Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2022, 11(3), 278-287
Published Online August 2022 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2022.113029

Hans汉斯

孪生神经网络在目标跟踪中的算法研究

李铭涵

北方工业大学计算机系，北京

收稿日期：2022年6月23日；录用日期：2022年8月5日；发布日期：2022年8月15日

摘　要

目标跟踪算法是对给定目标的位置进行预估与定位从而实现持续跟踪。随着硬件技术以及神经网络的发展，目标跟踪在精度与速度上远超传统算法，基于孪生神经网络的跟踪算法是当前众多学者主要研究的方向之一。本文主要对孪生网络结构以及相关算法进行介绍。首先介绍孪生结构原理，其次根据改进方向对存在的算法进行阐述，随后介绍经典数据集，最后对现有算法发展进行总结与展望。

关键词

目标跟踪，孪生神经网络

Algorithm Research of Siamese Neural Network in Target Tracking

Minghan Li

Computer Department, North China University of Technology, Beijing

Received: Jun. 23rd, 2022; accepted: Aug. 5th, 2022; published: Aug. 15th, 2022

Abstract

Target tracking algorithm is to estimate and locate the position of a given target to achieve continuous tracking. With the development of hardware technology and neural network, target tracking is far more accurate and faster than traditional algorithms. Tracking algorithm based on siamese neural network is one of the main research directions of many scholars at present. This paper mainly introduces the siamese network structure and related algorithms. Firstly, the principle of siamese structure is introduced, then the existing algorithms are described according to the improvement direction, then the classical data sets are introduced, and finally, the development of existing algorithms is summarized and prospected.

文章引用: 李铭涵. 孪生神经网络在目标跟踪中的算法研究[J]. 人工智能与机器人研究, 2022, 11(3): 278-287.
DOI: 10.12677/airr.2022.113029

Keywords

Target Tracking, Siamese Neural Network

1. 引言

目标跟踪是当前计算机视觉领域的热点问题之一[1] [2] [3] [4]，众多领域中都会涉及到目标跟踪，例如视频安防监控系统[5] [6]、自动驾驶[7]、医疗诊断[8]、军事安全[9]等。按照跟踪对象分为单目标跟踪与多目标跟踪，本文主要针对单目标跟踪进行研究。单目标跟踪的主要任务是在给定视频初始目标状态后，目标跟踪算法能够对视频后续帧中该目标的位置进行预估与定位。由于视频中目标会受不同因素的影响而发生变化，例如目标形状变化、环境遮挡目标本身、光照影响等，因此设计一个能够稳定高效准确的跟踪器是十分具有挑战性的任务。

传统的目标跟踪算法较为粗糙，主要基于光流法、粒子滤波以及均值漂移实现的，这类方法很容易受到目标尺度变化、背景复杂等因素的干扰，并且由于计算过程较为复杂，导致执行速度较慢无法有本质上的提升[10]。随着通信领域的发展，将相关滤波引入到了目标跟踪中，基于相关滤波的目标跟踪算法层出不穷，相比于传统的目标跟踪算法，跟踪速度大幅提升并且出现了很多跟踪性能优良的跟踪器，一定程度上解决了跟踪速度与性能问题[11]。

随着深度学习的发展，卷积神经网络在提取特征方面有较好的鲁棒性，逐渐替代了传统手工设计的特征[12] [13]。基于深度学习的跟踪方法主要分为两类：一类是与相关滤波相结合的算法，另一类是神经网络通过端对端的训练完成目标的特征提取与定位[3]。其中孪生神经网络既可较好地提取目标特征，在跟踪速度上也较为优秀。

2. 孪生网络结构

孪生结构最早在 1993 年由 Bromley J [14]等人提出，用于验证签名的一致性。2010 年 Hinton [15]等人用孪生网络来验证人脸。2015 年 Sergey Zagoruyko [16]等人对孪生结构进行了改进，极大地提高了网络判别的准确度[17]。随着神经网络的发展，孪生网络也慢慢出现在计算机视觉领域的目标跟上，孪生结构的原理如图 1。

其中，神经网络 1 和神经网络 2 是权值共享的神经网络，这两个结构完全相同的网络将不同分输入分别映射到新的空间中，形成新的表示，通过损失的计算评价两个输入的相似度。

3. 历史发展与现状

SINT [18]算法最早将孪生结构用于目标跟踪，通过学习匹配函数返回后续帧中与目标最相似的 patch 从而实现较精准的定位。同年，Bertinetto L [19]等人在跟踪部分使用相似度量的思想提出 SiamFC 算法，采用孪生神经网络结构如图 2，将第一帧的目标作为模板图像，后续帧作为搜索图像，对模板图像和搜索图像进行放缩填充后，输入到相同的骨干网络中，生成各自的特征图，以模板图像的特征图作为卷积核在搜索图像的特征图上滑动互卷积，进行相似性判断生成对应的置信图，得分最高的子窗口即为预测目标所在位置。

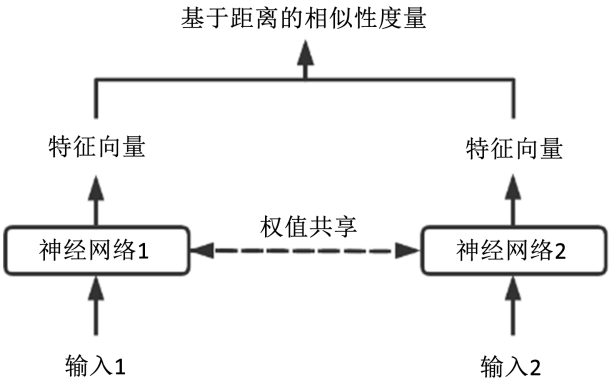

Figure 1. Structure of siamese network
图 1. 孪生网络结构

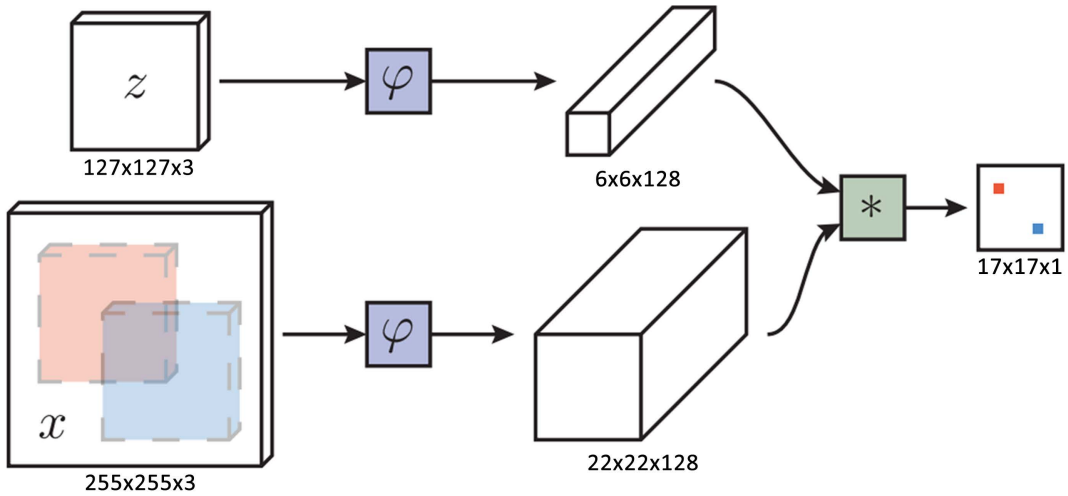

Figure 2. Structure diagram of SiamFC
图 2. SiamFC 结构图

SiamFC 结构较为简单，在当时跟踪实时性较好，但在面对目标形变较大、背景较为复杂、相似外观等情况下跟踪效果较差。后续很多孪生神经网络的目标跟踪算法都是在 SiamFC 的基础上进行改进的，主要从主干网络结构、引入回归、结合注意力机制等方面进行改进。

3.1. 主干网络结构

SiamFC 的提出将孪生网络结构作为目标跟踪领域中一个新分支，借助神经网络的发展，后续的算法在其结构上不断地优化，从而实现更高效准确的跟踪器。利用神经网络进行特征提取时通常使用 AlexNet 作为骨干网络，网络层数较浅且较为轻量是其受欢迎的原因。但网络层数较浅就会导致特征提取的准确性较低，蒲磊等[20]设计了基于高层语义嵌入的孪生网络跟踪算法，在依旧使用 AlexNet 网络的前提下，在模板分支中构造语义嵌入模块，训练过程中可以将深层语义转换到较浅层的特征，进一步优化整个网络学习特征的能力，增强特征的表达，通过实验发现，在计算量不变的情况下，增加语义嵌入模块能在精度和成功率方面有更好的表现。Li [21]等人使用修改后的 VGG16 网络作为骨干网络，更深的网络结构让其在大数据集上训练效果更好。陈[22]等人在 Li 的基础上以 VGG16 作为骨干网络设计了抗遮挡的孪生网络跟踪算法，通过对比网络输出的置信图和连通域进行目标遮挡情况的判断，从而提高算法的抗遮挡性。

张志鹏[23]等在文中将骨干网络更换为更深层的网络如 ResNet、Inception，发现特征提取的效果没有上升反而下降了。主要原因是网络加深导致感受野增加，使用 padding 会导致网络在学习过程中出现位置偏差，从而影响了准确度。为了解决上述问题，文中提出了 CIR 残差块用来减弱 padding 带来的偏差影响。邵[24]等人在 SiamFC 的基础上通过残差链接融合模板分支网络提取不同层的特征，通过特征融合增大模型的表征能力，结合注意力机制模块充分挖掘目标语义信息。通过实验发现效果较前文有更好的表现。

3.2. 引入回归

面对跟踪目标的尺度发生变化时，SiamFC 中使用搜索域的方法增加多尺度计算，但这样做增加计算量的同时，无法适应尺度的变化。为了解决这个问题，Li [25]等人借鉴 RPN (区域建议网络)的思想，提出了 SiamRPN 算法，结构如图 3，将跟踪问题转化为分类和回归两个分支，一个用于区分前景和背景，一个用于边框回归。后续引入回归的跟踪算法也层出不穷，其中又分为基于锚框的跟踪算法和无锚框的跟踪算法。

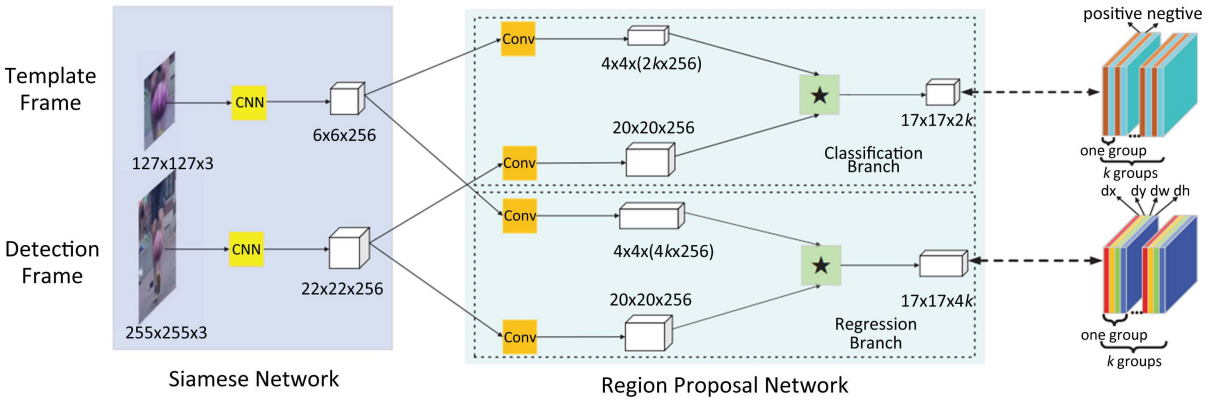

Figure 3. Structure diagram of SiamRPN
图 3. SiamRPN 结构图

3.2.1. 有锚框的跟踪算法

基于区域建议网络的思想是在特征提取生成的特征图上，基于锚点生成指定尺寸的锚框，根据锚框与真实目标框的 IOU 值来确定目标是否在锚框中。在跟踪算法的回归过程中根据最佳锚框与真是目标框的距离计算出来的损失调整网络参数。

SiamRPN 算法使用的骨干网络为 AlexNet 网络，如前文提到，使用加深的主干网络会使性能大幅衰减，针对这种情况 Li [26]等对其进行改进设计了 SiamRPN++，在训练过程中加入位置均衡的采样策略，以此来进一步环节网络在训练过程中出现的位置偏离问题，同时利用了多层融合，将浅层特征的细节信息和深层网络的语义信息相结合从而进一步提升跟踪性能。

区域建议网络会产生大量的锚框，从而影响跟踪的精度和速度。尚[27]等人针对冗余锚框的问题提出了导向锚框网络，根据孪生网络中提取高层语义特征中的目标位置和形状分布来学习锚框的形状，锚框的位置预测通过使用 1*1 的卷积核与输入特征图进行卷积得到置信图，再通过激活函数获取与输入特征图相同大小的概率图。将概率与设置的阈值确定可能存在的位置，这样可以过滤掉 90%的区域，大幅提高网络效率。

引入区域建议网络能够规避图像金字塔对跟踪产生的影响，但是大多数算法中生成的锚框为平行坐标轴的矩形框，当跟踪目标旋转时会，锚框的方向不会发生变化，锚框中过多的背景信息会导致跟踪效

果的下降。为了解决这个问题，姜[28]等人，提出了旋转区域提议的孪生神经网络跟踪算法，引入了AO-RPN 网络并结合残差网络结构，在多个特征提取层使用 AO-RPN 网络进行提取，特征融合后通过分类预测实现端到端的训练。

3.2.2. 无锚框的跟踪算法

基于锚框的跟踪算法能够获得目标边界，在训练过程中对分类预测的中心进行训练，如果分类预测的中心出现偏差，就会导致回归框出现偏差；并且基于锚框的跟踪算法会引入大量参数与计算，从而影响跟踪结果，无锚框（anchor-free）的算法也因此被提出。借鉴目标检测中的 FCOS [29]算法，在特征图对应点上直接进行回归操作，预测该点与目标上下左右侧的距离进而预测目标位置。SiamCar [30]中引入了中心度得分图，通过比较分类分支的得分图和中心度得分图，选择出最佳的目标中心点，根据中心点和真实边框的距离得到预测框。SiamBAN [31]引入空洞卷积原理，提升跟踪器的性能。针对之前孪生网络跟踪器的不合理性，SiamFC++ [32]增加了回归分类分支以及质量评估分支。无锚点使得算法泛化能力更强，无先验知识使得整体结构更加简洁，从而使得运行速度有一定的提升。

基于锚框与无锚框本质的区别在于基于锚框的跟踪是具有先验知识的，能够较好的感受到物体的尺度变化从而进一步回归，而无锚框本身不具有尺度的感受能力，于是 Ocean [33]算法中提出了 feature combination 模型如图 4，在特征编码时不同方向使用不同的 dilated strides 得到不同尺度的感受能力。为了更好的学习特征，将边缘点感受部分的预测框的偏差通过 Object-aware 模块调整，从而增强预测的准确性。

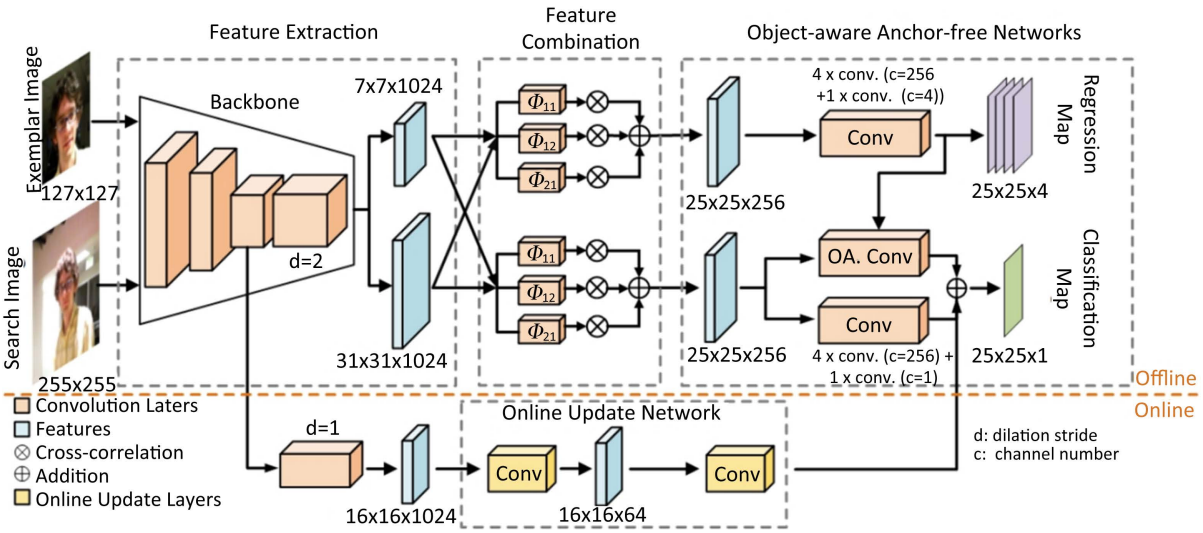

Figure 4. Structure model of Ocean
图 4. Ocean 算法模型

3.3. 结合注意力机制

注意力机制最早应用于自然语言处理的任务中[34]，目的是解决长序列带来的遗忘现象。Volodymyr [35]等人将其应用在视觉领域，后 Ashish Vaswani [36]等人提出了 Transformer 结构后，注意力机制在自然语言处理、计算机视觉等相关领域的网络设计上被广泛应用。在计算机视觉中，注意力机制的核心思想是基于原有的数据找到其之间的关联性，突出某些重要特征。RASNet [37]将注意力机制模型引入到孪生神经网络结构的跟踪问题上，文中提出通用注意力机制、残差注意力机制和通道注意力机制，缓解深度网络的过拟合问题，提升网络的判断能力和适应能力。Yu Y 等人[38]提出了可变形孪生注意力网络 SiamAttn 来增强孪生网络跟踪器的特征学习能力。包括可变形的自注意力机制和互注意力机制两部分。

自注意力机制通过空间注意力和通道注意力可学习到强大的上下文信息，并选择性地增强通道特征之间的相互依赖；而互注意力机制则可以有效地聚合与沟通模板和搜索区域之间丰富的信息。这种注意力机制为跟踪器提供了一种自适应模板特征隐式更新方法。F. Du 等人[39]提出使用像素方向相关引导的空间注意模块和通道方向相关引导的通道注意模块，利用目标模板和感兴趣区域之间的关系突出焦点区域，增强感兴趣区域的特征进行角点检测，从而进一步提高边界框估计的准确性。

3.4. 新趋势

3.4.1. 基于 Transformer 的跟踪

Transformer [36]结构是第一个只基于注意力机制完成编码器－解码器功能的模型，在机器翻译的任务上相对于 RNN 与 CNN 来说有更好的效果。随后 Transform 逐渐应用到图像分类、图像分割等领域[40]，基于 Transformer 的孪生结构跟踪器也逐渐被提出。Transformer 的结构如图 5，左侧为编码器，右侧为解码器，Transformer 提出了多 Muti-Headed Attention 为了模仿卷积网络可以进行多通道识别多模式的效果。

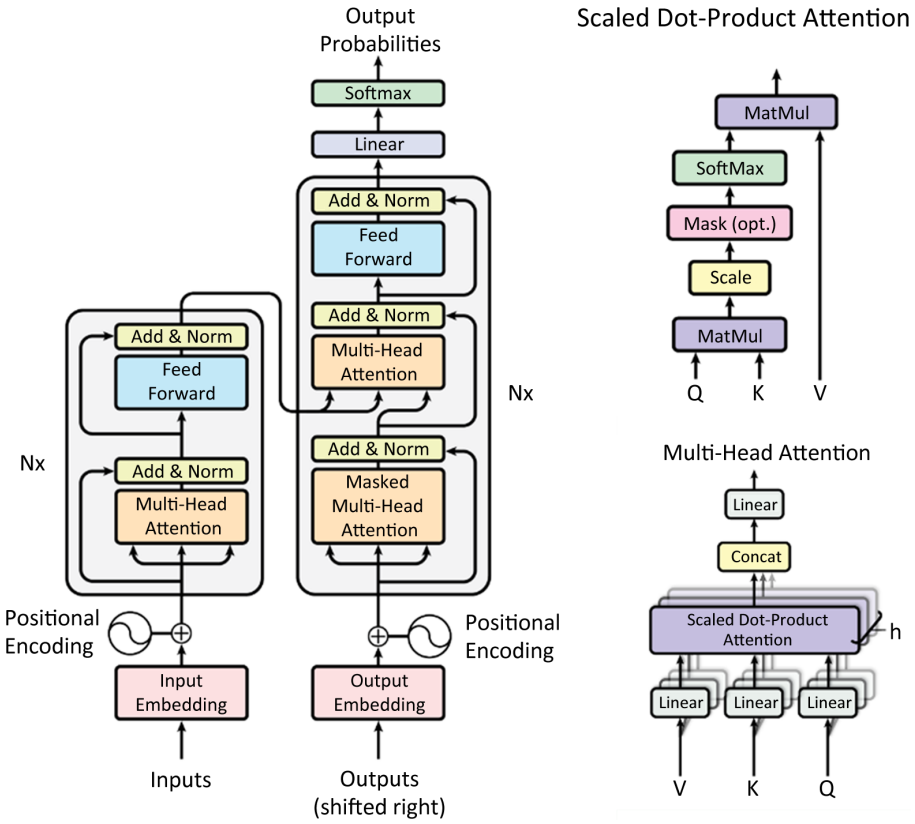

Figure 5. Structure diagram of Transformer
图 5. Transformer 结构图

在目标跟踪任务中，连续帧之间存储着较为丰富的时间信息，大多数算法忽略了这一特征，Wang N [41]等将 Transformer 结构引入到跟踪框架中进行辅助跟踪，不修改模板匹配方法，将编码器和解码器分离为两个并行分支，连接视频流中的隔离帧，在帧间传递时间上下文信息。TransT [42]中借助 Transformer 的思想引入上下文增强模块 ECA 与交叉特征增强模块 CFA 来完成特征融合的功能。Zhao M [43]等改进 Ocean 算法，使用 Transformer 结构完成孪生网络中的互相关操作，以此来获得全局和丰富的上下文相关性。

基于 Transformer 的许多算法在性能方面不如 CNN，直到 Swin-Transformer [44]的出现。Lin L [45]

等将特征提取网络直接换成 Swin-Transformer，提出了完全基于注意力机制的跟踪算法 SwinTrack，在孪生结构的基础上，直接使用 Transformer 结构进行特征提取与特征融合，在众多具有挑战的数据集上都处于领先位置。

3.4.2. 轻量型跟踪

为了追求跟踪精度，跟踪的模型越来越复杂，无法平衡的速度与精度，于是轻量型的跟踪算法也成为有价值的研究方向。LightTrack [46]使用神经架构搜索来设计更轻量高效的目标跟踪器，相比于之前性能较好的 Ocean 算法，参数量减少 90%以上，速度也快了 12 倍。E.T.Track [47]对 Transformer 架构进行轻量化操作，提出了 Exemplar Transformer 来代替卷积，相对于其他基于 Transformer 的跟踪算法相比速度快了 8 倍，并且在 CPU 上运行速度可达 47FPS。FEAR [48]中提出了两个轻量化模型，模型较 LightTrack 更小，跟踪准确率更高。

4. 经典数据集

在单目标跟踪方向较为经典的数据集(表 1)有 OTB 系列、VOT 系列以及 GOT-10K，OTB50 和 OTB100 提供 51 和 98 个视频序列，每个帧使用 11 个不同的属性和垂直边界框进行注释。VOT 针对多达 60 个视频序列提出了几个挑战。它引入了旋转边界框以及对对象跟踪注释的广泛研究。GOT-10K 包含 10,000 个视频序列以及手工标注的 150 万个边界框，使用 WordNet 英文词汇数据库作为骨架搭建。

基于神经网络的跟踪算法出现后，具有挑战性的跟踪评测数据集也随之增多。比较具有代表性的数据集有 UAV123、LaSOT、TrackingNet 等。

LaSOT 数据集包含 1400 个序列，是较大的密集注释跟踪基准，每个序列都包含来自野外的各种挑战。早期的数据集通常属于小型数据集，对于训练大型网络较为不利，并且短时基准的评估可能无法反映跟踪器在实际应用中的实际性能，LaSOT 的大规模数据集能够提供更为可靠的评估结果。

TrackingNet 数据集包含了 30,643 个视频片段，通过 YouTube 视频采样有更真实的场景，该数据集囊括了自然场景下的各种情形，包含了各种帧率，分辨率，上下文场景以及目标类别。

UAV 数据集引入了无人机拍摄视频，通过低空拍摄方式构建了包含 123 个高清视频，每个视频序列有完整的标签。在视频序列中目标的矩形框的长宽比随着无人机的运动变化较为明显，所以对于跟踪器的尺度自适应要求较高。测试序列中包括背景杂波、快速运动、完全遮挡等属性。

Table 1. Classic dataset
表 1. 经典数据集

数据集名称\指标	视频数量	最小帧数	最大帧数	总帧数	总时长	帧率	目标类别数
LaSOT	1400	100	11,397	3.52 M	32.5 hours	30 fps	70
TrackingNet	30,643	-	-	14 M	140 hours	30 fps	21
UAV123	123	109	3085	113 K	62.5 min	30 fps	9
UAV20L	20	1717	5527	59 K	32.6 min	30 fps	5
OTB2013	51	71	3872	29 K	16.4 min	30 fps	10
OTB2015	100	71	3872	59 K	32.8 min	30 fps	16
GOT-10k	10,000	-	-	1.5 M	-	10 fps	563
VOT-2014	25	164	1210	10 K	5.7 min	30 fps	11
VOT-2017	60	41	1500	21 K	11.9 min	30 fps	24

5. 研究展望

近年来，目标跟踪领域发展迅速，相关算法也层出不穷，但是目标跟踪依旧面临着诸多挑战。解决环境对目标跟踪的影响，更好地利用目标本身的特征以及平衡准确性、鲁棒性和实时性是具有挑战性的研究方向。

1) 环境变化影响跟踪

目标在跟踪过程中自身会发生形变、快速运动的变化，并且目标所在环境也会影响目标本身，例如光照变化、目标遮挡、背景干扰等，现有的算法通常只对其中部分影响做出调整，并且由于环境干扰，目标丢失后较难跟踪，因此防止环境变化影响目标是值得深入研究的问题

2) 更好地利用有效信息

目标跟踪是基于目标的特征的，目标周围以及背景中存在一些可以帮助判断物体位置的信息，充分利用目标背景信息与目标周围特征等辅助信息对实现目标的准确定位具有一定价值。

3) 平衡准确性、鲁棒性和实时性

在实际环境中对目标进行跟踪通常要在准确性和鲁棒性的基础上具备实时性，但为了提高跟踪的准确性，现在大多算法的模型较为复杂，导致跟踪速度相对较慢，平衡准确性、鲁棒性和实时性是具有研究价值的方向。

参考文献

[1] 李玺, 查宇飞, 张天柱, 崔振, 左旺孟, 侯志强, 卢湖川, 王菡子. 深度学习的目标跟踪算法综述[J]. 中国图象图形学报, 2019, 24(12): 2057-2080.

[2] 葛宝义, 左宪章, 胡永江. 视觉目标跟踪方法研究综述[J]. 中国图象图形学报, 2018, 23(8): 1091-1107.

[3] 卢湖川, 李佩霞, 王栋. 目标跟踪算法综述[J]. 模式识别与人工智能, 2018, 31(1): 61-76.

[4] 刘艺, 李蒙蒙, 郑奇斌, 秦伟, 任小广. 视频目标跟踪算法综述[J]. 计算机科学与探索, 2022, 16(7): 1504-1515.

[5] Jha, S., Seo, C., Yang, E. and Joshi, G.P. (2020) Real Time Object Detection and Tracking System for Video Surveillance System. *Multimedia Tools and Applications*, **80**, 3981-3996. https://doi.org/10.1007/s11042-020-09749-x

[6] 毛昭勇, 王亦晨, 王鑫, 沈钧戈. 面向高速公路的车辆视频监控分析系统[J]. 西安电子科技大学学报, 2021, 48(5): 178-189.

[7] 金立生, 华强, 郭柏苍, 谢宪毅, 闫福刚, 武波涛. 基于优化 DeepSort 的前方车辆多目标跟踪[J]. 浙江大学学报(工学版), 2021, 55(6): 1056-1064.

[8] 林梦馨. 基于孪生网络的颈动脉超声图像血管目标跟踪方法研究[D]: [硕士学位论文]. 上海: 华东师范大学, 2022.

[9] 何金刚, 徐林峰, 张金鹏, 余治民. 一种基于自适应网格机制的强机动目标滤波算法[J]. 航空兵器, 2021, 28(6): 40-45

[10] 鲁仁全, 罗勇民, 林明, 徐雍, 绕红霞. 基于轻量型孪生网络的降落跟踪控制方法和系统及无人机[P]. 中国专利, CN202110426555.2. 2021-07-16.

[11] 刘晓峰, 张春富, 唐鹏. 基于单目视觉的移动光斑跟踪定位方法[J]. 信息技术, 2020, 44(1): 48-53.

[12] 周俊宇, 赵艳明. 卷积神经网络在图像分类和目标检测应用综述[J]. 计算机工程与应用, 2017, 53(13): 34-41.

[13] 李旭冬, 叶茂, 李涛. 基于卷积神经网络的目标检测研究综述[J]. 计算机应用研究, 2017, 34(10): 2881-2886+2891.

[14] Bromley, J., Guyon, I., Lecun, Y., *et al*. (1993) Signature Verification Using a "Siamese" Time Delay Neural Network. In: Cowan, J., Tesauro, G. and Alspector, J., Eds., *Advances in Neural Information Processing Systems* 6.

[15] Nair, V., Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the* 27*th International Conference on International Conference on Machine Learning*, Haifa, 21-24 June 2010, 807-814.

[16] Zagoruyko, S. and Komodakis, N. (2015) Learning to Compare Image Patches via Convolutional Neural Networks. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 4353-4361. https://doi.org/10.1109/CVPR.2015.7299064

[17] 张卡, 宿东, 王蓬勃, 陈辉, 张珊, 叶龙杰, 赵娜. 深度学习技术在影像密集匹配方面的进展与应用[J]. 科学技

术与工程, 2020, 20(30): 12268-12278.

[18] Tao, R., Gavves, E. and Smeulders, A.W.M. (2016) Siamese Instance Search for Tracking. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 1420-1429. https://doi.org/10.1109/CVPR.2016.158

[19] Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A. and Torr, P.H.S. (2016) Fully-Convolutional Siamese Networks for Object Tracking. 2016 *European Conference on Computer Vision*, Amsterdam, 8-10 and 15-16 October 2016, 850-865. https://doi.org/10.1007/978-3-319-48881-3_56

[20] 蒲磊, 李海龙, 侯志强, 冯新喜, 何玉杰. 基于高层语义嵌入的孪生网络跟踪算法[J/OL]. 北京航空航天大学学报, 2022: 1-10. https://doi.org/10.13700/j.bh.1001-5965.2021.0319, 2022-07-09.

[21] Li, Y. and Zhang, X. (2019) SiamVGG: Visual Tracking Using Deeper Siamese Networks.

[22] 陈富健, 谢维信. 引入抗遮挡机制的 SiamVGG 网络目标跟踪算法[J]. 信号处理, 2020, 36(4): 562-571. https://doi.org/10.16798/j.issn.1003-0530.2020.04.010

[23] Zhang, Z. and Peng, H. (2020) Deeper and Wider Siamese Networks for Real-Time Visual Tracking. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4586-4595. https://doi.org/10.1109/CVPR.2019.00472

[24] 邵江南, 葛洪伟. 融合残差连接与通道注意力机制的 Siamese 目标跟踪算法[J]. 计算机辅助设计与图形学学报, 2021, 33(2): 260-269.

[25] Li, B., Yan, J., Wu, W., Zhu, Z. and Hu, X. (2018) High Performance Visual Tracking with Siamese Region Proposal Network. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8971-8980. https://doi.org/10.1109/CVPR.2018.00935

[26] Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J. and Yan, J. (2018) SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 4277-4286. https://doi.org/10.1109/CVPR.2019.00441

[27] 尚欣茹, 温尧乐, 奚雪峰, 胡伏原. 孪生导向锚框 RPN 网络实时目标跟踪[J]. 中国图象图形学报, 2021, 26(2): 415-424.

[28] 姜文涛, 崔江磊. 旋转区域提议网络的孪生神经网络跟踪算法[J/OL]. 计算机工程与应用, 2022: 1-11. http://kns.cnki.net/kcms/detail/11.2127.TP.20220618.1146.014.html, 2022-07-09.

[29] Tian, Z., Shen, C., Chen, H. and He, T. (2019) FCOS: Fully Convolutional One-Stage Object Detection. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 9626-9635. https://doi.org/10.1109/ICCV.2019.00972

[30] Guo, D., Wang, J., Cui, Y., Wang, Z. and Chen, S. (2020) SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, 13-19 June 2020, 6268-6276. https://doi.org/10.1109/CVPR42600.2020.00630

[31] Chen, Z., Zhong, B., Li, G., Zhang, S. and Ji, R. (2020) Siamese Box Adaptive Network for Visual Tracking. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6667-6676. https://doi.org/10.1109/CVPR42600.2020.00670

[32] Xu, Y., Wang, Z., Li, Z., Yuan, Y. and Yu, G. (2019) SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12549-12556. https://doi.org/10.1609/aaai.v34i07.6944

[33] Zhang, Z., Peng, H., Fu, J., Li, B. and Hu, W. (2020) Ocean: Object-Aware Anchor-Free Tracking. *European Conference on Computer Vision* 2020, Vol. 12366, Glasgow, 23-28 August 2020, 771-787. https://doi.org/10.1007/978-3-030-58589-1_46

[34] Bahdanau, D., Cho, K. and Bengio, Y. (2014) Neural Machine Translation by Jointly Learning to Align and Translate.

[35] Mnih, V., Heess, N., Graves, A. and Kavukcuoglu, K. (2014) Recurrent Models of Visual Attention. arXiv:1406.6247.

[36] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., *et al*. (2017) Attention Is All You Need. arXiv:1706.03762.

[37] Ni, Z.L., Bian, G.B., Xie, X.L., Hou, Z.-G., Zhou, X.-H. and Zhou, Y.-J. (2019) RASNet: Segmentation for Tracking Surgical Instruments in Surgical Videos Using Refined Attention Segmentation Network. 2019 41*st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (*EMBC*), Berlin, 23-27 July 2019, 5735-5738. https://doi.org/10.1109/EMBC.2019.8856495

[38] Yu, Y., Xiong, Y., Huang, W. and Scott, M.R. (2020) Deformable Siamese Attention Networks for Visual Object Tracking. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6727-6736. https://doi.org/10.1109/CVPR42600.2020.00676

[39] Du, F., Liu, P., Zhao, W. and Tang, X. (2020) Correlation-Guided Attention for Corner Detection Based Visual Tracking. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 6835-6844. https://doi.org/10.1109/CVPR42600.2020.00687

[40] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., *et al*. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

[41] Wang, N., Zhou, W., Wang, J. and Li, H. (2021) Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 1571-1580. https://doi.org/10.1109/CVPR46437.2021.00162

[42] Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X. and Lu, H. (2021) Transformer Tracking. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 8122-8131. https://doi.org/10.1109/CVPR46437.2021.00803

[43] Zhao, M., Okada, K. and Inaba, M. (2021) TrTr: Visual Tracking with Transformer. arXiv:2105.03817.

[44] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., *et al*. (2021) Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

[45] Lin, L., Fan, H., Xu, Y. and Ling, H. (2021) SwinTrack: A Simple and Strong Baseline for Transformer Tracking. arXiv:2112.00995.

[46] Yan, B., Peng, H., Wu, K., Wang, D., Fu, J. and Lu, H. (2021) LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 15175-15184. https://doi.org/10.1109/CVPR46437.2021.01493

[47] Blatter, P., Kanakis, M., Danelljan, M. and Van Gool, L. (2021) Efficient Visual Tracking with Exemplar Transformers. arXiv:2112.09686.

[48] Borsuk, V., Vei, R., Kupyn, O., Martyniuk, T., Krashenyi, I. and Matas, J. (2021) FEAR: Fast, Efficient, Accurate and Robust Visual Tracker. arXiv:2112.07957