

DOI:10.19651/j.cnki.emt.2519250

基于监督反馈和 Transformer 的目标跟踪算法^{*}符 强^{1,2} 殷奇晨^{1,2} 纪元法^{1,2,3} 任风华^{1,2}(1. 桂林电子科技大学信息与通信学院 桂林 541000; 2. 桂林电子科技大学广西精密导航技术与应用重点实验室 桂林 541000;
3. 时空信息与智能位置服务国际合作联合实验室 桂林 541000)

摘 要: 针对传统孪生网络在复杂场景下鲁棒性不足以及 Transformer 架构对计算资源依赖性强等问题,提出了一种基于监督反馈和 Transformer 的目标跟踪算法。首先,设计了监督反馈模块,在特征提取过程中引入任务相关的反馈信息,引导网络更加聚焦于目标区域,从而提升特征判别性并抑制背景干扰;其次,构建了轻量化的 Transformer 结构,在保持全局建模能力的基础上,有效降低计算复杂度和参数量,实现性能与效率的良好平衡;最后,提出自适应模板更新机制,结合当前帧的状态信息与场景变化,动态调整模板内容以应对目标外观变化,降低跟踪漂移风险。在多个主流公开数据集上的实验结果表明,所提出方法在鲁棒性和实时性方面均优于现有先进算法。

关键词: 目标跟踪;监督反馈;Transformer;模板更新

中图分类号: TN391.41 **文献标识码:** A **国家标准学科分类代码:** 520.604

Object tracking algorithm based on supervised feedback and Transformer

Fu Qiang^{1,2} Yin Qichen^{1,2} Ji Yuanfa^{1,2,3} Ren Fenghua^{1,2}

(1. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541000, China;

2. Guangxi Key Laboratory of Precision Navigation Technology and Application, Guilin University of Electronic Technology, Guilin 541000, China; 3. International Joint Laboratory of Spatiotemporal Information and Intelligent Location Services, Guilin 541000, China)

Abstract: To address the limited robustness of traditional Siamese networks in complex scenarios and the high computational demands of Transformer-based architectures, this paper proposes a novel object tracking algorithm based on supervised feedback and a lightweight Transformer. First, a supervised feedback module is designed to incorporate task-relevant feedback during feature extraction, guiding the network to focus more effectively on target regions, thereby enhancing feature discriminability and suppressing background interference. Second, a lightweight Transformer structure is constructed, which maintains strong global modeling capabilities while significantly reducing computational complexity and parameter overhead, achieving a favorable balance between performance and efficiency. Finally, an adaptive template update mechanism is introduced to dynamically adjust the template content based on the current frame's object state and scene variations, improving adaptability to target appearance changes and mitigating tracking drift. Experimental results on multiple mainstream public datasets show that the proposed method outperforms existing advanced algorithms in terms of both robustness and real-time performance.

Keywords: target tracking; supervised feedback; Transformer; template update

0 引 言

单目标跟踪是计算机视觉领域中一项具有基础性和实用性的重要研究方向,广泛应用于公共安全监控、自动驾驶系统、智能机器人导航以及人机交互等多个实际场景^[1-4]。

其核心任务是在给定视频序列的初始帧中手动指定目标位置后,利用跟踪算法在后续帧中持续预测目标的空间位置,从而实现目标的连续跟踪与定位。然而,单目标跟踪面临诸多挑战,如目标类别的任意性、外观的动态变化,以及实际环境中普遍存在的运动模糊、复杂背景、部分或完全遮

收稿日期:2025-07-01

^{*} 基金项目:国家自然科学基金(U23A20280,62161007,62061010)、广西科技厅项目(AD22080061,AA23062038,AB23026120)、广西精密导航与应用重点实验室基金(DH202308)项目资助

挡、光照变化等干扰因素,这些都显著增加了精确定位目标的难度,对算法的鲁棒性与适应性提出了更高要求。

近年来,随着深度学习的发展,基于孪生网络的跟踪算法成为研究热点。其中,SiamFC^[5]作为代表性模型,首次将孪生网络引入跟踪任务,通过简单的模板与搜索区域特征互相关实现高效匹配,具备良好的实时性。然而,SiamFC本身存在明显局限:其采用单一的初始模板且缺乏更新机制,难以适应目标在长时跟踪中的外观变化;特征提取网络较浅,表达能力有限;缺乏对全局上下文和尺度变化的建模能力,导致在复杂场景下易发生跟踪漂移。为克服上述缺陷,研究者提出了一系列基于 SiamFC 的改进方法。SiamRPN^[6]引入区域提议网络,增强了对目标尺度和形变的适应能力;SiamFFN^[7]融合尺度注意力机制以提升多尺度特征建模能力;SiamAttn^[8]通过空间与通道注意力增强特征表达能力;UnifiedTT^[9]联合卷积与自注意力机制,实现局部与全局特征的融合;SiamDMU^[10]则设计了动态模板更新策略,缓解了静态模板的退化问题。这些方法虽在一定程度上提升了 SiamFC 的性能,但其改进多集中于局部结构或特定模块,在全局依赖建模与动态适应能力方面仍有欠缺,且多数未系统性地解决模型复杂性与实际部署需求之间的平衡问题。

与此同时,Transformer 架构因其强大的全局建模能力被引入视觉跟踪任务,为目标外观建模与长程依赖捕获提供了新思路。AMTT^[11]提出了多尺度特征编码器与锚点驱动的解码结构,有效提升了跨尺度特征建模能力;DPT-Tracker^[12]构建双池化 Transformer 架构并融合时空注意力机制,强化了对动态场景中目标的建模效果;TransT^[13]设计了交叉相关层以提升特征交互效率;AtptTrack^[14]引入先验模板与多尺度嵌入机制;SSTrack^[15]融合 Swin Transformer^[16]与空间尺度注意力,显著提升了多尺度建模精度。然而,标准 Transformer 的计算复杂度随序列长度呈二次增长,对计算资源要求高,限制了其在资源受限场景下的应用。

针对传统 SiamFC 模型在复杂场景下鲁棒性不足,以

及典型 Transformer 跟踪器对算力依赖性强的问题,提出了 3 项创新改进策略,旨在兼顾性能与效率。首先,设计了监督反馈模块,通过在特征提取过程中引入任务相关的反馈机制,增强对目标区域的关注并抑制背景干扰。与 SiamAttn 等仅依赖前向传播的注意力方法不同,本模块通过反馈路径显式优化特征表示,提升判别能力。其次,构建了轻量化 Transformer 结构,在保持全局建模能力的同时显著降低计算复杂度,使其更适合边缘部署。相较于 TrSiam 等直接采用标准 Transformer 架构的方法,本研究结构更注重效率与性能的平衡。最后,提出自适应模板更新机制,根据目标状态与干扰情况动态调整模板,提升对目标外观变化的适应性。

在实验设计上,选取了 SiamFC、DaSiamRPN、SiamBAN 等代表性非 Transformer 方法作为对比基准,以验证本研究模块在传统 Siamese 框架下的有效性;同时,也对比了 PrDiMP、TrSiam 等使用 Transformer 架构的跟踪器,以说明轻量化设计在性能与效率上的综合优势。通过系统性的实验分析,验证了本研究方法在提升跟踪鲁棒性、适应性与可部署性方面的先进性。

1 本研究算法

本研究所提的基于监督反馈和 Transformer 的目标跟踪算法由特征提取网络 ResNet-50、监督反馈模块、轻量化 Transformer、预测头和自适应模板更新机制组成,算法整体框图如图 1 所示。在特征提取阶段,利用预训练的 ResNet-50 网络对输入的搜索图像与模板图像进行高维特征编码,提取出具有判别性的表示。引入监督反馈模块以增强特征之间的语义联系与任务相关性,从而提升对目标的感知能力。接着,轻量化 Transformer 结构被用于进一步建模搜索区域与模板之间的长距离依赖关系,并增强时序上下文建模能力,提升跟踪的稳定性与鲁棒性。在预测阶段,预测头负责输出目标的位置信息,而自适应模板更新机制则用于动态调整模板特征,以应对目标在外观、尺度及姿态等方面的变化。

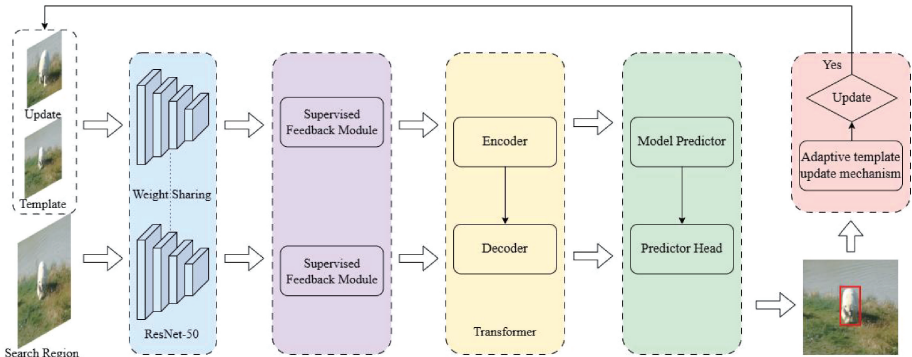


图 1 系统框图

Fig. 1 System chart

1.1 监督反馈模块

在基于深层神经网络的目标跟踪算法中,主干特征提取网络通常仅依赖于单一监督信号(如目标分类与边界框回归)进行训练。然而,这种单路径监督策略在处理复杂背景、遮挡干扰、尺度变化等真实场景时,容易使模型陷入局部最优,导致所提取特征的判别性与鲁棒性不足,进而影响整体的跟踪精度与跨场景泛化能力。为缓解上述问题,引入了一个监督反馈模块,作为主任务之外的辅助监督路径,为特征学习过程提供额外的训练信号。该模块通过引导网络从全局语义层面学习更具判别性和稳定性的特征表示,有效正则化主干网络的训练过程,从而提升特征的鲁棒性与泛化能力,进一步间接增强主任务的性能表现。

如图 2 所示,结构上首先采用一个 3×3 的卷积层对输入的高维特征图进行局部上下文建模,提取空间结构和局部上下文关系,并在不改变通道数的情况下为后续处理提供更紧凑的特征表示。然后使用批标准化层用于规范化特征分布,缓解梯度不稳定问题,加速网络收敛。并且使用 ReLU 激活函数为网络提供非线性的建模能力,表达式为:

$$F = \text{ReLU}(\text{BN}(\text{Conv}(f))) \quad (1)$$

其中, f 为输入特征图,为了进一步压缩空间维度并突出全局信息,采用自适应平均池化将特征图聚合为 1×1 的全局向量,从而简化分类决策过程。最后通过一个全连接层和 Softmax 激活函数,将特征映射为源域和目标域的二分类概率分布,形成辅助监督分支。

$$Y_i = \text{Softmax}(\text{FC}(\text{Pool}(F))) \quad (2)$$

其中, Y_i 表示第 i 个样本在辅助监督分支上的预测结果,可为模板或搜索区域的输出特征。整个结构不仅提升了特征的判别性和稳定性,还确保在训练阶段对主干网络起到正则化作用,而不会在推理阶段引入额外计算开销。

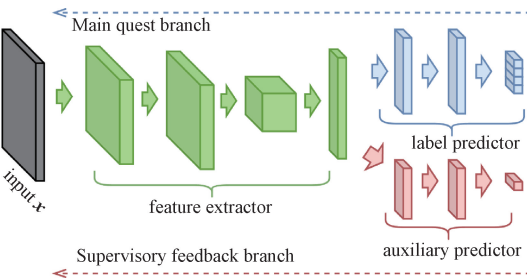


图 2 监督反馈结构

Fig. 2 Supervisory feedback structure

为使监督输出层能够学习到更具判别性的高层表示,对该辅助监督分支输出添加交叉熵损失:

$$L_{\text{aux}} = - \sum_i y_i \log(Y_{\text{aux},i}) \quad (3)$$

其中, $y_i \in \{0,1\}$ 表示输入样本在辅助分类任务中的标签。需要说明的是,该辅助任务并不直接参与最终的目

标预测输出,而是作为特征级正则化项存在,引导 backbone 在训练阶段学习到更具稳定性和鲁棒性的特征,从而间接提升主任务性能。

主任务本身的损失函数包括目标分类损失和边界框回归损失,目标分类损失采用交叉熵损失度量跟踪目标类别的预测与真实标签之间的差异,表达式为:

$$L_{\text{cls}} = - \sum_i y_{\text{cls},i} \log(Y_{\text{cls},i}) \quad (4)$$

边界框回归损失通过平滑 L1 损失函数衡量预测框和真实框之间的误差。

$$L_{\text{bbox}} = - \sum_i \text{SmoothL1}(b_i - \hat{b}_i) \quad (5)$$

最终,整个网络的总损失函数表达式为:

$$L_{\text{total}} = L_{\text{cls}} + L_{\text{bbox}} + \lambda_1 L_{\text{aux_search}} + \lambda_2 L_{\text{aux_temp}} \quad (6)$$

训练阶段,模型接收模板图像和搜索图像的输入,经过骨干网络和监督反馈模块处理后,分别输出主任务结果与辅助任务监督。整体网络通过联合损失进行优化,从而实现主任务精度与特征鲁棒性的双重提升。

在推理阶段,辅助模块不再参与前向计算,保证推理效率不受影响,模型只基于融合特征输出最终的分类结果与边界框预测。

1.2 轻量化 Transformer

近年来,Transformer 架构凭借其卓越的全局建模能力,已在自然语言处理和计算机视觉等领域取得了突破性进展。然而,标准的 Transformer 结构存在计算复杂度高、参数量大的问题,特别是在处理高分辨率视觉特征时,其计算和存储开销成为主要瓶颈。为了解决上述问题,提出了一种轻量化的 Transformer 架构,整体包括编码器和解码器两个部分。

编码器由 N 个堆叠的编码器层构成,每层通常包括一个多头自注意力模块和一个前馈网络。为减轻计算负担,本研究中省略了前馈模块,仅保留多头自注意力部分,这一设计不仅显著降低计算复杂度,也有助于保持空间结构信息完整性。其结构如图 3 所示。设输入的原始特征张量为 $\mathbf{X} \in \mathbf{R}^{B \times C \times H \times W}$,其中, B 为批次大小, C 为通道数, H 和 W 分别为图像的高度和宽度。为满足 Transformer 对序列输入的要求,首先将输入张量按空间维度展平,转化为一维序列形式:

$$\mathbf{X}_{\text{flat}} = \text{Reshape}(\mathbf{X}) \in \mathbf{R}^{L \times B \times C} \quad (7)$$

为增强模型的空间建模能力,引入相对位置编码,将位置信息直接融入注意力权重中,从而提升了模型对局部结构和位置信息的敏感度。将其与序列特征相加输出:

$$\mathbf{X}_{\text{seq}} = \mathbf{X}_{\text{flat}} + \mathbf{P} \quad (8)$$

对于每个编码器层,输入特征首先进入多头自注意力模块。多头注意力机制将输入分别映射为查询(Query)、键(Key)和值(Value)。

$$\text{Attention}(\mathbf{T}) = \text{Softmax}\left(\frac{\mathbf{QK}^T}{\sqrt{d}}\right)\mathbf{V} \quad (9)$$

其中, $\sqrt{d}^{-1}$ 是缩放因子, 避免内积的值过大。为了避免梯度消失问题并加速收敛, 在每个多头自注意力机制之后加入残差连接。为了保持图像的空间结构并减少空间信息损失, 引入了 L2Norm 归一化操作, 该操作通过计算每个实例的 L2 范数对每个通道进行归一化, 使得每个像素的特征幅度得到标准化。

$$\mathbf{Y} = \text{L2Norm}(\mathbf{X}_{\text{seq}} + \text{Attention}(\mathbf{T})) \quad (10)$$

最后通过 reshape 输出表达式为:

$$\mathbf{Y}_{\text{out}} = \text{Reshape}^{-1}(\mathbf{Y}) \in \mathbf{R}^{B \times C \times H \times W} \quad (11)$$

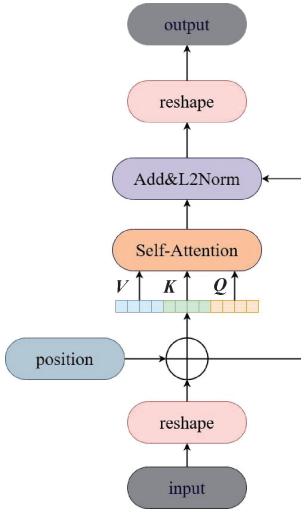


图 3 编码器结构

Fig. 3 Encoder structure

解码器由 M 个堆叠的解码器层构成, 每层包含一个稀疏多头自注意力模块和一个交叉多头自注意力模块。该结构的目的是结合编码器输出信息, 实现更有效的特征重构与目标定位, 其结构如图 4 所示。设输入的原始特征张量为 $\mathbf{X}' \in \mathbf{R}^{B \times C \times H \times W}$, 对输入张量展平并加入相对位置编码。

$$\mathbf{X}'_{\text{seq}} = \text{Reshape}(\mathbf{X}') + \mathbf{P} \quad (12)$$

然后引入稀疏多头自注意力模块, 该模块与标准多头注意力类似, 但在注意力计算中引入了稀疏性限制, 从而降低计算复杂度并提升局部建模能力。其注意力机制表达式为:

$$\mathbf{S} - \text{Attention}(\mathbf{T}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d^r}} + \mathbf{M}\right)\mathbf{V} \quad (13)$$

其中, $\mathbf{Q} = \mathbf{X}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}^K$, $\mathbf{V} = \mathbf{X}\mathbf{W}^V$, $\mathbf{M}$ 为稀疏掩码矩阵。模块输出与输入进行残差连接后, 采用 L2 归一化操作以保持数值稳定性。

$$\mathbf{Y}' = \text{L2Norm}(\mathbf{X}'_{\text{seq}} + \mathbf{S} - \text{Attention}(\mathbf{T})) \quad (14)$$

接下来, 输出特征 $\mathbf{Y}'$ 将作为查询输入交叉注意力模块。与此同时, 来自编码器输出的序列特征作为键和值参与交叉注意力计算。该模块用于引入编码器的全局上下文信息, 指导目标位置的精确回归或分类。

$$\mathbf{C} - \text{Attention}(\mathbf{T}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d^r}}\right)\mathbf{V} \quad (15)$$

最后同样采用残差连接并进行 L2 归一化, 得到该层解码器的输出:

$$\mathbf{Y}'' = \text{L2Norm}(\mathbf{Y}' + \mathbf{C} - \text{Attention}(\mathbf{T})) \quad (16)$$

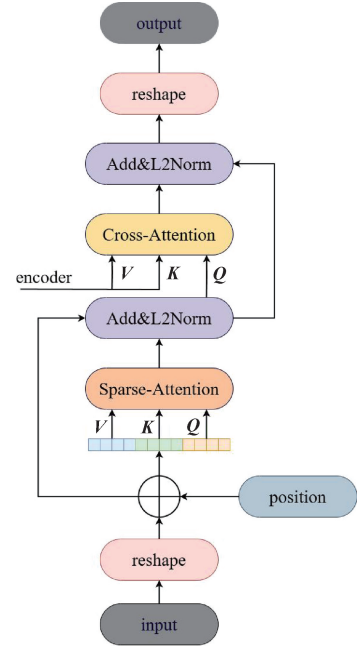


图 4 解码器结构

Fig. 4 Decoder structure

上述过程在每一层解码器中重复进行, 输出的序列特征可在后续模块中进一步重构为二维图像格式。通过上述轻量化设计, 本研究所提出的 Transformer 结构在保证建模能力的同时, 显著降低了参数量与计算成本, 为高效视觉建模任务提供了一种有效的解决方案。

1.3 自适应模版更新机制

目标的外观信息经常受到视角变化、尺度变化、部分遮挡及光照波动等因素影响, 导致初始模板无法长期有效描述目标特征。如果模板更新不及时, 可能引发跟踪漂移; 而更新不当, 则可能造成模板污染。提出了一种高效且鲁棒的自适应模板更新策略。该策略以 IoU 预测模块对比为核心, 通过严格的更新决策规则与加权融合机制, 实现目标模板的动态优化。与传统基于时间或固定阈值的更新方法相比, 本策略更具适应性与稳定性。

为了量化当前帧与模板的匹配程度, 设计了一种轻量的 IoU 预测模块, 该模块通过学习目标的空间与外观关系, 有效预测当前检测框与历史模板之间的 IoU 分数。结合了当前帧的图像特征与前一帧预测结果的位置结构信息。该模块通过多模态融合与 MLP 回归网络, 有效捕捉目标外观与空间布局之间的关系, 实现精确而稳定的 IoU 估计, 具体表达式为:

$$\begin{cases} \mathbf{Z}_t = \text{concat}(\mathbf{F}_t, T_{\text{old}}) \\ IoU_{\text{old}} = \text{MLP}(\mathbf{Z}_t; \mathbf{F}_{t-1}) \\ IoU_{\text{new}} = \text{MLP}(\mathbf{Z}_t; \mathbf{F}_t) \end{cases} \quad (17)$$

其中, $\mathbf{F}_t, \mathbf{F}_{t-1}$ 分别表示当前帧候选区域的特征和上一帧候选区域的特征, T_{old} 表示上一帧候选区域的归一化特征编码, 该向量融合了当前帧目标外观与其历史空间动态信息, 为 IoU 估计提供上下文支撑。

根据预测的两组 IoU 值, 构造差值指标, 表达式为:

$$\Delta IoU = IoU_{\text{new}} - IoU_{\text{old}} \quad (18)$$

当该值大于设定阈值 δ , 说明当前帧特征比旧模板更加准确和贴近目标状态, 可考虑进入模板更新流程。该 IoU 预测机制通过引入上一帧的空间信息, 实现了外观特征与目标动态的联合建模, 使得模板更新策略更具时序一致性与预测稳定性。

2 实验结果与分析

2.1 实验细节

实验采用 PyTorch 1.10.0 深度学习框架, 在 Ubuntu 20.04 操作系统上运行。模型训练编程语言为 Python3.7, 模型测试编程语言使用 Python 和 Matlab。硬件配置包括一块 NVIDIA GeForce RTX 4070 GPU, 搭配 12 GB 显存, 32 GB 内存, 以及一个 Intel Core i7-12700KF 处理器。

整个网络使用 GOT-10k、LaSOT 和 CO-CO 数据集进行离线训练。训练过程中采用指数衰减的学习率策略, 初始学习率设定为 0.01, 每 15 个 epoch 衰减一次, 衰减系数为 0.2。训练轮数共计 50 个 epoch, batch_size 设为 16, num_workers 设置为 8。

2.2 评估指标

为了全面评估所提出目标跟踪算法的性能表现, 本研究在多个主流基准数据集上采用各具特色的评价指标体系, 涵盖了空间精度、区域重合度以及模型的泛化能力等多维度的性能考量。

在 OTB50 与 OTB100 两个经典基准数据集上, 主要采用精度与成功率作为评估标准。精度指标通过计算每帧中预测边界框中心与真实目标中心之间的欧氏距离, 并统计其中小于设定阈值的帧所占比例, 从而反映跟踪器的空间定位准确性。成功率则基于预测框与真实边界框之间的交并比 (intersection over union, IoU) 计算而得, 即 IoU 被定义为预测框与真实框交集面积与并集面积之比, 通过绘制不同 IoU 阈值下的成功率曲线并计算其曲线下面积 AUC, 评估跟踪器的整体区域覆盖能力。

在 GOT-10k 数据集中, 为避免评估结果对特定类别产生偏倚, 该数据集引入了类间零重叠的 “one-shot” 评估策略, 即保证训练与测试类别之间完全不重复, 从而强化对跟踪算法泛化能力的考察。该数据集采用的核心评估指标包括平均重叠率 (average overlap, AO) 和成功率 (success rate, SR), 其中 AO 衡量所有帧中预测框与真实

标注框 IoU 值的平均水平, SR 则衡量 IoU 超过给定阈值 (如 0.50 或 0.75) 的帧所占比例, 分别记为 $SR_{0.50}$ 和 $SR_{0.75}$ 。这一指标体系不仅注重精度, 也充分体现了模型在不同目标类别上的鲁棒性。

UAV123 数据集侧重于无人机视角下的跟踪性能评估, 采用一次通过评估策略, 通过重叠评分与中心定位误差两个指标进行综合考察。其中, 重叠评分依旧基于 IoU 定义。中心定位误差则反映预测边界框中心与真实中心点之间的距离偏差, 并以指定阈值为界, 评估跟踪器的空间精确性。

在 LaSOT 数据集中, 由于其任务强调长时序目标跟踪的稳定性与泛化能力, 因此采用的评估体系进一步引入了标准化精度指标。LaSOT 中的评估包括成功率和精度。其中, 成功率同样以 IoU 为基础, 统计在不同 IoU 阈值下预测框覆盖真实框的比例, 并以曲线下面积 AUC 作为最终评价指标。精度计算方法与 OTB 系列数据集一致, 通过预测中心点与真实中心点之间的欧几里得距离进行衡量。

2.3 定量分析

为系统验证提出的目标跟踪算法在不同场景下的性能表现与泛化能力, 在多个权威公开数据集上开展了广泛且严谨的实验评估, 并与一系列代表性的先进跟踪方法进行了对比分析。

OTB 系列是目标跟踪领域广泛使用的标准数据集。OTB50 包含 50 个短视频序列, 而 OTB100 扩展至 100 个视频。它们涵盖了多种挑战场景, 如遮挡、尺度变化、快速运动等, 适合用于跟踪算法在典型环境下的评估。如表 1 和 2 所示, 在 OTB50 和 OTB100 数据集上, 分别采用精度与成功率作为评估指标。并且本研究算法还与 DROL-RPN、SiamBAN^[17]、DaSiamRPN、SiamRPN 以及基准 SiamFC 算法进行对比实验, 这些对比算法涵盖了基于相关滤波、传统孪生网络结构、区域提议网络 (RPN) 改进、以及更先进的端到端目标跟踪方法, 能够全面反映不同技术路线下的跟踪性能与发展水平。通过与这些方法进行比较, 不仅可以验证本研究算法在整体性能上的优劣, 还能够从多角度评估所提模块对跟踪准确性、鲁棒性及泛化能力的实际贡献。

表 1 OTB50 数据集对比结果

Table 1 Comparison results of the OTB50 dataset

跟踪器	成功率	精度
SiamFC	0.519	0.693
SiamRPN	0.583	0.806
DaSiamRPN	0.604	0.829
SiamBAN	0.611	0.891
DROL-RPN	0.673	0.898
本文	0.676	0.921

表 2 OTB100 数据集对比结果

Table 2 Comparison results of the OTB100 dataset

跟踪器	成功率	精度
SiamFC	0.583	0.765
SiamRPN	0.626	0.842
DaSiamRPN	0.654	0.873
SiamBAN	0.691	0.902
DROL-RPN	<u>0.701</u>	<u>0.914</u>
本文	0.704	0.920

实验结果表明,所提出算法在这两个基准数据集上均获得了显著的性能优势。在 OTB50 上,相较于基准算法 SiamFC,成功率提升了 30.3%,而精度则提升了 30.9%;在 OTB100 上,成功率和精度分别实现了 20.8%与 20.3%的增幅。

GOT-10k 是一个大型目标跟踪数据集,特别强调算法的泛化能力。它包含超过 10 000 个目标视频,其中训练与测试类别完全不重叠,要求跟踪器能够适应新类别的目标,避免过拟合。该数据集的挑战性在于目标多样性、运动模式复杂以及极端的环境变化。如表 3 所示,实验采用平均重叠率 AO 与成功率 SR 作为评价指标,并且本研究算法还与 PrDiMP、DiMP-50、ATOM、SiamRPN、SiamDW、SiamFCv2 以及 SiamFC 算法进行对比实验,这一对比方案具有明确的层次性与系统性:SiamFC 作为性能基准,体现了在原始孪生网络框架下的起点;SiamFCv2 与 SiamDW 等算法展示了通过增加网络深度与宽度以提升模型表达能力的改进路径;SiamRPN 则反映了引入区域提议网络在应对目标尺度变化方面的进展;而 PrDiMP、DiMP 与 ATOM 代表了一系列基于判别式学习框架的先进方法,它们在模型在线适应与目标背景判别方面具有显著优势。

实验结果显示所提算法在 AO 指标上从基准模型 SiamFC 的 0.348 显著提升至 0.658,提升了 31%。在 $SR_{0.50}$ 和 $SR_{0.75}$ 两个阈值下,分别提升了 41.2%与 47.1%,在多类目标与多变环境下展现了优越的适应能力。

表 3 GOT10k 数据集对比结果

Table 3 Comparison results of the GOT10k dataset

跟踪器	AO	$SR_{0.50}$	$SR_{0.75}$
SiamFC	0.348	0.353	0.098
SiamFCv2	0.374	0.404	0.144
SiamDW	0.429	0.483	0.147
SiamRPN	0.563	0.549	0.253
ATOM ^[18]	0.556	0.634	0.402
DiMP-50 ^[19]	0.611	0.717	0.492
PrDiMP ^[20]	<u>0.634</u>	<u>0.738</u>	<u>0.543</u>
本文	0.658	0.765	0.569

UAV123 专注于无人机视角下的目标跟踪,涵盖了多达 123 个视频序列,适用于测试算法在高速移动和复杂背景中的表现。该数据集的特点是高分辨率视频、各种场景与天气条件,以及无人机特有的视角和运动模式。并且与 DiMP-50、ATOM、SiamBAN、SiamRPN 以及 SiamFC 算法进行了对比,如表 4 所示,本研究算法在成功率方面相较于基准方法提升了 32.9%,在精度指标上也实现了 20%的提升,验证了其在空中视角下的有效性与稳定性。

表 4 UAV 数据集对比结果

Table 4 Comparison results of the UAV dataset

跟踪器	成功率	精度
SiamFC	0.498	0.719
SiamRPN	0.540	0.735
SiamBAN	0.631	0.833
ATOM	0.642	0.857
DiMP-50	<u>0.653</u>	<u>0.858</u>
本文	0.662	0.863

LaSOT 是一个大规模、长序列的目标跟踪数据集,旨在评估跟踪器在长时间跟踪任务中的表现。该数据集包含超过 1 000 个视频,涉及多种目标类别和长时序追踪,适用于测试算法的稳定性和持续跟踪能力。如表 5 所示,分别评估了成功率和精度两项指标。并且与 TrSiam、SiamTPN、DiMP-50、DaSiamRPN、ATOM、SiamMask、SiamDW 以及 SiamFC 算法进行了对比,为全面验证本研究方法在长序列跟踪任务中的性能,选取了 TrSiam、SiamTPN、DiMP-50 等先进算法进行对比,同样引入 Transformer 结构的最新方法对比,直接论证了轻量化 Transformer 设计在平衡模型性能与计算效率方面的先进性,在获得高性能的同时,更具备实际部署的潜力。

表 5 LaSOT 数据集对比结果

Table 5 Comparison results of the LaSOT dataset

跟踪器	成功率	精度
SiamFC	0.498	0.420
SiamDW	0.347	0.437
SiamMask	0.467	0.552
ATOM	0.499	0.570
DaSiamRPN	0.515	0.605
DiMP-50	0.560	0.642
SiamTPN ^[21]	0.581	0.683
TrSiam ^[22]	<u>0.624</u>	<u>0.712</u>
本文	0.636	0.727

实验结果表明,所提出方法在长序列跟踪任务中仍具备较强的鲁棒性和持续性能表现,成功率相较于主流方法提升了 27.7%,精度指标提升 73.1%,在高遮挡与目标退化

等极端条件下保持了较高的准确性。

加粗表示排名第 1,下划线表示排名第 2。

2.4 属性分析

OTB100 数据集包含了 11 类影响跟踪性能的关键视觉属性,分别为:光照变化(illumination variation, IV)、尺度变化(scale variation, SV)、形变(deformation, DEF)、遮挡(occlusion, OCC)、运动模糊(motion blur, MB)、快速运动(fast motion, FM)、平面内旋转(in plane rotation, IPR)、平面外旋转(out of plane rotation, OPR)、超出视野(out of view, OV)、背景干扰(background clutters, BC)和低分辨率(low resolution, LR)。这些属性全面模拟了真实应用中可能出现的复杂场景,有助于揭示算法在细粒度挑战下的性能瓶颈。

LaSOT 数据集通过标注 14 类关键视觉属性,全面刻画了目标跟踪任务中常见的挑战因素,涵盖长宽比变化(aspect ratio change, ARC)、低分辨率(low resolution, LR)、超出视野(out of view, OV)、快速运动(fast motion, FM)、完全遮挡(full occlusion, FOC)、尺度变化(scale variation, SV)、视角变化(viewpoint change, VC)、背景干扰(background clutter, BC)、旋转(rotation, ROT)、相机运动(camera motion, CM)、运动模糊(motion blur, MB)、形变(deformation, DEF)、部分遮挡(partial occlusion, POC)及光照变化(illumination variation, IV) 14 类复杂场景。这些属性从目标外观变化、空间位置变动到外部环境干扰等多个维度对跟踪算法提出挑战,有助于在更具现实意义的场景中评估算法的鲁棒性和泛化能力。

在本实验中,分别在 OTB100 与 LaSOT 数据集上,对所提方法与多种主流跟踪算法进行了在不同属性场景下的性能对比。图 5 展示了在 OTB100 数据集 11 种属性上的成功率对比结果,图 6 给出了 LaSOT 数据集中 14 种挑战属性下的对比性能。图中括号左边表示最小值,右边表示最大值。

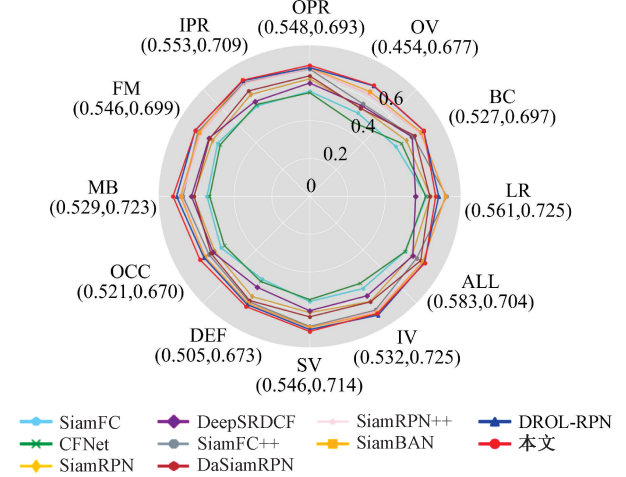


图 5 OTB100 数据集的属性分析结果

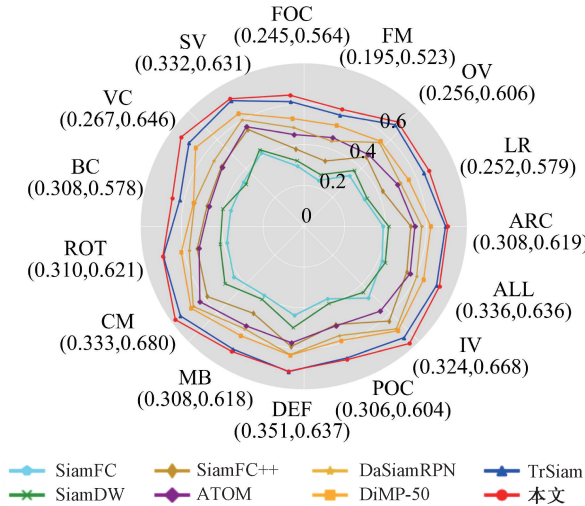


图 6 LaSOT 数据集的属性分析结果

Fig. 6 The attribute analysis results of the LaSOT dataset

2.5 消融实验

为寻找自适应模板更新机制的最佳阈值,在 LaSOT 数据集上进行了测试,比较了相应的成功率和帧率,如表 6 所示。实验表明,当阈值设置为 0.1 时,算法在准确性与实时性之间达到了较优的平衡,获得了最佳的综合性能表现。

表 6 不同阈值设置的实验结果

Table 6 Experimental results with different threshold settings

阈值 δ	成功率	帧率
0.30	0.619	48
0.20	0.628	47
0.10	0.636	45
0.05	0.632	38

为了进一步的验证监督反馈模块、轻量化 Transformer、自适应模版更新机制的实际效益,在 OTB100 数据集上开展了消融实验,详细结果如表 7 所示。其中,第 1 个模型为基准模型 SiamFC,后续模型依次引入不同模块,以清晰展示各组件对整体跟踪性能的影响。实验结果表明,监督反馈模块、轻量化 Transformer 与自适应模板更新机制均在不同程度上提升了跟踪器的准确性与鲁棒性,验证了其在提升跟踪性能方面的有效性与可行性。表中“√”和“×”分别表示采用和未采用相应项。

2.6 定性分析

为了更加直观展示算法的可视化跟踪结果,进一步分析算法的性能,从 OTB100 数据集中选取了 5 个具有代表性的视频序列,并在每个序列中截取 5 个关键帧,与 SiamBAN、SiamFC++、DaSiamRPN、DeepSRDCF、SiamFC 进行定性评估比较,结果如图 7 所示。

表 7 OTB100 数据集上的消融实验

Table 7 Ablation experiments on the OTB100 dataset

ResNet50	监督反馈	Transformer	轻量化 Transformer	模板更新	成功率	帧率
×	×	×	×	×	0.583	86
✓	×	✓	×	×	0.678	26
✓	✓	✓	×	×	0.690	25
✓	✓	×	✓	×	0.685	43
✓	✓	×	✓	✓	0.704	39

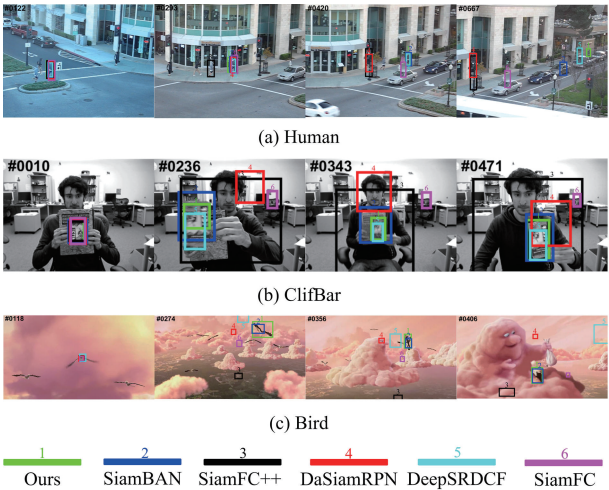


图 7 6 种算法的可视化跟踪对比结果

Fig.7 Visual tracking comparison results of six algorithms

在 Human 序列中,综合了尺度变化、低分辨率、频繁遮挡及快速运动等多重挑战,最后在第 667 fps,除本研究算法外,其余方法均未能保持对目标的有效跟踪,验证了本研究算法在复杂动态场景下的鲁棒性。

在 ClifBar 序列中,目标在与背景高度相似的区域中完全遮挡并移动,测试了算法在背景干扰条件下的判别能力。结果显示,在第 343 fps,仅有 DeepSRDCF 与本研究算法成功捕捉到目标位置,其余方法均发生跟踪混乱。

在 Bird 序列中,目标经历了长时间的完全遮挡与剧烈形变。大多数算法在遮挡结束后无法重新定位目标,仅有 SiamBAN 和本研究算法成功恢复跟踪,充分体现出其在处理形变与遮挡条件下的强鲁棒性。

3 结 论

在复杂场景下,基于深度学习的孪生网络目标跟踪方法已取得显著进展,但其在鲁棒性和适应性方面仍存在一定不足,尤其在处理遮挡、形变和外部干扰等挑战时性能易受影响。为提升跟踪算法在复杂环境下的稳定性与实用性,设计并引入了 3 项关键改进策略:监督反馈模块、轻量化 Transformer 结构以及自适应模板更新机制,以增强模型的特征表达能力和动态适应能力。为全面评估所提出方法的有效性,在多个主流基准数据集上开展了系统测

试。实验结果表明,本研究方法在复杂多变的跟踪场景中仍能保持较高的鲁棒性,同时具备良好的运行效率,满足实际应用中的实时性需求。

参考文献

[1] 王文胜,何君尧,黄民,等.融合梯度改进 YOLO 和 KCF 模型的无人机目标识别跟踪算法[J].仪器仪表学报,2025,46(2):221-233.
WANG W SH, HE J Y, HUANG M, et al. Fusion gradient improved YOLO and KCF models for UAV target recognition and tracking algorithm[J]. Chinese Journal of Scientific Instrument, 2025, 46 (2): 221-233.

[2] 郭秋蕊,李建良,田焱,等.基于改进 KCF 算法和多特征融合的车辆跟踪研究[J].电子测量与仪器学报,2022,36(4):231-240.
GUO Q R, LI J L, TIAN Y, et al. Research on vehicle tracking based on improved KCF algorithm and multi-feature fusion [J]. Jurnal of Electronic Measurement and Instrumentation, 2022, 36 (4): 231-240.

[3] 陈琳,刘允刚.面向无人机的视觉目标跟踪算法:综述与展望[J].信息与控制,2022,51(1):23-40.
CHEN L, LIU Y G. UAV visual target tracking algorithms: Review and future prospect [J]. Information and Control, 2022, 51(1): 23-40.

[4] 苑玉彬,吴一全,赵朗月,等.基于深度学习的无人机航拍视频多目标检测与跟踪研究进展[J].航空学报,2023,44(18):6-36.
YUAN Y B, WU Y Q, ZHAO L Y, et al. Research progress of UAV aerial video multi-object detection and tracking based on deep learning [J]. Acta Aeronautica et Astronautica Sinica, 2023, 44 (18): 6-36.

[5] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]. Computer Vision-ECCV 2016 Workshops: Amsterdam, 2016: 850-865.

[6] LI B, YAN J J, WU W, et al. High performance visual tracking with siamese region proposal network[C]. 2018

- IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018; 8971-8980.
- [7] BAO J H, YAN M L, YANG Y R, et al. SiamFFN: Siamese feature fusion network for visual tracking[J]. Electronics, 2023, 12(7): 1568.
- [8] YU Y CH, XIONG Y L, HUANG W L, et al. Deformable siamese attention networks for visual object tracking[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; 6727-6736.
- [9] YU P, DUAN ZH L, GUAN S J, et al. UnifiedTT: Visual tracking with unified transformer[J]. Journal of Visual Communication and Image Representation, 2024, 99: 104067.
- [10] LIU J, WANG H, MA CH, et al. SiamDMU: Siamese dual mask update network for visual object tracking[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2024, 8(2): 1656-1669.
- [11] ZHENG Y T, DENG H G, XU Q G, et al. AMTT: An end-to-end anchor-based multi-scale transformer tracking method[J]. Electronics, 2024, 13(14): 2710.
- [12] FANG Y, XIE B L, KHAIRUDDIN U, et al. DPT-tracker: Dual pooling transformer for efficient visual tracking [J]. CAAI Transactions on Intelligence Technology, 2024, 9(4): 948-959.
- [13] CHEN X, YAN B, ZHU J W, et al. Transformer tracking [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; 8126-8135.
- [14] ZHANG H, PIAO Y, QI N, et al. AtptTrack: Asymmetric transformer tracker with prior templates[J]. IEEE Access, 2024, 12: 10172-10185.
- [15] MU Q, HE Z H, WANG X Q, et al. SSTrack: An object tracking algorithm based on spatial scale attention[J]. Applied Sciences, 2024, 14(6): 2476.
- [16] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021; 10012-10022.
- [17] CHEN Z D, ZHONG B N, LI G R, et al. Siamese box adaptive network for visual tracking [C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; 6667-6676.
- [18] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: Accurate tracking by overlap maximization [C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019; 4660-4669.
- [19] BHAT G, DANELLJAN M, GOOL L V, et al. Learning discriminative model prediction for tracking [C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019; 6182-6191.
- [20] DANELLJAN M, GOOL L V, TIMOFTE R. Probabilistic regression for visual tracking[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020; 7183-7192.
- [21] XING D T, EVANGELIOU N, TSOUKALAS A, et al. Siamese transformer pyramid networks for real-time UAV tracking [C]. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, 2022; 2139-2148.
- [22] WANG N, ZHOU W G, WANG J, et al. Transformer meets tracker: Exploiting temporal context for robust visual tracking [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021; 1571-1580.

作者简介

符强, 硕士, 正高级实验师, 硕士生导师, 主要研究方向为图像处理、卫星导航和定位。

E-mail: 2310396462@qq.com

殷奇晨, 硕士研究生, 主要研究方向为计算机视觉跟踪。

E-mail: yin011220@163.com

纪元法(通信作者), 博士, 教授, 博士生导师, 主要研究方向为卫星导航。

E-mail: 1963096905@qq.com

任风华, 硕士, 副教授, 硕士生导师, 主要研究方向为图像处理。

E-mail: 643749271@qq.com