1	85.1	90.8
6	GRM-L30	66.2	79.7	86.8
7	LiteTrack-B9	67.7	81.5	88.2
8	LoMTrack	**70.8**	**86.9**	**92.5**

3.3 消融实验

为了展示所提方法的有效性, 本节从专家数量以及 top-k, 负载均衡机制, LoRA 应用范围, 不同模块冻结等多个方面进行消融实验.

3.3.1 专家数量以及 top-k

为确定 TR²-MoE 中专家数量与激活专家数的最优配置, 本文在精度与推理速度之间进行权衡实验. 在 TR^2-MoE 中, 专家数量与激活专家数量是影响检测精度和推理速度的关键设计参数: 为在二者之间寻求更具实用价值的平衡, 本文将测试范围限定为最多 4 个专家和最多激活 2 个专家, 并系统测试多种参数组合下的精度表现. 具体实验结果见表 5, 随着专家数量与激活专家数量的增加, 检测精度呈现逐步上升的趋势, 但这一过程也会伴随推理速度的下降, 这体现精度与速度的权衡特性, 而本文所限定的测试范围, 正是为了在精度提升的边际效益与速度损耗之间, 锁定更具实用价值的参数区间.

表 5 专家数量与 top-k 数量分析

#	专家	top-k	AO	$SR_{0.5}$	$SR_{0.75}$
1	2	1	74.2	84.2	72.6
2	4	1	74.7	84.5	73.1
3	4	2	**75.7**	**85.8**	**73.3**

3.3.2 解码器交替结构

为验证解码器中 DSSMamba 与交叉注意力非对称设计的必要性, 本文在 LaSOT 上对比五种层数组合. 不同 DSSMamba 与交叉注意力层数组合在 LaSOT 上的消融结果如表 6 所示. 全 DSSMamba 方案 AUC 仅 70.1%, 说明纯状态空间建模缺乏充分的模板-搜索交互, 定位精度受限. 全交叉注意力方案精度最低, 表明全局注意力在长序列上存在注意力弥散的问题, 损害长程建模能力. 随着 DSSMamba 层数减少, 性能先升后降, 在两层 DSSMamba 与两层交叉注意力时 AUC 达到最优的 71.2%, 同时保持 31.9 FPS 的推理速度. 这表明浅层 DSSMamba 以线性复杂度高效聚合全局上下文, 深层交叉注意力专注于精确匹配, 二者分工协作, 在长程建模能力与计算效率之间取得了最佳平衡.

表 6 解码器交替结构的消融实验

DSSMamba	交叉注意力	AUC	P_{Norm}	P	FPS
4	0	70.1	80.2	76.9	25.7
3	1	70.7	80.9	77.7	29.6
2	2	**71.2**	**81.5**	**78.3**	31.9
1	3	70.3	80.4	77.2	33.9
0	4	69.9	80.3	76.9	36.2

3.3.3 验证三重正则化机制

为验证 TR^2-MoE 中三重正则化机制的各自贡献, 本文在 GOT-10k 数据集上进行消融实验, 结果见表 7. 完整配置取得 75.7% 的 AO, 三项指标均为最优. 移除路由扰动后, AO 降至 74.0%, 表明路由扰动有助于增强门控泛化性、避免对单一专家组合的过度依赖. 移除隐空间正则化后, AO 下降至 73.7%, 验证了隐层正则化对抑制专家间过度共适应的关键作用. 移除负载均衡损失后, AO 降至 73.9%, 说明均衡的专家利用率对维持训练稳定性不可或缺. 上述结果表明, 三种正则化机制分别从路由泛化、专家独立性与负载均衡性三个维度互补协同, 共同保障了 TR^2-MoE 的动态路由能力.

表 7 TR^2-MoE 模块中正则化机制的消融实验

路由扰动	隐空间正则化	负载均衡	AO	$SR_{0.5}$	$SR_{0.75}$
√	√	√	**75.7**	**85.8**	**73.3**
×	√	√	74.0	84.3	71.9
√	×	√	73.7	84.3	72.2
√	√	×	73.9	84.1	72.1

3.4 可视化与定性分析

3.4.1 不同属性场景下的 AUC 性能评估

本节对不同场景属性下的跟踪性能进行细致评估, 进一步结果如图 6 所示. 该雷达图直观呈现了 LoMTrack 与当前主流先进跟踪器在该数据集各属性场景下的 AUC 性能对比. 在相机运动、光照波动、目标形变、低分辨率、运动模糊、目标暂离视野、快速运动等多类高挑战场景中, LoMTrack 均展现出显著优势, 其性能明显优于其它对比算法. LoMTrack 在多项挑战性场景中的卓越性能, 不仅体现了其对复杂动态环境的理解能力, 更彰显在真实多变场景中保持稳定跟踪的强适应性与鲁棒性.

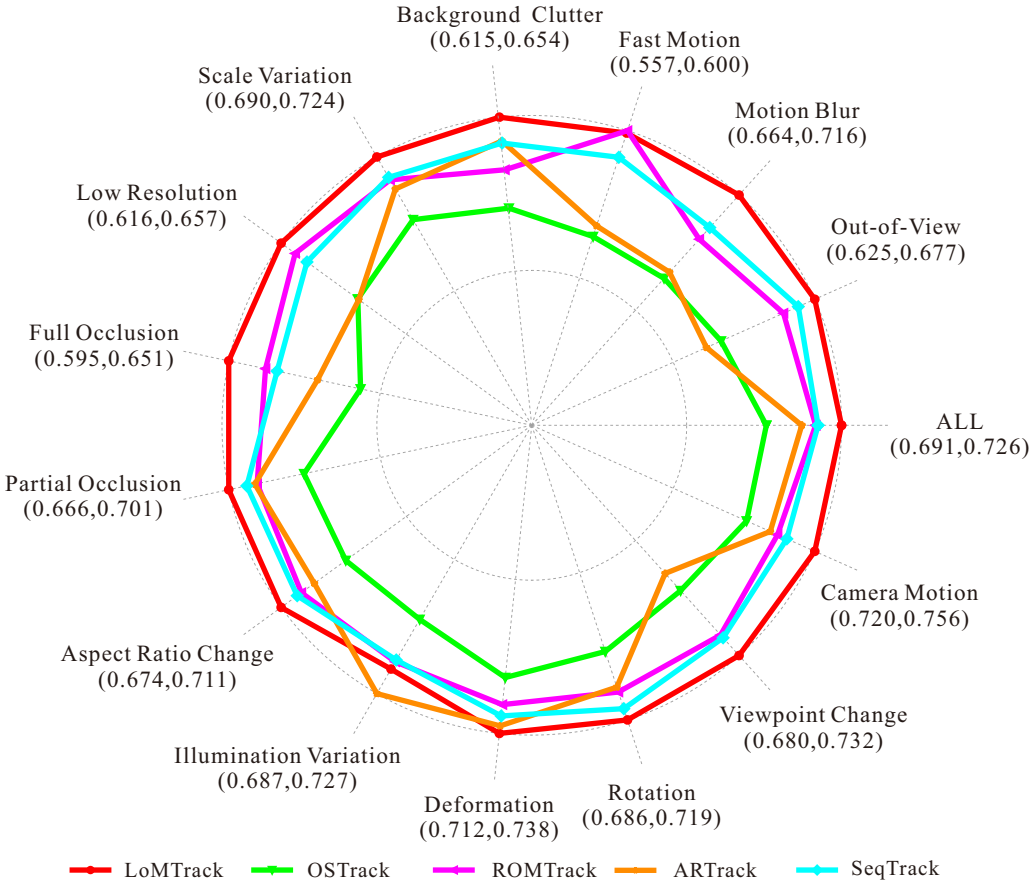


图 6　LaSOT[32] 数据集上不同属性的 AUC 评分

3.4.2　LaSOT 数据集可视化展示

在 LaSOT 长时目标跟踪各类典型挑战场景的对比结果如图 7 所示, LoMTrack 展现出了相较其它算法的全面优越性. 无论面对相似外观带来的干扰、复杂背景造成的杂乱影响、目标形态的动态形变, 还是各类额外元素的干扰, LoMTrack 的跟踪框都能始终精准且稳定地贴合目标主体, 既能适配目标的状态变化, 也能有效抵御外界干扰因素的影响. 而其它算法在这些场景下, 常出现框位偏移、目标包裹不完整或是受干扰发生漂移的情况.

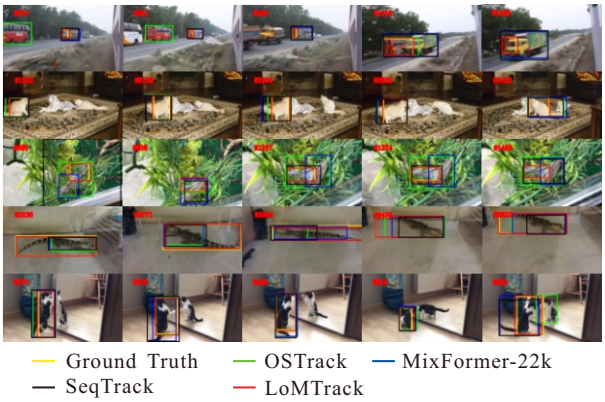


图 7　LaSOT[32] 测试集可视化对比结果

4　结　论

针对视觉目标跟踪任务中模型微调参数量大、易发生灾难性遗忘及场景适应性弱等挑战, 本文提出一种高效跟踪框架 LoMTrack. 在编码器端, 采用 LoRA 进行低秩自适应, 在冻结预训练权重、保留通用视觉知识的同时, 将可训练参数量降低数个数量级, 有效规避灾难性遗忘. 在解码器端, 设计了 TR^2-MoE 机制, 通过路由阶段的随机正则化、专家内部的结构化约束以及负载均衡的自适应归一化, 实现场景自适应的专家稀疏激活与稳定协作, 显著增强模型在多样化跟踪挑战下的泛化稳健性与训练鲁棒性. 在多个基准数据集上的广泛实验表明, LoMTrack 不仅在性能上超越多种先进方法, 更从实验上验证了其在抗遗忘能力与场景泛化性方面的综合优势, 证明了所提方案的有效性与先进性.

参考文献 (References)

[1]　Ye B T, Chang H, Ma B P, et al. Joint feature learning and relation modeling for tracking: A one-stream framework[C]. Computer Vision – ECCV 2022. Cham: Springer, 2022: 341-357.

[2]　Xu C L, Zhong B N, Liang Q H, et al. Less is more: Token context-aware learning for object tracking[PP/OL]. 2025, arXiv: abs/2501.00758.

[3]　Xie J X, Zhong B N, Mo Z Y, et al. Autoregressive queries for adaptive tracking with spatio-temporal transformers[C]. IEEE/CVF Conference on Computer

Vision and Pattern Recognition. Seattle, 2024: 19300-19309.

[4] Zheng Y Z, Zhong B N, Liang Q H, et al. ODTrack: Online dense temporal token learning for visual tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(7): 7588-7596.

[5] 牟琦, 张正义, 王雪倩, 等. 基于时空建模的多分支agent transformer目标跟踪算法[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2026.0024.
(Mou Q, Zhang Z Y, Wang X Q, et al. Multi-Branch agent transformer visual tracking with spatiotemporal modeling[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2026.0024.)

[6] Chen X, Peng H W, Wang D, et al. SeqTrack: Sequence to sequence learning for visual object tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 14572-14581.

[7] Cai W R, Liu Q J, Wang Y H. HIPTrack: Visual tracking with historical prompts[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 19258-19267.

[8] Li X H, Zhong B N, Liang Q H, et al. Mambalct: Boosting tracking via long-term context state space model[C]. Sponsored by the Association for the Advancement of Artificial Intelligence. Philadelphia, 2025: 4986–4994.

[9] 柳长源, 范培栋, 兰朝凤. 双模板跨模态交互与前景选择的高效RGB-T目标跟踪[J]. 控制与决策, 2025, 40(12): 3725-3733.
(Liu C Y, Fan P D, Lan C F. Efficient RGB-T object tracking network with dual-template crossmodality interaction and foreground selection[J]. Control and Decision, 2025, 40(12): 3725-3733.)

[10] Edward H J, Shen Y L, Phillip W, et al. Lora: Low-rank adaptation of large language models[PP/OL]. 2021, ArXiv: abs/2106.09685.

[11] Huang L H, Zhao X, Huang K Q. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562-1577.

[12] Fan H, Bai H X, Lin L T, et al. Lasot: A high-quality large-scale single object tracking benchmark[J]. International Journal of Computer Vision, 2021, 129: 2.

[13] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[14] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 4277-4286.

[15] Xu Y D, Wang Z Y, Li Z X, et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12549-12556.

[16] Luca B, Jack V, Joao F H, et al. Fully-convolutional siamese networks for object tracking[Z]. 2021.

[17] 陈志旺, 高一铭, 吕昌昊, 等. 基于空域抑制与轨迹关联优化的目标跟踪算法[J]. 控制与决策, 2026, 41(1): 165-174.
(Chen Z W, Gao Y M, (Lü/lv/lu/lyu) C H, et al. Object tracking algorithm based on spatial inhibition and trajectory correlation optimization[J]. Control and Decision, 2026, 41(1): 165-174.)

[18] 侯志强, 郭凡, 杨晓麟, 等. 基于混合注意力的Transformer视觉目标跟踪算法[J]. 控制与决策, 2024, 39(3): 739-748.
(Hou Z Q, Guo F, Yang X L, et al. Transformer visual object tracking algorithm based on mixed attention[J]. Control and Decision, 2024, 39(3): 739-748.)

[19] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C]. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2021: 3045-3059.

[20] Houlsby N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]. International Conference on Machine Learning. Long Beach, 2019: 2790-2799.

[21] Liu H T, Li C Y, Li Y H, et al. Improved baselines with visual instruction tuning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 26286-26296.

[22] Lin L T, Fan H, Zhang Z P, et al. Tracking meets LoRA: Faster training, larger model, stronger performance[C]. Computer Vision – ECCV 2024. Cham: Springer, 2025: 300-318.

[23] Liu Y X, Zhang K, Li Y, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models[PP/OL]. 2024, arXiv: abs/2402.17177.

[24] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79-87.

[25] David R, Sam R, Blake A R, et al. Mixture-of-depths: Dynamically allocating compute in transformer-based language models[PP/OL]. 2024, arXiv: abs/2404.02258.

[26] 李昱洁, 吴晗, 孟丹, 等. 面向开放世界持续学习的任务敏感提示驱动混合专家模型[J]. 软件学报, 2026, 37(4): 1531-1547.
(Li Y J, Wu H, Meng D, et al. Task-aware prompt-driven mixture-of-experts model for open-world continual learning[J]. Journal of Software, 2026, 37(4): 1531-1547.)

[27] Tang Z Y, Xu T Y, Wu X J, et al. Revisiting RGBT tracking benchmarks from the perspective of modality validity: A new benchmark, problem, and solution[J]. IEEE Transactions on Image Processing, 2025, 34: 7235-7249.

[28] Cai W R, Liu Q J, Wang Y H. SPMTrack: Spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern

Recognition. Nashville, 2025: 16871-16881.

[29] Zhang X S, Tian Y J, Xie L X, et al. HiViT: A simpler and more efficient design of hierarchical vision transformer[C]. International Conference on Learning Representations. Online, 2023.

[30] Rezatofighi H, Tsoi N, Gwak J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 658-666.

[31] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]. Computer Vision – ECCV 2016. Cham: Springer, 2016: 445-461.

[32] Fan H, Lin L T, Yang F, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 5369-5378.

[33] Wang X, Shu X J, Zhang Z P, et al. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark[C]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, 2021: 13763-13773.

[34] Muller M, Bibi A, Giancola S, et al. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild[C]. *Proceedings of the European Conference on Computer Vision*. Munich, 2018: 300-317.

[35] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]. *Computer Vision – ECCV 2014: 13th European Conference*. Zurich, 2014: 740-755.

[36] Huang K, Chu J, Leng L, et al. TATrack: Target-aware transformer for object tracking[J]. Engineering Applications of Artificial Intelligence, 2024, 127: 107304.

[37] Wei X, Bai Y F, Zheng Y C, et al. Autoregressive visual tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, 2023: 9697-9706.

[38] Cui Y T, Jiang C, Wu G S, et al. MixFormer: End-to-end tracking with iterative mixed attention[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4129-4146.

[39] Cai Y D, Liu J, Tang J, et al. Robust object modeling for visual tracking[C]. IEEE/CVF International Conference on Computer Vision. Paris, 2024: 9555-9566.

[40] Shi L T, Zhong B N, Liang Q H, et al. Explicit visual prompts for visual object tracking[C]. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*. Vancouver, 2024: 4838-4846.

[41] Xie F, Wang Z D, Ma C. DiffusionTrack: Point set diffusion model for visual object tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 19113-19124.

[42] Wang W W, Lv M B, Zhu L, et al. Siamese visual tracking with multi-parallel interactive transformers[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(10): 10353-10366.

[43] Zhao J, Chen X, Yuan Y S, et al. Efficient motion prompt learning for robust visual tracking[PP/OL]. 2025, arXiv: abs/2505.16321.

[44] Han Y M, Cai M Y, Wu J, et al. Visual object tracking with multi-frame distractor suppression[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(3): 2556-2569.

[45] Zhang H L, Fu W Q, Qi R, et al. Target–background interaction modeling transformer for object tracking[J]. Knowledge-Based Systems, 2025, 315: 113230.

[46] Liu N, Zhang Y. A transformer based visual tracker with restricted token interaction and knowledge distillation[J]. Knowledge-Based Systems, 2025, 307: 112736.

[47] Hong L Y, Yan S L, Zhang R R, et al. OneTracker: Unifying visual object tracking with foundation models and efficient tuning[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2024: 19079-19091.

[48] Hu K, Yang W J, Huang W R, et al. Sequential fusion based multi-granularity consistency for space-time transformer tracking[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(11): 12519-12527.

[49] Li N, Zhong B N, Zheng Y Z, et al. Robust tracking *via* combing top-down and bottom-up attention[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 9774-9785.

[50] Xiong J B, Ling Q. Mask-guided Siamese tracking with a frequency-spatial hybrid network[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(1): 103-117.

作者简介

刘大千 (1992−), 男, 教授, 博士, 主要研究方向为智能无人系统、跨域无人平台组织协同, E-mail: liudaqian@lntu.edu.cn;

张嘉豪 (2000−), 男, 硕士生, 主要研究方向为图像处理、目标跟踪, E-mail: jiahaozhang0290@gmail.com;

费博雯 (1991−), 女, 副教授, 博士, 主要研究方向为无人系统资源组织优化与协同, E-mail: feibowen@lntu.edu.cn.