

文章编号:1007-2780(2026)05-0754-15

基于目标提示令牌的轻量化 Transformer 单目标跟踪算法

朱浩然^{1,2,3}, 王 铮⁴, 张海峰^{1,3*}, 陈卫宁^{1,3}, 何泽禹^{1,2,3}, 吴雄志^{1,2,3}, 梁思荣^{1,2}

(1. 中国科学院 西安光学精密机械研究所, 陕西 西安 710119;

2. 中国科学院大学, 北京 100049;

3. 西安飞行器光学成像与测量技术重点实验室, 陕西 西安 710119;

4. 北京宇航系统工程研究所, 北京 100076)

摘要: Transformer架构跟踪器凭借其优异的全局建模能力,在准确性方面表现突出,但仍存在模型规模庞大与计算延迟较高等问题。针对该问题,本文提出了一种基于目标提示令牌的轻量化 Transformer单目标跟踪方法 TPTTrack。该方法引入目标提示令牌,将其与目标模板和搜索区域特征共同输入轻量级 Transformer骨干网络,利用目标先验引导跨区域特征交互,提升跟踪鲁棒性。同时,通过分层注意力解耦机制在浅层完成特征提取与融合、在深层削减冗余自注意力计算,以降低计算开销。结合轻量自回归序列预测模块,实现对目标状态的动态建模与高效估计。在 GOT-10k 和 LaSOT 数据集上的对比实验表明, TPTTrack 在 GOT-10k 上的平均重叠率较最优轻量化基线模型提升 1.4%, 在 LaSOT 上的精度提升 1.7%。同时,其参数量和 FLOPs 仅为 5.6M 和 1.82G, GPU 和 CPU 推理速度分别达到 109.2 FPS 和 20.43 FPS, 其中 CPU 速度较最优轻量化基线模型提升 9.1%。综上,本文方法与现有轻量化跟踪算法相比,在精度、速度及部署友好性方面均表现优异,为资源受限场景下的单目标跟踪任务提供了一种轻量化方案。

关键词: 单目标跟踪; Transformer; 轻量化模型; 目标提示令牌; 分层注意力解耦

中图分类号: TP391.4 文献标识码: A doi:10.37188/CJLCD.2026-0058 CSTR:32172.14.CJLCD.2026-0058

Target prompt token-based lightweight transformer for single object tracking algorithm

ZHU Haoran^{1,2,3}, WANG Zheng⁴, ZHANG Haifeng^{1,3*}, CHEN Weining^{1,3},
HE Zeyu^{1,2,3}, WU Xiongzhì^{1,2,3}, LIANG Sirong^{1,2}

(1. Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China;

3. Xi'an Key Laboratory of Aircraft Optical Imaging and Measurement Technology, Xi'an 710119, China;

4. Beijing Institute of Astronautical Systems Engineering, Beijing 100076, China)

Abstract: Although Transformer-based trackers have demonstrated strong accuracy owing to their superior global modeling capability, they still suffer from large model size and high computational latency. To address

收稿日期:2026-04-03;修订日期:2026-04-15.

基金项目:陕西省重点研发计划(No. 2025CY-YBXM-076)

Supported by Shaanxi Provincial Key Research and Development Program (No. 2025CY-YBXM-076)

*通信联系人, E-mail: Zhanghf@opt.ac.cn

these issues, this paper proposes TPTTrack, a lightweight Transformer-based single object tracking method built upon target prompt tokens. Specifically, a target prompt token is introduced and jointly fed with the target template and search-region features into a lightweight Transformer backbone, where target priors are exploited to guide cross-region feature interaction and enhance tracking robustness. In addition, a hierarchical attention decoupling mechanism is designed to perform feature extraction and fusion in shallow layers while reducing redundant self-attention computation in deep layers, thereby lowering the overall computational cost. A lightweight autoregressive sequence prediction module is further incorporated to enable dynamic target-state modeling and efficient state estimation. Comparative experiments on the GOT-10k and LaSOT benchmarks show that TPTTrack improves Average Overlap on GOT-10k by 0.9% and precision on LaSOT by 1.4% over the strongest real-time baseline. Meanwhile, the model contains only 5.5M parameters and 1.79G FLOPs, and achieves inference speeds of 111.6 FPS on GPU and 20.76 FPS on CPU, with the CPU speed being 10.8% higher than that of the strongest real-time baseline. Overall, compared with existing lightweight tracking algorithms, the proposed method achieves superior performance in terms of accuracy, speed, and deployment friendliness, providing an effective lightweight solution for single object tracking in resource-constrained scenarios.

Key words: single object tracking; transformer; lightweight model; target prompt token; hierarchical attention decoupling

1 引言

目标跟踪旨在从视频序列中持续定位特定目标,是自动驾驶、无人机导航、移动增强现实和智能监控等实时视觉任务的关键技术^[1-2]。随着深度学习模型不断发展,主流跟踪器在基准数据集上取得了较高精度,但复杂网络结构也带来了较大的计算负载和内存消耗,限制了其在移动端、嵌入式平台和物联网设备等资源受限场景中的实时部署^[3]。目前,目标跟踪模型的实际应用主要面临3方面挑战:复杂Transformer^[4]和大规模CNN^[5]结构导致推理延迟较高;模型参数数量和计算复杂度较大,难以适配低功耗设备;在线更新机制可能引入额外延迟。因此,设计兼顾精度、速度和模型规模的轻量化实时跟踪模型具有重要研究意义。

单目标跟踪是指在首帧给定目标边界框的条件下,持续预测后续帧中同一目标的位置和尺度。早期方法多基于相关滤波器,通过在频域中执行相关操作获得目标响应图,具有较高计算效率,代表性方法包括MOSSE^[6]、KCF^[7]和CSK^[8]等。随着深度学习的发展,基于深度网络的跟踪方法逐渐成为主流。其中,孪生网络跟踪器通过双分支结构分别提取模板和搜索区域特征,并利用相似性

度量完成目标定位,代表性方法包括SiamFC^[9]、SiamRPN^[10]、Ocean^[11]和SiamRPN++^[12]等,它们有效结合了端到端训练与快速推理机制,在精度与效率间取得了良好平衡。近年来,基于Transformer的跟踪方法受到广泛关注^[13]。Transformer能够建模空间区域间的长程依赖关系,捕捉全局上下文信息,从而提升复杂背景和快速运动场景下的鲁棒性。TransT^[14]、OSTrack^[15]、STARK^[16]、SwinTrack^[17]和SeqTrack^[18]等方法已在GOT-10k^[19]、LaSOT^[20]和TrackingNet^[21]等基准上取得较好性能。此外,机器学习与Transformer结构在光学智能计算、复杂特征编码和遥感目标感知等相关领域也得到了广泛研究。Fang等利用轨道角动量辅助机器学习实现了高精度模式特征编码^[22];Wang等提出了可重构多功能集成光子计算芯片,为高效智能计算提供了新的硬件实现思路^[23];李大湘等针对航拍图像小目标检测任务提出了位置敏感Transformer模型,通过增强位置感知能力提升复杂航拍场景下的目标检测效果^[24]。这些研究表明,高效特征表达、位置敏感建模与轻量化计算已成为复杂视觉感知任务中的重要研究方向。

然而,现有Transformer跟踪器通常存在参数量大、计算开销高和实时部署困难等问题。为

此,本文提出一种轻量化 Transformer 单目标跟踪方法 TPTTrack(Target Prompt Token Tracking)。该方法引入目标状态提示令牌,将其与模板令牌和搜索区域令牌共同输入 Transformer 骨干网络,以目标状态先验引导特征交互。同时,将跟踪任务建模为自回归序列预测问题,并设计分层解耦注意力机制,对不同层级中的注意力计算进行合理分配,在保持跟踪精度的同时降低计算开销,使模型更适合实时和资源受限场景。

2 轻量化目标跟踪算法及不足

近年来,多种轻量化追踪器相继被提出。LightTrack^[25]采用神经架构搜索框架,通过深度可分离卷积和倒残差模块构建轻量级搜索空间,对骨干网络和头部结构进行联合优化,在资源受限设备上具有较好的部署潜力,但其仍依赖传统卷积算子,长程依赖建模能力有限。FEAR^[26]引入单参数双模板更新机制和像素级特征融合模块,具有较高效率,但在遮挡和快速运动场景下稳定性不足。MixFormerV2^[27]采用纯 Transformer 架构,并通过可学习预测令牌替代传统检测头实现端到端预测,但统一注意力策略容易产生冗余计算,且高性能仍依赖较大的模型规模。HiT^[28]通过桥接模块融合层级特征,并利用双图位置编码增强空间关系建模能力,但骨干结构和计算图仍较复杂。MVT^[29]采用 MobileViT^[30]作为骨干网络融合模板与搜索区域信息,具备较好的轻量化特性,但模块间精细化交互不足,限制了鲁棒性的进一步提升。

针对上述问题,本文提出 TPTTrack,一种融合 Transformer 表达能力与轻量化计算策略的单目标跟踪框架。本文主要贡献如下:

(1)针对模板区域与搜索区域中特征冗余和目标先验不足的问题,引入高效通道特征增强模块与目标状态提示令牌。前者通过轻量级通道重标定,增强目标相关特征响应并抑制背景干扰;后者通过编码紧凑目标状态信息参与注意力计算,引导模型聚焦目标相关区域,从而提升特征匹配鲁棒性。

(2)设计分层解耦注意力机制。在浅层采用混合注意力以增强特征表达,在深层削减冗余自注意力计算,仅保留关键区域交互建模,从而在

保持性能的同时降低计算量。

(3)引入轻量化自回归序列预测模块,对目标状态变化进行动态建模,提升模型在复杂运动、遮挡和尺度变化场景下的跟踪精度与稳定性。

3 TPTTrack 模型结构与训练

3.1 模型架构

TPTTrack 的整体结构如图 1 所示,采用轻量化 Transformer 单流跟踪框架。模型主要由图像块嵌入模块、高效通道特征增强模块、目标提示令牌生成模块、Transformer 编码器以及轻量级自回归预测模块组成。在训练阶段,首先将模板图像与搜索图像按照固定大小 $P \times P$ 进行非重叠图像块划分,并通过线性投影映射到统一特征空间,分别得到模板分支与搜索分支的视觉令牌: $Z \in R^{B \times m \times C}$, $X \in R^{B \times n \times C}$ 。其中, B 表示批大小, m 和 n 分别表示模板区域与搜索区域的令牌数量, C 表示特征维度。

为增强输入视觉特征的判别性,本文在图像块嵌入之后设置高效通道特征增强模块(Efficient Channel Feature Enhancement, ECFE)。该模块对模板特征和搜索特征进行轻量级通道重标定,以增强目标相关通道响应并抑制冗余背景信息。首先将视觉令牌按照空间结构恢复为二维特征图: $F_z \in R^{B \times C \times H_z \times W_z}$, $F_x \in R^{B \times C \times H_x \times W_x}$ 。其中, $H_z \times W_z = m$, $H_x \times W_x = n$ 。对于任意输入特征图 $F \in R^{B \times C \times H \times W}$, ECFE 首先通过全局平均池化获得通道描述向量:

$$y_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W F_c(i, j). \quad (1)$$

随后利用一维卷积建模局部跨通道依赖关系,并通过 Sigmoid 函数生成通道注意力权重:

$$s = \sigma(\text{Conv1D}_k(y)). \quad (2)$$

最终对原始特征进行通道重标定:

$$\hat{F} = F \otimes s, \quad (3)$$

其中, $\otimes$ 表示按通道逐元素相乘。经过通道重标定后,模板特征和搜索特征分别重新展平为增强视觉令牌 $\hat{Z}$ 和 $\hat{X}$, 作为后续特征交互的输入。ECFE 仅包含全局平均池化、一维卷积和逐通道乘法操作,能够以较低的数量和计算开销完成通道级特征筛选。

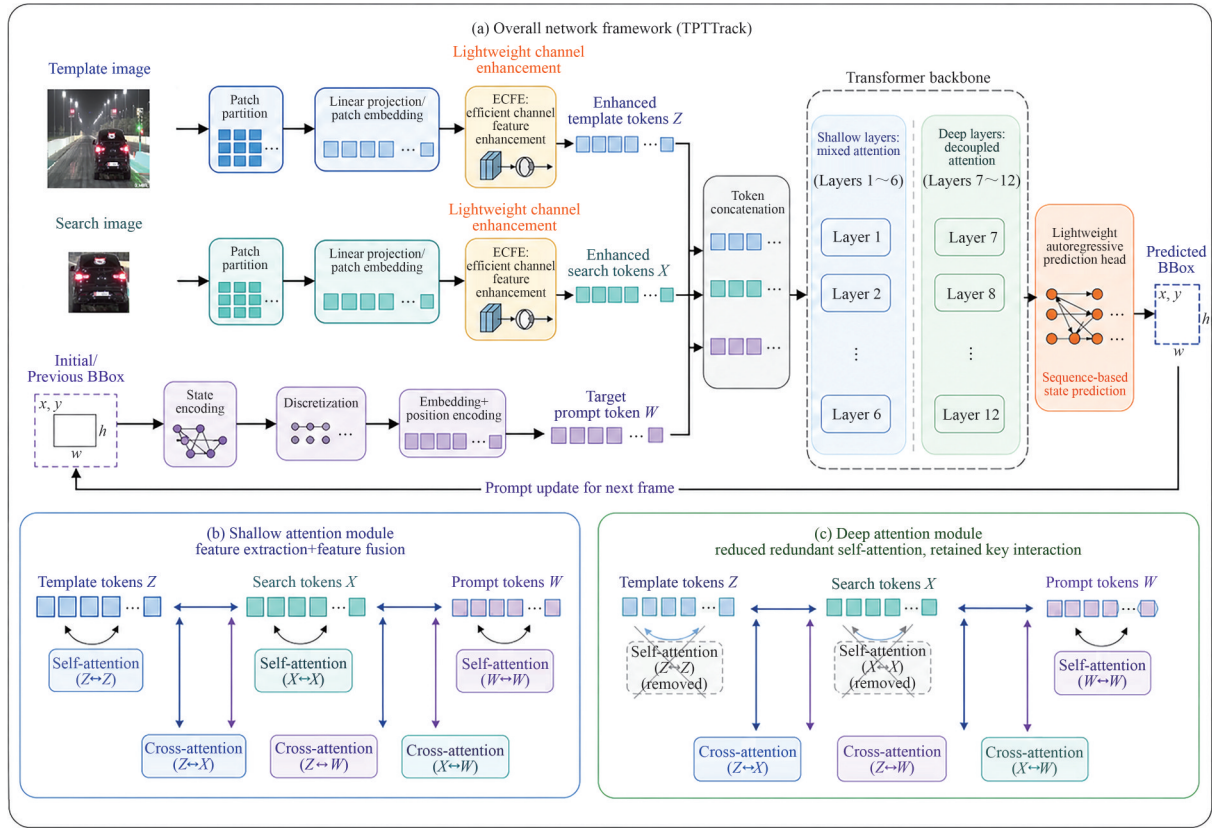


图1 TPTTrack模型。(a) TPTTrack网络框架;(b) 浅层注意力模块;(c) 深层注意力模块。

Fig. 1 TPTTrack model. (a) TPTTrack network framework; (b) Shallow attention module; (c) Deep attention module.

在视觉特征增强的基础上,本文进一步构造目标提示令牌。目标提示令牌由目标边界框状态生成,用于向Transformer特征交互过程显式注入目标位置和尺度先验。设第 t 帧目标边界框状态为:

$$b_t = (x_t, y_t, w_t, h_t), \quad (4)$$

其中: (x_t, y_t) 表示目标框左上角坐标, w_t 和 h_t 分别表示目标框宽度和高度。首先将其转换为角点形式,并在预设范围内进行约束:

$$\tilde{b}_t^i = \text{clip}(b_t^i, r_{\min}, r_{\max}). \quad (5)$$

随后将连续坐标离散映射为4个状态索引: $q_t = (q_t^1, q_t^2, q_t^3, q_t^4)$ 。其中,第 i 个状态索引为: $q_t^i = \left\lfloor \frac{\tilde{b}_t^{i,i} + r}{2r} N \right\rfloor$ 。式中: N 表示每一维坐标的离散bin数, r 表示坐标平移范围, $\lfloor \cdot \rfloor$ 表示取整操作。随后,通过目标状态Token将4个离散状态索引映射为固定长度 $l=4$ 的目标提示令牌序列。设状态嵌入矩阵为 $E_s \in \mathbb{R}^{N \times d}$,坐标位置编码为 $P_s = (p_1, p_2, p_3, p_4) \in \mathbb{R}^{4 \times d}$,则第 i 个状态令牌可表示为:

$$w_t^i = E_s(q_t^i) + p_i, i = 1, 2, 3, 4. \quad (6)$$

因此,目标状态Token定义为:

$$W_t = \mathcal{T}(b_t) = [w_t^1; w_t^2; w_t^3; w_t^4], \quad (7)$$

其中, $\mathcal{T}(\cdot)$ 表示由边界框状态到目标提示令牌序列的映射函数。由此,4维边界框状态被编码为固定长度 $l=4$ 的目标提示令牌序列,每个令牌对应一个边界框坐标分量,并通过坐标位置编码区分其语义位置。

经过视觉特征增强和目标状态提示编码后,将增强后的模板令牌、搜索令牌与目标提示令牌进行拼接,得到Transformer编码器的输入序列:

$$T_t = [\hat{Z}; \hat{X}; W_t], \quad (8)$$

随后输入Transformer编码器进行特征交互:

$$F_t = \Phi(T_t), \quad (9)$$

其中: $\Phi(\cdot)$ 表示Transformer编码器, F_t 表示编码器输出的融合特征。ECFE从通道维度筛选视觉特征,目标提示令牌从状态维度提供先验约束,Transformer编码器进一步建模模板区域、搜索区域与目标提示令牌之间的关联关系。

在第 t 帧跟踪过程中,模型由融合特征提取

状态回归令牌,并通过轻量级自回归预测模块输出当前帧离散状态预测结果,再通过反映射恢复连续边界框坐标:

$$\hat{b}_t = \phi(\hat{q}_t), \quad (10)$$

其中: $\hat{q}_t$ 表示自回归预测模块输出的离散状态索引, $\phi(\cdot)$ 表示由离散索引恢复连续边界框坐标的反映射函数。当前帧预测结果进一步用于更新下一帧搜索区域和目标提示令牌,其递推过程可表示为:

$$W_{t+1} = g(\hat{b}_t), \quad (11)$$

$$X_{t+1} = \Omega(I_{t+1}, \hat{b}_t), \quad (12)$$

其中: $g(\cdot)$ 表示由预测边界框生成目标提示令牌的编码过程, $\Omega(\cdot)$ 表示根据预测结果在下一帧图像中裁剪搜索区域的操作。

3.2 高效通道特征增强模块

为降低模型复杂度,本文采用轻量级 Transformer 骨干网络 ViT-Tiny^[31]进行特征建模。然而,轻量化结构在减少参数数量和计算开销的同时,也可能削弱对目标细节、弱纹理区域和复杂背景中判别性特征的提取能力。为此,本文在图像块嵌入之后设置高效通道特征增强模块(Efficient Channel Feature Enhancement, ECFE),对模板区域和搜索区域特征进行通道重标定,以增强目标相关响应并抑制冗余背景信息。

ECFE通过全局平均池化获取通道描述信息,并利用一维卷积建模局部跨通道依赖关系,最终生成通道注意力权重,其核心过程可表示为:

$$\hat{F} = F \otimes \sigma(\text{Conv1D}_k(\text{GAP}(F))), \quad (13)$$

其中: F 表示输入特征图, $\text{GAP}(\cdot)$ 表示全局平均池化, $\text{Conv1D}_k(\cdot)$ 表示卷积核大小为 k 的一维卷积, $\sigma(\cdot)$ 表示 Sigmoid 函数, $\otimes$ 表示按通道逐元素相乘。

该模块不引入复杂空间注意力分支或多尺度融合结构,参数数量和计算开销较低,适合嵌入轻量化跟踪框架。通过通道级特征筛选,ECFE能够提升输入视觉特征的判别性,为后续目标提示令牌引导下的跨区域特征交互提供更加稳定的特征表示。

3.3 分层注意力

在以往的目标跟踪模型中,MixFormer^[32]提出混合注意力模块,以同时实现特征提取与特征融合。本文在此基础上引入目标提示令牌,并进一步设计分层解耦注意力机制,以降低冗余注意力计算。

在注意力模块中,从模板帧的 Z_{token} 生成 Q_z, K_z, V_z ,从搜索帧的 X_{token} 生成 Q_x, K_x, V_x ,从目标提示令牌 W_{token} 生成 Q_w, K_w, V_w 。最终拼接得到 $Q = [Q_z; Q_x; Q_w]$, $K = [K_z; K_x; K_w]$, $V = [V_z; V_x; V_w]$,则大多数单流跟踪器中的注意力计算过程可表示为:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V. \quad (14)$$

为进一步分析混合注意力机制,上述公式可改写为式(15)形式:

$$A([Q_z; Q_x; Q_w], [K_z; K_x; K_w], [V_z; V_x; V_w]) = \text{softmax}\left(\frac{[Q_z; Q_x; Q_w][K_z; K_x; K_w]^T}{\sqrt{d}}\right) \cdot [V_z; V_x; V_w] = \text{softmax}\left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & Q_z K_x^T & Q_z K_w^T \\ Q_x K_z^T & Q_x K_x^T & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & Q_w K_w^T \end{bmatrix}\right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (15)$$

其中: $Q_z K_z^T, Q_x K_x^T$ 和 $Q_w K_w^T$ 分别表示模板区域、搜索区域和目标提示令牌的自注意力; $Q_i K_j^T (i, j \in \{z, x, w\}, i \neq j)$ 表示不同区域之间的交叉注意力,用于建模模板、搜索区域和目标提示令牌之间的关联关系。

设模板区域令牌长度为 m ,搜索区域令牌长度为 n ,目标提示令牌长度为 $l=4$,各注意力项的复杂度为:

$$\begin{cases} Q_z K_z^T: O(m^2) \\ Q_x K_x^T: O(n^2) \\ Q_w K_w^T: O(16) \\ Q_z K_x^T \text{ 和 } Q_x K_z^T: O(mn) \\ Q_z K_w^T \text{ 和 } Q_w K_z^T: O(4m) \\ Q_x K_w^T \text{ 和 } Q_w K_x^T: O(4n) \end{cases}. \quad (16)$$

由于通常满足 $n \gg m \gg l$,搜索区域内部自注意力 $Q_x K_x^T$ 带来的计算开销较大。因此,本文

将Transformer网络划分为浅层和深层两个阶段。设网络总层数为12,前 k 层为浅层阶段,后 $12-k$ 层为深层阶段。浅层阶段保留混合注意力以增强特征表达,深层阶段省略模板区域和搜索区域内部的冗余自注意力,仅保留跨区域交互以及目标提示令牌相关注意力。

设标准全注意力单层计算复杂度为 $C_{\text{full}} = (m+n+l)^2$,其中, m 、 n 和 l 分别表示模板区域令牌长度、搜索区域令牌长度和目标提示令牌长度。浅层阶段单层计算复杂度记为 $C_{\text{shallow}} = m^2 + n^2 + l^2 + mn + 2ml + 2nl$,深层阶段单层计算复杂度记为 $C_{\text{deep}} = l^2 + 2mn + 2ml + 2nl$ 。因

此,当网络总层数为12,且前6层为浅层阶段、后6层为深层阶段时,相对于12层标准全注意力结构,本文分层解耦注意力机制的总体计算量减少比例可表示为:

$$\eta = \frac{12C_{\text{full}} - 6C_{\text{shallow}} - 6C_{\text{deep}}}{12C_{\text{full}}} = \frac{m^2 + n^2 + mn}{2(m+n+l)^2}. \quad (17)$$

在实验设置中,取 $k=6$,模板区域令牌 Z 的长度为 $m=128$,搜索区域令牌 X 的长度为 $n=256$ 。基于上述设置,与式(3)相比,注意力计算量减少了约38.3%

浅层与深层阶段的注意力计算公式如下:

$$A_{\text{Shallow}} = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & 0 & Q_z K_w^T \\ Q_x K_z^T & Q_x K_x^T & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & Q_w K_w^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}. \quad (18)$$

$$A_{\text{Deep}} = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} 0 & Q_z K_x^T & Q_z K_w^T \\ Q_x K_z^T & 0 & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & Q_w K_w^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}. \quad (19)$$

浅层注意力模块与深层注意力模块的计算细节如图2所示。在浅层阶段,模板区域和搜索区域仍需保留一定的区域内自注意力,以提取局部纹理和低级语义特征;在深层阶段,网络更侧重目标与背景之间的高层语义交互,此时继续保留模板和搜索区域内部自注意力会带来较多冗

余计算。因此,本文在深层阶段去除 $Q_z K_z^T$ 和 $Q_x K_x^T$,仅保留跨区域交互以及目标提示令牌相关注意力项。由于目标提示令牌长度较短,且承担状态先验引导作用,所有涉及 Q_w 和 K_w 的注意力项均予以保留。该分层解耦设计能够在降低计算复杂度的同时保持关键特征交互能力。

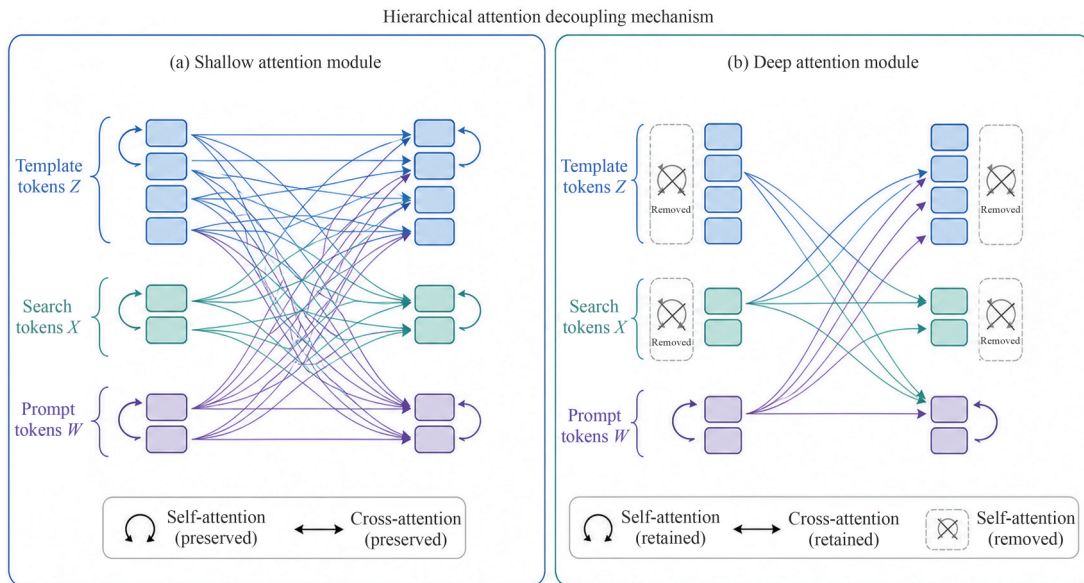


图2 浅层注意力模块与深层注意力模块的计算细节

Fig. 2 Computational details of the shallow attention module and the deep attention module

3.4 轻量级自回归预测模块

为实现目标状态的动态建模,并避免传统检测头带来的额外计算开销,本文在 Transformer 编码器后引入轻量级自回归预测模块,用于连续估计视频序列中的目标边界框状态。该模块将编码器输出的目标相关特征视为时序隐状态表示,并在历史目标状态与当前视觉特征条件下,以自回归方式预测当前帧目标边界框。

相比于一次性回归完整边界框参数的静态预测方式,自回归建模能够更好地刻画目标状态的时间连续性,有助于提升模型在快速运动、遮挡和尺度变化等复杂场景下的稳定性。同时,该模块无需构建额外特征金字塔、密集候选框或复杂后处理过程,而是直接基于 Transformer 编码后的紧凑状态表示完成目标状态预测,因此具有较低的参数开销和计算复杂度。

其自回归序列的概率表达形式为:

$$p = (Y^t | Y^{t-1}, (W^t, Z^t, X^t)), \quad (20)$$

式中: Z^t 和 X^t 分别表示 t 时刻的模板特征与搜索特征; Y^{t-1} 表示上一时刻的目标状态; W^t 由 Y^{t-1} 重新编码得到,并作为显式状态先验参与当前帧特征交互。自回归模块输出的目标状态用于当前帧边界框预测,同时在下一时刻重新编码为新的目标提示令牌,继续参与后续帧估计。

由此,目标提示令牌与自回归预测模块形成闭环状态更新机制:目标提示令牌提供显式目标状态先验,自回归预测模块逐帧估计并更新该先验。该机制使模型不仅能够在单帧内获得判别性特征表示,还能够时间维度上持续建模目标运动变化,从而提升复杂动态场景下的跟踪鲁棒性。

3.5 训练损失

对于自回归边界框生成模型,本文采用 L_1 损失与 GIoU 损失共同构建总体损失函数,其定义如下:

$$L = \lambda_1 L_{L1} + \lambda_2 L_{GIoU}, \quad (21)$$

式中, λ_1 和 λ_2 为权重系数,用于平衡 L_1 损失与 GIoU 损失在总体损失函数中的贡献。

L_1 损失用于度量预测边界框与真实边界框在参数空间中的差异,其定义为:

$$L_{L1} = \|B_{\text{pred}} - B_{\text{gt}}\|_1, \quad (22)$$

其中: $B_{\text{pred}} = (x, y, w, h)$ 表示预测边界框, $B_{\text{gt}} =$

(x^*, y^*, w^*, h^*) 表示真实边界框。

为增强预测框与真实框之间的整体空间匹配能力,本文进一步引入 GIoU 损失:

$$L_{GIoU} = 1 - \text{GIoU}(B_{\text{pred}}, B_{\text{gt}}), \quad (23)$$

其中, GIoU 定义为:

$$\text{GIoU} = \text{IoU} - \frac{|C - (B_{\text{pred}} \cup B_{\text{gt}})|}{|C|}, \quad (24)$$

式中: IoU 表示预测框与真实框之间的交并比, C 表示同时包围二者的最小外接矩形区域, $|\cdot|$ 表示对应区域面积。 L_1 损失侧重边界框参数的局部回归精度, GIoU 损失侧重整体空间重叠质量,二者结合能够同时约束位置尺度误差与几何匹配关系,从而提升边界框预测的准确性与稳定性。

4 实 验

4.1 训练

本研究的跟踪器基于 Python 3.6 和 PyTorch 1.7.0 实现。GOT-10k 结果仅使用 GOT-10k 训练集训练获得, LaSOT 结果则基于 GOT-10k 与 LaSOT 训练集联合训练获得。训练过程中采用水平翻转和亮度扰动等数据增强策略,模板图像与搜索图像尺寸分别设为 128×128 和 256×256 。模型在 NVIDIA GeForce RTX 3060 GPU 上训练 300 个周期,批次大小为 32,初始学习率为 4×10^{-4} ,并在第 240 个周期后采用 10^{-4} 的权重衰减。优化器采用 AdamW^[33],以实现模型的高效训练与稳定收敛。

4.2 在 GOT-10k 和 LaSOT 数据集上的评估

GOT-10k 是单目标跟踪领域常用基准之一,基于 WordNet 构建类别体系,强调开放世界条件下的泛化评估。该数据集覆盖 563 个目标类别,包含约 1.5M 帧逐帧标注数据,并采用严格的训练、验证和测试划分。其训练集与测试集类别不重叠,能够较好地检验跟踪器对未知类别的迁移与泛化能力。

LaSOT 是面向长期单目标跟踪的大规模基准数据集,包含 1 400 条视频序列、超过 3.5M 帧标注数据,平均序列长度超过 2 500 帧。该数据集覆盖 70 个目标类别,包含遮挡、快速运动、光照变化、背景干扰和视角变化等复杂因素,适合评估跟踪器在长序列场景下的鲁棒性与时间稳定性。

为保证不同方法的可比性,本文采用 GOT-10k

和LaSOT的标准评估协议。设第 t 帧预测框为 B_t ,真实框为 G_t ,总帧数为 N ,交并比IoU定义为:

$$\text{IoU}_t = \frac{|B_t \cap G_t|}{|B_t \cup G_t|}. \quad (25)$$

成功率用于统计预测框与真实框重叠程度超过给定阈值的帧比例:

$$\text{SR}(\alpha) = \frac{1}{N} \sum_{t=1}^N q(\text{IoU}_t > \alpha), \quad (26)$$

其中: $\alpha \in [0, 1]$ 为阈值, $q(\cdot)$ 为指示函数。GOT-10k中常用 $\text{SR}_{0.5}$ 和 $\text{SR}_{0.75}$ 分别衡量一般定位成功率和高精度定位能力。

LaSOT中的AUC定义为成功率曲线在不同阈值下的面积:

$$\text{AUC} = \int_0^1 \text{SR}(\alpha) d\alpha. \quad (27)$$

AO表示所有帧IoU的平均值,用于衡量整体重叠质量,是GOT-10k的核心指标之一:

$$\text{AO} = \frac{1}{N} \sum_{t=1}^N \text{IoU}_t. \quad (28)$$

设预测框中心点为 $c_t = (x_t, y_t)$,真实框中心点为 $g_t = (x_t^*, y_t^*)$,中心误差定义为:

$$\text{CLE}_t = \|c_t - g_t\|_2. \quad (29)$$

精度 P 统计中心误差小于给定像素阈值 τ 的帧比例,通常取 $\tau = 20$ 像素:

$$P = \frac{1}{N} \sum_{t=1}^N q(\text{CLE}_t < \tau). \quad (30)$$

为降低目标尺度差异对中心误差的影响,LaSOT进一步采用归一化中心误差:

$$\text{NCLE}_t = \sqrt{\left(\frac{x_t - x_t^*}{w_t}\right)^2 + \left(\frac{y_t - y_t^*}{h_t}\right)^2}, \quad (31)$$

其中, w_t 和 h_t 为真实框宽度和高度。归一化精度 P_{norm} 定义为:

$$P_{\text{norm}} = \frac{1}{N} \sum_{t=1}^N q(\text{NCLE}_t < \tau_n). \quad (32)$$

推理速度用于衡量模型在固定硬件环境下的处理效率。设连续处理 N 帧所需时间为 T ,则:

$$\text{FPS} = \frac{N}{T}. \quad (33)$$

模型参数量用于衡量模型规模。设模型包含 L 个可训练参数张量 $\{W_i\}_{i=1}^L$,则参数量表示为:

$$\text{Params}(M) = \frac{\sum_{i=1}^L |W_i|}{10^6}. \quad (34)$$

综上,本文采用AO、SR、AUC、 P 和 P_{norm} 从重叠精度和中心定位误差两方面评价跟踪性能;同时结合FPS、参数量和FLOPs衡量模型的运行效率与计算开销,从而全面分析算法的精度、速度和轻量化表现。

4.2.1 定量分析

将本文算法TPTTrack与现有跟踪方法在GOT-10k和LaSOT数据集上进行性能、模型复杂度和运行速度评估。表1展示了评估结果。其中LightTrack、FEAR-XS、HCAT、MVT、HiT-B、MixFormerV2和SMAT是轻量化跟踪器,TransT、DIMP-50、Ocean和SiamRPN++是非轻量化跟踪器。表1中的速度结果均在统一硬件平台下测试获得,具体实验环境为CPU:13th Gen Intel Core i7-13700和GPU:NVIDIA GeForce RTX 3060。

从表1可以看出,在GOT-10k数据集上,TPTTrack的AO达到65.0%,不仅优于LightTrack、FEAR-XS、HCAT、MVT、HiT-B、MixFormerV2和SMAT等轻量化方法,而且仅比非轻量化方法TransT低0.3%。在LaSOT数据集上,TPTTrack的AUC达到61.7%,高于LightTrack、FEAR-XS、HCAT、MVT、HiT-B、MixFormerV2和SMAT;精度达到68.8%,同样在轻量化方法中处于较优水平。同时,TPTTrack的GPU和CPU速度分别达到109.2 FPS和20.43 FPS,虽然GPU速度略低于HiT-B,但明显快于TransT、DIMP-50、Ocean和SiamRPN++等非轻量化方法。

从轻量化方面分析。TPTTrack的参数量为5.6M,明显低于TransT、DIMP-50、Ocean和SiamRPN++等非轻量化方法;FLOPs为1.82G,也明显低于上述非轻量化方法,说明本文方法在计算复杂度上具有明显的轻量化特征。与现有轻量化方法相比,TPTTrack的参数量和FLOPs虽然并非最小,但整体仍处于较低水平,同时与现有轻量化方法相比在精度指标上取得了更优或接近最优的结果。

如图3所示,不同轻量化跟踪器在精度、速度和模型复杂度之间呈现出明显差异。TPTTrack在AO、 $\text{SR}_{0.5}$ 和 $\text{SR}_{0.75}$ 等精度指标上整体处于较高水平,说明其具有较好的目标定位能力和重叠

表 1 在 GOT-10k 和 LaSOT 基准上与当前最先进方法的对比结果

Tab. 1 Comparison of our model TPTTrack with state-of-the-art methods on the GOT-10k and LaSOT benchmark

	Trackers	GOT-10K			LaSOT			Speed/FPS		Parameters/M	FLOPs/G
		AO	SR _{0.5}	SR _{0.75}	AUC	P _{Norm}	P	GPU	CPU		
Lightweight	LightTrack	58.2	66.8	44.2	52.3	62.3	55.6	81.58	10.32	2.0	0.53
	FEAR-XS	57.3	68.1	45.5	53.5	64.1	54.5	104.26	12.84	1.4	1.59
	HCAT	64.1	72.5	55.7	59.3	68.7	61.0	86.25	13.85	4.3	2.19
	MVT	62.4	73.0	54.6	55.3	64.1	58.3	90.8	11.4	5.5	0.77
	HiT-B	64.0	72.1	56.1	60.2	72.1	67.1	120.67	18.73	6.6	1.42
	MixFormerV2	63.6	71.9	54.7	60.6	69.9	60.4	99.49	17.54	4.8	1.56
	SMAT	64.3	72.8	55.8	60.7	71.1	64.6	115.3	16.51	3.78	1.63
	TPTTrack(ours)	65.5	73.9	56.8	61.7	71.6	68.8	109.2	20.43	5.6	1.82
Non-lightweight	TransT	65.8	74.8	58.5	62.9	71.8	67.0	40	4	26.1	32.9
	DIMP-50	61.1	69.7	49.2	58.1	65.6	60.4	35.5	7.5	23.0	21.5
	Ocean	61.1	70.1	47.3	57.3	66.2	61.9	30.5	6.2	44.3	26.0
	SiamRPN++	61.8	71.9	51.8	58.6	65.8	63.7	21.4	3.8	54.0	48.9

注:加粗数字表示轻量化方法中最优结果。GOT-10k 结果对应仅使用 GOT-10k 训练集训练,LaSOT 结果对应使用 GOT-10k 与 LaSOT 训练集联合训练。

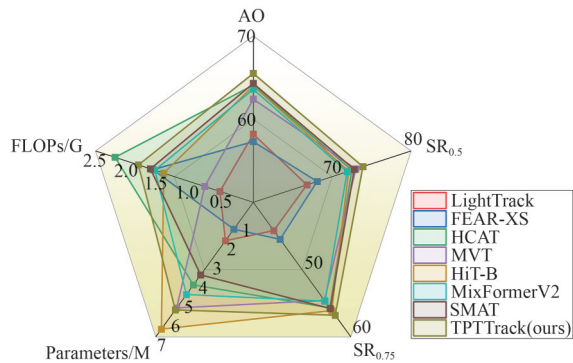


图 3 不同轻量化跟踪器在 AO、SR、FLOPs 和 Parameter 指标上的雷达图对比。

Fig. 3 Radar chart comparison of various lightweight trackers on AO, SR, FLOPs and Parameter metrics.

精度。同时,TPTTrack在推理速度方面也保持较优表现,表明模型具有较高的运行效率。相比部分精度较高但计算开销较大的方法,TPTTrack在参数数量和FLOPs方面更具轻量化优势;相比部分速度较快但精度不足的方法,TPTTrack能保持更稳定的跟踪性能。综合来看,TPTTrack在跟踪精度、运行速度和模型复杂度之间取得了较好的平衡,体现出较强的轻量化实时跟踪能力。

如图 4 所示,TPTTrack 在 4 类典型困难场景下的成功率曲线整体优于其余轻量化跟踪器。在相似目标干扰和低对比度场景中,TPTTrack

在中高重叠阈值下仍保持较高成功率;在遮挡重现场景中,其成功率曲线优势最明显,随阈值升高下降较为缓慢;在尺度变化场景中,TPTTrack 在大多数重叠阈值范围内保持领先,仅在局部区间略低于 HiT-B。总体来看,TPTTrack 在不同挑战场景下具有较高的重叠精度和稳定性。

如图 5 所示,TPTTrack 在 4 类场景下的精度曲线整体保持较高水平。在相似目标干扰场景中,其在较大误差阈值下仍能维持稳定的中心定位能力;在低对比度场景中,TPTTrack 在中高阈值区域与其他方法的差距更加明显;在遮挡重现场景中,其精度曲线在整个阈值范围内均处于领先地位;在尺度变化场景中,TPTTrack 在大部分阈值范围内保持较高精度,仅在少数区间与 HiT-B 接近。综合来看,成功率曲线和精度曲线结果表明,TPTTrack 在复杂场景下具有较好的目标重叠精度和中心定位精度。

4.2.2 定性分析

如图 6 所示,为更直观地说明目标提示令牌在跟踪过程中的作用,本文选取 3 个具有代表性的跟踪序列,对模型的特征响应热度图进行可视化分析。图中采用颜色强度表征模型对不同区域的响应程度,其中红色或高亮区域表示模型对该区域具有更强的关注,蓝色或低亮区域表示响应较弱。通过对比可以看出,在引入目标提示令

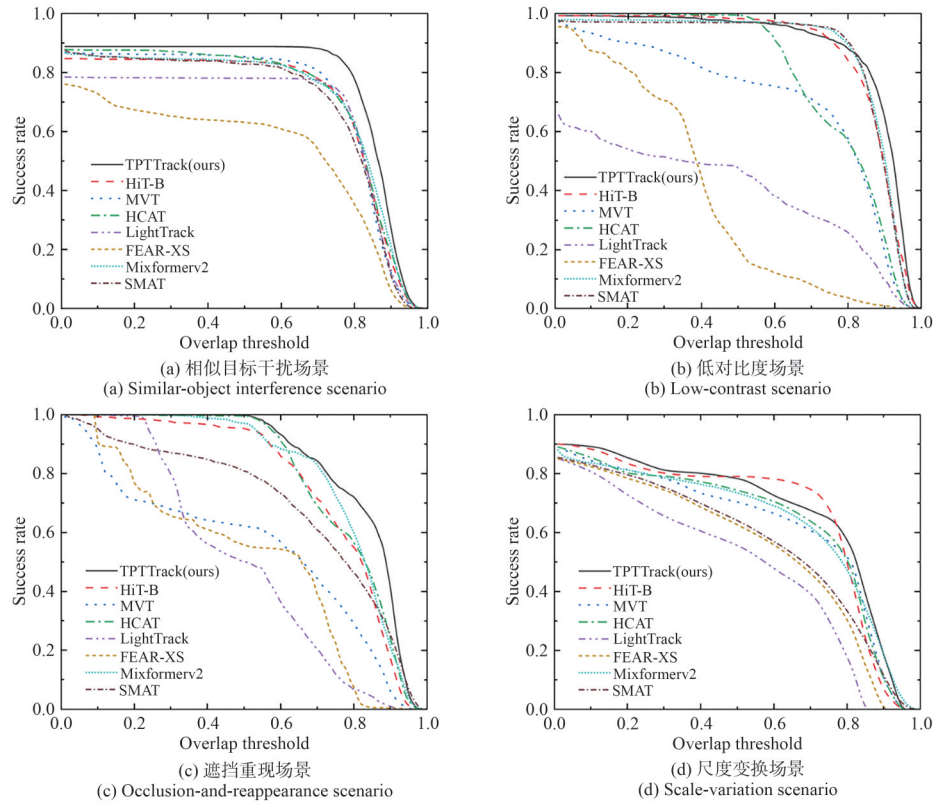


图 4 不同算法在不同场景下的跟踪成功率曲线

Fig. 4 Tracking success rate curves of different algorithms in different scenarios

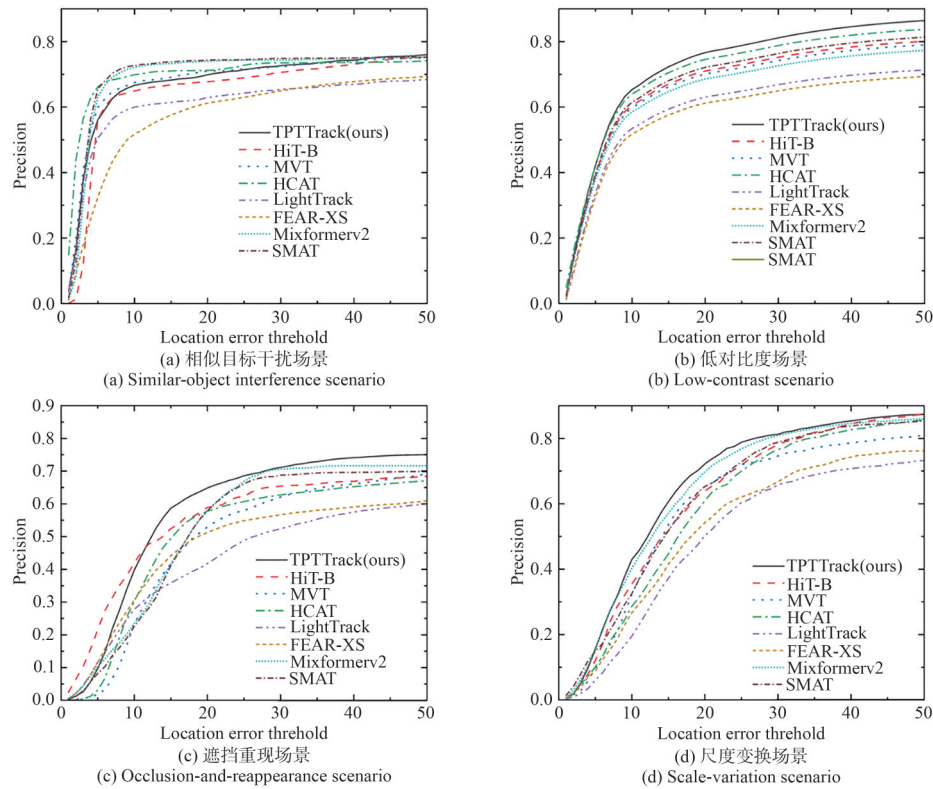


图 5 不同算法在不同场景下的跟踪精度曲线

Fig. 5 Tracking precision rate curves of different algorithms in different scenarios

牌后,模型对目标区域的响应更加集中,且对背景及无关干扰区域的响应明显减弱。这表明目标提示令牌能够为模型提供更明确的目标先验

信息,从而引导其在特征交互过程中更有效地聚焦于目标相关区域,进一步提升目标定位的准确性与跟踪稳定性。

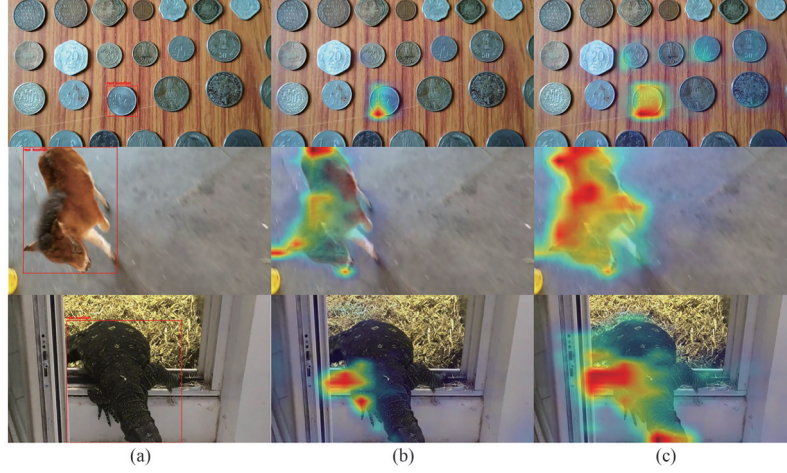


图6 热度图。(a) 原图;(b) 无目标提示令牌;(c) 有目标提示令牌。

Fig. 6 Heat maps. (a) Original image; (b) Without target prompt token; (c) With target prompt token.

为了验证本文算法在不同挑战场景下的跟踪性能,将 TPTTrack 与其他 7 种主流轻量化算法在 4 类典型场景中进行定性对比,结果如图 7 所示。在相似目标干扰场景中,目标周围存在外观相近的干扰物,容易导致误匹配和跟踪漂移。如图 7(a)所示,部分算法出现目标误跟踪或尺度估计不准确,而 TPTTrack 通过目标提示令牌引入目标状态先验,能够更稳定地聚焦目标区域,保持较准确的跟踪结果。在低对比度场景中,目标与背景之间亮度和纹理差异较小,易造成定位偏移。如图 7(b)所示,其他算法出现不同程度的跟踪漂移,尤其 FEAR-XS 的定位偏移较为明显。相比之下,TPTTrack 依靠特征增强和分层注意力机制强化目标相关区域建模,能够较稳定地覆盖目标主体。在遮挡重现场景中,目标受到遮挡后重新出现,容易导致跟踪中断或恢复困难。如图 7(c)所示,HCAT、MVT 和 HiT-B 等方法在遮挡后出现不同程度的框偏移,而 TPTTrack 通过自

回归预测模块对目标状态变化连续建模,在遮挡和重现过程中保持了较好的鲁棒性。在尺度变化场景中,目标尺寸和形态发生明显变化,对尺度适应能力提出了更高要求。如图 7(d)所示,部分跟踪器出现尺度估计不准确的问题,而 TPTTrack 结合目标提示令牌和自回归状态建模,能够随目标形态变化调整预测框,表现出较好的尺度适应能力。

综上,TPTTrack 在相似目标干扰、低对比度、遮挡重现和尺度变化等场景下均表现出较好的定位精度和跟踪稳定性,说明本文方法能够有效增强目标辨识能力、抑制背景干扰,并提升复杂场景下的鲁棒跟踪能力。

4.2.3 消融实验

为验证 TPTTrack 模型中分层注意力策略的有效性,模型在 GOT-10k 基准上评估了不同注意力机制组件的影响。公式(35)~(40)展示了各类注意力变体的计算方法:

$$A_1 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & 0 & Q_z K_w^T \\ Q_x K_z^T & Q_x K_x^T & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & Q_w K_w^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (35)$$

$$A_2 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & 0 & Q_z K_w^T \\ 0 & Q_x K_x^T & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & Q_w K_w^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (36)$$

$$A_3 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} 0 & Q_z K_x^T & Q_z K_w^T \\ Q_x K_z^T & 0 & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & 0 \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (37)$$

$$A_4 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} 0 & 0 & Q_z K_w^T \\ Q_x K_z^T & 0 & Q_x K_w^T \\ Q_w K_z^T & Q_w K_x^T & 0 \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (38)$$

$$A_5 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & 0 \\ Q_x K_z^T & Q_x K_x^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}, \quad (39)$$

$$A_6 = \text{softmax} \left(\frac{1}{\sqrt{d}} * \begin{bmatrix} Q_z K_z^T & 0 \\ 0 & Q_x K_x^T \end{bmatrix} \right) \cdot \begin{bmatrix} V_z \\ V_x \\ V_w \end{bmatrix}. \quad (40)$$

为验证分层注意力解耦策略的有效性,本文在 GOT-10k 基准上设计了不同注意力配置的对比实验。首先,以全注意力机制 A_{full} 作为基准,即网络所有层均采用标准注意力计算;随后,将解

耦注意力应用于不同阶段,浅层阶段采用 A_1 或 A_2 ,深层阶段采用 A_3 或 A_4 ,由此得到 A_{13} 、 A_{14} 、 A_{23} 和 A_{24} 四种组合。实验中浅层与深层比例设为 1:1,即 $k=6$,结果如表 2 所示。

由表 2 可见,配置 A_{13} 在 AO、 $\text{SR}_{0.5}$ 和 $\text{SR}_{0.75}$ 上均取得较优结果,同时保持了 111.6 FPS 的较高速度,体现出较好的精度与效率平衡。相比 A_{14} 、 A_{23} 和 A_{24} , A_{13} 在精度上更稳定,说明浅层保留区域内自注意力、深层侧重跨区域交互的设计



图7 8种轻量化跟踪器在4种场景上的定性对比

Fig. 7 Qualitative comparison of eight lightweight trackers across four scenarios

更有利于目标与背景的判别建模。虽然全注意力机制 A_{full} 的精度略高,但速度仅为 69.1 FPS,说明其引入了较多冗余计算,不利于模型的轻量化。

表 2 在 GOT-10k 测试集上的结果
Tab. 2 Performance of the model on GOT-10k test set

配置方案	AO	SR _{0.5}	SR _{0.75}	Speed/FPS
1(A_{13})ours	65.5	73.9	56.8	109.2
2(A_{14})	64.1	72.3	55.9	112.1
3(A_{23})	64.7	73.5	56.1	111.5
4(A_{24})	62.8	70.6	54.4	114.9
5(A_{full})	65.6	74.5	56.9	68.4

注:加粗数字表示注意力组合方式中最优结果

为验证高效通道特征增强模块对模型性能的影响,本文在 GOT-10k 基准上进行了消融实验。在该实验中,保持目标提示令牌、分层解耦注意力机制和轻量级自回归预测模块等,其余结构不变,仅对是否引入 ECFE 模块进行对比分析,实验结果如表 3 所示。

表 3 在 GOT-10k 测试集上的结果
Tab. 3 Performance of the model on GOT-10k test set

ECFE	AO	SR _{0.5}	SR _{0.75}
Not used	65.0	73.7	56.4
Used	65.5	73.9	56.8

注:加粗数字表示最优结果

为验证轻量级自回归预测模块的作用,本文在保持 Transformer 主干网络、目标提示令牌和分层解耦注意力机制一致的条件下,对不同边界框预测策略进行对比。实验包括传统角点预测头、弱自回归预测和完整自回归预测 3 种配置,结果如表 4 所示。

由表 4 可见,完整自回归预测模块在各项指标上均取得最优结果。相比角点预测头,自回归方式能够更好地建模目标状态变化;相比弱自回

表 4 在 GOT-10k 测试集上的结果
Tab. 4 Performance of the model on GOT-10k test set

Configuration	AO	SR _{0.5}	SR _{0.75}
Corner point prediction head	64.1	72.0	55.3
Weak autoregression	64.5	72.8	55.9
Autoregression	65.5	73.9	56.8

注:加粗数字表示最优结果

归结构,完整自回归进一步利用上一时刻状态信息作为条件输入,从而提升了目标在连续帧中的定位稳定性。

为验证目标提示令牌的有效性,本文构建了去除目标提示令牌的对照实验。该变体保留骨干网络、分层解耦注意力机制和自回归预测模块,仅去除目标状态提示信息。结果如表 5 所示。

表 5 在 GOT-10k 测试集上的结果
Tab. 5 Performance of the model on GOT-10k test set

Target prompt token	AO	SR _{0.5}	SR _{0.75}
Not used	64.8	73.1	55.7
Used	65.5	73.9	56.8

注:加粗数字表示最优结果

由表 5 可见,引入目标提示令牌后,模型在 AO、SR_{0.5} 和 SR_{0.75} 上均有所提升,说明目标状态提示信息能够为特征交互提供有效先验,引导模型更稳定地关注目标相关区域,从而增强跟踪鲁棒性。

5 结 论

针对基于 Transformer 架构的单目标跟踪器存在模型规模偏大、注意力计算冗余导致推理延迟较高,从而难以在资源受限场景中部署的问题,本文提出了一种轻量化 Transformer 单目标跟踪方法 TPTTrack。通过引入目标提示令牌,将目标状态先验显式融入特征交互过程,以增强目标相关区域表征并抑制背景干扰。通过分层注意力解耦机制,在不同网络层次上合理分配注意力计算,减少冗余运算开销。同时结合轻量级自回归预测模块,对目标状态进行逐帧连续建模,提高了模型在复杂动态场景下的适应能力。实验结果表明,TPTTrack 在 GOT-10k 上 AO 达到 65.5%,在 LaSOT 上精度达到 68.8%,在 GPU 与 CPU 上速度达到 109.2 FPS 与 20.43 FPS,模型参数量为 5.6M,计算复杂度为 1.82G。在保持较低参数量和计算复杂度的前提下,TPTTrack 仍展现出较高的跟踪精度和较好的实时性。这说明,本文方法在轻量化设计与性能保持之间实现了较合理的平衡。尽管本文方法在 CPU 平台上达到了 20.43 FPS,体现出一定的轻量化部署潜力,但对于高动态视频流场景下的严格实时要求

而言,当前速度仍存在一定差距。后续将进一步结合模型压缩、推理加速与硬件协同优化,并在真实边缘设备上补充帧率与延迟测试,以更全面评估其实际部署能力。

参 考 文 献:

- [1] 闵志方,杜虎,朱雪琼,等. 单目标跟踪研究综述[J]. 光学与光电技术,2023,21(4):1-14.
MIN Z F, DU H, ZHU X Q, *et al.* Survey of single target tracking research [J]. *Optics & Optoelectronic Technology*, 2023, 21(4): 1-14. (in Chinese)
- [2] 陈志旺,杨天宇,曹索航,等. 基于稀疏注意力的孪生网络目标跟踪算法[J]. 控制与决策,2024,39(12):4017-4026.
CHEN Z W, YANG T Y, CAO S H, *et al.* Siamese network object tracking algorithm based on sparse attention [J]. *Control and Decision*, 2024, 39(12): 4017-4026. (in Chinese)
- [3] ZHANG M, ZHANG F E, LANE N D, *et al.* Deep learning in the era of edge computing: challenges and opportunities [M]//ZOMAYA A, ABBAS A, KHAN S. *Fog Computing: Theory and Practice*. Hoboken: John Wiley & Sons, Inc., 2020: 67-78.
- [4] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need [C]//*Proceedings of the 31st International Conference on Neural Information Processing System*. Long Beach: Curran Associates Inc., 2017: 6000-6010.
- [5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]. 26th Annual Conference on Neural Information Processing Systems 2012. Lake Tahoe: Curran Associates Inc., 2012: 1106-1114.
- [6] BOLME D S, BEVERIDGE J R, DRAPER B A, *et al.* Visual object tracking using adaptive correlation filters [C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 2544-2550.
- [7] HENRIQUES J F, CASEIRO R, MARTINS P, *et al.* High-speed tracking with kernelized correlation filters [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(3): 583-596.
- [8] HENRIQUES J F, CASEIRO R, MARTINS P, *et al.* Exploiting the circulant structure of tracking-by-detection with kernels [C]. 12th European Conference on Computer Vision. Florence: Springer, 2012: 702-715.
- [9] BERTINETTO L, VALMADRE J, HENRIQUES J F, *et al.* Fully-convolutional Siamese networks for object tracking [C]//*Proceedings of the Computer Vision*. Amsterdam: Springer, 2016: 850-865.
- [10] 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. Salt Lake City: IEEE, 2018: 8971-8980.
- [11] ZHANG Z P, PENG H W, FU J L, *et al.* Ocean: object-aware anchor-free tracking [C]. 16th European Conference on Computer Vision. Glasgow: Springer, 2020: 771-787.
- [12] LI B, WU W, WANG Q, *et al.* SiamRPN++: evolution of Siamese visual tracking with very deep networks [C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 4277-4286.
- [13] KUGARAJEEVAN J, KOKUL T, RAMANAN A, *et al.* Transformers in single object tracking: an experimental survey [J]. *IEEE Access*, 2023, 11: 80297-80326.
- [14] CHEN X, YAN B, ZHU J W, *et al.* Transformer tracking [J/OL]. *arXiv*, 2021: 2103.15436.
- [15] YE B T, CHANG H, MA B P, *et al.* Joint feature learning and relation modeling for tracking: a one-stream framework [C]. 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 341-357.
- [16] YAN B, PENG H W, FU J L, *et al.* Learning spatio-temporal transformer for visual tracking [C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021: 10428-10437.
- [17] LIN L T, FAN H, ZHANG Z P, *et al.* SwinTrack: a simple and strong baseline for transformer tracking [C]//*Proceedings of the 36th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2022: 1218.
- [18] CHEN X, PENG H W, WANG D, *et al.* SeqTrack: sequence to sequence learning for visual object tracking [C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 2023:

- 14572-14581.
- [19] 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.
 - [20] FAN H, LIN L T, YANG F, *et al.* LaSOT: a high-quality benchmark for large-scale single object tracking [C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 5369-5378.
 - [21] MÜLLER M, BIBI A, GIANCOLA S, *et al.* TrackingNet: a large-scale dataset and benchmark for object tracking in the wild [C]. 15th European Conference on Computer Vision. Munich: Springer, 2018: 310-327.
 - [22] FANG X Y, HU X N, LI B L, *et al.* Orbital angular momentum-mediated machine learning for high-accuracy mode-feature encoding [J]. *Light: Science & Applications*, 2024, 13(1): 49.
 - [23] WANG Y F, LIAO K, ZHANG K, *et al.* Reconfigurable versatile integrated photonic computing chip [J]. *eLight*, 2025, 5(1): 20.
 - [24] 李大湘, 辛嘉妮, 刘颖. 位置敏感Transformer航拍图像目标检测模型[J]. *光学精密工程*, 2024, 32(5): 727-739.
LI D X, XIN J N, LIU Y. Position-sensitive Transformer aerial image object detection model [J]. *Optics and Precision Engineering*, 2024, 32(5): 727-739. (in Chinese)
 - [25] YAN B, PENG H W, WU K, *et al.* LightTrack: finding lightweight neural networks for object tracking via one-shot architecture search [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 15175-15184.
 - [26] BORSUK V, VEI R, KUPYN O, *et al.* FEAR: fast, efficient, accurate and robust visual tracker [C]. 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 644-663.
 - [27] CUI Y T, SONG T H, WU G H, *et al.* MixFormerV2: efficient fully transformer tracking [C]//*Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2023: 2561.
 - [28] KANG B, CHEN X, WANG D, *et al.* Exploring lightweight hierarchical vision transformers for efficient visual tracking [C]//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Paris: IEEE, 2023: 9578-9587.
 - [29] GOPAL G Y, AMER M A. Mobile vision transformer-based visual object tracking [C]. 34th British Machine Vision Conference. Aberdeen: BMVA Press, 2023: 800-805.
 - [30] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer [C]. The 10th International Conference on Learning Representations. OpenReview.net, 2022.
 - [31] WU K, ZHANG J N, PENG H W, *et al.* TinyViT: fast pretraining distillation for small vision transformers [C]. 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 68-85.
 - [32] CUI Y T, JIANG C, WANG L M, *et al.* MixFormer: end-to-end tracking with iterative mixed attention [C]//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 13598-13608.
 - [33] ZHOU P, XIE X Y, LIN Z C, *et al.* Towards understanding convergence and generalization of AdamW [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(9): 6486-6493.

作者简介:



朱浩然, 男, 硕士研究生, 2023 年于湖南大学获得学士学位, 主要从事图像处理、目标跟踪的研究。E-mail: zhuhaoran23@mails.ucas.ac.cn



张海峰, 男, 硕士, 副研究员, 2007 年于暨南大学获得硕士学位, 主要从事高可靠光电成像、航天器图像与视频信息处理等方面的研究。E-mail: Zhanghf@opt.ac.cn