祝现染, 康凯, 张彬,等. 基于自然语言描述的追踪方法研究综述[J]. 智能计算机与应用,2026,16(4):116-125. DOI:10.20169/j.issn.2095-2163.24062801

基于自然语言描述的追踪方法研究综述

祝现染, 康 凯, 张 彬, 李国胜
(华东光电集成器件研究所, 安徽 蚌埠 233000)

摘 要: 由于传统的基于包围盒(Bounding Box, BBox)的跟踪方式在处理视频序列目标跟踪时存在较大的模糊性和局限性,研究者开始探索使用高层语义信息来指导目标跟踪的新方法。为此,基于自然语言描述的跟踪技术被提出,旨在通过视频序列中目标对象的语言描述来准确定位目标。该方法考虑了将局部搜索和全局搜索结合的策略,其中局部搜索继续采用传统的基于BBox 的跟踪方式,而全局搜索则涉及 Visual Grounding 任务,专注于根据自然语言的描述定位图像中最相关的对象或区域。本文详细阐述了基于自然语言描述的追踪算法,并对全局搜索和局部搜索算法进行了详尽的解析。本文涵盖了 Visual Grounding 任务、目标追踪任务以及基于自然语言描述追踪的历史发展,同时评估了相关数据集和评判标准。通过对这些算法的归纳、分析和总结,本文不仅凝练了基于自然语言描述的追踪技术,还对其进行了综合性的分析与讨论,为该领域的发展做出了重要贡献。

关键词: 目标追踪; Visual Grounding; 自然语言描述; 深度学习
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0116-10

Review of tracking methods based on natural language descriptions

ZHU Xianran, KANG Kai, ZHANG Bin, LI Guosheng
(East China Institute of Optoelectronic Integrated Devices, Bengbu 233000, Anhui, China)

Abstract: Due to the inherent ambiguity and limitations of traditional Bounding Box (BBox)-based tracking methods in video sequence target tracking, researchers have been prompted to explore new methodologies that utilize high-level semantic information for guiding target tracking. Consequently, tracking technology based on natural language descriptions has been introduced, which is designed to accurately locate targets within video sequences through verbal descriptions of the objects. This approach integrates a combined strategy of local and global searches; local search maintains the conventional BBox-based tracking method, while global search involves the task of Visual Grounding, focusing on identifying the most relevant objects or regions in images based on natural language descriptions. This paper provides detailed descriptions of tracking algorithms that utilize natural language, and conducts comprehensive analyses of both global and local search methodologies. It covers the tasks of Visual Grounding and target tracking, and explores the historical development of tracking based on natural language descriptions, assessing relevant datasets and benchmarking standards. Through collating, analyzing, and synthesizing these algorithms, the paper not only refines the technology for tracking based on natural language descriptions but also offers a comprehensive analysis and discussion on the subject, contributing significantly to the field.

Key words: target tracking; Visual Grounding; natural language descriptions; deep learning

0 引 言

随着计算机视觉技术的飞速发展,目标追踪技术已成为机器视觉领域的核心研究课题之一。在无人机追踪、智能驾驶等众多应用场景中,目标追踪技术发挥着至关重要的作用。然而,传统的基于包围盒(Bounding Box, BBox)的目标追踪方法存在一定的局限性,特别是在处理目标外观变化和复杂场景干扰时,其性能往往受限。为克服这些局限,近年来,研究者们引入了自然语言描述来辅助视觉信息,提出了基于自然语言描述的目标追踪方法(Tracking by Natural Language, TNL)。该方法结合了自然语言处理和视觉目标追踪的优势,通过解析语言中的描述来指导目标的定位与跟踪,能够更精

基金项目: 国家科技创新 2030"新一代人工智能"项目(2018AAA0103100)。
作者简介: 祝现染(1989—),男,硕士,工程师,主要研究方向:人工智能的新型感知器件和芯片技术。Email:zhuxianran@126.com。
收稿日期: 2024-06-28

准地处理目标的动态变化和环境因素的干扰。

最初基于自然语言描述的视觉目标追踪方法是基于手工设计的模板匹配和颜色直方图的方法，由于手工特征的局限性，这些方法无法处理复杂场景和目标外观的变化。近年来，随着深度学习技术的发展，基于深度学习的方法逐渐成为主流。例如，通过将自然语言描述和视觉信息相结合，使用卷积神经网络（CNN）提取视觉特征，再结合循环神经网络（RNN）或注意力机制来建模语义信息和上下文信息，从而实现基于自然语言描述的视觉目标追踪。在此基础上，部分研究者提出了基于跨模态学习的方法来进一步提高基于自然语言描述的视觉目标追踪的性能。这些方法利用多模态信息（如视觉信息和语义信息）来提高追踪的准确性和鲁棒性。此外，基于强化学习的方法也被提出，可自动化选择最佳追踪策略，进一步优化追踪过程的效率和效果。

本文对传统的视觉追踪和视觉定位方法以及近几年基于自然语言描述的目标追踪方法进行了详细介绍和分析，并对基于自然语言描述的追踪技术进行凝练，展开综合性分析与讨论。

1　基于 Bounding Box 的单目标追踪概述

目标追踪技术是机器视觉中一个重要的研究课题，被广泛应用于民事和军事活及无人机追踪，智能驾驶等领域。目标追踪在机器视觉任务中占比很大，其任务简单来说就是给定任意物体的初始状态，要求在后续的视频或者图像序列中追踪物体的运动轨迹。这里的初始状态通常用包含追踪物体大部分信息的 BBox 表示，也叫做模板区域。后续追踪过程通过对比模板区域与以模板区域中心为中心、但面积更大的搜索区域的相似度，从而找到物体的运动轨迹。

目标追踪技术从光流法等经典算法到基于相关滤波（correlation filter）和深度学习（Deep Learning）的追踪算法经历了长时间的发展。

1.1　相关滤波算法的发展历程

相关滤波追踪算法的流程简单来说就是通过初始帧的信息（Initial frame）构造一个相关滤波器，然后使用此相关滤波器与搜索区域进行相关操作找寻出最大响应的区域，将该区域作为下一帧的目标继续进行跟踪。该思想源于信号处理，类似于判断两个信号的相似程度。2010 年，基于 MOSSE（最小输出平方误差和）的目标追踪算法[1]横空出世，该算法将时间域的计算转化为频域上的计算，大大减少了计算量，其追踪速度达到了 669 fps/s，是当时主流追踪算法速度的 20 倍，将目标追踪速度推向了一个新的高度，为跟踪领域的实时性要求提供了有效的解决方案。但是作为相关滤波的开山算法，MOSSE 算法在特征表示，模板尺寸选择以及边界效应的影响等方面，还是存在很多不足。此后研究者对相关滤波追踪算法从各个方面进行了更加深入的研究。

文献[2]将 MOSSE 中提出的损失函数进行了改进，加入正则项（或惩罚项）以防止函数过拟合，其次引入了比较重要的循环矩阵（通过循环采样生成）和核函数。循环矩阵将原始数据进行循环移位拓展，然后将循环移位后的数据放在矩阵中的每一行，形成循环矩阵，这样一来就可以使用循环矩阵的频域性质，进一步提升追踪速度。核函数本质是一种映射问题，通过非线性函数对数据进行映射，进而描述一些线性函数无法表述的特征。

文献[3]提出的核相关滤波（DCF）采用了尺度不变的 HOG 特征（方向梯度直方图特征），其与 MOSSE 的不同是在核函数的使用不同。文献[4]在之前的基础上提出了 CN 特征（多通道颜色特征），将之前的 RGB 特征映射到多个颜色通道，在特征表示上进行了极大的改善。文献[5]使用 HOG 特征和 CN 特征共同描述追踪的目标，在各个特征模板上分别得到最后的响应得分图，最后使用两个得分图的线性结合求解出目标所在的位置。结合之后，不仅追踪的准确性效果得到了加强，帧率也达到了 80 fps/s。

DSST[6]算法在 KCF 的基础上引入特征融合机制，采用有方向梯度直方图特征、灰度特征以及 CN 特征等多种特征，算法简洁，并且引入了图像金字塔对物体尺度这一问题进行了改进。此外，还有一些算法也对相关滤波的边界效应、遮挡问题进行了解决，例如引入正则化系数矩阵、加入新的响应检测指标等。这些方法归纳起来就是对损失函数进行创新和整改，从而提高追踪算法的性能。

自从相关滤波被引进了目标追踪领域之后，该算法在很长一段时间内统治了目标追踪领域，因为相关滤波算法具有比当时算法更好的准确性，更重要的是具备优异的实时性。同时为将通信领域一些有关信号处理的知识迁移到追踪领域，为追踪问题的解决提供了新的解决办法。直到后来深度学习和神经网络的出现和蓬勃发展，目标追踪领域才渐渐出现多样性，但是相关滤波的追踪思想一直影响追踪算法的发展，同时也受到很多人的关注，特别是在

工程应用领域。

1.2 深度学习追踪算法的发展历程

最近几年,随着深度学习的发展,特别是神经网络和注意力机制的提出,深度学习被广泛的应用在图像分类、目标追踪、图像分割等视觉领域,特别是神经网络结构的多样性造就了其在目标追踪领域的强大功能。卷积神经网络(Convolutional Neural Network, CNN)的出现很大程度上解决了追踪目标特征提取问题,在此之前,大多数的机器视觉任务采用的特征大多是人工设计特征,这些特征仅依据某一规律或者概念,比较单一和浅显。但 CNN 凭借着其深层复杂的结构可以进行多次卷积从而获取更多深层特征。脉冲神经网络(Spiking Neural Network, SNN)则具有在线更新策略、抑制背景、一次性本地检测任务的特点。循环神经网络(Recurrent Neural Network, RNN)可以利用上下文的信息来处理时间维度上的信息,更好地利用语义环境,解决类似物体遮挡,物体形变等挑战。生成式对抗网络[7](Generative Adversarial Networks, GAN)最强大的特点就是其可以增强正样本的数量,极大程度上解决训练样本分布不平衡的问题,同时可以通过生成与真实数据分布相似的样本,有效扩展训练数据集,提高目标跟踪模型的泛化能力,例如,Kim 和 Park 提出了用于目标跟踪的孪生对抗网络(SANT),使用带有 Siamese 判别器的相似性学习,并利用 SAN 损失逐步训练时空生成网络,从而提升跟踪器的性能[8]。最近几年的注意力机制配合编码器解码器结构的兴起与应用使得基于深度学习的追踪方法得到进一步的加强。

文献[9]将深度学习算法引入到了目标追踪领域,其创新之处在于使用卷积神经网络提取追踪目标的运动特征并且将这种特征添加到目标追踪过程当中,强力的特征提取能力也让该算法取得了 2015 年 VOT 的冠军。也有很多的算法融合多个卷积神经网络结构来发现追踪物体的多模态信息。文献 SiamFC[9]将孪生网络(Siamese network)引入到了目标追踪领域,开辟了经典的基于深度学习的追踪框架(pipeline),该框架分为两条支路(模板支路和搜索支路),两条支路通过共享权重的神经网络提取特征,然后经过相似性函数计算两者特征的相似度,最终得到得分图,通过观察得分图上的响应来判断追踪位置,通过步长将响应图上的位置映射回原来的搜索图像内。总的来说,SiamFC 的提出在追踪模型的准确率方面得到了很大的提升,同时也可以保证一定的实时性,该算法的提出引发了行业内的广泛关注。后面的很多文献在此基础上进行改进,例如,文献 SiamRPN[10]在孪生网络的基础之上引入了感兴趣区域提取网络(Region Proposal Network, RPN),更好地解决了孪生网络对于跟踪物体形状调节的问题,同时 RPN 网络也引入了分类支路和回归支路,巧妙的构造成单样本检测任务,使得结果更加的准确。同时,该网络还可以进行端到端的训练,换句话说可以进行数据驱动,作者使用大量的标注数据集进行训练,取得了较好的跟踪性能。随后在此基础上又衍生出很多 SiamRPN++, DaSiamRPN, SiamMask 等算法,更好地实现不同计算机视觉任务之间的相互补充。文献[11]引入了注意力机制,最早应用在机器翻译领域,后来,VIT(Vision Transformer)的出现将 Transformer 应用在图像分类领域。Transformer 也凭借其简单的公式,广阔的上下文关注视野以及其取得的良好效果,优异的可拓展性迅速在视觉任务中走红。2021 年,在目标追踪任务的顶会中,涌现出了一大批基于 Transformer 的追踪器,并且取得了很大的成功。文献[12]引入 Tranformer 框架,使得追踪器能够更好地捕获视频序列中空间和时间信息,建立全局的依赖项,生成有区分的时空特征。动态更新模板特征,使得追踪鲁棒性进一步提高。文献[13]提出了一个新型的辅助跟踪框架,将 Transformer 融入视觉跟踪,同时考虑特征提取和注意力转化,并且修改了经典的 Transformer 结构,更好的适应追踪任务。除此之外,还有一些策略使得追踪器变的更加的轻量化,方便在嵌入式设备上部署,例如通过神经结构搜索(NAS)来作为一种信号,监督追踪器的性能。

综上所述,基于深度学习的目标追踪算法可谓是百花齐放,神经网络结构和注意力机制使得追踪效果逐渐提高,特别是注意力机制的引入使得追踪感受野变大,更容易关注时序上的信息,能够更好地解决之前尺度变换、遮挡等追踪过程中困难的挑战。

1.3 测试基准集及追踪算法评判标准

1.3.1 测试基准集

在目标追踪领域,随着技术的不断进步,追踪算法的场景需求呈现出上升的趋势。各测试基准集也层出不穷,每个数据集中包含了大量的人工标注的追踪示例,方便网络进行学习。此外,数据集的包含类型也各式各样,有通用集合和鸟瞰图等,方便各个追踪场景的实验应用。本文采用了 4 种基准集,分别是 NFS30、GOT10k-test、UAV123、OTB100。

NFS30[14]：NFS(Need for Speed)数据集具有 30 帧和 240 帧两种视频序列。NFS30 是 30 帧的视频序列,是由 240 帧的视频序列中每隔 8 帧采样提取的。一共包含 100 个序列,该数据集的标注形式为[$x1$, $y1$, $x2$, $y2$],其中($x1$, $y1$) 代表人工标注框(ground-truth)的左上角坐标位置,(x2, y2)代表人工标注框(ground-truth)的右下角坐标。

GOT10k-test[15]:GOT10k 又名通用目标追踪基准集,是用于单目标追踪的大型数据集,分为训练集,测试集,验证集三大部分,其中包含大量人工标注数据信息。本文选取 GOT10k 测试集,测试集包含 180 个视频序列,平均长度约为 150 帧。标注形式为[x, y, w, h]。(x, y) 代表人工标注框(ground-truth)的左上角坐标位置, w, h 分别代表边界框(bounding box)的宽和高。该数据集拥有自己的检测网站,需要将实验生成的数据提交至其官网进行测试与评估。

UAV123[16]:该数据集是由无人机拍摄的追踪场景,追踪视角属于鸟瞰图的形式,适合一些高空追踪的任务。该数据集所包含的视频序列有的非常长,经过视频分解之后形成了 123 个追踪示例,每一个视频序列平均有 915 帧。标注形式为[x, y, w, h]。

OTB100[17]:此数据集的源压缩文件中只有 98 个视频序列,其中有 3 个为特殊的视频序列,有的序列会包含两组人工标注信息。长久以来,大家统一的处理方式是将这 3 个特殊序列中的 2 个分别拆分为 2 个独立的视频序列,所以,最后处理后的结果包含 100 个视频序列,因此成为 OTB100。标注的形式也为[x, y, w, h]。

1.3.2　算法评判标准

(1)准确性指标

精确率(Precision rate):用于衡量追踪算法预测边界框(BBox)与数据集标注文件(anno)中人工标注真实框(ground-truth)的中心点之间的距离。给定一个阈值,判断两者距离在阈值之内的帧数占所有视频帧数的百分比。不同阈值有不同的精确率结果,最后多个阈值可以画出精确率曲线(Precision plot),使得最后的效果更加明显。

成功率(Success rate):该指标以重合分数(Overlap Score,OS)为核心,即预测边界框(BBox)与人工标注真实框(ground-truth)之间的交并比(Intersection over Union, IoU)为核心计算依据。假设预测框的范围是 S_1,真实框的预测范围是 S_2,则交并比的计算公式如下:

$$\mathrm{IoU} = \frac{S_1 \cap S_2}{S_1 \cup S_2}$$

给定一个阈值 s,判断重合分数在 s 之上的帧数占所有视频帧数的比例。不同的阈值对应不同的结果,最后多个阈值(一般是 0-1)可以画出成功率曲线(Success plot),表示效果更加明显。

本文中 OTB,UAV,NFS 三个数据集采用了上述两种判断标准,并且绘制了相应的追踪效果曲线图使追踪效果更加明显,对于 GOT10k 数据集,因为测试过程的特殊性,没有可视化效果的曲线,从官网提供的反馈数据来看,其测试主要使用了平均重合分数,并且使用 0.5 和 0.75 的 success rate 进行判断。通过以上的衡量指标,可以很好地判断追踪器的追踪效果的准确性。

(2)速度指标

每秒帧数(Frame Per Second,FPS):跟踪算法每秒内追踪算法处理的视频序列的帧数,通过该指标可以衡量追踪算法的实时性。本文计算该指标的方法为:在追踪过程中记录追踪器处理每一帧的时间;对于每一个视频序列,将所有视频帧的处理时间累加然后除以当前视频序列所包含的帧数;对于每一数据集,将该数据集包含的所有的视频序列的计算结果取平均值。

2　视觉定位(Visual Grounding)任务

近年来,随着自然语言处理和多模态模型的发展,衍生出了许多新兴的研究领域。多模态概念的出现意味着模型可以处理更多交叉任务,其中文本图像多模态的发展更是势头迅猛。许多研究都将 NLP 领域备受关注的 prompt 概念推广至图像领域。Visual Grounding 任务作为结合自然语言与文本信息和图像信息的下游任务,目标是将自然语言描述的物体在图像上标出来。简要来说,该任务的输入是图片(image)和对应的物体描述(sentence/caption/description),输出是描述物体的边界框(box)。该任务看似与目标检测非常类似,区别在于输入包含额外的语言信息,在对物体进行定位时,要先对语言模态的输入进行理解,并且和视觉模态的信息进行融合,最后利用得到的特征表示进行定位预测。Visual Grounding 按照是否要对语言描述中所有提及的物体进行定位,可进一步划分为两个子任务:短语定位(Phrase Localization) 和指代表达理解(Referring Expression Comprehension,REC)。

2.1 短语定位(Phrase Localization)

Phrase Localization 又称为 Phrase Grounding,如前所述,对于给定的句子(sentence),要定位其中提到的全部短语(phrase)对应的物体,在数据集中对于所有短语(phrase)都有对应原的边界框(box)标注(如图 1 所示)。

图 1 短语定位任务效果展示

Fig. 1 Phrase Localization task performance showcase

2.2 指代表达式理解(Referring Expression Comprehension)

Referring Expression Comprehension(也称为 Referring Expression Grounding)。每个语言描述(此处为 expression)仅指向一个物体,每句话即使有上下文物体,也仅对应一个指示物体的边界框(box)标注。图 2 展示了该任务的示例效果。

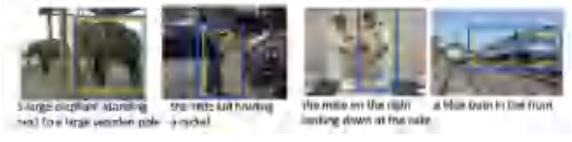


图 2 指代表达理解任务效果展示

Fig. 2 Referring Expression Comprehension task performance showcase

2.3 数据集及评价指标

Phrase Localization 常用的数据集即 Flickr30k Entities 数据集,包含 31 783 张图像(image),每张图会对应 5 个不同的描述(caption),所以总共 158 915 个 caption,以及 244 035 个 phrase-box 标注。每个 phrase 还细分为 people、clothing、body parts、animals、vehicles、instruments、scene、other 共 8 个不同的类别。另外很多 phrase localization 的工作还会在 ReferItGame 数据集(又称 RefCLEF)上进行实验,这个数据集严格来说应该属于 REC 任务。图片来自 ImageCLEF 数据集,包含 130 525 个表达式(expression),涉及 238 个不同的物体种类,有 96 654 个物体、19 894 张图像。其中的数据是通过一种名为 refer it game 的双人游戏进行标注的,Referring Expression Comprehension 常用的数据集有 3 个,RefCOCO[18], RefCOCO+[18], RefCOCOg[19]。

评价指标首先是 prediction box 和 Groud-Truth Box 的交并比(Intersection over Union,IoU)。大于 0.5 记为一次正确定位,以此来计算准确率(Accuracy)。最近的一些工作使用 Recall@k 指标,表示预测概率前 k 的预测值(prediction box)和真实框(Ground-Truth Box)的 IoU 大于 0.5 的定位准确率。此外还有 Pointing game 指标:选择最终预测的注意力掩码(attention mask)中权重最大的像素位置,如果该点落在真实框(ground-truth)区域内,记为一次正确定位。相比 Acc 指标更加宽松。

2.4 Visual Grounding 主流做法

目前 Visual Grounding 可以分为全监督(Fully-supervised)、弱监督(Weakly-supervised)、无监督(Unsupervised)3 种。

全监督(Fully-supervised):顾名思义,就是拥有物体-短语(object-phrase)的边界框(box)标注信息。

弱监督(Weakly-supervised):仅输入图像(image)和对应的句子(sentence),没有句子中的物体-短语标注。

无监督(Unsupervised):既图像-句子(image-sentence)的配对信息也无任务标注。

全监督中,现在的做法可以分为两阶段(two-stage)和单阶段(one-stage)两种做法。two-stage 就是第一个阶段先通过 RPN 或者传统的算法(Edgebox、SelectiveSearch)等提取候选的 proposals 及其特征(features),然后在第二个阶段进行详细的推理,例如常见的做法是把视觉特征和语言特征投射到一个公共的向量空间,计算相似度,选择最相近的 proposal 作为预测结果。单阶段(one-stage)方法则是基于目标检测领域的单阶段(one-stage)模型,例如 YOLO、RetinaNet 等。弱监督由于缺少 phrase 和 box 之间的映射关系(mapping),会额外设计很多损失函数,例如基于重构(reconstruction),引入外部知识(external knowledge),基于图像描述(image-caption)匹配设计损失等。无监督方法先将图像分割成不同的区域,然后根据这些区域的特征进行目标定位。例如,通过图像像素的相似性和连通性,自动生成图像中的目标区域,从而实现目标定位。

3 基于自然语言描述的追踪

受 Visual Grounding 任务的发展和启发,Li 等[20]提出了基于自然语言描述追踪的任务。相较于基于框的追踪任务,基于自然语言的追踪任务可以不受初始框信息的干扰。TNL2K[21]是针对基于自然语言追踪提出的新的数据集,同时也提出了一个简单的追踪框架。UNINEXT[22]是一种通用实例

感知模型，其可以完成但不仅限于完成基于自然语言的追踪。JointNLT[23]算法则是用于基于自然语言追踪的算法，融合了 Visual Tracking 和 Visual Grounding 的框架，使得一个框架可以实现两个任务。接下来将详细介绍以上内容。

3.1　TNL2K

3.1.1　数据集介绍

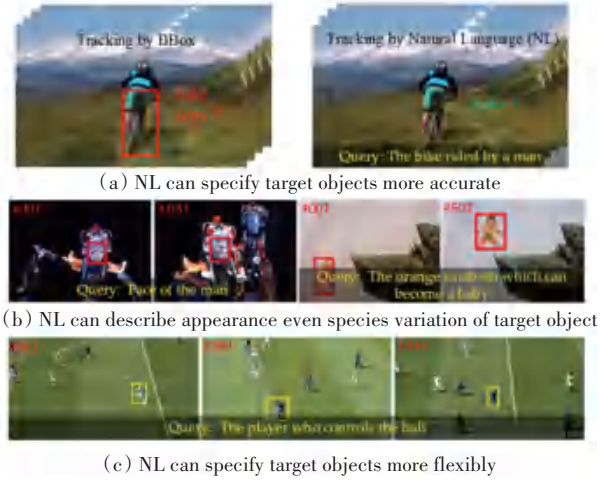
基于自然语言描述的跟踪是一个新兴的研究课题，其目的是根据视频序列中目标对象的语言描述定位目标对象。与传统的基于包围盒(BBox)的跟踪相比，该设置利用高层语义信息指导目标跟踪，解决了 BBox 的模糊性，将局部搜索和全局搜索有机地结合在一起。这些益处可以在实际场景中带来更灵活、鲁棒和准确的跟踪性能。然而，目前针对自然语言初始化跟踪器的研究较少，且相关基准数据集也较为匮乏，这使得语言在跟踪任务中的真正潜力未能充分发挥。为此，研究者提出了一个新的专门用于跟踪的语言方案，其中包含一个大规模的数据集、性能优异且多样化的基线方法。同时为该数据集新增了对抗样本和模态切换两个样本。

通用的基于 BBox 的跟踪有以下问题：

在实际场景中，具有 BBox 的第一帧中的目标对象不便于初始化。换句话说，初始化限制了现有 BBox 初始化跟踪器的广泛应用。

初始化的 BBox 对于目标对象的特征表示可能并非最优，这可能导致歧义。如图 3(a)中，跟踪器可能将目标误识别为跟踪自行车或行人的下半身。

当前基于 BBox 的跟踪器在面对目标对象的突然外观变化(如面部/衣服变化或物种变化)时，跟踪性能往往下降，如图 3(b)情况。



(a) NL can specify target objects more accurate

(b) NL can describe appearance even species variation of target object

(c) NL can specify target objects more flexibly

图 3　TNL2K 数据集展示

Fig. 3　TNL2K dataset showcase

这些问题的存在启发人们开始思考如何以更适用和准确的方式进行跟踪，一些研究人员试图引入自然语言描述进行追踪。与 BBox 相比，自然语言对于人类来说表达起来更加方便和直观。其可以从空间位置到属性、类别、形状以及与其他对象的结构关系等高级语义信息提供更精确的表达。该信息将有利于解决 BBox 的模糊性问题和目标对象的巨大外观变化。同时，语言还可以更灵活地指定目标对象，如图 3(c)中的“控制球的球员”。智能跟踪器应该专注于目标球员，即使球传给不同的人，而不必像视觉跟踪的标准设置那样重新初始化目标人。然而，这一研究课题受到的关注远远低于标准的目标跟踪。

在此基础之上，Wang[21]等提出了一个新的大规模基准数据集 TNL2K。该数据集包含 2 000 个视频序列，针对每个视频，研究人员密集地为每一帧标注目标对象的位置信息，并为整个视频标注一句英文描述。具体来说，自然语言包含描述的类别，形状，属性及空间位置，将提供丰富的细粒度的外观信息和高层次的语义信息跟踪。训练集包含了 1 300 个视频，剩下的 700 个视频作为测试集。遵循流行的跟踪基准集，TNL2K 定义了每个视频序列的多个属性，用于在每个挑战性因素下的评估。具有以下 17 个属性：CM(摄像机运动)、ROT(目标旋转)、DEF(DEFormation)、FOC(完全闭塞)、IV(照明变化)、OV(视图外)、POC(部分闭塞)、VC(视点变化)、SV(比例变化)、BC(背景杂波)、MB(运动模糊)、ARC(宽高比变化)、LR(低分辨率)、FM(快速运动)、AS(对抗样本)、TC(热交叉)、MS(模态切换)。值得注意的是，视频还反映了其他数据集不包含的两个额外属性，即：对抗样本和模态在 RGB 和热数据之间切换。

3.1.2　AdaSwitcher

本文提出了一种简单但性能优异的基线方法(称为 AdaSwitcher)为未来的作品进行比较，该方法可以自适应的切换局部跟踪算法和全局 Grounding 算法。

Visual Grounding 模块：在自然语言跟踪任务中，该方法首先定位目标对象。如图 4 所示，Visual Grounding 模块以视频帧和自然语言描述作为输入，使用深度卷积网络来提取视频帧特征，对于自然语言，使用预训练好的 BERT[24]来提取特征表示，这两种特征被馈送到两个全连接层进行进一步微调。将全局帧、重复语言特征和空间坐标的视觉特征图连

接在一起,并输入到核大小为 1×1 的卷积层进行信息融合。然后将输出的特征图发送到 Grounding 模块,Grounding 模块将输出目标对象的预测位置。

Visual Tracking 模块:前面提到的 Visual Grounding 可在初始阶段检测目标对象,然而,仅依靠 Grounding 对于高性能跟踪是不够的,因为其容易受到背景杂波的影响。在这项工作中,采用的架构是首先经过 Grounding 模块得到追踪的第一帧物体位置,然后使用 Visual Tracking 算法 SiamRPN++[25] 进行视觉追踪。

AdaSwitcher 模块:在给定视觉背景和视觉跟踪模块的情况下,可以分别从全局和局部视图捕获目标对象。仍然存在的一个棘手问题是,当本文使用视觉基础进行全局搜索(或视觉跟踪进行局部搜索)时,一种方法是基于跟踪器的置信度进行这种切换,然而,置信度分数并不总是可靠的,特别是在具有挑战性的场景中。受异常检测(也称为离群值检测)的启发,其目标是识别罕见的项目、事件或观察结果,这些项目、事件或观察结果与大多数数据有显著差异,从而引起怀疑。在这项工作中,本文把视觉跟踪的失败作为一种异常检测,并提出了一种新的 AdaSwitcher 模块来检测这样的失败。一旦检测到异常(来自 AdaSwitcher 的预测大于预定义阈值),就可以将候选搜索区域从视觉跟踪切换到视觉基础,以实现更鲁棒和更准确的跟踪。

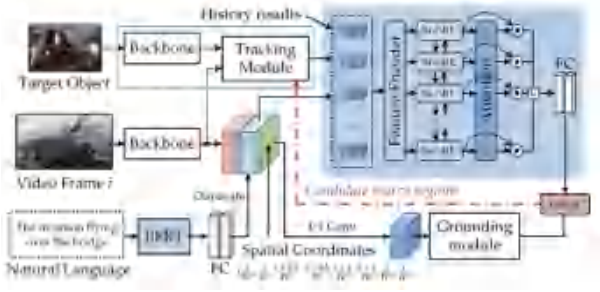


图 4 自适应框架概览图

Fig. 4 Overview diagram of adaptive framework

3.2 UNINEXT

所有的实例感知任务都旨在找到由某些查询指定的某些对象,例如类别名称、语言表达和目标注释,但这个完整的领域已被拆分为多个独立的子任务。在这项工作中,本文提出了一个通用的实例感知模型,称为 UNINEXT。如图 5 所示,UNINEXT 将不同的实例感知任务重新表述为统一的对象发现和检索范式,并且可以通过简单地改变输入提示来灵活地感知不同类型的对象。这种统一的范式带来以下益处:

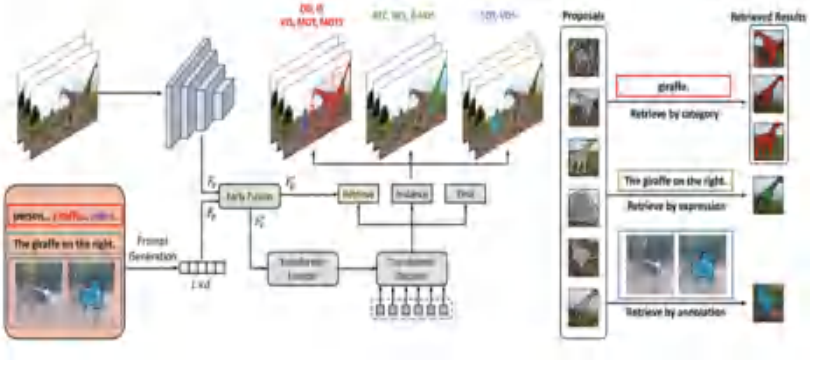


图 5 UNINEXT 框架

Fig. 5 UNINEXT framework

(1)可以利用来自不同任务和标签词汇表的大量数据来联合训练一般实例级特征表示,这对于缺乏训练数据的任务尤为有益。

(2)统一模型参数效率高,且在同时处理多个任务时可以避免冗余计算。UNINEXT 在 10 类实例级任务的 20 个具有挑战性的基准测试中表现出卓越的性能,包括经典的图像级任务(对象检测和实例分割),视觉和语言任务(指表达理解和分割),以及 6 类视频级对象跟踪任务。

不同的实例感知任务可通过输入的 prompt 的格式进行区分,主要分为类别标签(category)、描述语句(language expressions)、区域图像信息(reference annotations as prompts),UNINEXT 首先在提示符的引导下生成 N 个对象建议,然后根据“实例-提示符”的匹配得分从建议中检索最终实例。具体方法如下:

提示生成词(Prompt Generation):首先,采用提示生成模块,将原本多种多样的提示输入转化为统一的形式。对于输入的 prompt 信息,采用一个预训练好的语言编码器,对于 category 和 language expressions 提示,直接拼接后进行特征提取即可。对于 annotation-guided 任务,提取细粒度的视觉特

征，并充分利用目标注释，引入注释框掩码，提取多尺度特征后与另外两个任务的提示特征进行维度对齐。这样就实现了不同的 prompt 输入对应同一种特征形式。

图像-提示词特征融合（Image-Prompt Feature Fusion）：作为文本图像特征融合模块。该模块中图片特征与文本特征平行生成，具体流程是当前图像经过视觉编码器得到视觉特征。为了增强图像上下文的原始提示嵌入，并使原始视觉特征具有文本 prompt 信息，引入了早期融合模块。早期融合模块采用了交叉注意模块，从不同的输入检索信息，然后将检索到的表示融合到原始特征之中。

目标发现与检索（Object Discovery and Retrieval）：在得到具有区别性的视觉特征和提示特征后，下一个关键步骤是将输入特征转换为各种感知任务的实例。UNINEXT 采用了 Deformable DETR 中的编码器-解码器架构，其中编码器采用 Transformer 编码器，加入了一个辅助预测头来生成 N 个初始参考点作为解码器的输入。解码器采用增强的多尺度特征、N 个参考点和 N 个对象查询作为输入，利用可变形注意力机制来检索包含短语信息的视觉特征，并生成最终的实例预测结果。另外，文章探索了两种查询生成策略：静态查询和动态查询，实验结果表明，静态查询通常表现更优。最后，文章介绍了实例-短语匹配机制，通过加权矩阵来检索真实匹配的对象，进一步提高了模型的准确性和鲁棒性。

UNINEXT 首次将 10 个实例感知任务与提示引导的对象发现和检索范式进行统一。大量的实验表明，UNINEXT 仅通过单一框架，即可在 20 多个具有挑战的多任务数据集上取得超前的效果。

3.3　基于自然语言规范的联合视觉定位和跟踪

自然语言描述跟踪的目的是在基于自然语言描述的序列中定位所提及的目标。现有算法通过视觉定位（Visual Grounding）和视觉跟踪（Visual Tracking）两个步骤来解决这一问题，并相应地部署分离的 Visual Grounding 模型和 Visual Tracking 模型来分别实现这两个步骤。这种分离的框架忽略了 Visual Grounding 和 Visual Tracking 之间的联系，即自然语言描述为两个步骤的目标本地化提供了全局语义线索。此外，分离的框架很难进行端到端的训练。为了处理这些问题，这篇文章提出了一个联合的定位和跟踪框架（JointNLT），重新把定位和跟踪作为一个统一的任务：基于给定的视觉语言引用来定位目标。具体而言，文章提出了一个多源关系建模模块，以有效地建立视觉语言参考和测试图像之间的关系。此外，文章设计了一个时态建模模块，以全局语义信息为指导，为本文的模型提供时态线索，有效地提高了对目标外观变化的适应性。

图 6 展示了 JointNLT 的具体框架图，其主要由语言和视觉编码器、多源关系建模模块、目标解码器、语义引导的时间建模模块和定位头组成。具体流程是：给定输入 prompt 和测试图像，语言和视觉编码器首先将其嵌入到特定的特征空间中，产生输入单词或图像块的标记嵌入。然后，采用两个线性投影层的语言和视觉令牌嵌入投影到具有相同维度的潜在空间中。投影嵌入被送到多源关系建模模块，其对多源参考与测试图像之间的关系进行建模，以增强测试图像嵌入中的目标信息。在测试图像的增强嵌入后，将其送入目标解码器和定位头部，完成特征解码和边界框定位。对于 Visual Grounding 任务，因其输入比 Visual Tracking 任务少一个模板信息，所以为了统一格式，文章使用了零填充标记作为占位符，以填充与多源关系建模模块输入中对应模板的缺失嵌入。在视觉跟踪过程中，语义引导的时间建模模块将为目标解码器生成时间线索，使模型能够利用历史目标状态。

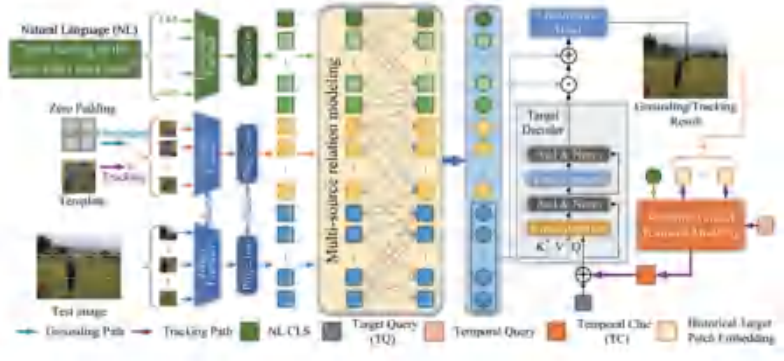


图 6　JointNLT 结构概览

Fig. 6　Overview diagram of the JointNLT structure

追踪过程可以被总结为:

(1)对于第一帧,模型把自然语言,零填充标记和测试图像作为输入,执行 Visual Grouding 任务,追踪所用到的模板信息就可以通过 Grounding 结果得到。

(2)在后续的每一帧中,模型将自然语言、模板信息和测试图像作为输入,通过相同框架执行基于自然语言的视觉追踪。

视觉和语言特征提取:对于文本编码器,选择经典的语言转换器模型 BERT [24]作为本文框架中的语言编码器。给定自然语言查询 Lq,首先将句子标记化,并分别在标记化的语言查询的开始和结束处附加 CLS 标记和 SEP 标记,从而产生标记序列。随后将标记序列嵌入到语言编码器中,最终嵌入到需要的维度。图像编码器,采用 Swin Transformer[26]来提取特征,该模型具备优异的图像特征学习能力。仅保留 Swin Transformer 的前 3 个阶段(因为最后一个阶段的输出特征分辨率过低)。对输出特征进行平坦化,以分别获得其令牌嵌入。

多源关系建模:采用 Transformer 编码器来进行关系建模,通过自注意力机制捕捉全局依赖性,并使用语言和视觉投影层来统一不同模态的嵌入维度。在视觉定位任务中,由于模板嵌入不可用,文章使用零填充张量作为占位符,并使用掩码值将其屏蔽,以避免干扰其他有用信息的计算。然后,将参考嵌入和测试图像嵌入进行拼接,并将其输入 Transformer 编码器进行关系建模,以实现多源关系建模。最后,文章介绍了为保留位置信息而添加的可学习位置嵌入。

目标解码器和定位头:为了进一步增强测试图像的目标信息,模型采用了目标解码器和目标查询来学习区分性目标嵌入。对于视觉定位任务,目标查询是预先学习的离线嵌入,包含潜在的目标信息;对于视觉跟踪任务,模型将语义引导的时间建模模块输出的时间线索与离线学习的嵌入相结合,得到一个包含最近目标外观先验的目标查询。为了实现视觉定位和跟踪的统一,模型采用了共享的定位头进行边界框预测。定位头首先计算目标解码器和测试图像嵌入之间的相似度,然后通过残差连接来增强目标相关区域,并最终预测目标边界框。

语义引导的时间建模:由于视觉定位和跟踪任务之间存在一些差异,模型提出了 SGTM 模块,利用历史目标外观信息进行跟踪。SGTM 模块主要由 Transformer 编码器与解码器组成,通过自然语言中的语义引导学习近期目标外观的时间线索。在每帧的跟踪过程中,模型对预测的目标边界框进行 RoI 池化,提取目标区域特征,并展平,得到历史目标补丁嵌入。另外,模型还考虑了自然语言多源关系建模模块输出的全局语义表示,并将其与历史目标补丁嵌入进行拼接,输入编码器进行特征增强。随后,模型利用解码器计算增强后的历史目标补丁嵌入与可学习的时间查询的交叉注意力,以学习未来跟踪的鲁棒历史目标表示。

模型在 NVIDIA 3090 GPU 上测试,跟踪速度约为 39 FPS。模型使用 OTB99、LaSOT 和 TNL2K 数据集进行评估,采用成功率和精度来衡量跟踪性能,当中心位置误差阈值设置为 20 像素时的成功率曲线下面积(AUC)和精度得分。此外,模型在 RefCOCOg 的 Google-split val 验证集上进行了视觉定位评估,并使用标准协议报告 Top-1 准确率,其中与真实标注的 IoU 大于 0.5 的预测框被视为正确预测。

综上,模型提出联合视觉定位与跟踪框架,通过统一多源引用和测试图像之间的关系建模来连接两个任务,包括跨模态(视觉和语言)关系和跨时间(历史目标补丁和当前搜索帧)关系。此外,模型提出了一个语义引导的时间建模模块,利用全局语义信息来建模历史目标状态,有效提高了跟踪性能。在 3 个自然语言跟踪数据集和一个视觉定位数据集上,本文的方法实现了优于现有算法的良好性能。

4 结束语

本文全面分析了基于自然语言描述的目标追踪方法,特别是其在视觉定位和目标追踪中的应用。文章首先综述了传统追踪技术的局限,然后深入探讨了自然语言描述带来的新机遇和挑战,最后详细介绍了几种创新的追踪算法,并对这几种算法在不同数据集上的表现进行了比较。主要结论如下:

(1)自然语言描述可以显著提高目标追踪的准确性和鲁棒性,尤其是在面对复杂场景和动态变化时。

(2)通过结合自然语言处理和视觉跟踪技术,研究者们开发出的新模型比传统模型表现更好,能够处理更多复杂的追踪任务。

(3)尽管已有显著进展,但当前方法在处理极端遮挡、高速移动目标等情况时仍面临挑战。

尽管本文的研究取得了一定的成果,但基于自然语言的目标追踪技术仍处于发展阶段。未来研究

需要进一步优化算法的实时性和准确性，特别是在多目标追踪和跨场景适应性方面。此外，如何有效整合不同类型的语言描述和视觉信息，以及如何减少对大量标注数据的依赖，也是未来研究的重要方向。

参考文献

[1] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2544-2550.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]// Proceedings of the European Conference on Computer Vision. Cham: Springer, 2012: 702-715.

[3] 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, 2014, 37(3): 583-596.

[4] DANELLJAN M, SHAHBAZ K F, FELSBERG M, et al. Adaptive color attributes for real-time visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2014: 1090-1097.

[5] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: Complementary learners for real-time tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 1401-1409.

[6] DANELLJAN M, HÄGER G, KHAN F, et al. Accurate scale estimation for robust visual tracking[C]//Proceedings of the British Machine Vision Conference. BMVA, 2014: 10.

[7] 王坤峰，苟超，段艳杰，等. 生成式对抗网络 GAN 的研究进展与展望[J]. 自动化学报，2017，43(3)：321-332. DOI：10.16383/j.aas.2017.y000003.

[8] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4293-4302.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]// Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 850-865.

[10] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8971-8980.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. NeurIPS, 2017: 5998-6008.

[12] YAN B, PENG H, FU J, et al. Learnings patio-temporal transformer for visual tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10448-10457.

[13] WANG N, ZHOU W, WANG J, et al. Transformer meets tracker: Exploiting temporal context for robust visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1571-1580.

[14] KIANI G H, FAGG A, HUANG C, et al. Need for speed: A benchmark for higher frame rate object tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2017: 1125-1134.

[15] 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, 2019, 43(5): 1562-1577.

[16] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 445-461.

[17] WU Y, LIM J, YANG M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2411-2418.

[18] YU L, POIRSON P, YANG S, et al. Modeling context in referring expressions[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 69-85.

[19] MAO J, HUANG J, TOSHEV A, et al. Generationand comprehension of unambiguous object descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 11-20.

[20] LI Z, TAO R, GAVVES E, et al. Tracking by natural language specification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6495-6503.

[21] WANG X, SHU X, ZHANG Z, et al. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 13763-13773.

[22] YAN B, JIANG Y, WU J, et al. Universal instance perception as object discovery and retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15325-15336.

[23] ZHOU L, ZHOU Z, MAO K, et al. Joint visual grounding and tracking with natural language specification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 23151-23160.

[24] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv, 1810.04805, 2018.

[25] LI B, WU W, WANG Q, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4282-4291.

[26] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012-10022.