光学学报

深度学习单目标跟踪方法的基础架构研究进展

许廷发[1,2*]，王颖[1]，史国凯[3]，李天昊[1]，李佳男[1**]

[1]北京理工大学光电学院光电成像技术与系统教育部重点实验室，北京 100081；

[2]北京理工大学重庆创新中心，重庆 401120；

[3]北方自动控制技术研究所，山西 太原 030006

摘要 单目标跟踪是计算机视觉领域重要的分支，旨在对视频序列中的指定目标进行连续跟踪。近年来，基于深度学习的单目标跟踪方法发展迅猛，其中基于孪生网络的双流跟踪方法和基于Transformer的单流跟踪方法是两种基础架构。本文从原理、组成结构、局限性及未来发展方向等角度对这两种架构进行了全面介绍与分析。另外，数据集是方法训练及评测的基石，本文汇总了当前主流的深度学习单目标跟踪数据集，详细阐述了跟踪方法在数据集上的评测方式及评测指标，并总结了多种方法在数据集上的表现。最后，从宏观角度分析了深度学习目标跟踪方法的未来发展趋势，以期为相关研究人员提供参考。

关键词 深度学习目标跟踪；单目标跟踪；深度学习；孪生网络；Transformer

中图分类号 TP121 **文献标志码** A **DOI**：10.3788/AOS230746

1 引 言

单目标跟踪(SOT)是计算机视觉的基础问题之一，因其在智能视频监控、人机交互、自动驾驶、目标分析等领域具有重要应用而受到国内外学者和业界的广泛重视。对于给定的视频序列，单目标跟踪方法须根据初始帧中待跟踪目标的状态(多为目标边界框)，对后续视频序列中目标的状态(位置及尺寸)实现实时、准确的预测。与目标检测不同，目标跟踪任务中的跟踪目标无指定类别，且跟踪场景复杂多变，存在目标尺度变化、目标遮挡、运动模糊、目标消失等诸多问题。因此，对目标进行实时、精准、鲁棒的跟踪是一项极具挑战的任务。

主流的单目标跟踪方法按照结构成分可分为三类：判别式相关滤波方法[1]、基于孪生网络的方法[2-3]和基于Transformer的方法[4]。首先，判别式相关滤波方法通过优化岭回归模型在线学习相关滤波器，实现目标与背景的分离。采用手工特征的相关滤波方法速度快、易部署，但方法的精度及鲁棒性仍有很大提升空间；2016年，Bertinetto等[2]提出基于深度学习的SiamFC方法。该方法首次将双分支的孪生网络概念引入目标跟踪任务中，通过共享参数的骨干网络分别提取目标模板和搜索帧特征，并将跟踪任务构建为两特征之间的相似度匹配问题。同时，该方法可利用监督数据集进行端到端的离线训练，且跟踪速度快，所以一经提出即受到广泛关注。总的来说，基于孪生网络的目标跟踪方法可分为三个基本组成模块：特征提取、特征融合及跟踪头。后续工作通过引入深度骨干网络[5]、优化特征融合[6]、设计跟踪头部及损失函数[7]等方法进一步提升方法性能；Transformer[8]是一种基于注意力机制的编码-解码器结构，最初用于机器翻译任务[8]。作为Transformer结构的核心，注意力机制模块是非常灵活的，具有全局和动态的建模能力，且无需数据及任务先验信息，在图像分类[9]、目标检测[10]、语义分割[11]等视觉任务中有出色的表现。基于Transformer的目标跟踪方法即在单目标跟踪任务中引入Transformer结构，该结构善于捕获特征的长程依赖，非常适用于成对匹配任务。在对特征融合进行改进的过程中，TransT方法[12]首先采用堆叠的Transformer结构替代传统的交叉卷积结构实现模板特征与搜索特征之间的信息交互，该方法性能远超同期方法。

然而，Transformer的能力远不止于此。2022年，SBT[13]、OsTrack[14]等方法则完全移除卷积操作，提出基于Transformer的全注意力机制目标跟踪方法。区别于孪生网络中双流的骨干网络结构，该方法采用单流结构，将拼接后的目标模板和搜索帧一同输入基于Transformer的骨干网络中，同时完成特征的提取与融

收稿日期：2023-03-29；**修回日期**：2023-05-29；**录用日期**：2023-06-15；**网络首发日期**：2023-06-21

基金项目：国家自然科学基金青年科学基金(62101032)

通信作者：*ciom_xtf1@bit.edu.cn；**lijianan@bit.edu.cn

合。在此过程中，模板特征与搜索特征之间建立了双向信息流，进而可输出为跟踪目标定制的搜索特征。该方法简单、整齐、高效，在多个大规模数据集上的表现遥遥领先。同时，采用预训练技术得到的Transformer权重可用来初始化网络参数，提高网络收敛速度的同时进一步提高方法精度。

目前，基于深度学习的单目标跟踪方法为研究重点，故本文将对基于深度学习的目标跟踪方法进行回顾、归纳和整理。与既往的分类方式（即上述将基于深度学习的目标跟踪方法按照结构成分划为基于孪生网络的方法和基于Transformer的方法）不同，本文将跟踪方法按照架构类型分为两类：基于孪生网络的双流跟踪方法和基于Transformer的单流跟踪方法。首先，本文分别介绍这两类基础架构的基本原理并总结其发展脉络。其中，对双流架构的分析基于其三个基本模块：特征提取、特征融合和跟踪头。大规模跟踪数据集（如VOT[15]、GOT-10k[16]、LaSOT[17]、TrackingNet[18]等）的发展极大推动了基于深度学习的跟踪方法的进步。故本文后续介绍了深度学习方法中常用的目标跟踪数据集及其评价标准，并整理了代表性方法在这些数据集上的表现。最后，本文对目标跟踪方法的发展方向进行了展望。

目前已存在一些综述工作，例如，Li等[19]和Marvasti-Zadeh等[20]均从网络结构、网络函数以及训练方法等角度整理了跟踪方法的发展历程；Javed等[21]详细分析了基于判别式相关滤波的跟踪方法和基于孪生网络的跟踪方法的原理和挑战。尽管这些综述对跟踪方法进行了全面的总结，但均缺少对Transformer相关跟踪方法的介绍。本文则在既有工作的基础上，引入了领域最新成果，同时创新地从架构类型角度对主流目标跟踪方法进行分类，清晰地呈现了方法改进的原因及方向，以期为后续工作提供参考。

2 基于深度学习的单目标跟踪方法

近些年，深度学习技术的发展极大地推动了单目标跟踪方法的进步。通过在大规模跟踪数据集上进行离线训练，单目标跟踪方法能够提取更加鲁棒且语义信息丰富的特征，使得这些方法能够应对不同的复杂场景。多种网络结构及学习范式可用于完成单目标跟踪任务：文献[22]采用循环神经网络（RNNs）；文献[23]采用生成对抗网络（GANs）；文献[24]采用元学习方法（ML）；文献[25]采用自监督学习方法（SSL）等。

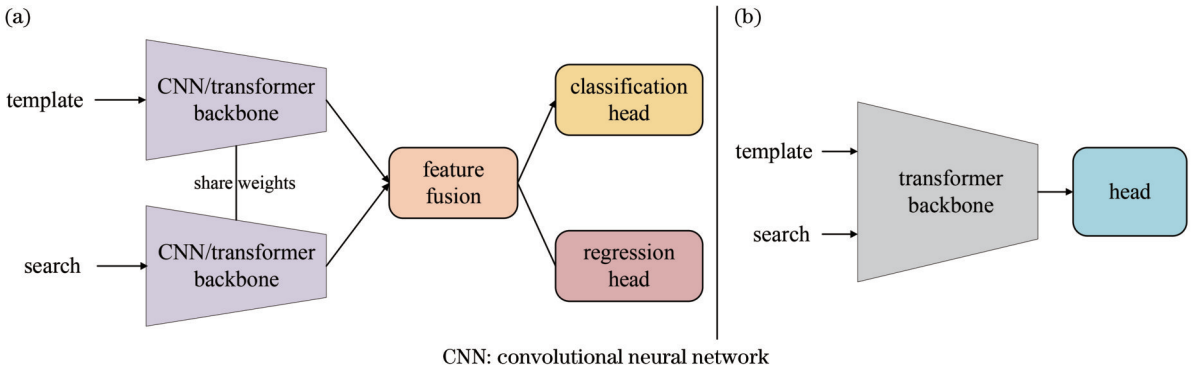

图1 单目标跟踪方法的两种基础架构。(a)基于孪生网络的双流跟踪方法；(b)基于Transformer的单流跟踪方法

Fig. 1 Two basic architectures of single object tracking methods. (a) Siamese network-based two-stream tracking method; (b) transformer-based one-stream tracking method

如图1所示，基于孪生网络的双流跟踪方法（以下简称双流方法）和基于Transformer的单流跟踪方法（以下简称单流方法）是单目标跟踪方法的两种基础架构。总的来说，两种架构均为端到端的离线训练网络。通常情况下，方法将带有标注的初始帧作为模板帧，后续帧作为搜索帧。通过对模板与搜索帧进行截取、数据增强等操作，提取固定尺寸的模板-搜索图像对。方法以图像对作为输入、输出目标的状态估计。而两种架构的不同之处在于：双流方法采用共享参数的双流骨干网络分别提取模板特征与搜索特征，并采用特征融合模块将模板信息注入搜索特征；而单流方法直接采用一个骨干网络同时处理模板特征与搜索特征，输出为模板定制的搜索特征。下文将分别详细介绍两种架构的基本原理及发展历程。图2按时间顺序总结了基于深度学习的单目标跟踪方法的标志性进展。其中，方法名称上方的标记用于表明方法的主要贡献构成。

2.1 双流方法

2016年，Bertinetto等[2]提出双流方法SiamFC，开启了孪生网络结构在跟踪方法中的应用。方法将跟踪任务构建为目标模板与搜索帧之间的相似度匹配问题，由特征提取、特征融合和跟踪头三个基本模块构成。具体来说，对于模板$Z\in\mathbb{R}^{3\times H_z\times W_z}$和搜索帧$X\in\mathbb{R}^{3\times H_x\times W_x}$，共享权重的双流骨干网络$\varphi$用于提取模板特征$\varphi(z)$和搜索特征$\varphi(x)$。两特征经过融合后，输出表征模板特征与搜索特征相似度的矩阵：

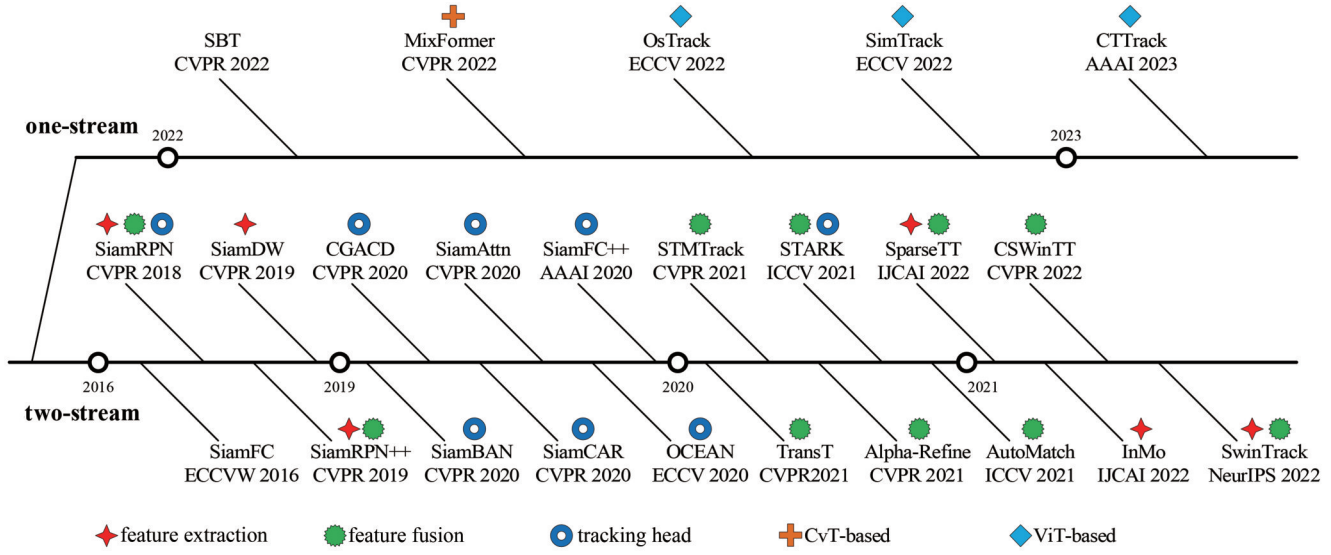

图 2 基于深度学习的单目标跟踪方法的标志性进展

Fig. 2 Landmark advances in deep learning-based single object tracking methods

$$f(z,x)=\varphi(z)\otimes\varphi(x), \tag{1}$$

式中，$\otimes$ 表示特征融合操作。随后，跟踪头（分类分支/回归分支）输出目标状态估计矩阵（目标存在概率热图/坐标矩阵）。在后续的改进方法中，特征提取网络由浅至深，特征融合模块由粗至细，目标跟踪头由繁至简，该方法在复杂场景下的表现逐渐提高。以下将分别分析并总结各模块的发展进程。

2.1.1 特征提取

特征提取部分采用流行的视觉骨干网络结构，按结构类型可分为基于卷积神经网络（CNN）的骨干网络和基于 Transformer 的骨干网络。从采用 AlexNet 网络[26]的 SiamFC 方法[2]到采用 Swin-Transformer[27]的 SwinTrack[28]方法，跟踪方法中使用的骨干网络结构由浅至深，提取到的特征更加鲁棒，目标的表观模型也具有更强的判别能力。

1）基于 CNN 的骨干网络

SiamFC[2]和 SINT[29]采用传统的、无填充的 AlexNet[26]作为骨干网络。同时期，结构更深更宽的现代神经网络，如 VGG[30]、Inception[31]和 ResNet[32]已经在多种视觉任务中具有出色的表现，但难以应用于跟踪任务。实验证明，简单地将 AlexNet[26]更换为深度网络会导致效果的下降。探究如何在跟踪方法中充分利用现代神经网络结构成为了亟待解决的难题。SiamDW[33]中通过实验分析得出，神经元感受野大小、网络步幅和特征填充是影响跟踪精度的三个重要因素。感受野和步幅直接影响输出特征图的分辨性，继而影响方法精度；而特征填充则在结构中引入了潜在的位置偏差，严重影响边缘特征的质量。同时，SiamDW[33]提出两种可能的解决方法：一是删除网络中的填充操作；二是扩大输入模板及搜索区域的大小，并裁剪出受填充影响的外围特征。众所周知，无填充的骨干网络会加速特征图尺寸的减小，进而导致网络结构无法加深。故 SiamFC++[7]中采用第二种方法，成功应用带特征填充的 InceptionV3[31]作为骨干网络。随后，SiamRPN++[34]方法中提出，特征填充打破了深度网络的平移不变性是导致网络无法加深的根本原因。该方法通过在训练过程中引入一个合适的随机移位因子处理训练数据，成功引入了 ResNet[32]骨干网络，效果得到了显著的提升。自此，InceptionV3[31]和 ResNet[32]等深度网络广泛应用于跟踪方法。

2）基于 Transformer 的骨干网络

继 ViT[35]、DeiT[36]等视觉 Transformer 方法大获成功后，基于 Transformer 的骨干网络被引入跟踪任务，极大提升了跟踪方法的性能。SparseTT[37]和 SwinTrack[28]采用 Swin-Transformer[27]作为特征提取的骨干网络，输出紧凑且语义丰富的特征。Swin-Transformer[27]采用了类似 CNN 的层次化特征构建，有利于其在目标检测、目标跟踪、实例分割等任务上的应用。InMo[38]在 Swin Transformer[27]骨干网络的不同阶段引入孪生网络双流间的特征互动，增强了目标表征能力及干扰目标的鉴别能力。

2.1.2 特征融合

特征提取骨干网络输出模板特征与搜索特征。特征融合模块则用于建模模板特征与搜索特征之间的相似性，可进一步分为基于交叉卷积的特征融合和基于注意力的特征融合。图 3 汇总了 5 种常用的特征融合方式。

1）基于交叉卷积的融合方式

SiamFC 方法[2]首先采用 Naïve 交叉卷积[Naïve cross correlation，如图 3(a)所示]进行特征融合，即以模板特征为卷积核，对搜索特征做步长为 1 的滑动卷积计算。输出的单通道响应图直接反应模板特征与搜

索特征的相似程度，响应值越高表示特征越相似。后续，SiamRPN[39]采用通道提升卷积扩展输出特征通道以嵌入高层次信息。但由于其结构繁重导致网络参数量严重分布不均，进而导致网络训练难度的提升及跟踪效果的降低。因此，SiamRPN＋＋[34]方法中提出了深度交叉卷积(depth-wise cross convolution)，如图 3(b)所示，即模板特征与搜索特征逐通道进行相关操作。此时输出特征的通道数与模板特征和搜索特征的通道数相同。实验证明，该方法丰富了特征的语义信息，降低了模型计算量，同时稳定了训练过程。

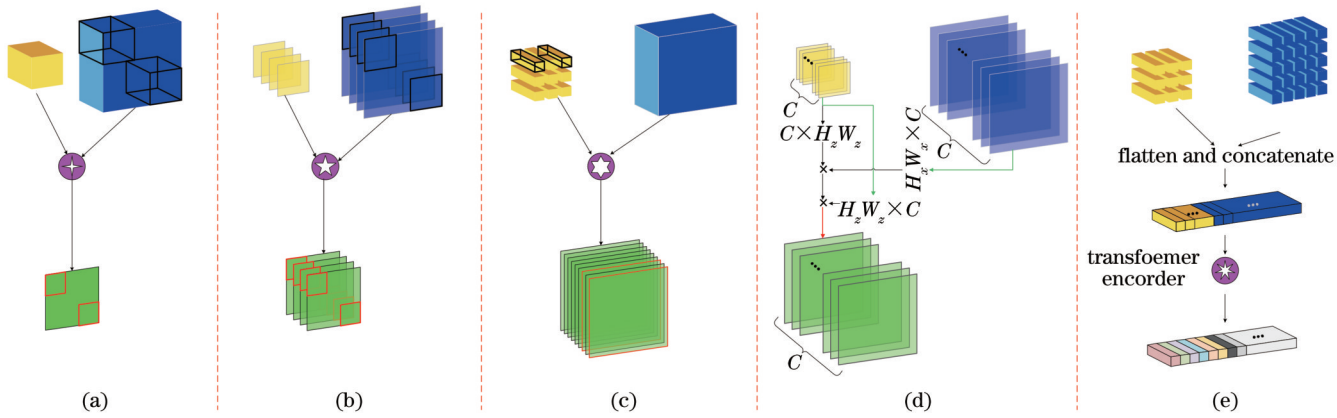

图 3 单目标跟踪方法中的特征融合方式。(a)基于交叉卷积的融合；(b)基于深度交叉卷积的融合；(c)基于像素相关卷积的融合；(d)基于互相关的融合；(e)基于拼接的融合

Fig. 3 Feature fusion methods in single object tracking methods. (a) Naïve cross correlation-based fusion; (b) depth-wise cross convolution-based fusion; (c) pixel-wise correlation-based fusion; (d) cross-attention-based fusion; (e) concatenation-based fusion

实际上，上述交叉卷积均以完整的模板特征作为卷积核，对搜索特征进行卷积操作。但在整个跟踪过程中，固定且描述模板整体的卷积核难以应对目标在快速移动中产生的外观变化。Alpha-Refine[40]方法中提出了像素相关卷积(pixel-wise correlation)，如图 3(c)所示，该方法在空间维度上逐像素切分目标特征得到一组卷积核，并将该组卷积核与搜索特征逐一进行Naïve交叉卷积。输出特征的通道数与卷积核个数相同。实验证明，像素相关卷积更好地保留了特征的空间信息，提升了相关特征的细粒度。在此基础上，PCDHV[41]提出金字塔卷积，通过引入注意力和池化操作优化卷积核组，同时利用组卷积加深特征维度，丰富了输出特征的信息。

2)基于注意力的融合方式

上述基于交叉卷积的特征融合方式为跟踪方法带来了显著的性能提升。但不可否认的是，该类方法仅计算了模板特征与搜索特征间的局部相似性，而忽略了全局信息。近年来，基于注意力的方法在视觉任务中较流行。注意力机制能够建立丰富的全局上下文依赖，善于提取边缘特征及区分相似特征，因此十分适用于单目标跟踪任务。STMTrack[42]方法基于NonLocal[43]结构提出时空记忆模块融合历史模板特征及搜索特征；TransT[12]中采用基于互相关的融合(cross-attention-based fusion)方式，如图 3(d)所示，即采用自注意层完成特征增强，并采用交叉注意层完成特征交互。图 3(d)展示了交叉注意层的实现方式。SparseTT[37]在此基础上引入稀疏多头注意力机制提升方法的前景-背景判别能力，减轻了目标边缘区域的模糊性。

另外一种基于注意力的融合方式为基于拼接的特征融合(concatenation-based fusion)，如图 3(e)所示。STARK[44]中将模板特征和搜索特征串联，并采用多个自注意层完成特征间的信息交互与融合。与基于互相关的特征融合相比，基于拼接的特征融合方式结构对称，能够通过运算减少计算消耗，通过权重共享减少模型参数，同时提升了方法的表现。SwinTrack[28]在此基础上引入运动标志记录目标历史轨迹，提升方法的鲁棒性。CSWinTT[45]提出一种基于多尺度循环移动窗口的注意力模块，将像素级别的注意力提升至窗口级别。该方法有助于维持目标的整体性，并保留更多的位信息。

值得一提的是，SparseTT[37]、SwinTrack[28]的特征提取与特征融合部分均采用Transformer结构，即为基于Full-Transformer的跟踪方法。相较于Full-CNN结构方法[5,7]和CNN-Transformer[4,44]结构方法，Full-Transformer的网络容量大，其输出特征具有更强的表征能力，且鲁棒性更强。

另外，AutoMatch[46]中没有采用传统的特征融合方式或其变体，而是设计了一个模型搜索方法，根据跟踪场景自动选择合适的特征融合算子。将该模块整合到现有跟踪器中可实现显著的性能提升，同时几乎不会降低跟踪速度。

2.1.3 跟踪头

1)基于锚框的跟踪头

跟踪头用于输出目标状态估计(通常用目标包围

框表示)。图 4 对常用跟踪头进行了汇总。SiamFC[2]采用基于分类的头部结构,该方法的特征融合模块直接输出“模板特征与搜索特征的相似度响应图。通过获取响应图的最大值可将目标从背景中分离,获得目标位置。同时,该方法采用多尺度搜索方法获得目标尺度估计,但输出结果的精度难以满足需求。进而,在区域提议网络(RPN)[47]的启发下,SiamRPN[39]在分类分支的基础上引入基于锚框的目标回归分支,如图 4(a)所示。具体来说,分类分支用于前背景分类,输出用于位置估计的置信图;回归分支用于锚框修正,预测与置信图逐像素对应的目标边界框信息,即每个预定义锚框的偏移坐标。分类分支和回归分支共同作用,有效实现了目标状态的精准估计。后续众多方法如 SiamRPN++[34]、SiamDW[33]、SiamAttn[48]等均采用该类跟踪头结构。

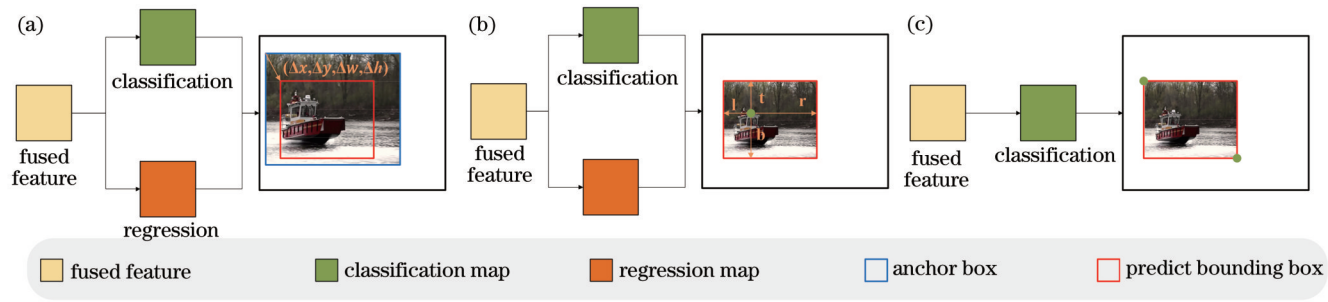

图 4 单目标跟踪方法中的跟踪头。(a)基于锚框的跟踪头;(b)无锚跟踪头;(c)基于角点的跟踪头

Fig. 4 Tracking head in single object tracking methods. (a) Anchor-based head; (b) anchor-free head; (c) corner-based head

基于锚框的回归具有较多缺点:a)该方法对超参数极其敏感,需要使用精细的调参技巧才能获得理想效果;b)锚的尺寸及长宽比是固定的,难以应对目标的快速形变及快速姿态变化;c)该方法需要获取数据分布的先验信息,违背了单目标跟踪任务的基本原则。基于以上观察,越来越多的研究工作致力于构建无锚的分类-回归跟踪头。

2) 无锚跟踪头

SiamBAN[49]充分利用了全卷积网络的表达能力,采用分类分支进行前背景分类,采用回归分支直接回归出对应位置的包围框坐标,如图 4(b)所示。该方法简单高效,且性能优于基于锚框的跟踪头。IoUNet[50]认为分类分支与回归分支存在原理上的不对齐。基于此观察及“中心区域附近特征质量较高”的假设,SiamCAR[51]、SiamFC++[7]在分类-回归双分支的基础上引入质量评估分支,该分支输出则作为权重调整分类分支输出。无锚跟踪头克服了基于锚框的跟踪头的缺点,广泛应用于后续方法中。

3) 基于角点的跟踪头

与双/三分支无锚方法不同,CGACD[52]和 Alpha-Refine[40]采用基于角点的单分支跟踪头输出目标位置,如图 4(c)所示。融合特征经过堆叠的卷积层后输出分别表示目标角点(左上点和右下点)的存在概率图。对存在概率图进行 Soft-argmax 计算可直接得出目标坐标。基于角点的跟踪头更好地利用详细的空间信息,同时单支结构避免了分类-回归双分支方法优化时遇到的分歧。STARK[44]在上述跟踪头的基础上融入了注意力机制,明确地建模了坐标估计中的不确定性,为单目标跟踪给出了更准确和稳健的预测。后续 MixFormer[53]、SimTrack[54]等方法直接调用了该跟踪头结构。

2.2 单流方法

尽管双流方法已经取得了极大的成功,但仍存在以下问题:1)由于双流网络的参数通过离线训练获得,且骨干网络的双分支之间不存在信息交互,故网络提取的搜索特征与目标无关,降低了模型的判别能力;2)特征融合模块的计算量繁重,严重影响了方法的速度。因此文献[14, 53-54]等基于 Transformer 强大的建模能力,提出了单流方法。

单流方法首先将目标模板与搜索帧分别分块和映射后拼接为一个特征。在该特征中嵌入位置编码后输入一个 Transformer 骨干网络,可以同时完成特征提取与特征融合。具体来说,目标模板 $z\in\mathbb{R}^{3\times H_z\times W_z}$ 和搜索区域 $x\in\mathbb{R}^{3\times H_x\times W_x}$ 首先被拆分和拉伸成序列块 $z_p\in\mathbb{R}^{N_z\times(3\cdot P^2)}$ 和 $x_p\in\mathbb{R}^{N_x\times(3\cdot P^2)}$,其中 $P\times P$ 表示每一个块的分辨率,$N_z=H_zW_z/P^2$ 和 $N_x=H_xW_x/P^2$ 分别表示目标模板和搜索区域包含块的个数。z_p 和 x_p 经线性映射 $E\in\mathbb{R}^{(3\cdot P^2)\times D}$ 投影为 D 维向量后分别与位置编码 $P_z\in\mathbb{R}^{N_z\times D}$ 和 $P_x\in\mathbb{R}^{N_x\times D}$ 相加,输出模板特征 $H_z\in\mathbb{R}^{N_z\times D}$ 和搜索特征 $H_x\in\mathbb{R}^{N_x\times D}$。将 H_z 与 H_x 拼接得到 $H=[H_z;H_x]\in\mathbb{R}^{(N_z+N_x)\times D}$ 并输入注意力模块中。线性映射后的 query、key、value 分别可表示为 $Q=[Q_z;Q_x]$、$K=[K_z;K_x]$、$V=[V_z;V_x]$。进而注意力 A 可表示为

$$A = A_{\text{map}} V = \text{Softmax}\left(\frac{QK^{\text{T}}}{\sqrt{d_k}}\right)V = \text{Softmax}\left(\frac{[Q_z;\ Q_x][K_z;\ K_x]}{\sqrt{d_k}}\right)[V_z;\ V_x], \tag{2}$$

式中，A_{map}表示注意力权重。A_{map}可继续拆分为

$$A_{\text{map}} = \text{Softmax}\left(\frac{[Q_z;\ Q_x][K_z;\ K_x]}{\sqrt{d_k}}\right) = \text{Softmax}\left(\frac{[Q_z K_z^{\text{T}}, Q_z K_x^{\text{T}};\ Q_x K_z^{\text{T}}, Q_x K_x^{\text{T}}]}{\sqrt{d_k}}\right) \triangleq [W_{zz}, W_{zx};\ W_{xz}, W_{xx}], \tag{3}$$

式中：Softmax()表示归一化指数函数；W_{zx}表示模板特征与搜索特征的相似性，其余同理。则A可以一步表示为

$$A = [W_{zz}, W_{zx};\ W_{xz}, W_{xx}][V_z;\ V_x] = [W_{zz}V_z + W_{zx}V_x;\ W_{xz}V_z + W_{xx}V_x], \tag{4}$$

式中：$W_{zz}V_z$和$W_{xx}V_x$分别表示模板特征与搜索特征各自的特征增强；$W_{zx}V_x$表示模板特征与搜索特征间的信息交互。经过堆叠的注意力模块处理后，最后一层的搜索特征$W_{xz}V_z + W_{xx}V_x$即可作为网络输出。特征融合操作在整个骨干网络中持续进行，促使网络输出为模板定制的搜索特征。相较于双流网络（分别进行特征提取与特征融合），单流方法结构简单，且不需要关于任务的先验。该与任务无关的网络更有助于构建适用多种任务的通用神经网络架构。

SBT方法中提出，EoC单元为同一物体时产生一致的特征，为相似目标时产生对比特征。其中，EoC-SA基于自注意力机制，用于丰富特征表征以实现更好的匹配，EoC-CA基于交叉注意力机制，用于逐层过滤掉与目标无关的特征。堆叠的EoC单元构成了方法的整个骨干网络。网络首先配合分类头在ImageNet[55]上进行预训练，得到的模型参数用于跟踪模型初始化。MixFormer[53]基于CvT[56]结构，采用一系列混合注意力模型（MAM）堆叠完成特征提取与交互。其中，CvT结构CNN的理想特性（如平移不变性等）引入到ViT架构中，故方法善于同时捕捉特征间的局部和全局依赖性。MAM也通过引入非对称结构降低计算开销。OsTrack[14]、SimTrack[54]则采用ViT[35]作为骨干网络，其中OsTrack[14]引入早期候选消除模块去除包含背景的标志；SimTrack[54]采用中心凹窗口技术增强模板信息。CTTrack[57]在ViT[35]的单流跟踪方法的基础上引入相关掩码解码器，采用掩码图像编码的训练技术增强网络中的信息交互。

另外，预训练技术进一步提升了单流方法的表现。由于跟踪数据集相对较小且跟踪任务复杂，跟踪方法中的骨干网络通常采用预训练模型进行参数初始化。高效的网络预训练方法能够显著提升下游任务的表现及模型训练的稳定性。长期以来，在大规模分类数据集ImageNet[55]上监督训练得到的预训练模型广泛应用于单目标跟踪任务。近些年，预训练技术发展迅速。相比于监督训练，采用对比学习（CL）[58]和掩码图像建模（MIM）[59]得到的预训练模型更利于跟踪任务。SimTrack[54]中对比了不同参数初始化（包含：DeiT[36]提供的监督预训练模型及Moco[58]、CLIP[60]、MAE[59]提供的自监督模型）对方法性能的影响。实验结果证明，基于MAE[59]的预训练模型取得了最好性能。OsTrack[14]、SwinTrack[28]中均采用基于MIM的预训练模型从而具有出色的表现。

2.3 分析与总结

表1对代表性的基于深度学习单目标跟踪方法的组成结构进行了汇总。经过分析可以得出方法各模块的发展趋势为：特征提取网络由浅至深；特征融合模块由粗略到精细；目标跟踪头则由繁至简。其中，单阶段

表1 代表性的基于深度学习的单目标跟踪方法的结构组成

Table 1 Structural composition of representative deep learning-based single object tracking methods

Tracker	Publication	Feature extraction	Feature fusion	Tracking head
SiamFC[2]	ECCVW 2016	AlexNet	Naïve cross correlation	—
SiamRPN[39]	CVPR 2018	AlexNet	Up-channel cross correlation	Anchor-based
SiamRPN++[34]	CVPR 2019	ResNet	Depth-wise cross convolution	Anchor-based
SiamFC++[7]	AAAI 2020	InceptionV3	Depth-wise cross convolution	Anchor-free
SiamCAR[51]	CVPR 2020	ResNet	Depth-wise cross convolution	Anchor-free
PCDHV[41]	ACCV 2021	InceptionV3	Pixel-wise cross convolution	Corner-based
TransT[12]	CVPR 2021	ResNet50	Cross-attention-based fusion	Anchor-free
STARK[44]	ICCV 2021	ResNet	Concatenation-based fusion	Corner-based
SwinTrack[28]	NeurIPS 2022	Swin transformer	Concatenation-based fusion	Anchor-free
MixFormer[53]	CVPR 2022	One-stage feature extraction and fusion		Corner-based
SimTrack[54]	ECCV 2022	One-stage feature extraction and fusion		Corner-based
OSTrack[14]	ECCV 2022	One-stage feature extraction and fusion		Anchor-free

跟踪范式(采用一个骨干网络同时完成特征提取与特征融合)结构简单,具有强大的学习及建模能力,是未来单目标跟踪方法的研究趋势。

3 单目标跟踪数据集及其评价指标

3.1 单目标跟踪数据集

单目标跟踪数据集为跟踪器的评估与比较提供了统一的平台,同时大规模数据集的发展极大地促进了基于深度学习单目标跟踪方法的进步。本文主要对6个常用的数据集进行比较,包括VOT-2018[15]、UAV123[61]、GOT-10k[16]、LaSOT[17]、TrackingNet[18]和TNL2K[62]。表2对上述数据集的信息进行了汇总。

1) VOT-2018[15]

VOT数据集最初在2013年发布于VOT竞赛,并随着竞赛的举办不断扩增数据集容量及标注。数据集包含了60个测试序列,无训练序列。序列中目标类别包含人体、动物、车辆、飞行器等,视频场景包含室内、天空、公路、水域等。每一帧的标注数据除目标包围框外,还包含6种视觉属性,如光照变化、尺度变换等。

表2 常用单目标跟踪数据集

Table 2 Common single object tracking datasets

Dataset	Sequence	Attribute	Average duration /s	Minimum frame	Maximum frame
VOT-2018[15]	60	12	355.90	41	1500
UAV123[61]	123	12	30.48	109	3085
GOT-10k[16]	Total: 9935 Training: 9335 Validating: 180 Testing: 420	6	15.00	51	920
LaSOT[17]	Total: 1400 Training: 1120 Testing: 280	14	83.57	1000	11397
TrackingNet[18]	Total: 30643 Training: 30132 Testing: 511	15	16.70	96	2368
TNL2K[62]	Total: 2000 Training: 1300 Testing: 700	17	20.74	21	18488

2) UAV123[61]

该数据集包含123个由低空无人机采集的遥测视频序列,拟用于探究及推进空中视觉跟踪技术的进步。数据集包含了123个测试序列,无训练序列。由于视频由无人机拍摄,视频序列中会出现目标完全消失、目标剧烈抖动等现象,进一步增添了跟踪难度。

3) GOT-10k[16]

该数据集为具有挑战性的大规模数据集。数据集覆盖真实世界中的563种常见户外目标,包含约1万个训练序列和420个测试序列。目标包围框数量超过150万个,且均为人工标注。数据集的训练集和测试集之间无类别重叠,故该数据集提出ont-shot协议:待测试跟踪器应严格保证仅使用该数据集的训练集进行训练。如此,测试结果即能直接反应方法的泛化能力,即跟踪未知目标的能力。

4) LaSOT[17]

该数据集为具有挑战性的大规模长时跟踪数据集,共包含1400个视频序列(1120个训练序列及280个测试序列)。数据集的平均视频长度超过2500 frame,且逐帧采用手工标注。跟踪器在该数据集上的表现能直接反应目标的长时跟踪能力。

5) TrackingNet[18]

该数据集为目前数据量最大的单目标跟踪数据集,包含超过3万个标注视频。数据集分为12个训练子集,每个子集包含2511个序列。数据集广泛覆盖了不同背景下的多种目标,为大容量网络的训练提供了监督数据的保障。

6) TNL2K[62]

该数据集为包含自然语言标注的多模态跟踪数据集。数据集共包含2000个视频序列和663个单词。除目标边界框外,每个视频序列还包含一条关于目标的描述语句,旨在通过自然语言消除目标的歧义行,实现对目标的精准定位。

以上数据集包含多种挑战属性,包含部分遮挡(POC)、全部遮挡(FOC)、尺度变化(SV)、长宽比变化(ARC)、快速运动(FM)、光照变化(Ⅳ)、低分辨(LR)、出视野(OV)、相机运动(CM)、背景杂波(BC)、视角转变(VC)、小目标(SOB)、形变(DEF)、运动模糊(MB)、平面内旋转(IPR)、平面外旋转(OPR)、生成目标影响(AS)、红外相似强度目标(TC)和红外可见多模态(MS)。表3汇总了GOT-10k[16]、UAV123[61]、LaSOT[17]、TrackingNet[18]和TNL2K[62]的属性类别。

表 3 常用单目标跟踪数据集的视频属性

Table 3 Video properties of common single object tracking datasets

No.	Attr	Description	GOT-10k	UAV123	LaSOT	TrackingNet	TNL2K
1	POC	Partial occlusion	√	√	√	√	√
2	FOC	Full occlusion		√	√	√	√
3	SV	Scale variation	√	√	√	√	√
4	ARC	Aspect ratio change	√	√	√	√	√
5	FM	Fast motion	√	√	√	√	√
6	IV	Illumination variation	√	√	√	√	√
7	LR	Low resolution	√	√	√	√	√
8	OV	Out-of-view		√	√	√	√
9	CM	Camera motion		√	√	√	√
10	BC	Background clutter		√	√	√	√
11	VC	Viewpoint change		√	√		√
12	SOB	Similar object				√	
13	DEF	Deformation			√	√	√
14	MB	Motion blur			√	√	√
15	IPR	In-plane rotation			√	√	√
16	OPR	Out-of-plane rotation				√	√
17	AS	Influence of adversarial samples					√
18	TC	Two targets with similar intensity cross each other					√
19	MS	Video contain both color and thermal images					√

3.2 单目标跟踪方法评价指标

当前，单目标跟踪数据集的真值标注及跟踪器的输出结果多为轴对齐的包围框。跟踪器预测结果与真值之间的差异可以反应跟踪器的性能。按照评测原理，可将评价指标分为成功率(S)和准确率(P)两大类别，每一类可引申出不同的表示形式。表 4 对单目标跟踪方法的评价指标进行了整理。

表 4 单目标跟踪方法的评价指标

Table 4 Evaluation metrics of single object tracking methods

Category	Principle	Evaluation metric	Applicable dataset
S	Intersection over union (IoU) between tracking results and groundtruths	AO (average overlap)	GOT-10k
		SR (success rate)	GOT-10k
		Success plot	LaSOT
		AUC (area under the curve)	LaSOT, TrackingNet
P	Pixel distance between centers of tracking results and groundtruths	Precision plot	LaSOT, TrackingNet
		Precision	LaSOT, TrackingNet
		Normalized precision	LaSOT, TrackingNet

1) 成功率

成功率为预测包围框与真值包围框之间像素级别的交并比(IoU)，能够反映跟踪器估计目标尺寸的性能，即

$$S = I_{\mathrm{oU}}(p, g) = \frac{p \cap g}{p \cup g}, \tag{5}$$

式中：I_{oU}表示交并比；p表示预测边框；g表示真实边框。平均重叠率(AO)可以反映方法在整个测试集上的性能，即

$$O_{\mathrm{A}} = \sum_{i=0}^{n} S \Big/ n, \tag{6}$$

式中，O_{A}表示平均重叠率。重叠率可以反映方法是否在当前帧中成功跟踪目标物体，如设定重叠率大于某个阈值(例如 0.5)表示该帧中成功跟踪目标。故可采用某个阈值下的成功比率(SR)评估方法的性能。常用指标有 $\mathrm{SR}_{0.50}$ 和 $\mathrm{SR}_{0.75}$，分别表示视频序列中 $S > 0.50$ 和 $S > 0.75$ 的帧数比重。以阈值 T 作为横坐标，视频序列中满足 $S > T$ 条件的帧数比重作为纵坐标，可绘制成功率曲线(success plot)。将众多跟踪器的成功率曲线进行对比，根据成功率曲线定义可知，“右上方”的曲线对应的跟踪器性能最优。进而，曲线下面积(AUC)可以用于反映方法性能，AUC 值越大，方法性

能越优。

2) 精确率

精确率为预测包围框与真值包围框中心位置的欧氏距离，能够反映跟踪器对定位目标的性能，即

$$P=\sqrt{\|(p_x,p_y)-(g_x,g_y)\|},\tag{7}$$

式中：(p_x,p_y)表示预测包围框的中心坐标；(g_x,g_y)表示真值包围框的中心坐标。当方法失去对目标对象的跟踪时，其输出的目标位置是随机的，因此平均精确率无法正确衡量方法性能。相反，可采用精确率在某一特定阈值（通常采用 20 个像素）内的帧数比重衡量跟踪器性能，直接记作精确率（P）。以阈值 T 为横坐标，视频序列中满足 $P>T$ 条件的帧数比重作为纵坐标，可绘制精确率曲线（precision plot）。与成功率曲线相反，“左上方”的曲线对应的跟踪器性能最优。精确率对跟踪序列的分辨率、目标包围框的尺寸非常敏感。进而归一化精确率（normalized precision）可精细化性能评价：

$$P_{\text{norm}}=\|W((p_x,p_y)-(g_x,g_y))\|,\tag{8}$$

式中，

$$W=\text{diag}(g_x,g_y),\tag{9}$$

式中，diag( )表示构造对角矩阵运算。

3.3 性能评估

表 5 对当前主流单目标跟踪方法在 GOT-10k[16]、LaSOT[17]和 TrackingNet[18]数据集上的性能进行了汇总与比较，其中加粗字体为最优性能。可以看出，相对于双流方法，基于 Full-Transfomer 的 SwinTrack[28]方法的表现取得了绝对的领先。而单流方法的性能整体优于双流方法。

表 5 单目标跟踪方法在 GOT-10k、LaSOT 和 TrackingNet 数据集上的性能比较

Table 5 Performance comparison of single object tracking methods on GOT-10K, LaSOT, and TrackingNet datasets

Type	Tracker	Publication	GOT-10k[16]			LaSOT[17]			TrackingNet[18]		
			AO	SR_{50}	SR_{75}	AUC	P_{norm}	P	AUC	P_{norm}	P
Two-stream	SiamFC[2]	ECCVW 2016	39.2	42.6	13.5	—	—	—	57.1	65.4	53.3
	SiamRPN[29]	CVPR 2018	48.1	58.1	27.0	—	—	—	—	—	—
	SiamRPN++[34]	CVPR 2019	51.7	61.6	32.5	49.6	56.9	49.1	73.3	80.0	69.4
	SiamBAN[49]	CVPR 2020	—	—	—	51.4	59.8	52.1	—	—	—
	CGACD[52]	CVPR 2020	—	—	—	51.8	62.6	—	71.1	80.0	69.3
	SiamCAR[51]	CVPR 2020	56.9	67.0	41.5	50.7	60.0	51.0	—	—	—
	SiamAttn[48]	CVPR 2020	—	—	—	56.0	64.8	—	75.2	81.7	—
	SiamFC++[7]	AAAI 2020	59.5	69.5	47.9	54.4	62.3	54.7	75.4	80.0	70.5
	Ocean[63]	ECCV 2020	61.1	72.1	47.3	56.0	65.1	56.6	—	—	—
	TransT[12]	CVPR 2021	67.7	76.8	60.9	64.9	73.8	69.0	81.4	86.7	80.3
	STMTrack[42]	CVPR 2021	64.2	73.7	57.5	60.6	69.3	63.3	80.3	85.1	76.7
	AutoMatch[46]	ICCV 2021	65.2	76.6	54.3	58.2	—	59.9	76.0	—	72.6
	STARK[44]	ICCV 2021	68.8	78.1	64.1	67.1	77.0	—	82.0	86.9	—
	SparseTT[37]	IJCAI 2022	69.3	79.1	63.8	66.0	74.8	70.1	81.7	86.6	79.5
	CsWinTT[45]	CVPR 2022	69.4	78.9	65.4	66.2	75.2	70.9	81.9	**86.7**	79.5
	SwinTrack[28]	NeurIPS 2022	**72.4**	**80.5**	**67.8**	**71.3**	—	**76.5**	**84.0**	—	**82.8**
One-stream	SBT[13]	CVPR 2022	70.4	80.8	64.7	66.7	—	71.1	—	—	—
	MixFormer[53]	CVPR 2022	70.7	80.0	67.8	70.1	79.9	76.3	83.9	88.9	83.1
	OsTrack[14]	ECCV 2022	**73.7**	**83.2**	**70.8**	**71.1**	**81.1**	**77.6**	83.9	88.5	83.2
	SimTrack[54]	ECCV 2022	69.8	78.8	66.0	70.5	79.7	—	83.4	—	**87.4**
	CTTrack[57]	AAAI 2023	72.8	81.3	71.5	69.8	79.7	76.2	**84.9**	**89.1**	83.5

4 总结与展望

当前，单流方法展现出了强大的建模能力，是未来单目标跟踪方法的研究趋势。但尽管该跟踪方法的表现已十分出色，单目标跟踪技术的研究仍具有强烈的需求。本文基于上述分析，对单目标跟踪方法的发展趋势进行展望。

1) 多任务协作跟踪

多任务协作跟踪是采用同一个网络同时完成单目标跟踪任务及多种其他任务（如视频目标分割、多目标跟踪等）。该方法能够促进任务交互，达到协同效果。例如视频目标跟踪任务（VOT）与视频目标分割（VOS）的联合方法中，通过简单地增加目标状态估计分支，即可在提升方法精度的同时，弱化对昂贵的掩膜

标注的需求。Transformer网络在多种类型任务上已经取得了优秀的成绩，预训练技术的发展进一步提升了方法的表现。故Transformer网络在多任务协作方法中具有绝对优势。

2）多模态跟踪

多模态跟踪是当前的另一研究重点。多模态视觉跟踪任务中将2D可见光模态与红外模态、深度模态、事件模态、语言模态和3D模态等相配合，在一定程度上提升方法在杂乱背景和恶劣天气情况下、目标发生快速形变、遮挡及外观歧义等条件下的表现。

除此之外，无人机跟踪、反无人机跟踪等特定场景下的目标跟踪具有较强应用价值；对方法进行优化，使得其在边缘设备上实时准确跟踪具有较强的实际需求；与自然语言处理任务相比，视觉任务数据远远不足，扩展大规模跟踪数据集十分重要。同时，无监督训练技术与目标跟踪方法的结合能减弱对监督数据的依赖，势必成为未来研究的趋势。

5 结 论

近些年，深度视频单目标跟踪方法取得了极大进展。本文从单目标跟踪方法的两个基础架构——双流方法和单流方法，对单目标跟踪方法的发展及现状进行了总结与分析。同时，还总结了当前主流的单目标跟踪数据集及其评价指标，汇总了主流单目标跟踪方法的表现。最后，从宏观角度，总结了视频目标跟踪方法的发展趋势。实现真正意义上的通用、准确、实时、鲁棒的目标跟踪仍旧任重道远，相信在未来，视频目标跟踪技术将不断取得突破，极大地应用于人类的生产生活之中。

参 考 文 献

[1] 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.

[2] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]//Hua G, Jégou H. Computer vision - ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 850-865.

[3] 崔洲涓, 安军社, 张羽丰, 等. 面向无人机的轻量级Siamese注意力网络目标跟踪[J]. 光学学报, 2020, 40(19): 1915001.
Cui Z J, An J S, Zhang Y F, et al. Light-weight Siamese attention network object tracking for unmanned aerial vehicle[J]. Acta Optica Sinica, 2020, 40(19): 1915001.

[4] Xie Z, Geng Z, Hu J, et al. Revealing the dark secrets of masked image modeling[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14475-14485.

[5] Held D, Thrun S, Savarese S. Learning to track at 100 fps with deep regression networks[C]//Computer Vision - ECCV 2016: 14th European Conference, October 11-14, 2016, Amsterdam, The Netherlands. Berlin: Springer International Publishing, 2016: 749-765.

[6] Guo D, Shao Y, Cui Y, et al. Graph attention tracking[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 9543-9552.

[7] 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.

[8] 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, December 4-9, 2017, Long Beach, California, USA. New York: ACM, 2017: 6000-6010.

[9] Rao Y, Zhao W, Liu B, et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification[J]. Advances in Neural Information Processing Systems, 2021, 34: 13937-13949.

[10] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision - ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12346: 213-229.

[11] Wang H Y, Zhu Y K, Adam H, et al. MaX-DeepLab: end-to-end panoptic segmentation with mask transformers[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 5459-5470.

[12] Chen X, Yan B, Zhu J W, et al. Transformer tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 8122-8131.

[13] Xie F, Wang C Y, Wang G T, et al. Correlation-aware deep tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8741-8750.

[14] Ye B T, Chang H, Ma B P, et al. Joint feature learning and relation modeling for tracking: a one-stream framework[M]//Avidan S, Brostow G, Cissé M, et al. Computer vision - ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13682: 341-357.

[15] Kristan M, Leonardis A, Matas J, et al. The sixth visual object tracking vot2018 challenge results[C]//Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018: 3-53.

[16] 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.

[17] 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. New York: IEEE Press, 2020: 5369-5378.

[18] Müller M, Bibi A, Giancola S, et al. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision - ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11205: 310-327.

[19] Li P X, Wang D, Wang L J, et al. Deep visual tracking: review and experimental comparison[J]. Pattern Recognition, 2018, 76: 323-338.

[20] Marvasti-Zadeh S M, Cheng L, Ghanei-Yakhdan H, et al. Deep learning for visual tracking: a comprehensive survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(5): 3943-3968.

[21] Javed S, Danelljan M, Khan F S, et al. Visual object tracking with discriminative filters and Siamese networks: a survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(5): 6552-6574.

[22] Yang T Y, Chan A B. Recurrent filter learning for visual

tracking[C]//2017 IEEE International Conference on Computer Vision Workshops (ICCVW), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2018: 2010-2019.

[23] Song Y B, Ma C, Wu X H, et al. VITAL: visual tracking via adversarial learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8990-8999.

[24] Park E, Berg A C. Meta-tracker: fast and robust online adaptation for visual object trackers[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision - ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11207: 587-604.

[25] Zheng J L, Ma C, Peng H W, et al. Learning to track objects from unlabeled videos[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 13526-13535.

[26] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[27] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9992-10002.

[28] Lin L T, Fan H, Zhang Z P, et al. SwinTrack: a simple and strong baseline for transformer tracking[EB/OL]. (2021-12-02) [2023-02-03]. https://arxiv.org/abs/2112.00995.

[29] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2805-2813.

[30] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2023-02-06]. https://arxiv.org/abs/1409.1556.

[31] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 2818-2826.

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[33] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2020: 4586-4595.

[34] 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. New York: IEEE Press, 2020: 4277-4286.

[35] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2020-10-22) [2023-02-06]. https://arxiv. org/abs/2010.11929.

[36] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention[EB/OL]. (2020-12-23)[2023-02-02]. https://arxiv.org/abs/2012.12877.

[37] Fu Z H, Fu Z H, Liu Q J, et al. SparseTT: visual tracking with sparse transformers[EB/OL]. (2022-05-08) [2023-03-02]. https://arxiv.org/abs/2205.03776.

[38] Guo M Z, Zhang Z P, Fan H, et al. Learning target-aware representation for visual tracking via informative interactions [C]//Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, July 23-29, 2022, Vienna, Austria. California: International Joint Conferences on Artificial Intelligence Organization, 2022: 927-934.

[39] 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. New York: IEEE Press, 2018: 8971-8980.

[40] Yan B, Zhang X Y, Wang D, et al. Alpha-refine: boosting tracking performance by precise bounding box estimation[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 5285-5294.

[41] Wang Y, Xu T F, Li J N, et al. Pyramid correlation based deep Hough voting for visual object tracking[C]//Asian Conference on Machine Learning, November 17-19, Virtual Event. Copenhagen: MLR Press, 2021: 610-625.

[42] Fu Z H, Liu Q J, Fu Z H, et al. STMTrack: template-free visual tracking with space-time memory networks[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13769-13778.

[43] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[44] 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), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 10428-10437.

[45] Song Z K, Yu J Q, Chen Y P P, et al. Transformer tracking with cyclic shifting window attention[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 8781-8790.

[46] Zhang Z P, Liu Y H, Wang X, et al. Learn to match: automatic matching network design for visual tracking[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 13319-13328.

[47] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[48] Yu Y C, Xiong Y L, Huang W L, et al. Deformable Siamese attention networks for visual object tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6727-6736.

[49] Chen Z D, Zhong B N, Li G R, et al. Siamese box adaptive network for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6667-6676.

[50] Jiang B R, Luo R X, Mao J Y, et al. Acquisition of localization confidence for accurate object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision - ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11218: 816-832.

[51] Guo D Y, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6268-6276.

[52] Du F, Liu P, Zhao W, et al. Correlation-guided attention for

corner detection based visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6835-6844.

[53] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 248-255.

[54] 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. New York: IEEE Press, 2022: 13598-13608.

[55] Chen B Y, Li P X, Bai L, et al. Backbone is all your need: a simplified architecture for visual object tracking[M]//Avidan S, Brostow G, Cissé M, et al. Computer vision - ECCV 2022. Lecture notes in computer science. Cham: Springer, 2022, 13682: 375-392.

[56] Wu H P, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 22-31.

[57] Song Z K, Luo R, Yu J Q, et al. Compact transformer tracker with correlative masked modeling[EB/OL]. (2023-01-26)[2023-02-06]. https://arxiv.org/abs/2301.10938.

[58] Chen X L, Xie S N, He K M. An empirical study of training self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV), October 10-17, 2021, Montreal, QC, Canada. New York: IEEE Press, 2022: 9620-9629.

[59] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 18-24, 2022, New Orleans, LA, USA. New York: IEEE Press, 2022: 15979-15988.

[60] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[EB/OL]. (2021-02-26)[2023-02-03]. https://arxiv.org/abs/2103.00020.

[61] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[M]//Leibe B, Matas J, Sebe N, et al. Computer vision - ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 445-461.

[62] Wang X, Shu X J, Zhang Z P, et al. Towards more flexible and accurate object tracking with natural language: algorithms and benchmark[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 13758-13768.

[63] Zhang Z P, Peng H W, Fu J L, et al. Ocean: object-aware anchor-free tracking[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12366: 771-787.

Research Progress in Fundamental Architecture of Deep Learning-Based Single Object Tracking Method

Xu Tingfa[1,2*], Wang Ying[1], Shi Guokai[3], Li Tianhao[1], Li Jianan[1**]

[1]*Key Laboratory of Photoelectronic Imaging Technology and System, Ministry of Education, School of Optics and Photonics, Beijing Institute of Technology, Beijing* 100081, *China*;

[2]*Chongqing Innovation Center, Beijing Institute of Technology, Chongqing* 401120, *China*;

[3]*North Automatic Control Technology Institute, Taiyuan* 030006, *Shanxi, China*

Abstract

Significance Single object tracking (SOT) is one of the fundamental problems in computer vision, which has received extensive attention from scholars and industry professionals worldwide due to its important applications in intelligent video surveillance, human-computer interaction, autonomous driving, military target analysis, and other fields. For a given video sequence, a SOT method needs to predict the real-time and accurate location and size of the target in subsequent frames based on the initial state of the target (usually represented by the target bounding box) in the first frame. Unlike object detection, the tracking target in the tracking task is not specified by any specific category, and the tracking scene is always complex and diverse, involving many challenges such as changes in target scales, target occlusion, motion blur, and target disappearance. Therefore, tracking targets in real-time, accurately, and robustly is an extremely challenging task.

The mainstream object tracking methods can be divided into three categories: discriminative correlation filters-based tracking methods, Siamese network-based tracking methods, and Transformer-based tracking methods. Among them, the accuracy and robustness of discriminative correlation filter (DCF) are far below the actual requirements. Meanwhile, with the advancement of deep learning hardware, the advantage of DCF methods being able to run in real time on mobile devices no longer exists. On the contrary, deep learning techniques have rapidly developed in recent years with the continuous improvement of computer performance and dataset capacity. Among them, deep learning theory, deep backbone networks, attention mechanisms, and self-supervised learning techniques have played a powerful role in the development of object tracking methods. Deep learning-based SOT methods can make full use of large-scale datasets for

end-to-end offline training to achieve real-time, accurate, and robust tracking. Therefore, we provide an overview of deep learning-based object tracking methods.

Some review works on tracking methods already exist, but the presentation of Transformer-based tracking methods is absent. Therefore, based on the existing work, we introduce the latest achievements in the field. Meanwhile, in contrast to the existing work, we innovatively divide tracking methods into two categories according to the type of architecture, i.e., Siamese network-based two-stream tracking method and Transformer-based one-stream tracking method. We also provide a comprehensive and detailed analysis of these two basic architectures, focusing on their principles, components, limitations, and development directions. In addition, the dataset is the cornerstone of the method training and evaluation. We summarize the current mainstream deep learning-based SOT datasets, elaborate on the evaluation methods and evaluation metrics of tracking methods on the datasets, and summarize the performance of various methods on the datasets. Finally, we analyze the future development trend of video target tracking methods from a macro perspective, so as to provide a reference for researchers.

Progress Deep learning-based target tracking methods can be divided into two categories according to the architecture type, namely the Siamese network-based two-stream tracking method and the Transformer-based one-stream tracking method. The essential difference between the two architectures is that the two-stream method uses a Siamese network-shaped backbone network for feature extraction and a separate feature fusion module for feature fusion, while the one-stream method uses a single-stream backbone network for both feature extraction and fusion.

The Siamese network-based two-stream tracking method constructs the tracking task as a similarity matching problem between the target template and the search region, consisting of three basic modules: feature extraction, feature fusion, and tracking head. The method process is as follows: The weight-shared two-stream backbone network extracts the features of the target template and the search region respectively. The two features are fused for information interaction and input to the tracking head to output the target position. In the subsequent improvements of the method, the feature extraction module is from shallow to deep; the feature fusion module is from coarse to fine, and the tracking head module is from complex to simple. In addition, the performance of the method in complex backgrounds is gradually improved.

The Transformer-based one-stream tracking method first splits and flattens the target template and search frame into sequences of patches. These patches of features are embedded with learnable position embedding and fed into a Transformer backbone network, which allows feature extraction and feature fusion at the same time. The feature fusion operation continues throughout the backbone network, resulting in a network that outputs the target-specified search features. Compared with two-stream networks, one-stream networks are simple in structure and do not require prior knowledge about the task. This task-independent network facilitates the construction of general-purpose neural network architectures for multiple tasks. Meanwhile, the pre-training technique further improves the performance of the one-stream method. Experimental results demonstrate that the pre-trained model based on masked image modeling optimizes the method.

Conclusions and Prospects One-stream tracking method with a simple structure and powerful learning and modeling capability is the trend of future target tracking method research. Meanwhile, collaborative multi-task tracking, multi-modal tracking, scenario-specific target tracking, unsupervised target tracking methods, etc. have strong applications and demands.

Key words deep learning-based object tracking; single object tracking; deep learning; Siamese network; Transformer