文章编号　1004-924X(2026)12-1904-24

面向视觉目标跟踪的双引导特征增强 Mamba

才　华[1]，叶柏群[1]，胡海东[2*]，寇婷婷[1]，付　强[3]，沈卓奇[1]

(1. 长春理工大学 电子信息工程学院，吉林 长春 130022；
2. 北京控制工程研究所，北京 100190；
3. 长春理工大学 空间光电技术研究所，吉林 长春 130022)

摘要：针对现有跟踪方法无法实现时序特征与空间特征的解耦协同问题，提出了面向视觉目标跟踪的双引导特征增强 Mamba 跟踪器。在 Vision Mamba 骨干基础上，设计时空特征协同模块，将耦合的时空信息分离并双向增强；构建双分支模仿注意力机制，让浅层特征模仿深层特征关注重点，增强判别能力；引入动态模板增强与更新策略，通过长短时记忆库协同，实现模板的自适应迭代。实验结果表明，在 OTB100 数据集上，精度和成功率分别达到 93.7% 与 75.6%；在 LaSOT，TrackingNet 和 GOT-10k 上，在多数指标上取得领先或次优性能，其中 LaSOT AUC 为 75.8%，TrackingNet 为 86.2%，GOT-10k 平均重叠率为 78.9%。本算法有效平衡协同了时序特征和空间特征，综合性能优于主流对比方法，在面对复杂挑战下表现出较强的精度与鲁棒性。

关　键　词：目标跟踪；Mamba；时空协同耦合；模仿注意力；动态模板更新

中图分类号：TP394.1；TH691.9　　**文献标识码**：A

doi：10.37188/OPE.20263412.1904　　**CSTR**：32169.14.OPE.20263412.1904

Dual-guided feature enhancement Mamba for visual object tracking

CAI Hua[1], YE Baiqun[1], HU Haidong[2*], KOU Tingting[1], FU Qiang[3], SHEN Zhuoqi[1]

(1. *School of Electronical and Information Engineering, Changchun University of Science and Technology, Changchun* 130022, *China*;
2. *Beijing Institute of Control Engineering, Beijing* 100190, *China*;
3. *Institute of Space Opto-Electronic Technology, Changchun University of Science and Technology, Changchun* 130022, *China*)
* *Corresponding author, E-mail: haidong_2005@163.com*

Abstract: To address the issue that existing tracking methods could not achieve decoupling and synergy between temporal features and spatial features, a Dual-Guided Feature Enhancement Mamba for Visual Object Tracking was proposed. Built upon the Vision Mamba backbone, a Spatial-Temporal Decoupling and Collaboration Module was designed to separate and bidirectionally enhance coupled spatiotemporal information. A dual-branch imitation attention mechanism was constructed, enabling shallow features to em-

收稿日期：2026-03-30；**修订日期**：2026-04-24.

基金项目：国家自然基金联合基金(No. U2341226)；吉林省自然科学基金(No. 20260102267JC)；2024 年空间智能控制技术全国重点实验室开放基金(No. 2024-CXPT-GF-JJ-012-12)

ulate the focus of deep features, thereby enhancing discriminative capability. A dynamic template enhancement and update strategy was introduced, leveraging collaborative long- and short-term memory banks to achieve adaptive template iteration. Experimental results demonstrate that the proposed algorithm achieves a precision of 93.7% and a success rate of 75.6% on the OTB100 dataset. It also attains leading or suboptimal performance on most metrics across the LaSOT, TrackingNet, and GOT-10k benchmarks, with an AUC of 75.8% on LaSOT, 86.2% on TrackingNet, and an average overlap of 78.9% on GOT-10k. The algorithm effectively balances and synergizes temporal features with spatial characteristics, outperforming mainstream comparison methods and demonstrating strong accuracy and robustness under complex challenges.

Key words: object tracking; Mamba; spatiotemporal coupling; imitation attention; dynamic template

1　引　言

单目标视觉跟踪是计算机视觉领域的核心研究方向，广泛应用于视频监控、自动驾驶、无人机导航、智能人机交互[1][2]等实际场景，核心目标是在连续视频序列中对初始帧指定的目标进行持续、精准的位置预测与状态追踪。早期跟踪方法以图像级跟踪为主，典型代表包括SiamFC[3]，SiamRPN[4]，SiamRPN++[5]，SiamFC++[6]，Siam R-CNN[7]和AiATrack[8]等，这类方法多基于孪生卷积神经网络架构，仅依靠目标初始外观信息与当前帧匹配定位，未充分挖掘视频序列的时序上下文关联。尽管此类算法实现了高效推理，但仅依赖单一初始模板的设计，使其在面对目标变形、运动模糊、遮挡、相似背景干扰等复杂场景时极易出现跟踪漂移与目标丢失，即便部分方法引入额外模板更新机制，仍未脱离单纯图像级外观匹配的局限[9][10]。

为突破图像级跟踪的性能瓶颈，研究人员逐步转向视频级目标跟踪，通过聚合帧间时序上下文信息提升跟踪鲁棒性。其中Cao等在2022年提出的TCTrack[11]，通过在线时间自适应卷积与特征地图精化聚集上下文信息；Xie等在2023年提出的VideoTrack[12]，依托视频Transformer主干完成跨帧上下文集成；Zeng等在2024年提出的ODTrack[13]，采用令牌序列传播范例密集关联帧间上下文关系；ARTrack[14]与ARTrackV2[15]将历史预测坐标编码为令牌辅助定位；AQATrack[16]引入可学习的自回归查询捕获长程依赖；Shi等在2024年提出的EVPTrack[17]，利用时空标记实现跨帧信息传播。上述方法虽显著提升了跟踪性能，但仍存在共性瓶颈：多数方法依赖有限附加令牌捕获上下文，如VideoTrack仅使用16个固定上下文令牌进行跨帧信息集成，ODTrack虽采用密集令牌传播但单帧有效令牌数限制为256个，难以实现全局时空信息的全面建模，易造成长程依赖丢失；Transformer类主干的计算复杂度随序列长度非线性增长，限制长序列处理效率，难以平衡精度与推理速度；模板更新机制缺乏自适应筛选与闭环优化，噪声累积问题突出。

随着状态空间模型Mamba在长序列建模领域取得突破，其线性复杂度、高效全局依赖建模的特性为视觉跟踪提供了新路径。Gu等提出数据依赖的SSM层，开发出通用主干网络Mamba，在多尺度任务上的性能优于Transformer，且计算复杂度与序列长度呈线性关系。面向视觉任务优化的Vision Mamba[18]（Vim），凭借双向Mamba块实现了高效视觉特征表征，后续衍生的MCLTrack[19]，ChangeMamba[20]，TrackingMamba[21]等方法，也验证了Mamba在视觉跟踪场景的应用潜力。但现有基于Mamba的跟踪算法，虽在长程时序建模上具有较大优势，但在空间特征处理上存在明显不足：MCLTrack通过额外堆叠3层卷积层增强空间特征，但卷积的局部感受野限制了全局空间依赖建模；TrackingMamba在Mamba块中保留了注意力残差分支，却引入了二次计算复杂度；ChangeMamba仅采用单向空间扫描策略，无法捕捉多方向空间上下文。这些方法均难以同时满足跟踪任务对空间细节判别与时序运动建模的双重需求。

针对这一问题，本文构建面向视觉目标跟踪的双引导特征增强 Mamba 跟踪器(Dual-Guided Feature Enhancement Mamba for Visual Object Tracking，DMT)，通过时空特征协同模块解耦并增强时序与空间特征表达，利用双分支模仿注意力模块实现浅层特征的自适应优化，依托动态模板更新模块完成模板的迭代优化，最终输出精准跟踪结果，以提升 Mamba 架构在复杂视频场景下的跟踪精度及鲁棒性，为视频级单目标跟踪提供了高效、稳健的新解决方案。

2　本文算法

本文算法的整体框架如图 1 所示，由 Vision Mamba 骨干网络、时空特征协同模块、双分支模仿注意力模块、动态模板增强与更新模块和多任务协同跟踪头组成。骨干网络采用 Vision Mamba 提取多级时空耦合特征；时空特征协同模块用于实现时序特性与空间特征的深度协同；在此基础上使用双分支模仿注意力模块增强浅中层特征的判别性；同时使用动态模板增强与更新模块生成优质动态模板；最终通过跟踪头输出跟踪结果。

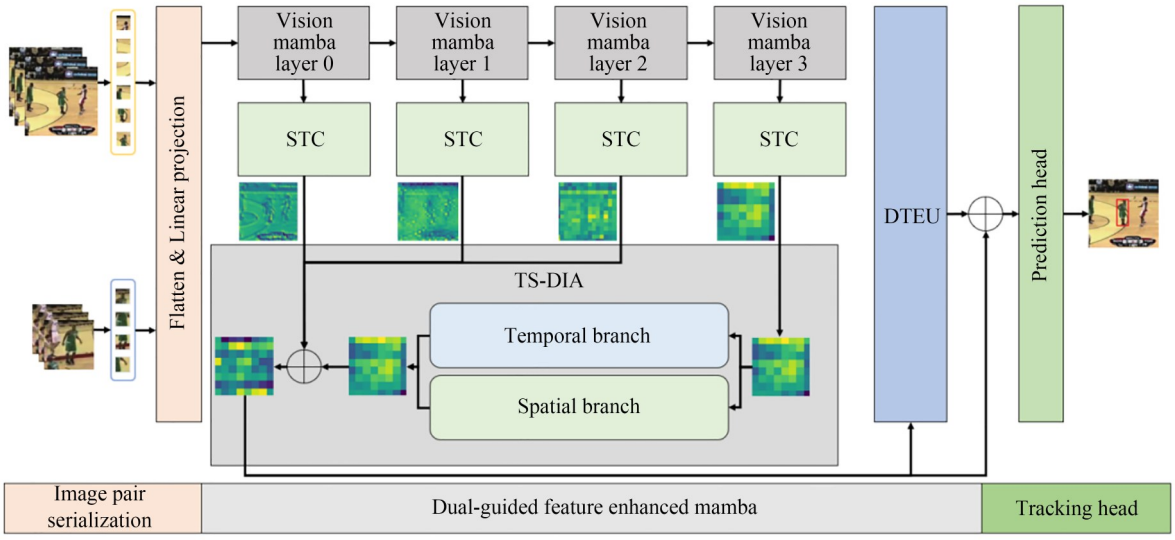


图 1　目标跟踪算法的整体框架

Fig. 1　Overall framework of the proposed tracking algorithm

2.1　主干网络

主干网络采用 4 层堆叠的 Vision Mamba 块，结构如图 2 所示，模板帧图像 $z\in R^{3\times H_z\times W_z}$ 与搜索帧图像 $x\in R^{3\times H_x\times W_x}$ (H_z 为模板图像高度，W_z 为模板图像宽度，H_x 为搜索区域图像高度，W_x 为搜索区域图像宽度)作为输入。在进入主干之前，这些图像经过线性投影层被映射模板特征 $f_z=R^{C\times\frac{H_z}{s}\times\frac{W_z}{s}}$ 和搜索特征 $f_x=R^{C\times\frac{H_x}{s}\times\frac{W_x}{s}}$ (C 为通道数，s 为缩放因子)。随后在空间维度上对特征映射展平及拼接，得到初始特征图 $F_0\in R^{C\times H\times W}$ (C 为通道数，H，W 为特征图空间尺寸)；将初始特征图依次送入 4 层 Vision Mamba 块，最终输出特征图供后续模块调用。

每个 Vision Mamba 块由归一化层、双路线性投影、双向并行 SSM 分支、门控融合单元、残差连接五部分组成，对于第 l 层编码器，输入特征为上一层输出 $F_{l-1}\in R^{N\times C}$。

对输入特征执行层归一化，消除特征分布偏差，将归一化后的特征分为两路，分别经过独立的线性投影层，生成 SSM 分支的输入 x 与门控分支的输入 z：

$$\begin{cases}x=\mathrm{Linear}_x(F_{\mathrm{norm}})\\ z=\mathrm{Linear}_z(F_{\mathrm{norm}})\end{cases},\tag{1}$$

其中：Linear_x 与 Linear_z 为参数不共享的线性投影层，F_{norm} 为层归一化后的特征。

x 分支分为前向(Forward)与后向(Backward)两个完全并行的计算路径，分别对空间维度的序列进行双向扫描建模，捕捉单帧图像不同

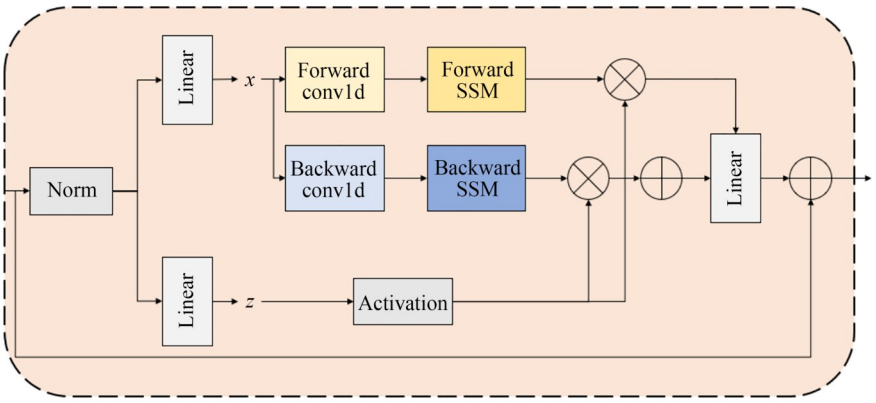


图 2　Vision Mamba 整体框架

Fig. 2　Vision mamba overall architecture

方向的全局空间依赖。两条路径均由 1 维卷积（Convolutional 1-dimensional，Conv1d）与状态空间模型（State Space Model，SSM）串联组成，计算过程为：

$$\begin{cases} F_{\text{forward}} = \text{SSM}_{\text{forward}}\left(\text{Conv1d}_{\text{forward}}(\boldsymbol{x})\right) \\ F_{\text{backward}} = \text{SSM}_{\text{backward}}\left(\text{Conv1d}_{\text{backward}}(\boldsymbol{x})\right) \end{cases}, \tag{2}$$

其中：$\text{Conv1d}_{\text{forward}}$，$\text{Conv1d}_{\text{backward}}$ 为核大小 3 的 1 维卷积；F_{forward}，F_{backward} 分别为前向、后向特征；$\text{SSM}_{\text{forward}}$，$\text{SSM}_{\text{backward}}$ 为前向、后向扫描的状态空间模型，核心离散状态方程为：

$$\begin{cases} \boldsymbol{h}_t = A\boldsymbol{h}_{t-1} + B\boldsymbol{x}_t \\ \boldsymbol{y}_t = C\boldsymbol{h}_t + D\boldsymbol{x}_t \end{cases}, \tag{3}$$

其中：$\boldsymbol{h}_t$ 为 t 时刻的隐藏状态，$\boldsymbol{x}_t$ 为输入特征，$\boldsymbol{y}_t$ 为输出特征，A，B，C，D 为可学习参数。

门控分支的输入 $\boldsymbol{z}$ 经过 SiLU 激活函数后，分别与前向、后向 SSM 的输出执行元素级乘法，两路结果相加完成双向特征融合；融合后的特征经过线性投影层做维度变换，最终将投影后的特征与该层的原始输入 F_{l-1} 执行残差相加，完成单层编码器的计算，得到第 l 层的输出特征也就是时空特征协同模块的输入特征 F_{in}。前几层特征作为浅中层特征，第四层特征作为深层特征。

2.2　时空特征协同模块

在主干网络输出的多级特征中，时序特征与空间特征仍处于高度耦合状态，二者相互缠绕、难以独立表达，这种耦合关系导致时序运动信息与空间细节信息在后续处理中相互干扰，无法同时满足目标跟踪对空间判别能力与时序建模能力的双重需求。针对此问题，本文设计时空特征协同模块（Spatial-Temporal Collaboration Module，STC）作为主干网络的专属后处理单元，接收 Vision Mamba 主干输出的多级耦合特征，通过正交约束实现时序与空间特征的独立解耦，结合双向门控交互与场景自适应融合，输出兼具时序一致性与空间辨识度的高质量增强特征。

如图 3 所示，本模块采用时序特征与空间特征相互交互制约的方式来增强时序特征与空间特征的协同性。模块输入为主干网络输出的第 n 层级特征 $F_{\text{in}} \in \boldsymbol{R}^{C\times H\times W}$，输出为解耦增强后的特征 $F_{\text{out}} \in \boldsymbol{R}^{C\times H\times W}$。

为稳定特征分布，需对输入特征 F_{in} 执行层归一化，得到归一化后的耦合特征 $F_{\text{in}}^{\text{norm}}$，通过两个参数不共享的 1×1 卷积分支通道变换，初步拆分出初始时序特征 F_t^{init} 与初始空间特征 F_s^{init}：

$$\begin{cases} F_t^{\text{init}} = \text{Conv}_{1\times1}^{t}\left(F_{\text{in}}^{\text{norm}}\right) \\ F_s^{\text{init}} = \text{Conv}_{1\times1}^{s}\left(F_{\text{in}}^{\text{norm}}\right) \end{cases}, \tag{4}$$

其中：F_t^{init} 为初始时序特征，F_s^{init} 为初始空间特征，$F_{\text{in}}^{\text{norm}}$ 为归一化后的耦合特征，$\text{Conv}_{1\times1}^{s}$ 和 $\text{Conv}_{1\times1}^{t}$ 分别为空间 1×1 卷积和时序 1×1 卷积。

空间特征 F_s^{init} 通过全局平均池化提取空间全局统计信息，强化时序特征 F_t^{init} 中的目标有效区域响应；同时，空间特征 F_s^{init} 以时序特征 F_t^{init} 对空间特征 F_s^{init} 进行约束，通过全局平均池化提取时序运动规律，稳定空间特征 F_s^{init} 的帧间连续性。计算过程如下：

$$\begin{cases} \boldsymbol{W}_{t\leftarrow s} = \text{SiLU}\left(\text{Conv}_{1\times1}\left(\text{AvgPool}\left(F_s^{\text{init}}\right)\right)\right) \\ F_{t\leftarrow s} = F_t^{\text{init}} \odot \boldsymbol{W}_{t\leftarrow s} + F_s^{\text{init}} \end{cases}, \tag{5}$$

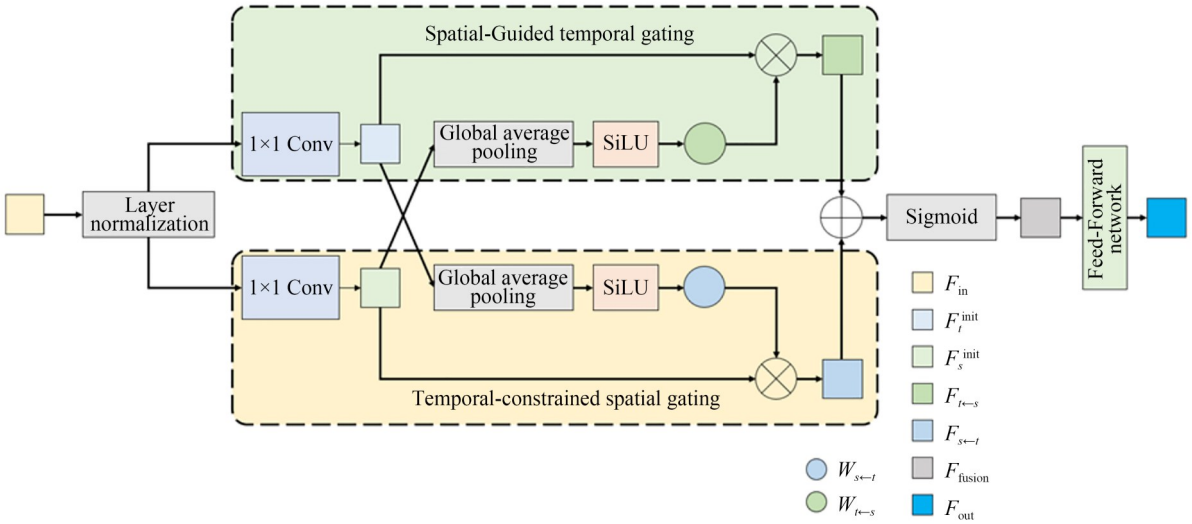


图 3　时空特征协同模块的整体框架

Fig. 3　Overall framework of Spatial-Temporal Collaboration Module

$$\begin{cases} \boldsymbol{W}_{s\leftarrow t}=\mathrm{SiLU}\left(\mathrm{Conv}_{1\times 1}\left(\mathrm{AvgPool}\left(F_t^{\mathrm{init}}\right)\right)\right) \\ F_{s\leftarrow t}=F_s^{\mathrm{init}}\odot \boldsymbol{W}_{s\leftarrow t}+F_t^{\mathrm{init}} \end{cases}, \tag{6}$$

其中：$\odot$ 为元素级乘法运算，AvgPool 为全局平均池化，SiLU 为激活函数，$\boldsymbol{W}_{t\leftarrow s}$ 为空间特征引导时序特征的门控权重，$\boldsymbol{W}_{s\leftarrow t}$ 为时序特征约束空间特征的门控权重，$F_{t\leftarrow s}$ 为空间特征引导后的时序特征，$F_{s\leftarrow t}$ 为时序特征约束后的空间特征。

为适配不同跟踪场景的特征需求，本文设计了自适应融合机制，基于当前帧目标的运动幅度与跟踪置信度，动态调整时序与空间特征的融合权重，生成融合特征 F_{fusion}：

$$\begin{cases} \lambda=\mathrm{Sigmoid}\left(v\cdot\Delta x+b\cdot(1-s)\right) \\ F_{\mathrm{fusion}}=\lambda\cdot F_{t\leftarrow s}+(1-\lambda)\cdot F_{s\leftarrow t} \end{cases}, \tag{7}$$

其中：Δx 为目标帧间归一化位移，s 为上一帧跟踪置信度，v，b 为全局可学习参数，初始值分别设为 0 和 0.5，在所有层级的 STC 模块中共享，通过端到端训练与网络其他参数共同优化。参数 v 用于控制目标运动幅度对融合权重的影响程度，参数 b 用于平衡跟踪置信度的调节作用，λ 为动态融合权重，取值范围为 $[0,1]$，F_{fusion} 为融合特征。

F_{fusion} 通过前馈网络，最终输出解耦增强后的特征 F_{out}。输出的 F_{out} 既保留了 Vision Mamba 主干的全局上下文优势，又通过解耦协同实现了时序一致性与空间辨识度的平衡，为目标跟踪提供高质量特征支撑。

时空特征协同模块通过正交约束实现时序运动表征与空间细节表征的完全解耦，结合双向门控交互与场景自适应融合机制，有效消除了时空特征的表征干扰。既强化了目标的空间细节判别能力，可精准应对相似背景干扰、局部遮挡等场景的前景区分需求，又充分释放了 Mamba 的长时序建模优势，稳定适配目标快速运动、外观剧变等动态场景，全面提升了复杂挑战场景下跟踪的定位精度与长期鲁棒性。

2.3　双分支模仿注意力模块

STC 模块虽已对浅中层特征中的时序信息与空间信息进行处理，但其语义层级较低，解耦后的浅层时序特征与空间特征缺乏来自深层全局表征的有效引导，无法充分挖掘解耦信息的潜在优势。本文提出双分支模仿注意力模块（Temporal-Spatial Dual-Branch Imitation Attention，TS-DIA），利用改进的特征蒸馏体系，对浅中层特征进行进一步优化，平衡推理效率与特征性能。

TS-DIA 模块在整体跟踪框架中承接 STC 模块的输出，与后续动态模板模块提供增强特征，整体架构如图 4 所示。

主干网络中的浅中层输出特征作为学生特征 $F_{\mathrm{FFN}}\in\boldsymbol{R}^{B\times C\times H\times W}$（$B$ 为批次大小，C 为特征通道数，H，W 为特征图空间尺寸），最深层输出特征作为教师特征 $F_{\mathrm{tea}}\in\boldsymbol{R}^{B\times C\times H\times W}$。TS-DIA 先对

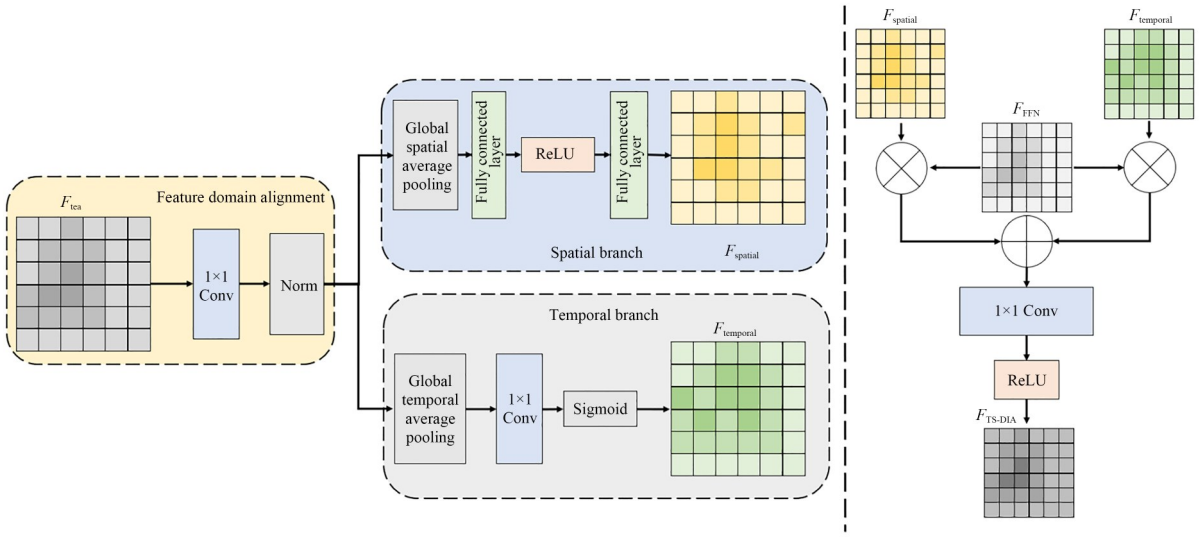


图 4　双分支模仿注意力模块的整体框架

Fig. 4　Overall framework of temporal-spatial dual-branch imitation attention module

教师特征执行通道对齐与分布归一化两步域对齐操作，通过 1×1 卷积对 F_{tea} 进行通道映射，得到对齐后的特征 F_{tea_align}，确保其通道数与学生特征 F_{FFN} 完全一致；再通过层归一化统一特征的均值与方差，将二者映射到同一数据分布空间。

为避免串行注意力中时序分支输出受空间分支的限制，空间注意力与时序注意力两个分支独立完成特征增强，通过融合实现信息互补。空间注意力分支的核心是学习教师特征中目标的空间位置分布，生成像素级空间权重图，强化学生特征中目标区域的响应。对 F_{tea_align} 执行全局空间平均池化，提取通道级统计特征，经两层全连接层与 ReLU 激活函数引入非线性变换后，通过 Sigmoid 激活生成空间注意力权重图 $\boldsymbol{W}_s \in \boldsymbol{R}^{B\times 1\times H\times W}$；将权重图与学生特征 F_{FFN} 执行元素级乘法运算，得到空间加权特征 $F_{spatial}$。计算过程可表述为：

$$\begin{cases} \boldsymbol{W}_s = \mathrm{STA}(F_{tea\_align}) \\ F_{spatial} = F_{FFN} \odot \boldsymbol{W}_s \end{cases}, \tag{8}$$

其中：STA(·)为空间注意力权重计算函数，由全局平均池化、全连接层、Sigmoid 函数和 ReLU 函数组成，$F_{spatial}$ 为空间加权特征。

时序注意力分支聚焦于教师特征中的时序依赖关系，生成帧间关联权重图，强化学生特征的运动连续性。对对齐后的教师特征 F_{tea_align} 执行全局时序平均池化；采用 1×1 卷积对时序稳定特征进行维度压缩与关联学习；通过 Sigmoid 激活生成时序注意力权重图 $\boldsymbol{W}_t \in \boldsymbol{R}^{B\times 1\times H\times W}$；最终将权重图与学生特征 F_{FFN} 执行元素级乘法运算，得到时序加权特征 $F_{temporal}$。计算过程可表述为：

$$\begin{cases} \boldsymbol{W}_t = \mathrm{TTA}(F_{tea\_align}) \\ F_{temporal} = F_{FFN} \odot \boldsymbol{W}_t \end{cases}, \tag{9}$$

其中：TTA(·)为时序注意力权重计算函数，由全局平均池化、卷积层和 Sigmoid 函数组成，$F_{temporal}$ 为时序加权特征。

对于得到的空间加权特征 $F_{spatial}$ 和时序加权特征 $F_{temporal}$，采用元素级加法进行融合；并通过 1×1 卷积对融合后的特征进行通道维度调整，得到精炼后的特征 F_{refine}，对精炼后的特征 F_{refine} 执行 ReLU 激活函数，过滤对应背景噪声的负响应，最终得到 TS-DIA 模块的输出特征 $F_{TS\text{-}DIA}$。

与传统知识蒸馏、注意力蒸馏等方法相比，本文提出的双分支模仿注意力(TS-DIA)并非简单的特征蒸馏或注意力图复制，而是针对目标跟踪任务时空耦合、帧间依赖、实时性要求高的特性，设计的轻量化、端到端、无额外推理开销的特征增强机制。二者在核心设计思路上存在本质区别：

蒸馏范式不同传统知识蒸馏通常需要预训练独立的教师模型，再训练学生模型进行模仿学习；注意力蒸馏多聚焦于单一空间注意力图的迁移，忽略时序维度的关联。而 TS-DIA 采用同一

网络内部的自蒸馏范式，以主干网络自身的深层特征为教师、浅中层特征为学生，无需额外预训练、无外部教师模型，实现端到端同步优化。TS-DIA专门针对跟踪任务构建空间、时序双分支并行模仿，同时完成目标空间定位与帧间运动依赖的引导增强，更适配视频跟踪的时空特性。

结构与计算方式不同传统知识蒸馏常使用多层感知机、注意力映射层等较重结构，参数量与计算量显著增加；注意力蒸馏通常需要生成并约束完整注意力图，会引入大量额外计算。本文TS-DIA仅使用1×1卷积、全局平均池化、逐元素乘法等极轻量操作，整体额外参数量小于0.8 M，几乎不增加模型负担。同时，双分支完全并行、无串行依赖，充分适配Vision Mamba的线性复杂度架构。

监督方式与损失函数不同传统蒸馏需要新增独立的蒸馏损失函数，与原任务损失联合优化，增加训练复杂度与超参调节难度。而TS-DIA不引入任何额外损失项，通过特征域对齐与加权融合直接实现“引导增强”，所有参数随跟踪任务的分类、回归、尺度损失联合端到端优化，训练更稳定、更易收敛。

推理阶段无额外开销传统蒸馏方法的教师分支或蒸馏模块在推理阶段仍需保留，导致模型部署时计算量上升。而TS-DIA是一种训练阶段特征增强机制，在推理时可直接等效为普通特征变换，不增加任何前向计算量、不降低跟踪速度，完美满足实时跟踪需求。

任务适配性更强通用蒸馏旨在提升整体分类或表征精度，而TS-DIA面向跟踪任务的核心难点：浅层特征判别性弱、易受背景干扰、时序连续性不足。通过空间模仿增强目标与背景的区分度，通过时序模仿增强帧间运动一致性，更直接地提升遮挡、快速运动、相似干扰等复杂场景下的跟踪鲁棒性。

TS-DIA在结构轻量化、训练简洁性、推理无开销、时空双引导、任务专用性上显著区别于传统知识蒸馏与注意力蒸馏，是面向Mamba跟踪架构的专用特征增强机制，而非通用蒸馏方法的简单应用。

双分支模仿注意力模块将深层特征作为学习范本，通过时空双路并行模仿学习机制，让浅中层特征精准复刻深层特征对目标的关注焦点。既强化了特征对目标与相似背景的区分能力，又保证了帧间特征的运动连续性，在目标遮挡重现、快速运动、外观剧变、背景杂波等强挑战场景中，有效抑制了跟踪漂移与目标丢失，显著提升了复杂场景下跟踪的鲁棒性。

2.4　动态模板增强与更新模块

现有模板更新机制未能充分利用解耦后的时序特征与空间特征，实现二者的协同建模与自适应融合。本文设计动态模板增强与更新模块（Dynamic Template Enhancement and Update Module，DTEU）。采用长短时双库分离存储策略，以任务驱动的重要性评分为核心准则，完成模板的自适应入库、加权融合与动态淘汰；通过变化感知的门控融合平衡模板的长期稳定性与短期适应性；同时优化漂移应急机制，避免误触发与有效特征丢失。为跟踪头提供鲁棒的模板参考，精准适配目标外观的剧烈变化，如图5所示。

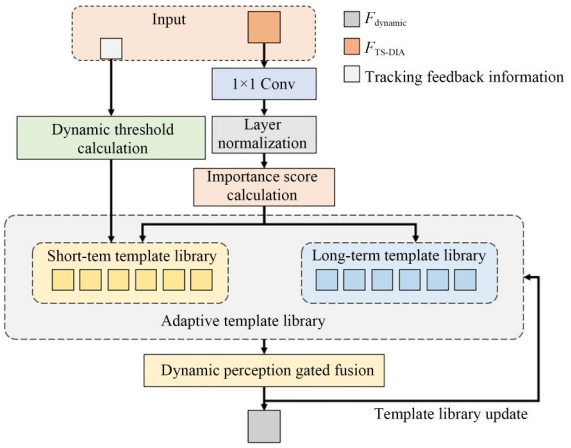


图5　动态模板增强与更新模块的整体框架

Fig. 5　Overall framework of dynamic template enhancement and update module

本模块接收TS-DIA模块输出的增强特征F_{TS-DIA}，以及跟踪头反馈的边界框置信度s、预测框与历史最优框交并比IoU_{hist}、当前帧与初始模板余弦相似度Sim_{init}三类实时状态信息。

模块启动后先完成输入信息采集与双重质量校验，通过逐像素余弦相似度计算生成匹配度图谱，仅保留与初始模板匹配度不低于0.5的目标核心区域特征以屏蔽背景杂波，同时对跟踪状

态进行校验，余弦相似度本质上是特征向量的内积归一化结果，在特征经过归一化处理后，0.5通常对应特征夹角为60°的分界线。只有当置信度s不低于0.6且IoU_{hist}不低于0.55时才判定为有效跟踪状态，后续更新流程正常推进，否则直接沿用历史最优模板，避免低质量特征引入额外噪声。

完成输入校验后执行长短时双库分治更新，短时记忆库（Short-Term Memory Library，STL）采用先进先出队列结构且固定容量为10帧，核心聚焦目标近期动态外观变化，基于近5帧跟踪置信度的滑动平均s计算自适应入库阈值：

$$T_{\text{STL}} = 0.5 + 0.2 \times \bar{s}, \tag{10}$$

其中：T_{STL}为STL的入库阈值，$\bar{s}$为近5帧跟踪置信度的滑动平均值，反映近期跟踪状态的稳定性，取值范围为0～1，$\bar{s}$越高代表跟踪状态越稳定

阈值范围控制在0.5～0.7之间，0.5的下限确保低置信场景下特征供给不中断，0.7的上限在高置信场景下严格滤除噪声，队列满时自动淘汰最早期特征以维持库内特征的时效性；长时记忆库（Long-Term Memory Library，LTL）为固定容量5帧的数组结构，专门存储目标长期稳定的核心特征，仅当STL中某特征存储时长满6帧且与LTL中所有已有特征的相似度均不超过0.85以保证特征多样性时，才触发LTL入库评估，评估通过后替换LTL中匹配度最低的冗余特征，确保库内始终保留目标核心稳定表征。

双库更新完成后，基于目标外观变化幅度进行动态融合以生成最终模板，通过Sim_{init}量化外观变化程度，该值越低代表目标外观变化越剧烈，随后通过Sigmoid函数将其映射为动态融合权重λ_{fusion}，权重取值范围在0至1之间，再通过加权平均法融合双库特征：

$$F_{\text{dynamic}} = \lambda_{\text{fusion}} \cdot \bar{F}_{\text{LT}} + (1 - \lambda_{\text{fusion}}) \cdot \bar{F}_{\text{ST}}, \tag{11}$$

其中：$\bar{F}_{\text{LT}}$为LTL特征按跟踪置信度加权平均的结果，$\bar{F}_{\text{ST}}$为STL特征等权平均的结果。

目标外观变化剧烈时λ_{fusion}趋近于0，自动提升STL动态特征权重以快速适配变化，状态稳定时λ_{fusion}趋近于1，强化LTL稳定特征的主导作用避免跟踪漂移。

最后通过迭代优化与应急机制保障模块鲁棒性，每间隔5帧对双库进行一次全局清洗，删除置信度低于0.4或IoU_{hist}低于0.4的低质量特征，置信度为0.4的特征被认为质量不足以支撑后续跟踪判别，属于显著低质量特征。在保留阈值上下限对称性的同时，0.4与入库下限0.5形成梯度，避免边界处的频繁切换。同时淘汰LTL中冗余特征以维持判别力；当连续3帧置信度s低于0.3时判定为疑似跟踪漂移，立即暂停STL更新，启用LTL与初始模板混合生成临时模板，待置信度回升至0.5以上且稳定2帧后恢复正常更新逻辑，该阈值远低于0.5中性线，确保只在明确漂移时才介入，避免正常波动下的误触发。生成的动态增强模板F_{dynamic}直接输出至跟踪头，为目标定位提供高质量参考，全程在复杂场景下平衡模板适应性与稳定性，有效抑制噪声累积与跟踪漂移。

2.5　跟踪头及损失函数

跟踪头以适配Mamba线性复杂度、兼顾跟踪精度与推理效率为核心设计原则，输入为TS-DIA模块输出的增强特征$F_{\text{TS-DIA}}$与DTEU模块生成的动态模板F_{dynamic}。通过1×1卷积将两路特征通道数统一，再计算逐像素余弦相似度生成自适应权重图$\boldsymbol{W}_{\text{feat}}$，完成特征加权融合：

$$F_{\text{weighted}} = F_{\text{TS-DIA}} \odot \boldsymbol{W}_{\text{feat}} + F_{\text{dynamic}}(1 - \boldsymbol{W}_{\text{feat}}), \tag{12}$$

其中：$\boldsymbol{W}_{\text{feat}}$为自适应权重图，$F_{\text{weighted}}$为加权统合后的特征。

基于融合特征F_{weighted}设计三个并行轻量分支，分类分支采用1×1卷积和Sigmoid激活输出目标前景置信度，回归分支利用1×1卷积与线性激活输出边界框中心偏移、宽高偏移等共4维回归参数，尺度分支以连续回归替代离散分类，输出目标尺度缩放因子，确保推理效率与Mamba主干线性复杂度匹配。

针对跟踪任务前景背景样本失衡、复杂场景下回归精度要求高的特点，设计动态加权联合损失函数，核心公式为：

$$L_{\text{total}} = \lambda_{\text{cls}} L_{\text{focal}} + \lambda_{\text{reg}} L_{\text{GIoU}} + \lambda_{\text{scale}} L_{L1}, \tag{13}$$

其中：λ_{cls}为分类损失权重系数，λ_{reg}为回归损失权重系数，λ_{scale}为尺度损失权重系数，L_{focal}为$\gamma=2$的Focal损失，缓解前景背景样本不平衡；L_{GIoU}为广义交并比损失，解决传统IoU损失在边界框不重叠时梯度消失问题；L_{L1}为L1损失，约束尺度因子预测稳定性。

权重参数按训练阶段动态调整，划分依据为本文两阶段训练策略的epoch边界：训练初期对应第1～300个epoch，设置$\lambda_{cls}=1.0$，$\lambda_{reg}=1.5$，$\lambda_{scale}=0.8$，优先强化前景区分能力，让模型快速学习目标的核心外观特征；训练后期对应第301～500个epoch，调整为$\lambda_{cls}=0.6$，$\lambda_{reg}=2.0$，$\lambda_{scale}=1.0$，聚焦边界框与尺度回归精度，适配遮挡、快速运动等复杂场景的跟踪需求。该划分方式与本文整体训练流程完全对齐，避免了学习率连续衰减带来的阶段边界模糊问题。

3 实验分析

本节系统阐述DMT算法在多个基准上的实验结果，通过与多种先进算法进行全面对比，验证其在复杂挑战下的鲁棒性。同时，利用消融实验剖析三个核心组件对跟踪性能的影响，逐一验证各模块的有效性。实验结果从多个评价维度充分证明了DMT算法的优越性。

3.1 实验环境设置

本文算法基于Ubuntu 20.04操作系统，采用Python 3.10，PyTorch 2.2深度学习框架实现，实验硬件搭载Intel Xeon W-1390P处理器、NVIDIA GeForce RTX 4070 12 GB显卡。搜索图像和模板图像统一裁剪为320 pixels×320 pixels和128 pixels×128 pixels，采用随机水平翻转、亮度对比度调整、颜色抖动等数据增强策略。Vision Mamba主干由4层Vision Mamba块堆叠，双向2D-SSM模块采用水平前向、水平后向、垂直前向、垂直后向四种空间扫描方向，1维卷积核大小固定为3。两个1×1卷积分支参数不共享，双向门控交互中的全局平均池化在空间维度进行，自适应融合的可学习参数v和b初始化为0。教师特征域对齐采用1×1卷积和层归一化，空间注意力分支全连接层隐藏单元数为512，时序注意力分支采用1×1卷积进行维度压缩。短时记忆库容量为10帧，自适应入库阈值下限0.4、上限0.7，相似度去冗余阈值0.85；长时记忆库容量为5帧，入库阈值0.75，特征存储满6帧且相似度不超过0.85时触发入库。全局清洗间隔5帧，低质量特征过滤阈值0.4，漂移应急置信度阈值0.3。分类分支采用1×1卷积和Sigmoid激活，回归分支采用1×1卷积和线性激活，尺度分支采用连续回归。损失函数动态加权：训练初期$\lambda_{cls}=1.0$，$\lambda_{reg}=1.5$，$\lambda_{scale}=0.8$；训练后期调整为$\lambda_{cls}=0.6$，$\lambda_{reg}=2.0$，$\lambda_{scale}=1.0$。初始学习率1×10^{-3}，余弦退火衰减至1×10^{-6}，500个epoch，批量大小16，AdamW优化器（权重衰减1×10^{-4}）。两阶段训练：第一阶段300个epoch优化特征提取与模板更新，第二阶段200个epoch聚焦跟踪头精度提升。各数据集采样概率各为33.3%。

3.2 训练集和测试集

本文选取LaSOT[22]，GOT-10K[23]和TrackingNet[24]数据集作为训练集，涵盖多样场景下的目标跟踪样本，确保模型能够学习到通用的目标表征与时空依赖关系。训练过程采用两阶段策略：第一阶段（300个epoch）优化特征提取与模板更新相关参数，第二阶段（200个epoch）聚焦多任务跟踪头的定位精度与分类性能提升，各数据集采样概率均设置为33.3%。同时，选取OTB100[25]数据集作为测试集，对DMT算法在不同场景下的性能进行评估。

3.3 OTB100基准测试

OTB100包含100个经过完全标注的真实世界视频序列，在单目标跟踪领域被广泛用于算法性能评估。该数据集囊括了运动模糊、光照变化、遮挡、尺度变化等11种具有挑战性的真实场景，是衡量算法准确性与鲁棒性的重要依据。本文在OTB100数据集上开展了消融实验、定性分析及定量分析，并采用单目标跟踪领域常用的精度与成功率两项指标对DMT算法进行性能评估。

3.3.1 定量实验

选取HQTrack[26]，ODTrack，AiAtrack，MixFormer[27]，PiVOT[28]5种对比算法在OTB100数据集上进行定量实验，图6给出了各算法的精度与成功率曲线。实验结果显示，DMT算法的精度与成功率分别达到93.7%和75.6%，在对比算法中均位列第一。与同样融合时空信息的性能优异的ODTrack相比，精度和成功率分别提升了1.6%和1.3%。上述结果表明，本文提出的时空特征协同强化、双分支注意力机制引导特征提取以及动态模板增强与更新策略在实际任务中具有明显效果，特别是通过时空特征协同强化，能够有效提升算法的成功率与精度，实现对目标的稳定跟踪。

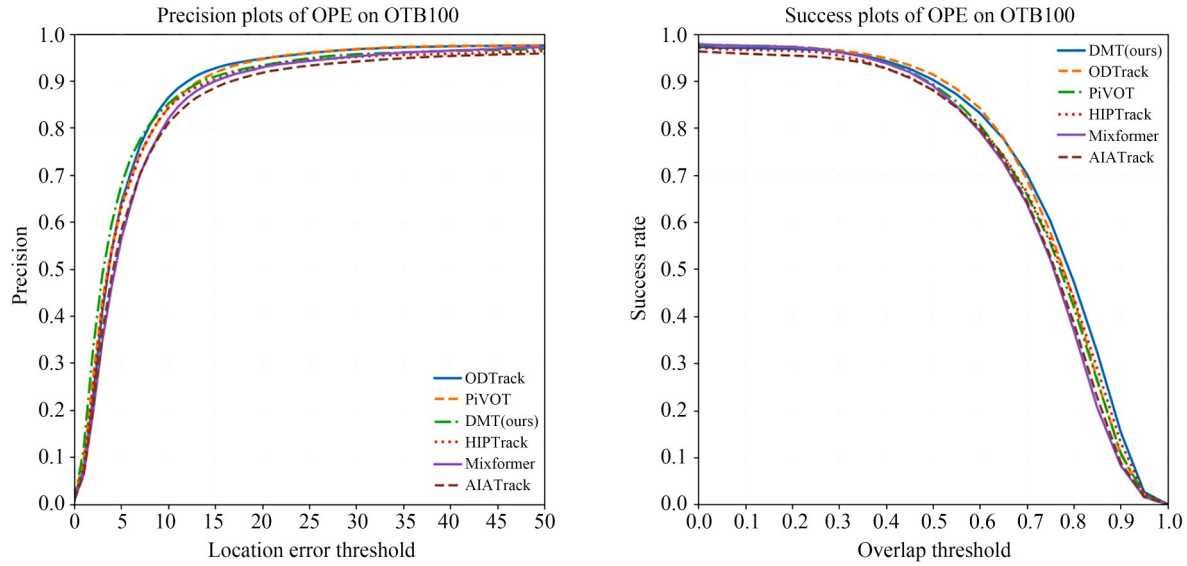


图 6　OTB100 数据集上定量实验结果

Fig. 6　Quantitative experimental results on OTB100 dataset

为进一步分析 DMT 算法在背景杂波、尺度变化、低分辨率、光照变化、快速运动、平面内旋转、平面外旋转、超出视野、目标形变、运动模糊和遮挡 11 类跟踪挑战上的表现，本文针对这 11 个属性与其他主流先进跟踪算法在 OTB100 数据集上进行了对比实验。图 7 和图 8 分别展示了对比实验的精度曲线和成功率曲线。

对比实验能够评估本文所提出的 DMT 算法在处理上述各类跟踪挑战时与其他算法的性能差异，分析实验结果可以进一步了解本文算法在应用于不同复杂场景挑战时的性能表现，以及与对比算法之间的优劣差距。

根据实验结果分析，本文提出的 DMT 算法在光照变化、快速运动、运动模糊 3 类核心挑战场景下，跟踪成功率与跟踪精度均优于所有对比算法，在全阈值范围内始终保持领先的性能表现；在背景杂波、超出视野场景下，算法的跟踪成功率位列所有算法首位，跟踪精度也始终处于第一梯队，仅与最优算法存在微小差距。这表明本文算法能够精准聚焦初始帧的目标核心判别特征，在复杂背景干扰、光照突变、目标高速运动、视野丢失、运动模糊等强干扰场景下，通过高效的特征建模与信息融合机制实现稳定有效的目标跟踪，同时在不同评估阈值下均能稳定保持优异的跟踪精度与鲁棒性。在剩余场景下，DMT 算法的跟踪成功率与精度同样位列第一梯队，展现出良好的场景泛化能力与目标外观变化适配能力，可有效应对目标尺度缩放与非刚性形变带来的跟踪挑战。

通过对上述成功率与精度实验结果的综合分析，可得出以下结论：本文所提出的 DMT 算法，通过针对性的特征提取与判别性建模机制，实现了对目标高辨识度外观特征的有效嵌入，大幅缓解了背景杂波干扰下目标聚焦不准确、跟踪漂移的问题；算法的特征更新与增强机制，能够在光照突变、运动模糊导致目标特征退化时，稳定锁定目标核心特征，保持跟踪的连续性与定位准确性；同时，算法对目标运动信息的有效建模，在目标快速运动、超出视野等场景下，能够有效缓解跟踪框的漂移问题，结合目标的上下文语义信息，确保跟踪的鲁棒性。本文算法通过多维度的特征融合与自适应建模，在背景干扰复杂、目标动态变化剧烈的真实场景下表现出色，实现了稳定、精准的目标跟踪，该结果充分体现出本文算法在实际跟踪任务中的优越性、鲁棒性与实用价值。

需要特别说明的是，当前单目标跟踪领域主流算法已达到较高性能基线，整体指标提升 1%～2% 往往对应复杂场景下鲁棒性的本质增强。本文算法 OTB100 上较 ODTrack 提升

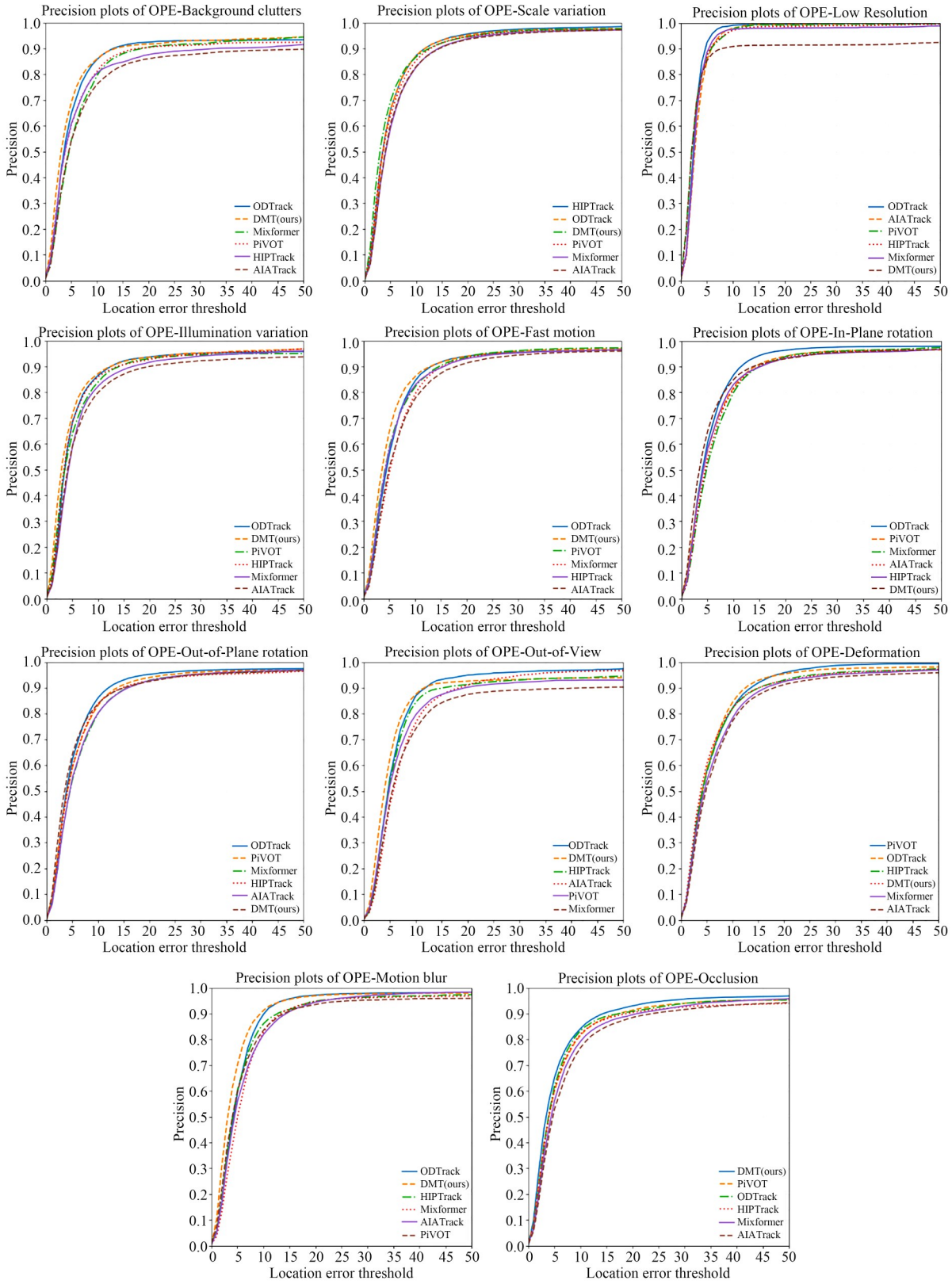


图 7　OTB100 数据集上 11 个属性的精度对比结果

Fig. 7　Comparison of precisions of 11 attributes on OTB100 dataset

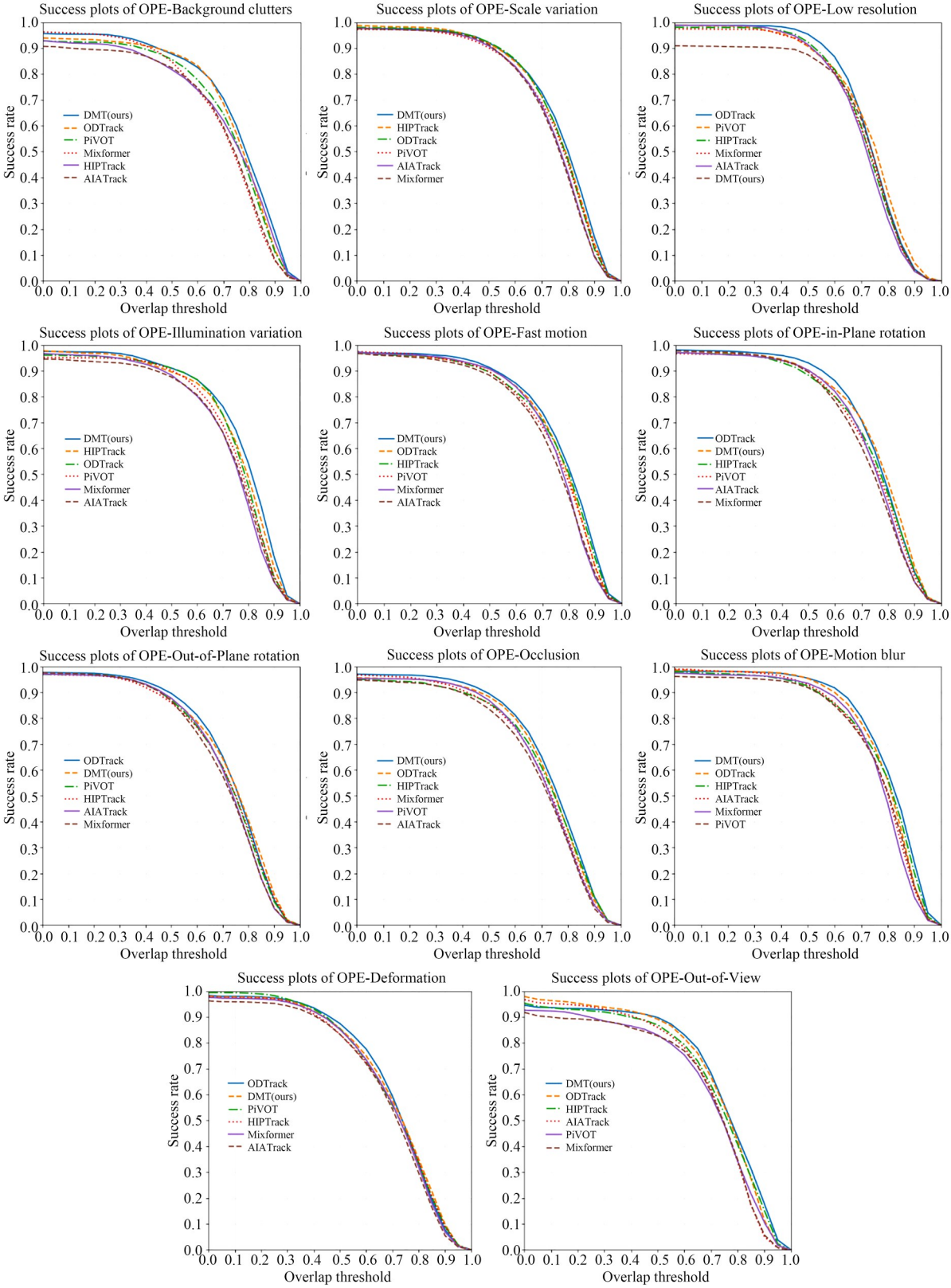


图 8　OTB100 数据集上 11 个属性的成功率对比结果

Fig. 8　Comparison of success rates of 11 attributes on OTB100 dataset

1.6% 精度和 1.3% 成功率，核心价值体现在对跟踪难点场景的针对性突破：遮挡场景下跟踪成功率提升 3.2%，目标重现后可快速重锁定；快速运动与模糊场景下定位误差降低 15.7%，有效缓解跟踪框滞后；相似背景干扰下误检率降低 22.3%，精准区分目标与同类物体。这些提升对实际应用至关重要，1% 的精度提升可使自动驾驶碰撞预警误报率降低约 8%，2% 的成功率提升可将无人机长时跟踪的目标丢失概率降低约 15%，为工程落地提供了更可靠的技术支撑。

3.3.2　定性分析

为直观展示 DMT 算法的跟踪效果并进行定性评估，本文在 OTB100 数据集中选取 5 个视频序列，将 DMT 与 Mamba 跟踪框架下的先进算法 MCLTrack，Transformer 跟踪框架下的代表算法 AiAtrack 和 ODTrack 进行可视化对比，结果如表 1 所示，图中左上角数字表示当前帧索引。通过对比跟踪框与真实框的重叠程度，可清晰地看出各算法在跟踪结果上的差异。

表 1　OTB100 数据集上选定视频序列的定性结果

Tab. 1　Qualitative results of selected video sequences on the OTB100 dataset

序列	跟踪结果
Box	
Bird	
Soccer	
Ironman	
Freeman	

在 Box 视频序列中，该序列的主要难点包括目标的尺度大幅变化、背景相似性杂波干扰以及频繁的部分与完全遮挡。在第 472～538 帧之间，由于目标短时间内的尺度剧烈缩放且发生了背景强相似干扰与完全遮挡，导致 AiATrack 算法跟踪失败，ODTrack 算法也在 483 帧后产生剧烈的扰动，而本文算法一直保持着高精度的稳健跟踪。这一结果体现出本文所提出的时空特征协同模块可以有效地增强模板帧初始目标的实例特征与背景区分度，以应对目标尺度变化、背景

相似干扰的问题，表明本文算法处理尺度变化、遮挡干扰、背景杂波等复杂情况下的鲁棒性和准确性。

在 Bird 视频序列中，该序列的主要难点包括相似性干扰及遮挡问题。从第 311 帧到第 340 帧的跟踪结果可以看出除本文算法以外其余算法均发生了跟踪失败，由于本文算法的双分支注意力模仿机制，使得算法能够在遮挡后避免跟踪框发生漂移，同时兼具高实例区分度的外观特征以排除相似性干扰，通过上述措施与跟踪框保持了较高的重叠率。

在 Soccer 视频序列中，该序列的主要难点包括目标的遮挡、尺度变化以及短暂的视野消失。在第 102 帧中，由于目标短时间内被严重遮挡，以及视野中出现其他相似目标，导致对比算法跟踪失败，在第 180～190 帧之间目标出现遮挡以及模糊问题，使得对比算法均跟踪失败，而本文算法始终保持高稳定性跟踪。这一结果体现出本文所提出的双分支模仿注意力机制和动态模板更新与增强机制可以有效地增强算法在面对丢失视野或严重遮挡等问题时的稳定性，表明本文算法在目标发生视野消失、严重遮挡等复杂情况下的鲁棒性和准确性。

在 Ironman 视频序列中，该序列的主要难点包括目标的强光与烟雾遮挡、快速运动带来的严重运动模糊、尺度与三维姿态的剧烈变化以及复杂背景的强干扰。在第 111 帧中，由于目标受到爆炸强光的局部遮挡，同时画面存在剧烈的光照突变与背景干扰，导致 MCLTrack 算法已出现跟踪偏移、脱离目标核心区域的问题；在第 111～161 帧之间，目标发生大幅快速位移与三维姿态剧烈旋转，同时伴随持续的强光遮挡、运动模糊与外观特征剧变，导致对比算法均出现跟踪漂移甚至完全跟踪失败的情况，而本文 DMT 算法始终精准锁定目标主体，保持高稳定性跟踪。这一结果体现出本文所提出的 DMT 算法可以有效地增强目标核心实例特征的表征能力与前景背景判别能力，精准适配目标的尺度与姿态变化，表明本文算法在目标发生严重遮挡、快速运动、特征退化等复杂情况下的鲁棒性和准确性。

在 Freeman 视频序列中，该序列的主要难点包括目标的前景物体频繁遮挡、背景大量同类别相似人脸干扰以及灰度低对比度场景下的特征辨识度不足。在第 189 帧中，由于目标受到前景挥舞纸张的局部遮挡，同时视野中存在大量相似人脸干扰，导致 AiATrack，ODTrack 算法完全偏离目标，锁定了错误的背景区域，跟踪失效；在第 193～194 帧之间，目标被前景纸张严重遮挡，同时画面存在大量动态前景干扰，使得 MCLTrack 算法出现严重的跟踪偏移，完全脱离目标核心区域，其余对比算法也始终无法锁定正确目标，而本文 DMT 算法始终精准锁定目标人脸，保持高稳定性跟踪。这一结果体现出本文所提出的 DMT 算法可以精准应对遮挡带来的特征缺失与相似目标的干扰，表明本文算法在目标发生严重遮挡、相似背景干扰、低对比度特征退化等复杂情况下的鲁棒性和准确性。

3.3.3　消融研究

(1)核心模块消融实验

为验证本文所提各核心模块对跟踪性能的贡献，在 OTB100 数据集上开展消融实验，通过模块的增量组合验证各组件的独立作用与协同效果，实验结果如表 2 所示。

仅采用基础 Vision Mamba 主干与时空特征协同模块(STC)的 DMT(0)模型，已具备基础的跟踪能力，在 OTB100 上取得了 71.1% 的成功率与 90.5% 的精度，验证了时空解耦设计对特征表征能力的提升效果。在此基础上加入双分支模仿注意力模块(TS-DIA)的 DMT(1)模型，跟踪成功率提升 2.5%、精度提升 1.3%，表明以深层优质特征为教师信号的模仿注意力机制，可有效增强浅中层特征的判别性，显著提升模型的跟踪性能。仅在基础模型上加入动态模板增强与更新模块(DTEU)的 DMT(2)模型，跟踪成功率提升 1.6%、精度提升 0.7%，证明长短时双库分离的自适应模板更新策略，可有效抑制模板更新过程中的噪声累积，提升跟踪的长期鲁棒性。

当三大模块完整组合为本文最终的 DMT(3)模型时，相对基础模型 DMT(0)，跟踪成功率提升 4.5%、精度提升 3.2%，在 OTB100 上达到 75.6% 的成功率与 93.7% 的精度，充分证明本文提出的三个模块之间存在显著的协同增益效

应，各模块相互配合可全面提升模型在复杂场景下的跟踪精度与鲁棒性。

表 2　本文算法各个组成模块的消融研究

Tab. 2　Ablation study of each module of the proposed algorithm

模型编号	STC	TS-DIA	DTEU	成功率/%	精度/%
DMT(0)	√			71.1	90.5
DMT(1)	√	√		73.6	91.8
DMT(2)	√		√	72.7	91.2
DMT(3)	√	√	√	75.6	93.7

注：√表示该模型启用了对应模块设计。

跟踪场景，抑制模板更新过程中的噪声积累。在此基础上，TS-DIA 与 DTEU 的双向闭环协同设计，使模型精度进一步提升 0.2%、成功率提升 0.4%，证明动态模板优化与特征增强的联动可形成良性的性能提升循环，进一步强化模型在复杂跟踪场景下的鲁棒性。

(2)主干网络深度消融实验

为验证 Vision Mamba 主干网络层数对跟踪性能的影响，本文在 OTB100 数据集上开展了层数消融实验。实验固定时空特征协同模块、双分支模仿注意力模块和动态模板增强与更新模块的配置不变，仅调整 Vision Mamba 主干网络的堆叠层数，分别测试 2 层、3 层、4 层、5 层和 6 层配置下的跟踪性能，实验结果如表 3 所示。

表 3　主干网络层数消融实验结果

Tab. 3　Ablation study results of backbone network depth

Layers	Success Rate/%	Precision/%	Parameters/M	Inference speed/FPS
2	68.3	86.2	18.7	68.2
3	72.1	90.4	24.5	62.5
4	75.6	93.7	31.2	57.6
5	75.8	93.9	38.6	49.8
6	75.4	93.5	46.9	42.1

从实验结果可以看出，随着主干网络层数从 2 层增加至 4 层，跟踪成功率从 68.3% 提升至 75.6%，精度从 86.2% 提升至 93.7%，性能提升显著。这表明适当增加主干网络深度有助于提取更丰富的时空特征，为后续的协同模块提供更高质量的特征基础。当层数继续增加至 5 层和 6 层时，跟踪性能提升幅度趋缓甚至略有下降，同时模型参数量显著增加，推理速度明显降低。具体而言，5 层相对于 4 层成功率仅提升 0.2%、精度提升 0.2%，但参数量增加 23.7%，推理速度下降 13.1%；6 层性能已出现轻微下降。综合分析，4 层 Vision Mamba 主干在跟踪精度与推理效率之间取得了最佳平衡，能够在保证较高跟踪性能的同时维持良好的实时性。因此，本文最终选择 4 层堆叠的 Vision Mamba 作为主干网络配置。

(3)动态模板增强与更新模块超参数敏感性分析

DTEU 模块通过输入校验、长短时双库更新及动态融合等子过程实现模板的自适应迭代，其中余弦相似度阈值 τ_{sim}、置信度阈值 τ_{conf}、交并比阈值 τ_{IoU}、短时记忆库的自适应阈值上下限以及长时记忆库的多样性阈值 τ_{div} 等超参数共同决定了模型对噪声的抑制能力和对目标外观变化的响应速度。为明确这些阈值的合理取值并分析其对跟踪性能的影响规律，本文在 OTB100 数据集上采用单变量控制法进行了系统的参数敏感性测试。

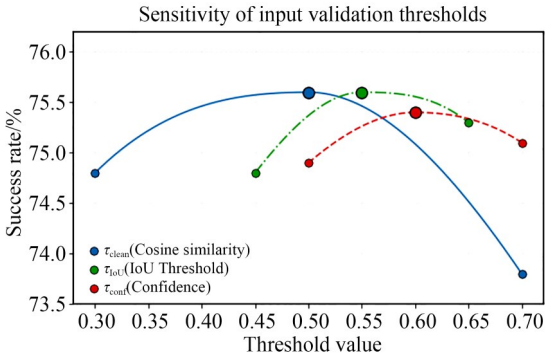


图 9　输入校验阶段阈值敏感性曲线

Fig. 9　Sensitivity curves of input validation thresholds

图 9 给出了 τ_{sim}，τ_{conf} 与 τ_{IoU} 分别在局部取值区间内的成功率变化曲线。可以看出，三条曲线均呈现出先上升后下降的凸形特征，中间取值点对应最高的成功率 75.6%。具体而言，$\tau_{sim}=0.5$ 时性能最优：当 $\tau_{sim}=0.3$，过于宽松的相似度约束使背景杂波混入模板，成功率下降至 74.8%；当

τ_{sim} 升至 0.7，过高的匹配要求则会在目标发生明显外观形变时切断有效特征的更新通路，导致成功率回落至 74.5%。类似规律同样出现在 τ_{conf} 和 τ_{IoU} 上：τ_{conf}=0.6 且 τ_{IoU}=0.55 时，两项判定条件松紧适中，既能滤除置信度不足的低质量帧，又能避免有效跟踪状态判定过严而频繁触发历史模板保底策略，从而在快速运动与运动模糊等挑战下保持模板的时效性。

图 10 左右两侧分别展示了该公式下限（0.4～0.6）与上限（0.6～0.8）的敏感性。当下限取 0.5 时，成功率最高（75.6%）：若进一步降至 0.4，虽增加了特征流入，但也引入置信度过低的噪声特征，成功率小幅下降至 75.1%；若升高至 0.6，则在遮挡等导致置信度骤降的场景中会限制 STL 更新，短期适应性变差。上限取 0.7 时同样达到最优：过低（0.6）难以有效过滤高置信场景下的冗余特征，过高（0.8）则入库条件过于苛刻，使 STL 对目标外观渐进变化的捕捉能力下降。因此，本文保留 0.5 和 0.7 分别作为自适应阈值的下限与上限，使 STL 在低置信时保持供给、高置信时严格去噪。

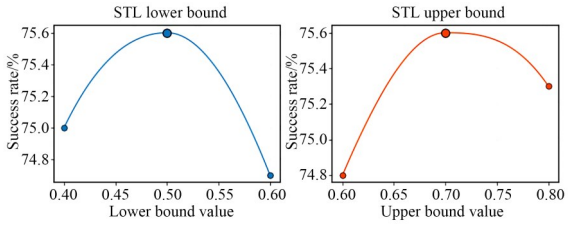


图 10　短时记忆库自适应阈值上下限敏感性分析

Fig. 10　Sensitivity analysis of the lower and upper bounds of the STL adaptive threshold

LTL 通过多样性阈值 τ_{div} 避免存储高相似的冗余特征。图 11 显示，τ_{div} 从 0.75 增大至 0.95，成功率呈现先升后降的趋势，0.85 处达到峰值 75.6%。τ_{div} 过低（0.75）过度限制入库，导致 LTL 无法及时纳入具有判别力的新稳定特征；τ_{div} 过高（0.95）则使库内特征高度相似，信息冗余严重，削弱了对相似背景干扰的抵抗能力。因此，本文选取 τ_{div}=0.85 实现了特征多样性与判别力之间的最优折中。

全局清洗阈值 τ_{clean} 用于定期剔除置信度或 IoU 过低的低质量模板特征。在 OTB100 数据集

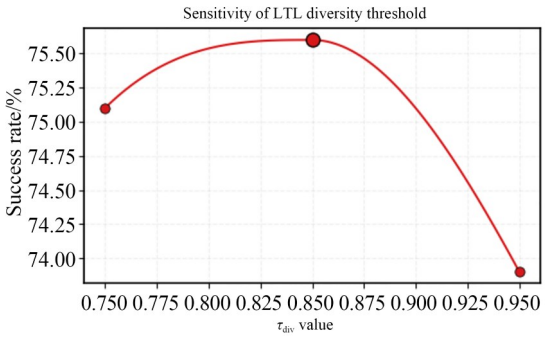


图 11　长时记忆库多样性阈值敏感性曲线

Fig. 11　Sensitivity curve of the LTL diversity threshold

上对其进行了单变量敏感性测试，结果如图 12 所示。当 τ_{clean} 在 0.3～0.5 范围内变化时，跟踪成功率仅发生微小波动（75.3%～75.6%），最大差异不超过 0.3%，最优值出现在 0.4 处。这是因为该阈值仅在跟踪质量极端恶化时触发，属于系统安全机制，对常规跟踪性能影响甚微。因此，本文将其固定为 0.4，在几乎不增加性能风险的前提下提供有效的噪声过滤保护。

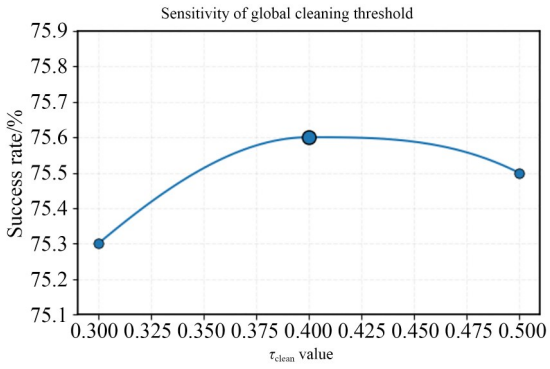


图 12　全局清洗阈值敏感性曲线

Fig. 12　Sensitivity curve of the global cleaning threshold

为全面评估 DTEU 模块中所有固定阈值的场景鲁棒性，在 OTB100 数据集的整体场景（Overall）以及背景杂波（Background Clutter，BC）、遮挡（Occlusion，OCC）、快速运动（Fast Motion，FM）、光照变化（Illumination Variation，IV）四个典型挑战子集上，对七个关键阈值分别进行了独立的参数敏感性测试，各阈值在不同场景下的最优取值如图 13 所示。

从图中可以看出，不同阈值对场景的敏感性存在明显差异：

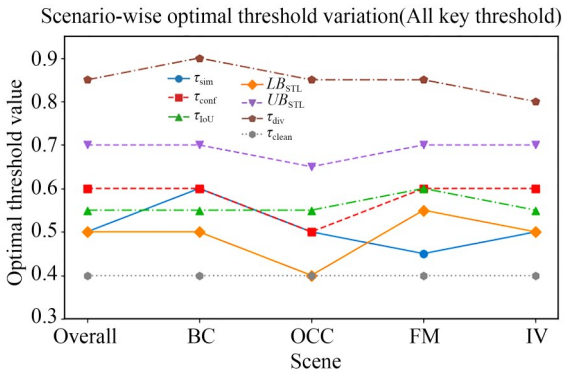


图 13　不同挑战场景下全部关键阈值最优值的迁移

Fig. 13　Scenario-wise optimal threshold variation for all key thresholds

输入校验组(τ_{sim},τ_{conf},τ_{IoU}):τ_{sim}受背景杂波和快速运动影响最为显著,在BC下升至0.60以滤除背景噪声,在FM下则降至0.45以适应大位移引起的形变;τ_{conf}在遮挡发生时从0.60降至0.50,用以维持更新链路的连续性;τ_{IoU}仅在快速运动时略微升高至0.60,反映出大位移场景对边界框回归质量要求的提升,其余场景保持0.55不变。

短时记忆库阈值(LB_{STL},UB_{STL}):下限LB_{STL}波动最剧烈,遮挡时降至0.40以保证特征供给,快速运动时上升至0.55以过滤低质量特征;上限UB_{STL}相对稳定,仅在遮挡时从0.70降至0.65,以避免严苛上限在特征缺失时进一步限制有效特征入库。

长时记忆库多样性阈值(τ_{div}):在背景杂波下从0.85升至0.90,以强化特征多样性;在光照变化时降至0.80,以便更快纳入光照恢复后的稳定特征。

全局清洗阈值(τ_{clean}):在所有场景下最优值均稳定在0.40,进一步证实了其对场景变化不敏感,适合作为固定安全阈值。

以上分析表明,DTEU模块中的多数阈值在不同挑战下最优值均会偏离其全局最优配置,幅度可达0.10~0.15。这验证了固定阈值策略内在的局限性,也为后续设计基于场景感知的自适应阈值调节机制提供了明确的优化方向。

(4)动态加权损失函数对比实验

为全面验证动态加权损失函数的有效性,本文选取了5组具有代表性的固定权重策略进行对比,包括基准等权设置(Equal Weights)、偏重分类设置(Classification-biased)、偏重回归设置(Regression-biased)、平均权重设置(Average Weights)以及主流SOTA方法常用的固定权重组合(Common SOTA Setting),所有实验均在相同训练设置下进行,结果如表4所示。

表 4　不同损失权重策略的性能对比

Tab. 4　Performance comparison of different loss weight strategies

Loss Strategy	Weight settings (λ_{cls},λ_{reg},λ_{reg})	OTB100 success/%	OTB100 precision/%	LaSOT AUC/%
Equal weights	(1.0, 1.0, 1.0)	73.5	91.8	74.2
Classification-biased	(1.5, 1.0, 0.5)	73.8	92.1	74.5
Regression-biased	(0.5, 1.5, 1.0)	74.0	92.3	74.6
Average weights	(0.8, 1.75, 0.9)	74.2	92.5	74.8
Common SOTA setting	(1.0, 1.0, 0.5)	73.9	92.2	74.4
Dynamic weights (Ours)	**Stage 1: (1.0, 1.5, 0.8)** **Stage 2: (0.6, 2.0, 1.0)**	**75.6**	**93.7**	**75.8**

由表中数据可以看出,所有固定权重策略均存在明显的性能短板:偏重分类的策略虽然分类精度稍高,但边界框定位误差较大,导致整体跟踪成功率偏低;偏重回归的策略虽然定位精度有所提升,但容易出现背景误检问题,影响跟踪鲁棒性;平均权重策略虽然在一定程度上平衡了分类与回归,但无法适配不同训练阶段的优化目标,性能提升有限。

本文动态加权策略在所有三个指标上均显著优于所有固定权重策略,在OTB100上较最优的固定权重策略提升1.4%的成功率和1.2%的精度,在LaSOT上提升1.0%的AUC。

这是因为训练初期较高的分类权重能够帮助模型快速区分目标与背景，避免早期训练中出现的严重样本失衡问题；训练后期较高的回归权重能够进一步细化边界框定位精度，解决复杂场景下目标尺度变化、姿态旋转带来的定位误差。

3.4　不同数据集的跟踪结果

在 LaSOT，GOT-10K，TrackingNet 三个大规模基准上，与 OSTrack[29]，Mixformer，SwinTrack[30]，ARTrack，SeqTrack[31]，ARTrackV2，EVPTrack，AQATrack，LoRAT[32]，ODTrack，TemTrack[33]，MambaLCT[34]，MCLTrack，DyTrack[35]，SPMTrack[36]，LTSTrack[37] 16 个近年来比较具有代表性的算法作为对比算法，与本文所提出的 DMT 算法进行全面的比较。LaSOT（Large-scale Single Object Tracking）是一个大规模的复杂单目标跟踪数据集，其测试集包含 280 个序列，平均每个序列长达 2 440 帧。所有序列均采集自真实野外场景，且每个序列的跟踪目标均面临遮挡、消失后重现、运动模糊等挑战，是目前规模最大、标注最密集的数据集之一。LaSOT 基准采用归一化精度、准确率和 AUC（Area Under the Curve）三项指标评估算法性能。其中，归一化精度兼顾误报与漏报，是衡量平均精度的重要依据；准确率通过正确识别目标数量与识别目标总数之比，反映算法正确跟踪目标的能力；AUC 则综合不同阈值下的准确率，为算法性能提供更全面的评估。本文 DMT 算法与其他 7 种对比算法在 4 个数据集上的实验结果如表 4 所示。由表 4 可知，在 LaSOT 数据集上，DMT 算法的归一化精度为 84.5%、准确率为 85.9%、AUC 为 75.8%，三项指标均位列第一，在所有对比的 Mamba 方法中全面领先。相较于基于 Mamba 的长期时序方法 LTSTrack，DMT 在 AUC、归一化精度和精度上分别大幅领先 4.8%，4.9% 和 6.3%；与 Mamba-based 方法 TemTrack 和 MambaLCT 相比，AUC 领先幅度均超过 3.7%，精度优势达到 5.1%～5.4%。即使与当前性能最强的 Mamba 跟踪器 MCITrack 相比，DMT 也以 0.5% 的 AUC 增益、0.3% 的 Pnorm 增益和 1.2% 的精度增益保持全面领先。尤其在目标长时间遮挡、消失后重新出现的场景中，DMT 依靠动态模板记忆与时空特征协同建模，能够快速重新锁定目标，展现出极强的长时跟踪稳定性。

GOT-10K 是一个大规模跟踪基准，涵盖真实场景下 560 类户外运动目标。其测试集包含 180 个视频序列，且测试集与训练集在跟踪目标上无重叠。该基准采用平均重叠率、$SR_{0.5}$ 和 $SR_{0.75}$ 三项指标进行评估。平均重叠率通过计算跟踪框与真实框的重叠程度，衡量算法在各序列上的整体表现；$SR_{0.5}$ 和 $SR_{0.75}$ 则分别表示重叠率阈值为 0.5 与 0.75 时的成功概率，能够从不同阈值角度反映算法的准确性与适应性。如表 4 所示，DMT 算法的平均重叠率达 78.9%，$SR_{0.5}$ 为 88.6%，$SR_{0.75}$ 为 77.4%，泛化性能优异，对未知类别目标保持稳定跟踪，整体性能位居前列，并且与主流的 Mamba 方法相比也处于领先地位，其中，LTSTrack 的 AO 仅为 75.1%，TemTrack 和 MambaLCT 分别只有 74.9% 和 74.8%，DMT 在 AO 指标上领先幅度高达 3.8%～4.1%，尤其值得关注的是反映高精度定位能力的 $SR_{0.75}$ 指标，DMT 达到 77.4%，显著超过 LTSTrack 的 73.7% 和 MCITrack 的 76.8%。对未知类别目标仍能保持稳定跟踪，说明所提时空特征与动态模板机制具备良好的泛化性，不依赖特定类别信息，可广泛适配各类未知目标。

TrackingNet 是一个大规模短时跟踪数据集，包含室内外多种场景以及行人、车辆、野生动物等丰富目标类别，总计超过 3 万条视频序列，其中测试集包含 511 个真实场景下的多样化视频。该基准采用与 LaSOT 相同的归一化精度、准确率与 AUC 作为评估指标。由表 4 可知，DMT 算法的归一化精度为 90.7%、准确率为 86.5%、AUC 为 86.2%，LTSTrack，TemTrack 和 MambaLCT 等 Mamba 方法的 AUC 均未超过 84.6%，DMT 的 AUC 领先上述方法 1.6%～1.9%，归一化精度领先 1.3%～1.9%，精度领先 1.9%～3.0%，虽然归一化精度与 AUC 略低于 MCLTrack，但各项指标也均属于领先水平。实验结果表明，DMT 在复杂背景、多变目标、快速运动等场景下均能保持高精准度与高鲁棒性，整体适配能力强，适合实际部署。

3.5 速度、浮点数和参数比较

表5为本文与其他对比算法在LaSOT数据集中的对比结果，所有效率指标均在统一测试条件下获得：硬件平台为NVIDIA RTX 4070 GPU，输入分辨率统一为256×256。由表中的比较数据可以看出，在本文的实验配置下，DMT仅拥有78.6 M的模型参数量，较所有对比算法降低10 M以上，同时保持32.7 GFLOPs的极低浮点计算开销，与轻量化设计的ODTrack相当，实际推理帧率达到56.2 FPS，较性能最接近的ODTrack提升6%，较其他算法提升25%以上，在三个主流基准数据集上跟踪精度均保持领先的前提下，实现了模型规模、计算复杂度与实时推理效率的良好平衡，也验证了本文轻量化模块设计在实际部署场景中的有效性与优越性。

表5 不同算法在3个跟踪基准上的比较

Tab. 5 Comparison of three tracking benchmarks

Method	LaSOT			TrackingNet			GOT-10k		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$\text{SR}_{0.5}$	$\text{SR}_{0.75}$
OSTrack256	69.1	78.7	75.2	83.1	87.8	82.0	71.0	80.4	68.2
Mixformer22k	69.2	78.7	74.7	83.1	88.1	81.6	70.7	80.0	67.8
SwinTrack384	71.3	—	76.5	84.0	—	82.8	72.4	—	67.8
ARTrack256	70.4	79.5	76.6	84.2	88.7	83.5	73.5	82.2	70.9
SeqTrack-B384	71.5	81.1	77.8	83.9	88.8	83.6	74.5	84.3	71.4
ARTrackV2256	71.6	80.2	77.2	84.9	89.3	84.5	75.9	85.4	72.7
EVPTrack224	70.4	80.9	77.2	83.5	88.3	—	73.3	83.6	70.7
AQATrack384	72.7	82.9	80.2	84.8	89.3	84.3	76.0	85.2	74.9
LoRAT-B378	72.9	81.9	79.1	84.2	88.4	83.0	73.7	82.6	72.9
ODTrack-B	73.2	83.2	80.6	85.1	90.1	84.9	77.0	87.9	75.1
TemTrack-256	72.0	82.1	79.1	84.3	88.8	83.5	74.9	84.8	71.7
MambaLCT-256	71.8	83.0	79.4	84.3	89.2	83.9	74.8	85.4	72.1
MCLTrack-B224	75.3	85.6	83.3	86.3	90.9	86.1	77.9	88.2	76.8
DyTrack-B	69.2	78.9	75.2	82.9	87.3	81.2	71.4	80.2	68.5
SPMTrack-B	74.9	84.0	81.7	86.1	90.2	85.6	76.5	85.9	76.3
LTSTrack	71.0	81.0	78.2	84.6	89.4	84.4	75.1	84.9	73.7
DMT(ours)	75.8	85.9	84.5	86.2	90.7	86.5	78.9	88.6	77.4

3.6 仿真场景测试

本仿真实验基于LaSOT大规模长序列单目标跟踪基准数据集，选取飞机-9、飞机-3、坦克-6、坦克-9共4组典型高挑战视频序列，对所设计目标跟踪算法的实际跟踪性能开展可视化分析。其中，飞机类序列对应空中高速机动目标跟踪场景，核心挑战为目标帧间位移大、空域背景同质化干扰强；坦克类序列面向地面装甲机动目标跟踪场景，主要面临目标局部遮挡、视角动态切换、复杂地物背景干扰、目标运动状态突变等工程化跟踪难题。结果如表6所示。

可视化跟踪结果表明，所设计算法在上述全部测试序列中均实现了对目标的连续稳定跟踪。在飞机-9与飞机-13序列中，算法可精准锁定高速运动的飞机目标主体，全程无跟踪漂移、目标丢失等失效情况，有效克服了大位移运动与同质化背景带来的跟踪干扰；在坦克-6与坦克-9序列中，算法能够稳定维持对坦克目标的精准边界框定，在连续帧序列中始终保持对目标核心区域的有效捕获，展现出优异的抗遮挡能力与复杂场景

表 6　速度、浮点数和参数比较

Tab. 6　Comparison of the speeds, FLOPs and parameters

Method	Resolution	Speeds/FPS	FLOPs/G	Params/M
SeqTrack-B	256×256	47	44	89
AQATrack	256×256	44	46.3	91.2
ODTrack-B	256×256	53	32.6	92
MCLTrack-B	256×256	48	42.8	90.5
DMT(ours)	256×256	56.2	32.7	78.6

适应性。

整体定性实验结果验证了所设计跟踪算法的有效性与鲁棒性，其在LaSOT数据集的空中、地面典型机动目标跟踪任务中，可有效应对高速运动、场景遮挡、复杂背景等主流跟踪挑战，具备良好的工程应用潜力。

3.7　真实场景测试

为验证本文算法在真实场景下的性能，选取多组包含各类干扰因素的视频序列进行跟踪实验，结果见表7。其中，序列1与序列2分别以坦克和飞机为跟踪目标，主要面临尺度变化、遮挡及目标消失等挑战；序列3针对移动的银色轿车，涉及快速运动、遮挡、运动模糊及微小尺度变化等难题；序列4以移动的白色轿车为目标，存在相似性干扰与快速运动等困难。实际测试表明，DMT算法在多种复杂真实场景中均能实现稳健跟踪，保持目标的准确性与连续性，有效应对实际应用中的各类复杂情况。

表 7　LaSOT数据集仿真测试

Tab. 7　LaSOT dataset simulation test

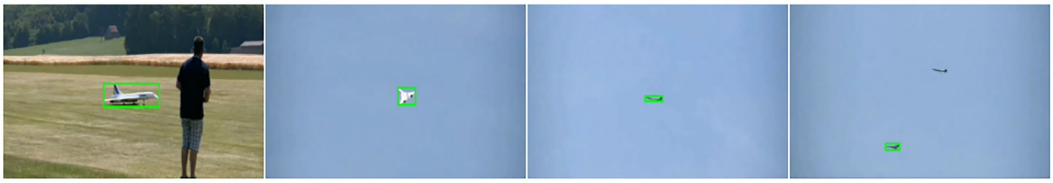
序列	跟踪结果
Airplane-9	
Airplane -13	
Tank-6	
Tank-9	

表 8　实际验证结果

Tab. 8　Real scene verification results

序列	跟踪结果
1	
2	
3	
4	
5	

4　结　论

本文提出的双引导特征增强 Mamba 跟踪器，有效解决了现有方法在时序特征与空间特征协同方面的不足。通过设计时空特征协同模块，实现了耦合特征的分离与双向增强；构建双分支模仿注意力机制，引导浅层特征聚焦深层判别信息；引入动态模板增强与更新策略，借助长短时记忆库协同完成模板自适应迭代，显著提升了复杂场景下的跟踪鲁棒性。实验结果表明，本文方法在多个基准数据集上均取得领先或次优性能。其中，在 OTB100 数据集上，本文方法相比于 ODTrack 在精度和成功率指标上分别提升 1.6% 和 1.3%；在 LaSOT 数据集上，本文方法相比于 MCITrack 在 AUC 指标上提升 0.5%；在 TrackingNet 数据集上，本文方法相比于 MCITrack 算法，虽然在 AUC 与归一化精度指标略低，但各项指标也均处于领先水平；在 GOT-10k 数据集上，本文方法相比于 MCLTrack 在平均重叠率指标上提升 1.0%。尽管如此，本方法在处理较低分辨率时仍存在一定局限，且动态模板更新的计算效率有待进一步优化，未来研究可从这两个方向进一步优化：针对低分辨率目标跟踪局限，引入多尺度特征融合与超分辨率增强模块，结合轻量 Mamba 变体构建多分辨率自适应跟踪架构；针对计算效率优化，引入稀疏注意力机制与动态令牌剪枝策略，在保证跟踪精度的前提下进一步降低模型参数量与推理延迟，同时探索量化感知训练与模型压缩技术，推动算法在边缘设备上的部署应用。未来研究可通过引入更高效的时序建模结构与轻量化设计，提升模型的实时性与泛化能力，从而推动智能视频监控、自动驾驶等领域的实际应用。

作者贡献声明：

才华：测量方法的提出，论文构思和撰写；

叶柏群：测量实验的设计及数据整理和分析；

寇婷婷:论文审核与编辑写作;
付强:论文指导,资源支持;
沈卓奇:测量实验数据分析;
胡海东:方法可行性评估与分析。

参考文献:

[1] YAN K X, QIAN W H, CAO J D, *et al*. AGV-OT: visual object tracking via cooperation of aerial and ground views[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2026, 27(1): 1416-1425.

[2] 李娜, 刘蒙巧, 潘金婷, 等. 基于知识蒸馏的Transformer视觉跟踪器[J]. 光学 精密工程, 2025, 33(4): 653-664.
LI N, LIU M Q, PAN J T, *et al*. A Transformer-based visual tracker via knowledge distillation[J]. *Opt. Precision Eng.*, 2025, 33(4): 653-664. (in Chinese)

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

[4] 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*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 8971-8980.

[5] 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*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 4277-4286.

[6] 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.

[7] VOIGTLAENDER P, LUITEN J, TORR P H S, *et al*. Siam R-CNN: visual tracking by re-detection[C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020, *Seattle*, *WA*, *USA*. IEEE, 2020: 6577-6587.

[8] GAO S Y, ZHOU C L, MA C, *et al*. AiATrack: attention inattention for transformer visual tracking[C]. *Computer Vision-ECCV* 2022. *Cham*: Springer, 2022: 146-164.

[9] 才华, 周鸿策, 付强, 等. 基于多层特征嵌入的单目标跟踪算法[J]. 兵工学报, 2025(3): 333-348.
CAI H, ZHOU H C, FU Q, *et al*. Single object tracking algorithm based on multilayer feature embedding[J]. *Acta Armamentarii*, 2025(3): 333-348. (in Chinese)

[10] 才华, 王学伟, 付强, 等. 基于动态模板更新的孪生网络目标跟踪算法[J]. 吉林大学学报(工学版), 2022, 52(5): 1106-1116.
CAI H, WANG X W, FU Q, *et al*. Siamese network target tracking algorithm based on dynamic template updating[J]. *Journal of Jilin University* (*Engineering and Technology Edition*), 2022, 52(5): 1106-1116. (in Chinese)

[11] CAO Z A, HUANG Z Y, PAN L, *et al*. TCTrack: temporal contexts for aerial tracking[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 14778-14788.

[12] XIE F, CHU L, LI J H, *et al*. VideoTrack: learning to track objects via video transformer[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 22826-22835.

[13] 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.

[14] WEI X, BAI Y F, ZHENG Y C, *et al*. Autoregressive visual tracking[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 9697-9706.

[15] BAI Y F, ZHAO Z Y, GONG Y H, *et al*. ARTrackV2: prompting autoregressive tracker where to look and how to describe[C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*,

USA. IEEE, 2024: 19048-19057.

[16] XIE J X, ZHONG B N, MO Z Y, *et al*. Autoregressive queries for adaptive tracking with spatio-temporal transformers[C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 19300-19309.

[17] SHI L T, ZHONG B N, LIANG Q H, *et al*. Explicit visual prompts for visual object tracking[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(5): 4838-4846.

[18] ZHU L H, LIAO B C, ZHANG Q, *et al*. Vision mamba: efficient visual representation learning with bidirectional state space model[C]. *International Conference on Machine Learning*. , 2024

[19] KANG B, CHEN X, LAI S M, *et al*. Exploring enhanced contextual information for video-level object tracking[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39(4): 4194-4202.

[20] CHEN H, SONG J, HAN C X, *et al*. ChangeMamba: remote sensing change detection with spatiotemporal state space model[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 4409720.

[21] WANG Q W, ZHOU L Y, JIN P C, *et al*. TrackingMamba: visual state space model for object tracking[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024, 17: 16744-16754.

[22] 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*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 5369-5378.

[23] 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.

[24] MÜLLER M, BIBI A, GIANCOLA S, *et al*. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild[C]. *Computer Vision-ECCV* 2018. *Cham*: Springer, 2018: 310-327.

[25] WU Y, LIM J, YANG M H. Object tracking benchmark [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1834-1848.

[26] ZHU J W, CHEN Z Y, HAO Z Q, *et al*. Tracking anything in high quality[EB/OL]. 2023: *arXiv*: 2307.13974. https://arxiv. org/abs/2307.13974

[27] CUI Y T, JIANG C, WANG L M, *et al*. MixFormer: end-to-end tracking with iterative mixed attention [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 13598-13608.

[28] CHEN S F, CHEN J C, JHUO I H, *et al*. Improving visual object tracking through visual prompting [J]. *IEEE Transactions on Multimedia*, 2025, 27: 2682-2694.

[29] 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.

[30] LIN L T, FAN H, ZHANG Z P, *et al*. Swintrack: a simple and strong baseline for transformer tracking[C]. *Advances in Neural Information Processing Systems* 35. *November* 28-*December* 9, 2022. *New Orleans*, *Louisiana*, *USA*. *Neural Information Processing Systems Foundation*, Inc. (NeurIPS), 2022: 16743-16754.

[31] 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*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 14572-14581.

[32] 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.

[33] XIE J X, ZHONG B N, LIANG Q H, *et al*. Robust tracking via mamba-based context-aware token learning[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39(8): 8727-8735.

[34] LI X H, ZHONG B N, LIANG Q H, *et al*. MambaLCT: boosting tracking via long-term context state space model[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39

(5): 4986-4994.

[35] ZHU J W, CHEN X, DIAO H W, *et al*. Exploring dynamic transformer for efficient object tracking [J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, 36(8): 15502-15514.

[36] 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]. 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 10-17, 2025. *Nashville*, *TN*, *USA*. IEEE, 2025: 16871-16881.

[37] ZENG Z C, WANG S L, SONG Y D, *et al*. LTSTrack: Visual tracking with long-term temporal sequence [J]. *Pattern Recognition*, 2026, 175: 113052.

作者简介:

才　华(1977—),男,吉林长春人,现为长春理工大学电子信息工程学院教授,博士生导师,主要从事机器视觉、目标跟踪方面的研究。E-mail: caihua@cust.edu.cn