动态综述

DOI：10.16818/j.issn1001－5868.20251105001

基于深度学习的视觉目标跟踪研究综述

罗　元，马文龙*，唐小平

（重庆邮电大学 电子科学与工程学院，重庆 400065）

摘　要： 目标跟踪是计算机视觉领域的重要研究方向，其任务是在视频序列中对感兴趣的目标进行持续检测与精确定位。随着深度学习技术的迅速发展，视觉目标跟踪在特征建模、时序关联以及端到端学习等方面取得了显著进展。文章系统回顾了基于深度学习的视觉目标跟踪研究现状，从单目标跟踪与多目标跟踪两个方向出发，分析了不同框架下的基本思想与实现机制，阐述了从传统手工特征方法向深度特征驱动模型的演化过程。还总结了常用的跟踪数据集与性能评价指标，并对当前研究面临的主要问题进行了讨论，包括长时间建模能力欠佳、复杂场景跟踪鲁棒性差以及跨场景泛化能力有限等。最后，展望了未来视觉目标跟踪的发展趋势，指出融合多模态信息与时空建模的统一深度框架将成为该领域的重要研究方向。

关键词： 深度学习；视觉目标跟踪；单目标跟踪；多目标跟踪；Transformer

中图分类号： TP242.6　**文章编号：** 1001－5868(2026)01－0013－15

Review of Visual Object Tracking Based on Deep Learning

LUO Yuan, MA Wenlong, TANG Xiaoping

(**School of Electronic Science and Engineering, Chongqing University of Posts and Telecommunications, Chongqing 40065, CHN**)

Abstract: Object tracking, an important research direction in the field of computer vision, aims to continuously detect and accurately locate objects of interest in video sequences. With the rapid development of deep learning technology, visual object tracking has made significant progress in feature modeling, temporal association, and end-to-end learning. This paper systematically reviews the current research status of deep learning-based visual object tracking and analyzes the fundamental concepts and implementation mechanisms under different frameworks from the perspectives of single object tracking and multi-object tracking. The evolution from traditional hand-crafted feature methods to deep feature-driven models is further elaborated. After summarizing commonly used tracking datasets and performance evaluation metrics, the paper discusses the main challenges in current research, including limited long-term modeling capabilities, weak robustness in complex scene tracking, and poor cross-scene generalization. Finally, future research trends in visual object tracking highlight the development of unified deep framework that integrates multimodal information and spatiotemporal modeling.

Key words: deep learning; visual target tracking; single-object tracking; multi-target tracking; Transformer

收稿日期：2025－11－05.

基金项目：校企合作项目（E020H2022009）.

* **通信作者：**马文龙

0 引言

目标跟踪(Object Tracking)是计算机视觉领域的重要研究方向之一,其任务是在视频序列中对感兴趣的目标进行检测[1]、识别与连续定位[2]。该技术在智能监控[3]、自动驾驶[4]、人机交互[5]、行为分析[6]、机器人导航[7]、医学成像[8]等领域具有广泛应用价值,因此长期以来一直是学术界与工业界关注的重点。根据跟踪目标数量的不同,目标跟踪可分为单目标跟踪[9](Single Object Tracking, SOT)与多目标跟踪[10](Multi-Object Tracking, MOT)两大类。SOT主要研究如何在给定初始位置的前提下准确、稳定地追踪单个目标,而MOT则需要同时检测并维持多个目标在时间序列中的身份一致性,具有更高的复杂性与应用难度。

目前,基于深度学习的视觉目标跟踪算法大量开源,为思想的碰撞创造了更多机会和可能。本文旨在对视觉目标跟踪技术的发展历程进行梳理,对基于深度学习的视觉目标跟踪技术中的经典算法进行模型解构、特点剖析并进行性能对比;对目标跟踪领域应用广泛的典型数据集进行简要介绍,最后对基于深度学习的视觉目标跟踪技术的研究进行展望和总结。

1 传统目标跟踪

传统目标跟踪主要包括单目标跟踪和多目标跟踪两类。传统单目标跟踪早期方法分为生成式与判别式两种框架:前者通过建立目标外观模型并计算候选区域与模板的相似度进行匹配,如Mean-Shift[11]和KCF[12]等;后者将跟踪视为目标与背景的分类问题,通过在线学习不断更新模型。部分研究引入卡尔曼滤波[13]或粒子滤波[14]等运动模型以提升时序稳定性,但整体仍依赖手工特征,难以适应光照变化、遮挡或形变等复杂场景。传统多目标跟踪的关键在于检测与数据关联。传统方法多采用Tracking-by-Detection框架[15],先利用检测器生成候选框,再依据外观与运动特征实现跨帧匹配。代表性方法包括多假设跟踪(MHT)[16]、联合概率数据关联(JPDAF)[17]及基于网络流的优化算法。运动预测通常采用线性卡尔曼滤波,外观特征主要依赖颜色或梯度信息。然而,这类方法在目标交互频繁或检测误差较大的场景中易发生身份切换与轨迹中断。

总体而言,传统目标跟踪依赖浅层特征与独立检测器,虽具有一定的实时性与可解释性,但缺乏强大的特征学习与全局时序建模能力。随着深度学习的兴起,特别是AlexNet[18],VGGNet[19]和ResNet[20]等卷积神经网络(CNN)的成功,目标跟踪研究迎来了革命性发展。近年来,Transformer[21]与状态空间模型的引入进一步推动了视觉跟踪由局部匹配向全局建模、由静态检测向时序融合的方向演进。

2 基于深度学习的视觉单目标跟踪算法

以往的基于深度学习的单目标跟踪算法主要可分为基于CNN的跟踪算法[22-27]和基于CNN与Transformer融合的跟踪算法[28-33],它们的共同点在于使用CNN主干进行特征提取,但这种方式缺乏对图像全局信息的充分利用。

近年来,为了应对视觉物体跟踪的挑战,研究者开始将Transformer[34]应用到计算机视觉领域。得益于注意力机制,Transformer能够在模板特征与搜索区域特征之间建立长距离依赖关系,实现对目标外观与背景信息的全局关联建模,拓展了目标跟踪中的全局建模能力,有效提升跟踪算法在目标外观变化、遮挡及复杂场景下的鲁棒性。本节将对经典的全Transformer跟踪算法模型进行分析介绍和对比。

2.1 双流Transformer跟踪算法:SwinTrack与SparseTT

基于双流特征提取的方法在以往SOT模型中的巨大成功,相关研究人员也将其用于全Transformer跟踪算法的特征提取阶段。

受Swin Transformer[35]在目标检测任务中以更少计算量获得更优性能的启发,Lin等提出一种基于双流Transformer视觉单目标跟踪模型SwinTrack[36],模型结构如图1所示。该方法采用Swin Transformer作为主干网络,选择在第三阶段输出模板图像和搜索图像的特征令牌(token),然后将获得的搜索图像令牌和模板图像令牌拼接后输入到基于自注意力机制的编码器中,实现模板内部、搜索区域内部以及两者之间特征交互同步处理;之后将编码器输出token和带有历史目标轨迹的运动令牌token进行拼接后送入基于交叉注意力机制的解码器,通过交叉注意力机制将时序信息整合到搜索区域特征中,提升模型在遮挡和快速运动等场景下的稳定性;最后,融合特征信息后

的 token 被输入至预测头，该模块包含分类分支和边界框回归分支，用于确定目标位置。

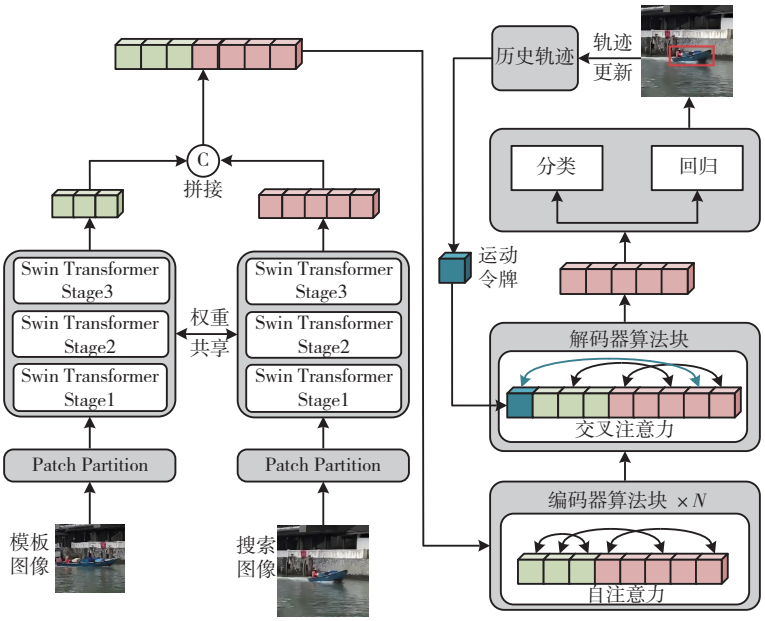


图 1　SwinTrack 模型结构

另外一种较为典型的基于双流 Transformer 跟踪算法是 Fu 等提出的 SparseTT[37]，模型结构如图 2 所示。相关研究人员发现，Transformer 的自注意机制虽然能够捕获长距离依赖，但在处理复杂背景时容易受到干扰，难以聚焦于特定目标。为此，研究人员提出利用稀疏注意模块在搜索区域中引入目标特定注意机制，从而增强对前景区域的关注。SparseTT 主要由基于 Swin Transformer 特征提取网络、基于编码器-解码器的目标聚焦网络和基于分类-回归的双头预测网络三个模块构成，利用 Swin Transformer 提取模板与搜索区域特征，然后将其输入到目标聚焦网络中。目标聚焦网络模块的编码器与一般 Transformer 类似，解码器则采用稀疏多头自注意机制让模型更加聚焦于前景区域。最后，增强后的特征被输入到双头预测网络中，用于预测目标的边界框位置，实现在复杂场景下更高的鲁棒性与精度。

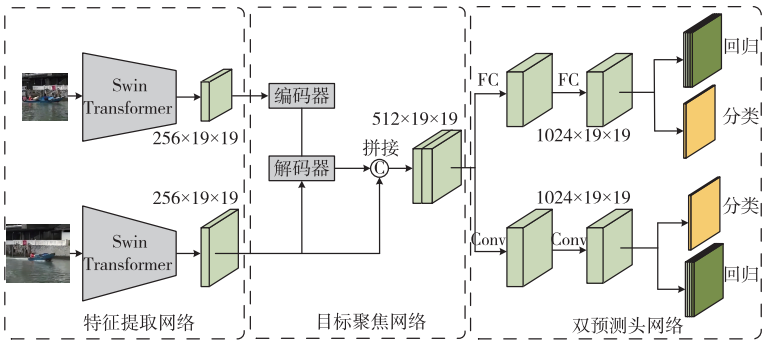


图 2　SparseTT 模型结构

SwinTrack 与 SparseTT 在 GOT-10K 数据集上的性能对比如表 1 所示，其中 AO 为衡量预测边界框与真实边界框的平均重叠率，$SR_{0.5}$ 和 $SR_{0.75}$ 为预测结果与真实标注的重叠度超过设定阈值（0.5 和 0.75）的帧所占比例。尽管 SwinTrack 在 GOT-10K 数据集上性能领先，但其跟踪算法未采用端到端架构。SwinTrack 通过预测响应图并引入汉宁窗惩罚等后处理，有效提升了跟踪的稳定性和收敛速度，使其取得了优异的性能指标，但代价则是牺牲了端到端模型的简洁性与全局优化潜力。

表 1　SwinTrack 与 SparseTT 在 GOT-10K 数据集上的性能

模型	*AO*	$SR_{0.5}$	$SR_{0.75}$
SparseTT	69.3	79.1	63.8
SwinTrack	72.4	80.5	67.8

总体而言，双流 Transformer 跟踪算法通过分别对模板图像和搜索图像采用 Transformer 主干网络进行特征提取，虽然在多个基准数据集上取得了优于传统 CNN 以及 CNN-Transformer 融合模型的性能表现，但这种“先分支提取，后特征融合”的结构导致模板与搜索图像之间的信息交互不足，未能充分发挥 Transformer 在全局建模方面的优势。基于此，后续研究逐渐转向单流特征提取的思路，并成为全 Transformer 架构的跟踪算法的主流结构。

2.2 单流 Transformer 跟踪算法：MixFormer，SimTrack 与 OSTrack

自 SiamFC[22] 以来，大多数单目标跟踪算法通常包含两个独立的特征提取分支，特征提取之后，再通过另一个网络进行特征融合，最后预测目标位置。与此不同，Cui 等发现将模板与搜索区域的特征提取和融合过程结合，可以更充分地捕获二者的关联与差异，从而提取更具判别力的目标特征，并基于此提出 MixFormer[38]。其模型采用单一流水线式的全 Transformer 结构，引入一组混合注意模块（MAM），同时对模板与搜索区域进行特征提取与交互融合。模型支持多模板与多搜索区域作为输入，最终通过一个基于卷积的预测头定位目标。MixFormer 在多个基准数据集上表现优异，但由于 MAM 的计算复杂度较高，其跟踪速度仍有待提升。

Chen 等提出另一种单阶段全 Transformer 跟踪算法 SimTrack[39]。该方法采用 Vision Transformer（ViT）[40] 作为骨干网络，将特征提取与特征融合统一在同一结构中完成。模板图像与搜索区域首先被划分为一组 token，并通过位置嵌入输入到 Transformer 主干中。由于模板在划分过程中可能包含背景区域，SimTrack 引入了中央凹窗口技术（Foveal Window Strategy），输入特征 token 除了搜索图像 token 和模板图像 token 以外，还加入了裁剪边缘背景信息后的模板图像 token 作为补充信息，来更精确地提取目标相关特征。

还有一种基于单流特征提取的典型跟踪算法，即 Ye 等提出的 OSTrack[41]。OSTrack 同样基于 ViT 主干网络实现特征提取与融合工作，模型结构如图 3 所示。与 SimTrack 处理模板图像的背景区域不同，Ye 等发现搜索图像中的部分 token 含有冗余的背景信息会影响注意力计算，基于此在 OSTrack 部分编码层中引入了早期候选模块，根据模板图像和搜索图像的注意力相关性筛除背景 token，从而提升了模型的效率与精度。

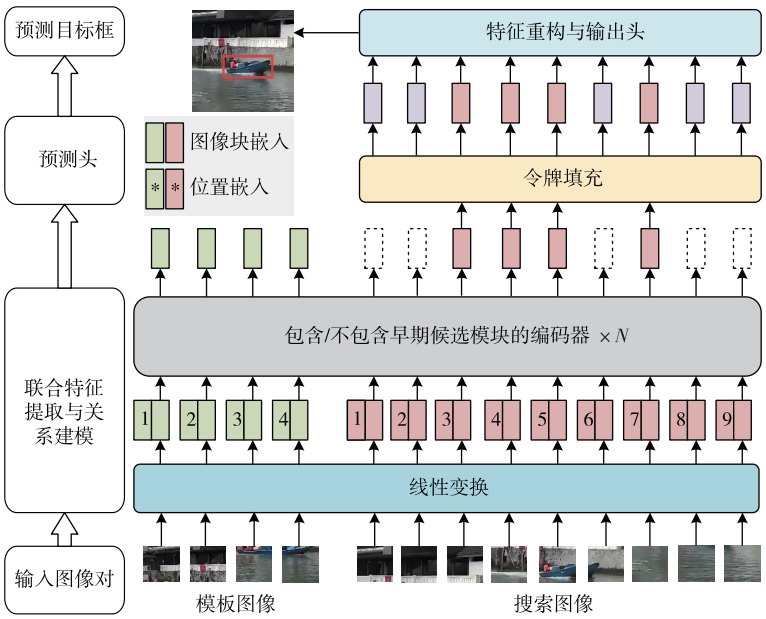


图 3 OSTrack 模型结构

MixFormer，SimTrack 与 OSTrack 在 GOT-10K 数据集上的性能如表 2 所示。虽然 MIXformer 性能优于 SimTrack，但其计算复杂度较高，跟踪实时性不足；OSTrack 基于视觉领域成熟的 ViT 预训练模型，利用早期候选模块消除搜索图像的背景 token，在显著减少计算复杂度的同时增大了有效特征信息的占比，使得搜索图像和模板图像的信息交互更加有效，提升了模型的效率与精度，成为后续许多单流 Transformer 跟踪算法（如 ODTrack[42]，ARTrack[43] 等）研究的参考对象。

表2　MixFormer，SimTrack与OSTrack在GOT-10K数据集上的性能

模型	AO	$SR_{0.5}$	$SR_{0.75}$
SimTrack	69.8	78.8	66
MIXformer	70.7	80	67.8
OSTrack	73.7	83.2	70.8

总体而言，MixFormer，SimTrack与OSTrack共同体现了Transformer在视觉跟踪任务中“统一特征建模”的发展趋势。它们通过将模板与搜索区域在同一Transformer流中处理，实现了更高效的特征交互与全局依赖建模，显著提升了目标跟踪在复杂场景下的表现。

2.3　基于序列学习的Transformer跟踪算法：SeqTrack，ARTrack与ARTrackV2

现有的跟踪方法通常采用分而治之的策略，即将跟踪问题分解为多个子任务，如对象尺度估计和中心点定位。Chen等提出一种全新的序列到序列视觉跟踪框架SeqTrack[44]，如图4所示。SeqTrack将视觉目标跟踪视为一个序列生成问题，并以自回归的方式预测目标边界框。该跟踪算法采用了简洁的编码器-解码器（encoder-decoder）Transformer架构：编码器使用ViT来提取模板图像与搜索图像的视觉特征；解码器则使用带有因果注意力掩码的Transformer解码器，以确保每个输出token仅依赖于前面出现过的token，从而实现序列式预测。为了更好地引导模型的序列生成过程，SeqTrack还在序列中引入了起始与结束标记（start/end token），作为序列生成过程的开始信号和结束信号。此外，该跟踪算法在在线跟踪阶段有效结合了窗口惩罚与在线更新机制，进一步提升模型的预测精度与稳定性。

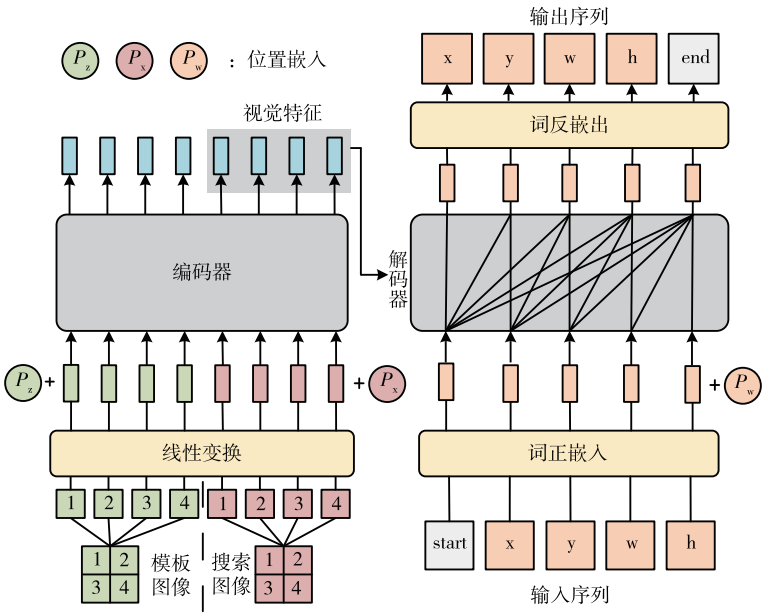


图4　SeqTrack模型结构

类似地，Wei等提出ARTrack与ARTrackV2[45]，将目标跟踪任务视为一个坐标序列解释问题。与SeqTrack仅侧重于帧内坐标生成不同，ARTrack强调跨帧的轨迹连续性，通过引入“时空提示”将历史轨迹信息传递至当前帧，从而在目标被遮挡或存在干扰时仍能保持稳定的跟踪。该方法在长时跟踪和复杂场景下表现出更强的鲁棒性，但其采用的帧内自回归解码机制在一定程度上限制了推理速度。

为了克服自回归解码机制的限制并更全面地建模目标演化，ARTrackV2提出一种联合轨迹与外观自回归的生成框架。ARTrackV2采用纯编码器架构，摒弃了帧内自回归，实现了所有token的并行处理，在保持精度的同时显著提升了推理速度。

SeqTrack，ARTrack与ARTrackV2在GOT-10K数据集上的性能如表3所示，ARTrack性能显著优于SeqTrack，体现出利用历史轨迹信息进行目标跟踪的有效性；ARTrackV2作为ARTrack的优化，在保持精度的同时优化了由于帧内自回归带来的实时性不足的限制。

总体来看，SeqTrack，ARTrack与ARTrackV2均体现出将目标跟踪转化为序列建模问题的创新思路，摆脱了传统匹配式跟踪框架的限制，为视觉跟踪任务的未来发展提供了参考。

表3 SeqTrack，ARTrack与ARTrackV2在GOT-10K数据集上的性能

模型	AO	$SR_{0.5}$	$SR_{0.75}$
SeqTrack	74.8	81.9	72.2
ARTrack	78.5	87.4	77.8
ARTrackV2	79.5	87.8	79.6

2.4 基于参数高效微调的Transformer跟踪算法：LoRAT与SPMTrack

基于全Transformer的单目标跟踪模型通常依赖于参数规模庞大的预训练骨干网络（如ViT，Swin Transformer等）。然而，对此类模型进行全参数训练不仅计算成本高，而且耗时长，已成为制约该领域进一步发展的重要瓶颈。为此，相关研究者们提出参数高效微调（Parameter-Efficient Fine-Tuning，PEFT）策略，通过冻结大部分预训练参数，仅对少量模块进行更新，在显著降低计算开销的同时保留模型在大规模预训练中所获得的通用表征能力。

Lin等提出的LoRAT[46]将低秩自适应方法（LoRA）应用到视觉跟踪任务中，其模型设计重点解决了两个关键挑战，如图5所示。首先，针对传统跟踪算法中模板与搜索区域使用独立位置嵌入会破坏预训练模型结构的问题，LoRAT提出一种解耦的位置嵌入设计，将位置信息解耦为共享空间嵌入（从预训练模型继承并冻结）和独立的可学习令牌类型嵌入（用于区分模板前景、模板背景和搜索区域）。其次，为克服卷积头固有归纳偏差对LoRA微调的阻碍，LoRAT采用了一个纯多层感知机（MLP）构成的无锚框预测头，该头部分为分类和回归两个分支，直接处理Transformer编码器的输出特征，实现了更优的收敛性和更低的计算开销。通过这些设计，LoRAT仅需微调LoRA矩阵、令牌类型嵌入和MLP头等少量参数，便能高效激发大型ViT骨干在跟踪任务上的性能。

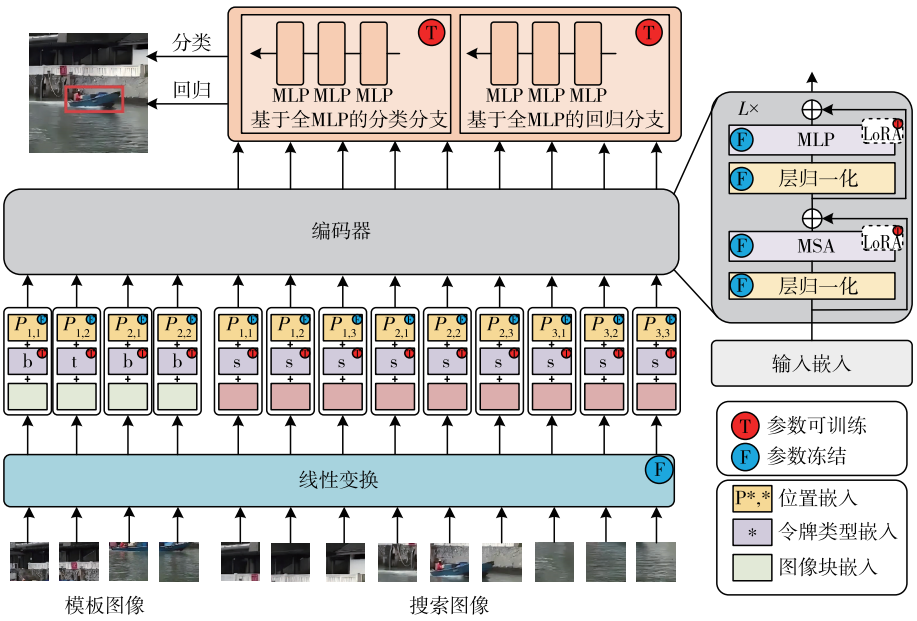


图5 LoRAT模型结构

另外一种利用参数高效微调技术的跟踪算法是Cai等提出的SPMTrack[47]。该算法为基于混合专家模型（Mixture of Experts，MoE），构建了一种专为视觉跟踪任务定制的混合专家模块（TMoE），并通过微调TMoE来减少可训练参数。TMoE被同时嵌入到ViT块的多头自注意力层和前馈网络层的线性投影中，取代了原有的线性层。在微调策略上，SPMTrack仅训练TMoE中的路由组件和预测头，而骨干网络的绝大部分参数保持冻结，在显著减少可训练参数量的同时，实现了更大的模型结构和更高的性能。

LoRaT与SPMTrack在GOT-10k数据集上的性能如表4所示，在GOT-10k等数据集上，SPMTrack性能显著优于LoRAT，其根本优势源于二者迥异的参数微调范式。LoRAT采用全局统一的低秩适应策略，虽具通用性却难以应对跟踪任务中多变的视觉模式。而SPMTrack引入的混合专家模块实现了关键突破：通过动态路由机制，模型能针对不同区域特性激活相应专家，专门处理边界模糊、背景干扰等特定挑战。这种“专才分工”的设计，结合其多帧时空建模

能力，使SPMTrack在相似参数效率下获得了更精细的关系建模力，从而在复杂场景中展现出显著优势。

表 4　LoRaT 与 SPMTrack 在 GOT-10k 数据集上的性能

模型	*AO*	$SR_{0.5}$	$SR_{0.75}$
LoRAT	78.9	87.8	80.7
SPMTrack	81	89.2	82.3

总体而言，基于参数高效微调的跟踪算法在保证高精度目标跟踪性能的同时，极大降低了模型训练与部署的资源消耗。这类方法不仅进一步提升了 Transformer跟踪算法的性能与泛化能力，也为视觉跟踪领域在更大规模模型的高效应用与扩展方面提供了新的思路与发展方向。

3　基于深度学习的视觉多目标跟踪算法

3.1　算法分类

多目标跟踪按照任务类型可以分为离线跟踪和在线跟踪两类。离线跟踪是指使用视频序列的全局信息对目标进行跟踪，这种方法可以认为是寻找全局最优解，通常可以达到较高的跟踪性能，但是由于其离线特性，故应用价值不如在线跟踪算法；在线跟踪是指在对目标进行处理时，只能使用当前帧以及之前帧的信息形成对当前帧中目标的跟踪。虽然离线跟踪方式在跟踪准确性以及鲁棒性上更好，但是由于在线跟踪方式更符合实际应用的特点，为本文主要关注的方向。

多目标跟踪按照是否依赖检测器可以分为基于检测的目标跟踪（TbD）算法和联合检测的目标跟踪（JDT）[48]算法。在下面几小节中，将按照这两种分类对经典多目标跟踪算法进行回顾，并对近年来新出现研究进展进行综述。

基于深度学习的视觉多目标跟踪任务经过众多研究者的多年努力，目前已有许多成果，下面在图 6 中对本文涉及的算法按照对检测器的依赖程度的区别进行分类并汇总。

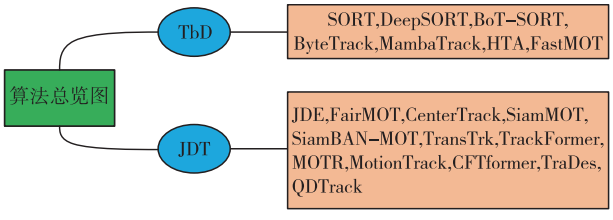


图 6　算法汇总图

3.2　基于检测的目标跟踪（TbD）

基于检测的跟踪是最早主导多目标跟踪研究的框架，其核心思想是“先检测，后跟踪”。该思路最早由 Breitenstein 等 2009 年提出并在随后的十余年中成为多目标跟踪的标准范式。在这种模式下，检测器负责在每一帧中定位目标，而数据关联模块通过外观特征与运动预测建立帧间联系，如图 7 所示为其算法框架图。深度学习的引入彻底改变了在存在遮挡、目标交互或检测误差时，容易出现轨迹中断与身份切换等问题的现象。

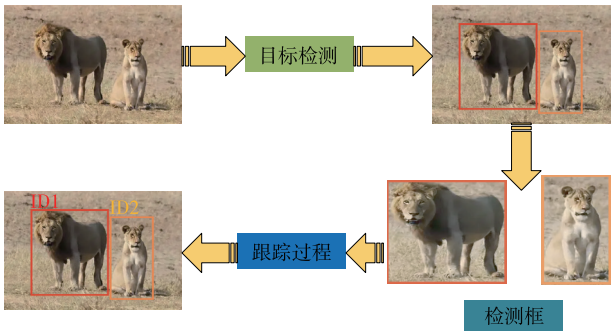


图 7　基于检测的目标跟踪算法框架

深度卷积神经网络使目标检测与特征提取性能获得质的飞跃。早期 Krizhevsky 等在 2012 年提出的 AlexNet，Simonyan 和 Zisserman，在 2014 年提出的 VGGNet 在 ImageNet 上的成功奠定了深层视觉特征学习的基础，推动了深度特征在 MOT 中的应用。SORT[49]是 TbD 框架的经典代表，利用检测结果和卡尔曼滤波预测目标状态，通过 IOU 匹配实现高效的在线关联。该算法结构极为简洁，具有实时性能，但缺乏外观建模能力，易在遮挡和相似目标情况下出现身份切换。为弥补这一缺陷，DeepSORT[50]在 SORT 的基础上引入了深度外观特征（ReID 特征）以增强匹配鲁棒性。其特征提取网络最初基于 VGGNet，后被更强的 ResNet-50[51]所取代，使得模型能够在外观空间中区分相似目标。DeepSORT 的出现标志着深度学习特征正式成为 TbD 模型的核心组成部分。此后，大量改进算法在 DeepSORT 框架基础上进一步提升性能。BoT-SORT[52]通过融合运动一致性与外观距离改进了关联策略，使匹配更加稳定；在运动建模方面，Xiao 等在 2024 年提出的 MambaTrack[53]首次将 Mamba 状态空间模型引入 TbD 框架，提出 MTP 与 TPM，以双向 Mamba 编码器学习非线性运动并自动修补轨迹断裂，在 DanceTrack[54] 与 SportsMOT[55] 上分别实现

HOTA56.8与72.6，显著超越基于线性预测的TbD模型，开创了学习型运动预测的新方向。而ByteTrack[56]则突破了传统的高置信度假设，提出对低置信度检测框进行二次匹配，从而减少漏检、提升召回率。ByteTrack虽然仍属于TbD框架，但在数据关联阶段充分利用检测器特征信息，实现了接近JDT水平的精度。混合轨迹关联（HTA）[57]则利用增量高斯混合模型（IGMM）建模历史外观距离分布，结合当前相似度计算混合代价函数，有效应对遮挡与长时间重现问题。

在TbD系列方法中，检测器的演进同样起到了关键作用。早期使用基于区域的检测框架Faster R-CNN[58]，其backbone网络多为ResNet或VGGNet，如图8所示其算法框架。随后，技术路线逐渐向更高效的单阶段检测器（YOLO[59]，RetinaNet[60]和anchor-free结构（CenterNet[61]）发展，显著提升了定位精度与速度。更重要的是，以DETR[62]为代表的端到端Transformer检测器为领域带来了深远影响，其摒弃手工设计组件（如anchor和NMS）、直接进行集合预测的全新范式，虽然其初始版本在实时性上存在挑战，但其思想启发了后续一系列更高效的改进模型，进一步丰富了检测器的设计选择。这些检测器的持续进步使TbD系统能够更好地满足实时应用需求。例如，FastMOT[63]结合YOLOv4检测器与DeepSORT跟踪模块，通过TensorRT加速实现1 080p视频的实时处理，体现了TbD在工程部署方面的优势。总体而言，TbD框架的成功得益于检测器的性能提升与深度特征的引入，其方法体系简洁高效，但检测与关联的解耦特性导致误差容易累积，限制了其在复杂场景下的表现。

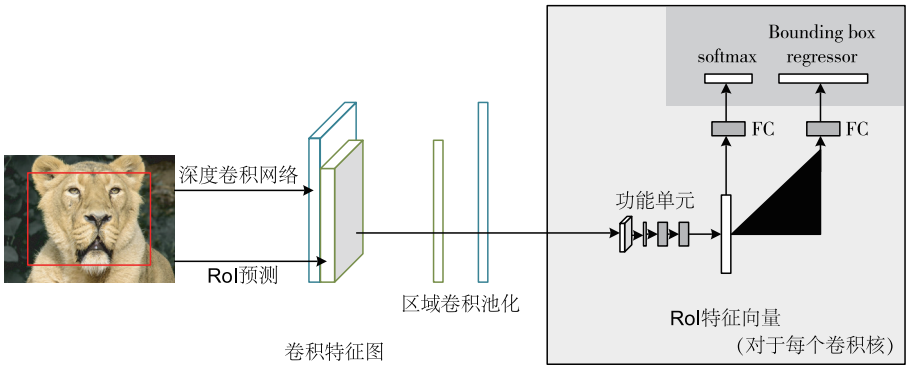


图8 Faster R-CNN目标检测器的算法框架

3.3 联合检测的目标跟踪（JDT）

与TbD不同，联合检测的目标跟踪方法旨在从网络层面实现检测与关联的统一建模。其基本思想是通过共享特征提取backbone，同时预测检测结果和目标身份嵌入，从而在同一前向过程中完成多目标的检测与跟踪，如图9所示的算法流程图。JDT解决了TbD框架中检测误差向后传播的问题，是多目标跟踪发展的重要方向。

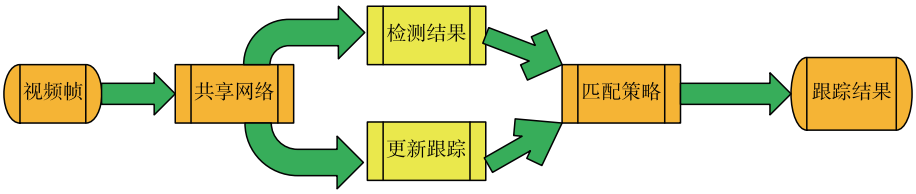


图9 联合检测的目标跟踪算法流程图

首个具有代表性的JDT方法是JDE[64]，该算法在单个神经网络中同时输出检测框与ReID特征。JDE通常采用ResNet-34[65]作为backbone，通过共享卷积特征实现检测与特征学习的联合优化，大幅减少了重复计算，提高了推理效率。随后，FairMOT[66]对JDE的结构进行了改进，提出“任务公平性”原则，认为检测和ReID的分支在低层特征共享上存在冲突。为解决这一问题，FairMOT采用anchor-free的CenterNet检测头，在同一高分辨率特征图上并行预测检测热力图和嵌入特征，并使用DLA-34 backbone保持空间分辨率。该设计在MOT16与MOT17[67]数据集上均取得了领先性能，成为JDT框架的经典基线。另一种代表性工作是CenterTrack[68]，它在CenterNet的检测框架上引入时间建模机制，在输入

中加入前一帧检测结果的热图，通过学习中心点偏移量实现检测与关联的联合预测。CenterTrack在无需独立ReID模块的情况下即可进行多目标跟踪，速度远超基于嵌入的模型，是典型的检测即跟踪方法。此外，QDTrack[69]提出一种准密集匹配策略，在检测分支输出后直接计算跨帧特征相似度，通过大量正负样本对进行端到端训练，从而在检测空间内实现隐式关联。QDTrack的backbone采用ResNet-50+FPN结构，通过在检测特征层内学习相似度矩阵，既保持了精度，又减少了额外的ReID计算开销。

在JDT框架中，Shuai等在2021年提出的SiamMOT[70]首次将孪生网络（Siamese Network）引入多目标跟踪。其孪生网络跟踪模块（Siamese Tracking Module，STM）同时输入历史轨迹模板与当前帧搜索区域特征，通过特征相关性计算实现跨帧关联，从而在检测阶段完成匹配预测。SiamMOT在MOT17上取得MOTA77%，实现了检测与时序建模的统一。随后，Hui等在2022年提出的SiamBAN-MOT[71]在Siamese[72]框架中引入RSFA模块，聚合历史模板信息以增强时序一致性，实现检测与关联特征的深度融合。

随着深度学习向时空建模方向发展，Transformer架构的引入进一步推动了JDT的演化。TransTrk[73]首次将Transformer引入MOT中，通过object queries和track queries实现新目标检测与历史目标关联的统一表示，摆脱了传统后处理匹配。TrackFormer[74]在DETR框架上通过自回归的track query更新实现tracking-by-attention，使模型能够在全局注意力空间中自动建立帧间对应关系。MOTR[75]则基于Deformable DETR提出连续查询传递（continuous query passing）机制，使每个目标的query在多帧间持续更新，实现了真正意义上的端到端跟踪。这类Transformer框架能够同时建模空间布局与时间依赖关系，在复杂遮挡和长时场景下表现尤为出色。近年来出现的TraDeS[76]（Track to Detect and Segment）与MotionTrack[77]等模型代表了JDT的进一步拓展方向。TraDeS提出“由跟踪反哺检测”的新理念，利用跟踪特征引导检测优化，通过代价体积关联（Cost Volume Association，CVA）与运动引导的特征扭曲（Motion-guided Feature Warping，MFW）模块利用运动线索增强检测特征，从而实现检测、跟踪与实例分割的协同训练。MotionTrack则提出同时学习短期与长期运动特征的混合建模框架，结合局部时间卷积与跨帧轨迹注意力机制，既能捕获快速运动的动态信息，又能在长期遮挡后恢复目标身份。这些算法不仅融合检测与跟踪，还在运动建模层面实现时空统一表示，成为Transformer时代的代表性工作。CFTformer[78]是另一类基于Transformer的端到端MOT方法，它通过引入跨帧注意力机制强化时序建模，有效减少遮挡或运动模糊导致的身份切换（ID switches）。该方法无需额外ReID模块即可实现检测与跟踪联合优化，在DanceTrack，MOT17等数据集上取得高精度与稳定的跟踪性能，展示了Transformer在连续帧全局建模中的优势。

4　数据集与评价指标

4.1　用于单目标跟踪的经典数据集及评价指标

4.1.1　OTB

OTB[79]数据集包括OTB2013和OTB2015。OTB2013是第一个用于视觉目标跟踪中训练和评估的基准数据集，它包括50个完全注释的视频序列，并使用11个属性对它们进行分类。数据集中的每个视频都根据其特征标记了属性。在OTB2013的基础上，OTB2015将数据集扩展到100个视频序列，是视觉目标跟踪中最经典的基准数据集之一。在性能评估上，OTB数据集提出了精度（Precision）和成功（Success）两个评价指标。

Precision是一种广泛使用的跟踪评估指标，通过被跟踪目标中心位置与真实框中心之间的平均欧式距离来评估目标中心的定位误差，被计算为整个视频序列中被跟踪对象预测框中心和真实框中心之间的平均欧式距离小于给定阈值的视频帧所占的百分比。

Success是根据交集与并集的比率来评估的。当预测框和地面真实框帧的交集超过并集比大于给定阈值时，认为视频帧中的对象被成功跟踪。成功的定义为成功跟踪的帧占视频序列中总帧数的百分比。成功率的曲线下面积（Area Under the Curve，AUC）用于对跟踪算法进行排序。

4.1.2　GOT-10k

GOT-10k[80]是由中国科学院基于WordNet层级结构构建的数据集，涵盖了超过560种运动目标

类别和87种运动模式，具有极强的通用性与多样性。该数据集包含超过10 000段视频片段，并提供了超过150万个人工标注的目标边界框，是目前规模庞大且标注精细的视觉目标跟踪数据集之一。

GOT-10k划分为训练集、验证集和测试集，在性能评估方面，GOT-10k采用平均重叠度（Average Overlap，AO）与成功率（Success Rate，SR）作为主要评价指标。其中，AO衡量预测边界框与真实边界框的平均重叠率，其统计意义等同于AUC，用于评估整体跟踪精度；SR则表示预测结果与真实标注的重叠度超过设定阈值（例如0.5，0.75）的帧所占比例，与传统的Success指标一致。

4.1.3 TrackingNet

TrackingNet[81]数据集包含超过30 000个视频序列和超过1 400万个对象边界框注释，使其成为迄今为止在规模、帧数和持续时间方面最大的视觉跟踪数据集。它涵盖了广泛的对象类别和多样化的环境背景，视频采样自YouTube，以确保现实世界场景的丰富性和多样性。每个视频都被标注了一系列属性，如尺度变化、光照变化、运动模糊等。这些属性便于分析不同挑战下的跟踪算法性能。

TrackingNet明确区分了训练集和测试集，除了采用Precision和Success作为评价指标外，它还引入了一个新的指标，即归一化精度（Normalized Precision），该指标利用真实目标框尺寸归一化精度来考虑目标尺寸变化，以获得正常精度。TrackingNet也使用AUC对跟踪算法进行排名。

4.1.4 LaSOT

LaSOT[82]数据集是近年来提出的一个高质量、超大规模的视觉目标跟踪数据集。该数据集共包含1 400段视频序列，被划分为70个类别，每个类别包含20段视频。其平均视频长度超过2 500帧，每个序列均来自复杂的真实场景，涵盖多种具有挑战性的环境因素。LaSOT特别注重长时跟踪（long-term tracking）场景的研究，其中目标可能在视频中出现消失、再出现等情况，以此模拟真实世界中目标持续时间不确定、遮挡频繁等复杂动态情形。此外，LaSOT对每个视频序列标注了14种视觉属性（如遮挡、快速运动、尺度变化、背景干扰等），便于研究者从多维度分析跟踪算法在不同挑战条件下的性能表现。

LaSOT隐性地划分有训练集与测试集，并采用Precision，Success和Normalized Precision三项指标对算法性能进行评估，同时通过AUC对各跟踪算法的综合性能进行排名。

4.1.5 VOT

VOT是为目标跟踪领域每年的VOT竞赛而准备的数据集，数据集中样本每年持续更新，且难度逐年增大。VOT数据集中均为彩色视频序列且视频序列的分辨率普遍较高。VOT涉及的性能测评指标包括准确率（Accuracy）、鲁棒性（Robustness）以及期望平均重叠度（Expected Average Overlap，EAO）作为主要评估指标。其中，Accuracy用于衡量跟踪算法预测边界框与真实边界框之间的重叠程度，通过计算二者的交并比（Intersection over Union，IoU）来反映跟踪精度；Robustness用于统计跟踪算法在跟踪过程中因偏离目标而需要重新初始化的次数，即衡量跟踪失败的频率；而EAO则综合考虑了准确率与鲁棒性两方面的因素，用于评估跟踪算法在整体性能上的平衡表现，是VOT评测体系中的核心指标之一。

除了上述数据集之外，UAV123，OxUvA，NFS和Temple-color等数据集也常用于视觉目标跟踪领域。

4.2 用于多目标跟踪的经典数据集及评价指标

4.2.1 数据集

近年来，许多大学、公司和研究团队发布了用于多目标跟踪（MOT）的公共数据集。这些数据集包括带注释的视频、标准化假设、标注和评估工具，有助于验证MOT方案并促进MOT技术的发展。

早期的MOT数据集主要聚焦于行人跟踪任务。最具代表性的为MOTChallenge系列，包括MOT15[83]，MOT16，MOT17和MOT20[84]。MOT15首次提出标准化的多目标行人跟踪评测框架，包含22段视频序列，总帧数超过11 000帧，标注了目标边界框、遮挡与可见比例，为后续算法比较提供了基础。随后推出的MOT16与MOT17在分辨率、帧数和检测源上显著提升，提供了三种检测器结果（DPM，Faster R-CNN和SDP），成为多目标跟踪研究的核心基准，绝大多数算法（如SORT，DeepSORT，JDE和FairMOT等）均在该数据集上进行评测。MOT20进一步引入高密度

人群场景，总帧数超过 13 万帧，目标数量是 MOT17 的两倍以上，用以验证算法在高遮挡与复杂背景下的稳健性。这一系列数据集奠定了行人跟踪任务的标准化基础，使 MOT 研究进入深度学习驱动的快速发展阶段。

随着算法能力的提升和应用需求的扩展，研究者逐渐从单一行人场景转向更复杂的多场景与多类别数据集。BDD100K[85]是首个大规模自动驾驶场景下的 MOT 数据集，包含 10 万段视频，覆盖城市、高速、夜间等多种天气与光照条件，标注了车辆、行人、自行车、交通信号灯等十余类目标，极大地拓宽了 MOT 的应用边界。随后推出的 DanceTrack 专为评估算法的时间一致性与身份保持能力而设计，数据集包含上百段舞蹈视频，总帧数超过十万，目标动作复杂、外观相似且存在频繁遮挡，成为测试轨迹连续性（如 HOTA，IDF1）的重要基准。SportsMOT 则面向体育赛事场景，涵盖篮球、足球等高速运动目标，考验算法的动态预测与遮挡恢复能力。与此同时，TAO[86]提供了 2 900 余段视频、833 个物体类别和超 400 万标注框，首次实现了跨类别、跨场景的通用跟踪评测，推动了从单类行人 MOT 向通用多目标跟踪（Generalized MOT）的演进。

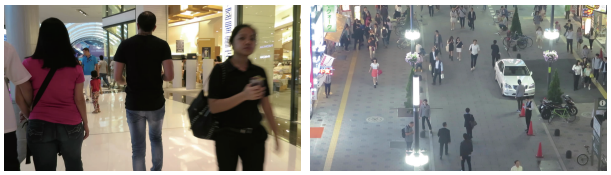
常用数据集中包含有两种挑战因素：目标自身因素和外部背景因素。自身因素主要是因跟踪目标自身变化所导致的问题如旋转、形变和运动模糊等；外部因素主要是由于跟踪场景环境因素所带来的影响，如遮挡、相似目标和相似背景干扰等。图 10 展示了来自基准数据集中的挑战因素。



（a）遮挡



（b）尺度变换



（c）运动模糊　　（d）相似目标干扰

图 10　常见挑战

4.2.2　评价指标

为了对模型进行客观评价，许多研究者结合 MOT 任务的性质制定了许多标准，这些评价指标分别侧重于算法不同性能，能够将不同模型通过统一的评价准则进行对比分析，对推动多目标跟踪的研究发展起到十分重要的作用，常用的评价指标及介绍如表 5 所示，表 6 和表 7 是本文所讨论的算法在不同数据集上表现的对比分析。

表 5　常用的评价指标及介绍

类别	评价指标	指标简述	类别	评价指标	指标简述
基于轨迹的评价指标	TP	模型预测为正，且数据标注为正样本的数量	基于轨迹的评价指标	HOTA	高阶跟踪精度用于统一评价检测、关联和定位 $HOTA=\frac{2\times DetA\times AssA}{DetA+AssA}$
	TN	模型预测为负，且数据标注为负样本的数量		DetA	检测精度，衡量检测框与真实框的重叠程度
	FP	模型预测为正，但数据标注为负样本的数量		AssA	关联精度，衡量轨迹关联的正确性
	FN	模型预测为负，但数据标注为正样本的数量	基于身份的评价指标	IDSW	身份切换次数，即目标身份在轨迹中被错误切换的次数
	MT	成功跟踪的帧数占比达 80% 以上的 GT 轨迹数量		IDF1	身份 F1 分数，衡量跟踪算法在身份分配上的准确性和召回率 $IDF1=\frac{2\times IDP\times IDR}{IDP+IDR}$
	ML	成功跟踪的帧数占比低于 20% 以上的 GT 轨迹数量		IDP	身份精度，表示正确分配的身份数量与分配的总身份数量的比值
	GT	序列中真实存在目标总数		IDR	身份召回率，表示正确分配的身份数量与真实身份数量的比值
	MOTA	衡量模型整体准确度 $MOTA=1-\frac{FN+FP+IDSW}{GT}$			

表6 算法在DanceTrack数据集上的对比分析

算法	所属分类	*HOTA*/%↑	*IDF1*/%↑	*AssA*/%↑	*MOTA*/%↑	*DetA*/%↑
DeepSORT	TbD	45.6	47.9	29.7	87.8	71.0
MOTR	JDT	54.2	51.5	40.2	79.7	73.5
FairMOT	JDT	39.7	40.8	23.8	82.2	66.7
TransTrk	JDT	45.5	45.2	27.5	88.4	75.9
TraDes	JDT	43.3	41.2	25.4	86.2	74.5
QDTrack	JDT	45.7	44.8	29.2	83.0	72.1
CenterTrack	JDT	41.8	35.7	22.6	86.8	78.1
SORT	TbD	47.9	50.8	31.2	91.8	72.0
ByteTrack	TbD	47.3	52.5	31.4	89.5	71.6
MambaTrack	TbD	56.8	57.8	39.8	90.1	80.1

表7 算法在MOT17数据集上的对比分析

算法	所属分类	*MOTA*/%↑	*IDSW*/%↓	*IDF1*/%↑	*MT*/%↑	*ML*/%↓
SORT	TbD	43.13	4 832	39.84	12.23	42.34
DeepSORT	TbD	78.0	1 821	74.5	32.8	18.2
BoT-SORT	TbD	80.60	1 257	79.50	——	——
ByteTrack	TbD	80.25	2 196	77.30	53.25	14.52
JDE	JDT	62.1	1 608	56.9	34.4	16.7
FairMOT	JDT	73.70	3 303	72.30	43.20	17.30
CenterTrack	JDT	61.5	2 583	59.6	——	——
SiamMOT	JDT	65.90	3 040	63.30	34.60	23.90
SiamBAN-MOT	JDT	64.80	——	60.90	——	——
TransTrk	JDT	74.50	3 663	63.90	46.80	11.31
TrackFormer	JDT	62.30	4 018	57.60	29.21	27.09
MOTR	JDT	71.91	2 115	68.4	——	——
MotionTrack	JDT	81.11	1 140	80.09	55.54	16.69
CFTformer	JDT	76.5	2 123	77.61	——	——
HTA	TbD	54.3	660	62.5	33.11	24.01

注：↑表示指标数值越高越好，↓则相反。

5 展望与总结

随着深度学习技术的迅猛发展，视觉目标跟踪已经从传统的手工特征建模阶段，演化为以端到端特征学习为核心的深度智能阶段。本文系统回顾了基于深度学习的视觉目标跟踪研究现状，分别从单目标跟踪与多目标跟踪两个方向进行了归纳与分析。单目标跟踪在特征表征、全局建模和时序更新等方面取得了显著突破，Transformer架构的引入使得模型能够捕获长距离依赖关系，实现了从局部匹配到全局上下文建模的转变。而多目标跟踪则在检测与关联的联合优化上取得长足进展，TbD与JDT两大框架不断融合，从最初的基于检测结果的外观与运动匹配，逐渐发展为通过共享特征提取与注意力机制实现统一建模的端到端系统。与此同时，状态空间模型（SSM）和多模态融合技术的引入，为运动预测与语义理解提供了新的方向，使视觉跟踪朝着更具智能化与自适应性的方向迈进。

尽管近年来的研究取得了显著成果，但视觉目标跟踪仍面临诸多挑战。首先，在复杂场景下的鲁棒性问题依然突出，例如长时遮挡、目标相似、光照变化和背景干扰等因素仍可能导致跟踪失败。其次，模型在跨场景与跨模态任务中的泛化能力不足，限制了其在真实应用中的推广。再次，Transformer等大规模模型虽然显著提升了精度，但计算代价较高，不利于嵌入式或实时系统部署。未来的研究应重点关注高效特征建模与轻量化网络设计，探索CNN，Transformer与状态空间模型（如Mamba）相结合的混合架构，以兼顾性能与效率。同时，多模态融合（视觉、语言、深度与语义信息）以及跨任务协同学习（检测、分割与跟踪一体化）将成为重要趋势。此外，构建更大规模、更具多样性的真实场景数据集，并完善评测体系，也是推动视觉目标跟踪研究向实际智能应用落地的关键。

总体来看，视觉目标跟踪正处于由“单一视觉建模”向“时空-语义融合建模”过渡的关键阶段。随着计算能力的提升与算法框架的持续创新，未来的视觉目标跟踪系统将具备更强的环境理解与自主决策能力，为智能交通、无人系统、人机交互和安防监控等领域带来更加可靠与高效的解决方案。

参考文献：

[1] 罗 元，王薄宇，陈 旭. 基于深度学习的目标检测技术的研究综述[J]. 半导体光电，2020，41(1)：1-10.
Luo Yuan，Wang Boyu，Chen Xu. Research progresses of target detection technology based on deep learning[J]. Semiconductor Optoelectronics，2020，41(1)：1-10.

[2] Su C，Li L，Wei H，et al. Enhanced target recognition and localization using binocular vision and infrared thermal imaging [J]. Multimedia Systems，2025，31(5)：393.

[3] Hussain A，Khan N，Khan Z A，et al. Edge-assisted framework for instant anomaly detection and cloud-based anomaly recognition in smart surveillance [J]. Engineering Applications of Artificial Intelligence，2025，160：111936.

[4] Pan C，Yaman B，Nesti T，et al. VLP：vision language planning for autonomous driving [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition，2024：14760-14769.

[5] MacKenzie I S. Human-Computer Interaction：An Empirical Research Perspective[M]. Holland：Morgan Kaufmann，2024.

[6] Breaux C A，Smith K. Assent in applied behaviour analysis

and positive behaviour support: ethical considerations and practical recommendations[J]. International Journal of Developmental Disabilities, 2023, 69(1): 111-121.

[7] Loganathan A, Ahmad N S. A systematic review on recent advances in autonomous mobile robot navigation[J]. Engineering Science and Technology, an International Journal, 2023, 40: 101343.

[8] Ma J, He Y, Li F, et al. Segment anything in medical images [J]. Nature Communications, 2024, 15: 654.

[9] Zhang Y, Wang T, Liu K, et al. Recent advances of single-object tracking methods: a brief survey[J]. Neurocomputing, 2021, 455: 1-11.

[10] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]//Computer Vision - ECCV 2020, 2020: 107-122.

[11] Choi S, Kang D, Cho M. Contrastive mean-shift learning for generalized category discovery [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 23094-23104.

[12] Maharani D A, Machbub C, Yulianti L, et al. Deep features fusion for KCF-based moving object tracking[J]. Journal of Big Data, 2023, 10(1): 136.

[13] Grewal M S. Kalman Filtering[M]. Heidelberg: Springer, 2025: 1285-1289.

[14] Yusuf M O, Hanzla M, Rahman H, et al. Enhancing vehicle detection and tracking in UAV imagery: a pixel labeling and particle filter approach[J]. IEEE Access, 2024, 12: 72896-72911.

[15] Breitenstein M D, Reichlin F, Leibe B, et al. Online multiperson tracking-by-detection from a single, uncalibrated camera [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1820-1833.

[16] Coraluppi S, Hunter A. Multiple-hypothesis tracking with unframed sensor measurements[C]//2025 IEEE Aerospace Conference, 2025: 1-8.

[17] Taplin T, Lyall A E, Dani A P. Multiple user intent prediction using interacting multiple model joint probabilistic data association filter [J]. IFAC-PapersOnLine, 2023, 56(2): 6946-6951.

[18] Eldem H, Ülker E, Işıklı O Y. Alexnet architecture variations with transfer learning for classification of wound images[J]. Engineering Science and Technology, an International Journal, 2023, 45: 101490.

[19] Song H, Lu H. Research on image recognition based on different depths of VGGNet[J]. Journal of Image Processing Theory and Applications, 2024, 7(1): 84-90.

[20] Liao J, Guo L, Jiang L, et al. A machine learning-based feature extraction method for image classification using ResNet architecture [J]. Digital Signal Processing, 2025, 160: 105036.

[21] Huang X, Ai Z, She C, et al. A CNN-Transformer fusion network for Diabetic retinopathy image classification[J]. Computerized Medical Imaging and Graphics, 2025, 126: 102655.

[22] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking [C]// Computer Vision - ECCV 2016 Workshops, 2016: 850-865.

[23] Li B, Yan 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.

[24] 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, 2019: 4277-4286.

[25] Hu W, Wang Q, Zhang L, et al. SiamMask: a framework for fast online object tracking and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3072-3089.

[26] Luo Y, Xiao H, Ou J, et al. SiamSMDFFF: Siamese network tracker based on shallow-middle-deep three-level feature fusion and clustering-based adaptive rectangular window filtering[J]. Neurocomputing, 2022, 483: 160-170.

[27] Fu C, Cao Z, Li Y, et al. Siamese anchor proposal network for high-speed aerial tracking[C]//2021 IEEE International Conference on Robotics and Automation, 2021: 510-516.

[28] Cao Z, Fu C, Ye J, et al. HiFT: hierarchical feature transformer for aerial tracking [C]//2021 IEEE/CVF International Conference on Computer Vision , 2021: 15437-15446.

[29] Gao S, Zhou C, Ma C, et al. AiATrack: attention inAttention forTransformer visual tracking [C]//Computer Vision-ECCV 2022, 2022: 146-164.

[30] Chen X, Yan B, Zhu J, et al. Transformer tracking[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,2021: 8126-8135.

[31] Song Z, Yu J, Chen Y P, et al. Transformer tracking with cyclic shifting window attention [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8781-8790.

[32] Xing D, 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: 1898-1907.

[33] Wang N, Zhou W, 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.

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[35] Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical

vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision, 2021: 9992-10002.

[36] Lin L, Fan H, Zhang Z, et al. Swintrack: A simple and strong baseline for transformer tracking [J]. Advances in Neural Information Processing Systems, 2022, 35: 16743-16754.

[37] Fu Z, Fu Z, Liu Q, et al. SparseTT: Visual tracking with sparse transformers [J]. arXiv preprint arXiv: 2205.03776, 2022.

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

[39] Chen B, Li P, Bai L, et al. Backbone is all your need: a simplified architecture for visual object tracking [C]// Computer Vision-ECCV 2022, 2022: 375-392.

[40] Dosovitskiy A. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[41] Ye B, Chang H, Ma B, et al. Joint feature learning andRelation modeling forTracking: a one-stream framework [C]//Computer Vision-ECCV 2022, 2022: 341-357.

[42] Zheng Y, Zhong B, Liang Q, 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.

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

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

[45] Bai Y, Zhao Z, Gong Y, et al. ARTrackV2: prompting autoregressive tracker where to look and how to describe[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 2024: 19048-19057.

[46] Lin L, Fan H, Zhang Z, et al. Tracking meets LoRA: faster training, larger model, stronger performance[C]//Computer Vision - ECCV 2024, 2025: 300-318.

[47] Cai W, Liu Q, Wang Y. 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, 2025: 16871-16881.

[48] Munjal B, Aftab A R, Amin S, et al. Joint detection and tracking in videos with identification features[J]. Image and Vision Computing, 2020, 100: 103932.

[49] Cao J, Pang J, Weng X, et al. Observation-centric SORT: rethinking SORT for robust multi-object tracking[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 9686-9696.

[50] Zhang C, Guo C. Research on improved DeepSort pedestrian detection algorithm[C]//2024 IEEE 6th Advanced Information Management, Communicates, Electronic and Automation Control Conference, 2024: 1768-1771.

[51] Suherman E, Rahman B, Hindarto D, et al. Implementation of ResNet-50 on end-to-end object detection (DETR) on objects[J]. SinkrOn, 2023, 8(2): 1085-1096.

[52] Arief M, Afdhal A, SaddamI K, et al. Real-time detection and tracking framework using extended Kalman filter BoT-SORT in uncertainty mixed-traffic[J]. IEEE Open Journal of Vehicular Technology, 2025, 6: 2804-2827.

[53] Xiao C, Cao Q, Luo Z, et al. MambaTrack: a simple baseline for multiple object tracking with state space model [C]//Proceedings of the 32nd ACM International Conference on Multimedia, 2024: 4082-4091.

[54] Sun P, Cao J, Jiang Y, et al. DanceTrack: multi-object tracking in uniform appearance and diverse motion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 20961-20970.

[55] Cui Y, Zeng C, Zhao X, et al. SportsMOT: a large multi-object tracking dataset in multiple sports scenes[C]//2023 IEEE/CVF International Conference on Computer Vision, 2023: 9887-9897.

[56] Zhang Y, Wang X, Ye X, et al. ByteTrackV2: 2D and 3D multi-object tracking by associating every detection box. arXiv 2023[J]. arXiv preprint arXiv:2303.15334.

[57] Lin X, Li C T, Sanchez V, et al. On the detection-to-track association for online multi-object tracking[J]. Pattern Recognition Letters, 2021, 146: 200-207.

[58] El Mallahi I, Riffi J, Tairi H, et al. Efficient vehicle detection and classification algorithm using faster R-CNN models [J]. Journal of Automation, Mobile Robotics and Intelligent Systems, 2024: 86-93.

[59] Song C, Wu F, Gao X, et al. Open environments-aware SLAM based on YOLO-enhanced open-vocabulary object detection[J]. Science China Technological Sciences, 2025, 68(11): 2120401.

[60] Lee S S, Lim L G, Palaiahnakote S, et al. Oil palm tree detection in UAV imagery using an enhanced RetinaNet[J]. Computers and Electronics in Agriculture, 2024, 227: 109530.

[61] Duan K, Bai S, Xie L, et al. CenterNet++ for object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(5): 3509-3521.

[62] Li B, Yang H, Xia C, et al. Adaptive DETR: a framework with dynamic sampling points and feature-guided adaptive attention updates [J]. Computer Vision and Image Understanding, 2025, 260: 104481.

[63] Yang Y. FastMOT: High-performance multiple object

tracking based on deep SORT and KLT[J]. Journal of Korea Multimedia Society, 2020, 20: 893-910.

[64] Jiang L, Wang C, Yang H. Joint detection and embedding of multi-object tracking with feature decoupling [J]. Signal, Image and Video Processing, 2025, 19(8): 655.

[65] Li Z, Wang J, Duan A, et al. An OCT retinal image classification model based on improved ResNet-34 network [J]. Biomedical Physics & Engineering Express, 2025, 11167: 41125104.

[66] Liu B ,Wang J ,Wang Z . Combined multilevel detection and trajectory prediction with FairMOT for pedestrian multi-object tracking[J]. IAENG International Journal of Computer Science, 2025, 52(4):1137.

[67] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A benchmark for multi-object tracking[J]. arXiv preprint arXiv: 1603.00831, 2016.

[68] Guan J, Cheng H Y, Wu Y P, et al. Multi-target tracking for star sensor based on CenterTrack deep learning model [J]. Scientific Reports, 2025, 15: 37125.

[69] Fischer T, Huang T E, Pang J, et al. QDTrack: quasi-dense similarity learning for appearance-only multiple object tracking [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15380-15393.

[70] Shuai B, Berneshawi A, Li X, et al. SiamMOT: Siamese multi-object tracking [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12367-12377.

[71] Hui B, Feng J, Yao Q, et al. Tracking multi-objects with anchor-free Siamese network [C]//Intelligence Science IV, 2022: 402-408.

[72] Zhai H, Luo N, Yang Y, et al. Multi-focus image fusion *via* multi-scale attention and Siamese networks[J]. Digital Signal Processing, 2026, 168: 105493.

[73] Sun P, Cao J, Jiang Y, et al. Transtrack: Multiple object tracking with transformer [J]. arXiv preprint arXiv: 2012.15460, 2020.

[74] Meinhardt T, Kirillov A, Leal-Taixé L, et al. TrackFormer: multi-object tracking with transformers [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8834-8844.

[75] Ma H, Zhao Y, Yin Z, et al. Automatic multiple-object tracking of farm livestock with improved MOTR[J]. Expert Systems with Applications, 2025, 281: 127605.

[76] Wu J, Cao J, Song L, et al. Track to detect and segment: an online multi-object tracker[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12347-12356.

[77] Qin Z, Zhou S, Wang L, et al. MotionTrack: learning robust short-term and long-term motions for multi-object tracking [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 17939-17948.

[78] Zhao J, Zhang M, Zhou Z, et al. CFFormer: a cross-fusion transformer framework for the semantic segmentation of multisource remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 63: 4401117.

[79] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2411-2418.

[80] Huang L, Zhao X, Huang K. 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.

[81] 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, 2018: 310-327.

[82] Fan H, Lin L, 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), 2019: 5369-5378.

[83] Leal-Taix′e L, Milan A, Reid I, et al. Motchallenge 2015: Towards a benchmark for multi-target tracking [J]. arXiv preprint, 2015, arXiv:1504.01942.

[84] Dendorfer P, Rezatofighi H, Milan A, et al. Mot20: A benchmark for multi object tracking in crowded scenes [J]. arXiv preprint arXiv:2003.09003, 2020.

[85] Yu F, Chen H, Wang X, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2633-2642.

[86] Dave A, Khurana T, Tokmakov P, et al. TAO: a large-scale benchmark for tracking any object [C]//Computer Vision – ECCV 2020, 2020: 436-454.

作者简介：

罗　元(1972—),女,湖北省宜昌市人,教授,硕导,博士,主要研究方向为移动机器人SLAM、机器视觉;

马文龙(2001—),男,山东省临沂市人,硕士研究生,主要研究方向目标跟踪。