·开发研究与工程应用·　　文章编号：1000-3428(2023)04-0281-08　　文献标志码：A　　中图分类号：TP391

结合卷积Transformer的目标跟踪算法

王春雷[1,2,3]，张建林[1,2]，李美惠[1,2]，徐智勇[1,2]，魏宇星[1,2]

（1.中国科学院光束控制重点实验室，成都 610209；2.中国科学院光电技术研究所，成都 610209；
3.中国科学院大学 电子电气与通信工程学院，北京 100049）

摘　要：现有基于Transformer的目标跟踪算法未充分利用Transformer的长距离依赖属性，导致算法提取的特征判别性不足，跟踪稳定性较差。为提高孪生网络目标跟踪算法在复杂场景中的跟踪能力，结合卷积与Transformer的优势，提出目标跟踪算法CTTrack。在特征提取方面，利用卷积丰富的局部信息和Transformer的长距离依赖属性，以卷积和窗口注意力串联的方式和层次化的结构构建一个通用的目标跟踪骨干网络CTFormer。在特征融合方面，利用互注意力机制构建特征互增强与聚合网络以简化网络结构，加快跟踪速度。在搜索区域选择方面，结合目标运动速度估计，设计自适应调整搜索区域的跟踪策略。实验结果表明，CTTrack在GOT-10k数据集上的平均重叠度为70.3%，相比基于Transformer的跟踪算法TransT和TrDiMP均提高3.2个百分点，在UAV123数据集上的曲线下面积为71.1%，相比TransT和TrDiMP分别提高2.0个百分点和3.6个百分点。在TrackingNet、LaSOT、OTB2015、NFS数据集上分别取得82.1%、66.8%、70.1%、66.3%的曲线下面积，并能以43帧/s的速度进行实时跟踪。

关键词：孪生网络；Transformer目标跟踪；窗口注意力；互注意力；运动估计；搜索区域

开放科学（资源服务）标志码（OSID）：

中文引用格式：王春雷，张建林，李美惠，等. 结合卷积Transformer的目标跟踪算法[J]. 计算机工程，2023，49(4)：281-288，296.

英文引用格式：WANG C L，ZHANG J L，LI M H，et al.Object tracking algorithm combining convolution and Transformer[J]. Computer Engineering，2023，49(4)：281-288，296.

Object Tracking Algorithm Combining Convolution and Transformer

WANG Chunlei[1,2,3]，ZHANG Jianlin[1,2]，LI Meihui[1,2]，XU Zhiyong[1,2]，WEI Yuxing[1,2]

（1.Key Laboratory of Beam Control，Chinese Academy of Sciences，Chengdu 610209，China；
2.Institute of Optics and Electronics，Chinese Academy of Sciences，Chengdu 610209，China；
3.School of Electronic，Electrical and Communication Engineering，University of Chinese Academy of Sciences，Beijing 100049，China）

【Abstract】The existing target object algorithms based on Transformer do not fully use Transformer's long-distance dependence attribute，resulting in insufficient discriminability of the features extracted by the algorithm and poor tracking stability.To improve the object tracking ability，a object tracking algorithm CTTrack is proposed for complex scenes，combining the advantages of convolution and Transformer.In terms of feature extraction，the algorithm combines the rich local information of convolution and long-distance dependence attribute of Transformer to construct a general object tracking backbone network CTFormer，by concatenating convolution and window attention in a hierarchical structure.In feature fusion，only the Cross-Attention Mechanism（CAM）is used to construct the feature mutual enhancement and aggregation networks，which simplifies the network structure and improves tracking speed.In search area selection，the tracking strategy of adaptive search area adjustment is designed based on object motion speed estimation.The experimental results show that the Average Overlap（AO）of CTTrack on GOT-10k dataset is 70.3%，which is 3.2 percentage points higher than that of TransT and TrDiMP，and the Area Under the Curve（AUC）on the UAV123 dataset is 71.1%，which is 2.0 and 3.6 percentage points higher than on TransT and TrDiMP，respectively.The AUC on the TrackingNet，LaSOT，OTB2015，and NFS datasets，are 82.1%，66.8%，70.1%，and 66.3%，respectively，with real-time tracking at a speed of 43 frames/s.

【Key words】siamese network；Transformer object tracking；window attention；cross-attention；motion estimation；search area

DOI：10. 19678/j. issn. 1000-3428. 0064096

基金项目：国家自然科学基金青年科学基金“基于交叉度量跨模态学习的多谱段目标跟踪方法研究”（62101529）。

作者简介：王春雷（1996—），男，硕士研究生，主研方向为目标跟踪；张建林（通信作者），研究员、博士、博士生导师；李美惠，博士；徐智勇，研究员、博士生导师；魏宇星，副研究员。

收稿日期：2022-03-04　　**修回日期：**2022-04-21　　**E-mail：**wangchunlei20@mails.ucas.ac.cn

0 概述

视频目标跟踪是计算机视觉领域中重要的方向，广泛应用于军事、医学、安防、无人驾驶等领域。但是在实际工程中经常存在目标姿态变化、背景干扰、遮挡、尺度变化等情况而影响目标跟踪效果[1-2]。此外，实时性也是评价跟踪算法实际应用的重要指标。因此，在满足实时性的前提下，提高算法在复杂场景中的跟踪精度具有重要意义。

近年来，基于孪生网络的跟踪算法因其具有精度高、速度快的特点而成为目标跟踪算法的主流方向。SiamFC[3]全面完整地将孪生网络引入目标跟踪中，将目标跟踪作为简单的相似性度量问题，使用浅层网络AlexNet提取特征，通过卷积度量两个分支的相似性，为后续算法的发展提供一个新的方向。SiamRPN[4]将检测领域中的区域提议网络（Region Proposal Network，RPN）引入到跟踪算法中，在一定程度上解决了SiamFC[3]的尺度问题，跟踪精度和速度得到有效提高，但是RPN的引入带来了部分超参数，使得网络对于超参数过于敏感。SiamRPN++[5]和SiamDW[6]通过深度分析孪生网络跟踪算法的特点，将骨干网络从浅层的AlexNet、GoogleNet等推广到深层的ResNet[7]，为后续算法的发展提供扎实的基础。研究人员提出的SiamFC++[8]和SiamCAR[9]算法再次将目标检测中的Anchor-Free策略引入到跟踪算法领域中，缓解超参数敏感的问题，提升跟踪精度。2021年主流的TransT[10]、STARK[11]、TrDiMP[12]等算法在孪生网络上引入Transformer[13]进行特征增强和融合，大幅提升算法的跟踪效果。

虽然现有基于Transformer[13]目标跟踪算法的性能获得显著提高，但是其本质仅简单使用Transformer进行特征的增强和融合，未充分利用Transformer的长距离依赖属性，无法完全发挥Transformer的优势。此外，Transformer相对于卷积神经网络具有更高的计算量，导致相关算法的网络过于臃肿，难以真正投入使用，而且因其长距离依赖属性导致在提取视觉特征时无法获取丰富的局部信息，然而，卷积神经网络能够提取丰富的局部特征且计算量较小。因此，为获得更优的跟踪效果和更快的跟踪速度，本文在现有算法的基础上，提出结合卷积Transformer的目标跟踪算法CTTrack。为充分利用卷积神经网络与Transfomer的特性，设计一个全新的目标跟踪骨干网络。利用互注意力设计简单的特征互增强与聚合网络，抛弃繁琐的编码-解码过程，降低计算量并加快跟踪速度。针对因跟踪过程中目标快速运动、目标丢失等存在搜索区域选择困难的问题，通过运动估计自适应动态调整搜索区域的策略，进一步提高跟踪精度。

1 相关工作

1.1 孪生网络目标跟踪算法

孪生网络目标跟踪算法具有结构简单、精度较优、速度较快特点。其中，SiamFC[3]普遍被认为是首个孪生跟踪网络，后续算法大多在此基础上从不同角度进行探索。SiamFC[3]网络结构如图1所示。

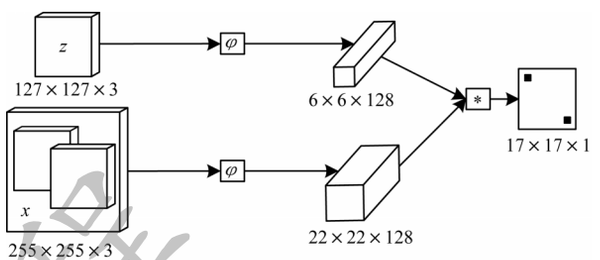


图1 SiamFC网络结构

Fig.1 Structure of SiamFC network

SiamFC[3]由模板分支和搜索分支组成。两个分支的输入模板图像z和搜索图像x通过共享权重的骨干网络φ提取特征，并以模板分支的特征图作为卷积核与搜索图像的特征图进行卷积，以得到响应图，响应图中响应值最高的位置对应着目标可能出现的位置，最后将响应图进行双三次插值定位目标位置，后续发表的孪生网络跟踪算法结构大致与此类似。

1.2 Transformer的应用

Transformer[13]于2017年被提出，最早被应用于机器翻译领域，使用注意力机制组成编码-解码的结构。后续研究发现基于Transformer[13]的模型在各种自然语言处理任务中表现良好，目前已经取代长短时记忆（Long Short-Term Memory，LSTM）[14]网络成为自然语言处理领域的首选框架。从2020年开始，Transformer被应用到计算机视觉领域，DETR[15]算法基于Transformer设计一个端到端的目标检测框架，在不增加任何先验知识的情况下，取得较优的效果。受DETR[15]的影响，Transformer在计算机视觉领域迅速发展。ViT[16]将图像拆分成不同的小块，设计一个完全无卷积的网络结构，在大规模数据集上获得优于ResNet[7]的性能，标志着完全无卷积的Transformer网络在计算机视觉领域具有较高的可行性，但是因Transformer具有较大的计算量，在下游任务中难以得到应用。

2 本文算法

针对孪生网络目标跟踪算法在复杂场景中跟踪漂移、鲁棒性不足、实时性较差等问题，本文提出结合卷积Transformer[13]的目标跟踪算法CTTrack，其网络结构如图2所示。该网络结合卷积与Transformer[13]的特性，设计通用的骨干网络CTFormer，仅采用互注意力机制（Cross-Attention Mechanism，CAM）构建一个简单的特征互增强与聚合网络（ECN）。针对在推理过程中搜索区域选择困难的问题，本文提出结合运动估计自适应调整搜索区域的策略AAS。

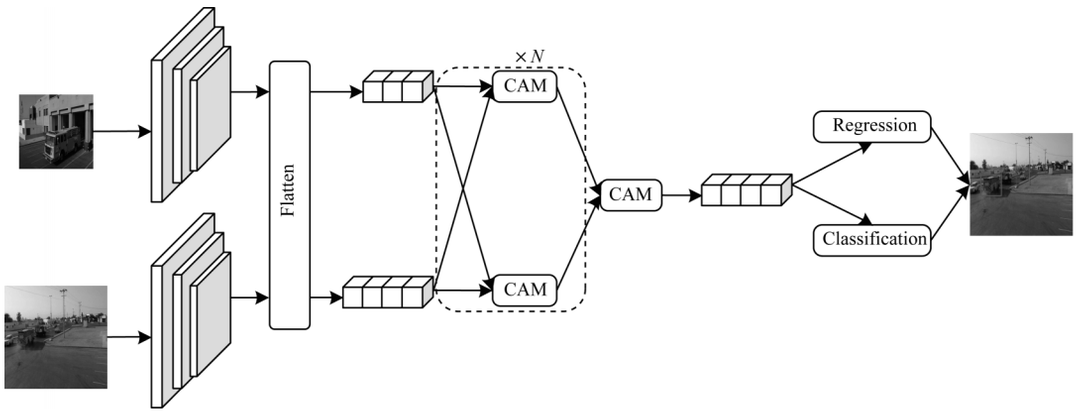


图 2　CTTrack网络结构

Fig.2　Structure of CTTrack network

CTTrack网络分为5个部分：

1）骨干网络，采用本文设计的CTFormer网络的前3个阶段，并去掉第3个阶段的池化层，共进行16倍下采样，2个分支权重共享维持孪生的结构。为获取鲁棒性更优的特征，本文同时对第2层和第3层的特征进行加权输出。

2）Flatten模块，通过Flatten模块调整骨干网络输出特征的维度和通道数。Flatten模块即为简单的卷积核为1×1的卷积，调整输出通道数为256。

3）特征互增强与聚合网络，分别对2个分支图像进行特征互增强与聚合，该网络主要由CAM模块组成。为获得更好的效果，通过重复4次实验验证该结构的有效性，在保证实时性的同时获得较优的性能。

4）相似性度量，采用CAM模块度量两个分支的相似性并生成响应图。CAM模块能进行像素级的逐点度量，相较于早期利用卷积进行全局相似性度量的方式具有更优的鲁棒性。

5）预测头网络，其设计参考DETR[15]算法，包括一个分类分支和一个回归分支。每个分支均由带有一个ReLU激活函数的多层感知机（Multi-Layer Perceptron，MLP）组成，对每个特征向量进行预测。分类分支预测每个特征向量的前景、背景分类结果；回归分支预测目标所在区域的归一化坐标，并采用分类分支的分类结果指导回归过程，即基于分类得分最高的值选取回归分支的最终唯一输出坐标。整个预测头采用Anchor-Free策略，完全抛弃基于先验知识的锚点框，使本文所提的网络结构更加简洁。

损失函数的设计也与DETR[15]算法类似，采用标准的二元交叉熵作为分类损失，如式（1）所示：

$$L_{\mathrm{cls}}=-\sum_{i=1}^{N}[y_i\log_a(p_i)+(1-y_i)\log_a(1-p_i)] \tag{1}$$

其中：y_i表示第i个样本的真实标签，1为前景，0为背景；p_i表示预测第i个样本为前景的概率。回归损失函数采用L1损失和GIOU损失的线性组合，如式（2）所示：

$$L_{\mathrm{reg}}=\sum_{i=1}^{N}[\lambda_G L_{\mathrm{GIOU}}(b_i,\hat{b})+\lambda_1 L_1(b_i,\hat{b})] \tag{2}$$

其中：b_i表示第i个预测的边界框；$\hat{b}$表示归一化的真实边界框；GIOU损失的系数λ_G为2；L1损失的系数λ_1为5。

2.1　CTFormer骨干网络

卷积神经网络被广泛应用于目标跟踪领域，从早期的AlexNet、GoogleNet到ResNet[7]，骨干网络一直向更深的网络发展。因此，骨干网络获得更优的特征表示对跟踪任务具有重要作用。但是，自从SiamRPN++[5]将ResNet[7]应用于跟踪任务中，受限于实时性的要求，骨干网络一直停留在ResNet[7]上。虽然纯Transformer[13]结构的ViT[16]已经在图像分类任务上获得远优于ResNet[7]的性能，但是极高的计算量使其难以真正应用于跟踪任务中。此外，卷积神经网络虽然提取特征的判别性不足、区域相关性较弱，但是在提取底层特征时获取局部信息方面具有较大的优势。Transformer[13]因其长距离依赖属性，更加擅长提取全局特征，因此，对卷积和Transformer[13]进行合理地结合可以有效地弥补各自缺陷并充分发挥各自优势。最近研究表明，Transformer[13]性能的强大不仅在于其频繁叠加的全局注意力，而且与其独特的结构密不可分。因此，为充分结合卷积与Transformer[13]的优势，本文设计结合卷积Transformer[13]的模块，命名为CTFormer，结构如图3所示。

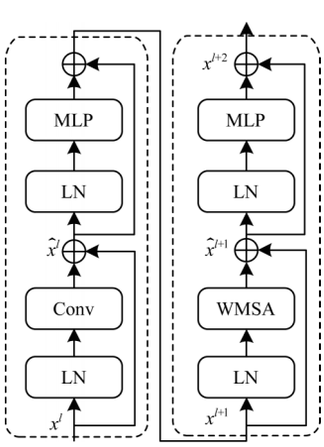


图 3　CTFormer模块结构

Fig.3　Structure of CTFormer module

该模块由归一化层(LN)、卷积层(Conv)、多层感知机(MLP)、窗口注意力层(WMSA)组成,其中,x^l为第l层的输入,x^{l+1}为第$l+1$层输入,也为第l层的输出,x^{l+2}为第$l+1$层输出。具体计算过程如式(3)~式(6)所示:

$$\hat{x}^l=\mathrm{Conv}(\mathrm{LN}(x^l))+x^l \tag{3}$$

$$x^{l+1}=M_{\mathrm{MLP}}(\mathrm{LN}(\hat{x}^l))+\hat{x}^l \tag{4}$$

$$\hat{x}^{l+1}=W_{\mathrm{WMSA}}(\mathrm{LN}(x^{l+1}))+x^{l+1} \tag{5}$$

$$x^{l+2}=M_{\mathrm{MLP}}(\mathrm{LN}(\hat{x}^{l+1}))+\hat{x}^{l+1} \tag{6}$$

使用窗口注意力代替原Transformer[13]中计算量庞大的全局自注意力。其中,窗口注意力层仅在固定尺寸为8的窗口内计算局部注意力,相对于全局注意力具有更少的计算量。虽然窗口注意力无法像全局注意力一样建模全局特性,但是在实际跟踪任务中使用局部注意力相较于全局注意力仅有细微的精度损失。为弥补精度的损失,本文在前端接入一个同样以卷积代替全局注意力的类Transformer[13]模块,将两者串联成对组成CTFormer模块。卷积的添加使不同的窗口间有了一定的信息交互,使得窗口注意力不仅局限于某个窗口内,而且能够获得鲁棒性更优的图像特征。

受PVT[17]和Swin Transformer[18]的启发,本文同样采用卷积神经网络的层次化结构构建网络,CTFormer网络结构如图4所示。

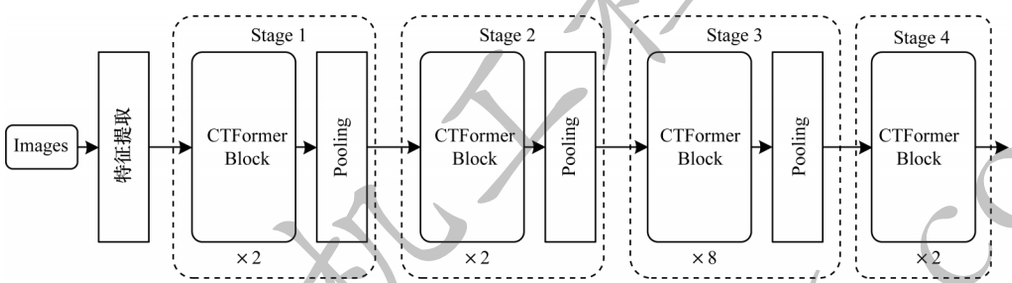


图4 CTFormer网络结构

Fig.4 Structure of CTFormer network

CTFormer网络由浅层特征提取层、CTFormer模块、池化层组成,分为4个阶段,各个阶段的CTFormer模块数量设置为{2,2,8,2}。其中,浅层特征提取层直接使用EfficientNetV2[19]网络的前3个阶段来提取底层特征,同时调整该层输出通道数为96,总步长为4,特征图分辨率降低1/4。

池化层为简单的2倍下采样,并调整输出通道数为输入的2倍,这样便构成典型的金字塔结构,特征图的分辨率随着不同阶段的网络深度逐渐减小,通道数逐渐增大。在ImageNet1k上对CTFormer网络进行预训练,最终获得83.1%的Top-1准确率,远超ResNet-50[7]的76.5%,后续实验结果表明,该网络更加适用于跟踪任务。

2.2 特征互增强与聚合

Transformer[13]的多层编码-解码结构广泛应用于目标跟踪领域,如STARK[11]、TrDiMP[12]等性能大幅度领先其他跟踪算法,但繁琐的编码-解码结构使网络过于臃肿,带来极大的计算量,难以真正投入使用。因此,为了在不产生过多计算量的情况下合理利用Transformer[13]的优势,本文参考TransT[10]的设计,仅截取Transformer[13]结构中计算互注意力的部分来融合不同分支的特征。CAM模块的结构如图5所示。

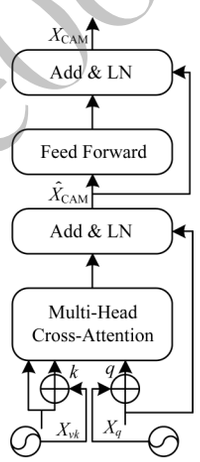


图5 互注意力机制模块结构

Fig.5 Structure of cross-attention mechanism block

CAM模块基于残差网络的思想,结合多头互注意(Multi-Head Cross-Attention,MHCA)、归一化、前馈神经网络设计而构建的,整个CAM模块的计算过程如式(7)和式(8)所示:

$$\hat{X}_{\mathrm{CAM}}=\mathrm{LN}(M_{\mathrm{MHCA}}(X_{kv}+P_{kv},X_q+P_q)+X_q) \tag{7}$$

$$X_{\mathrm{CAM}}=\mathrm{LN}(F_{\mathrm{FFN}}(\hat{X}_{\mathrm{CAM}})+\hat{X}_{\mathrm{CAM}}) \tag{8}$$

其中:X_q为本分支的输入;P_q为X_q的空间位置编码;X_{kv}为另一个分支的输入;P_{kv}为X_{kv}的空间位置编码,位置编码均由正弦函数生成。CAM模块通过多头互注意力获得两个分支的相似性后,结合残差连接及归一化获得初步聚合增强后的本分支特征$\hat{X}_{\mathrm{CAM}}$,

经过由2个线性变换和一个ReLU激活函数组成的前馈神经网络进行空间变换，以增强模型的表现能力，最终通过残差连接和归一化获得聚合增强后的本分支特征X_{CAM}。

CAM模块的交叉使用分别对2个分支的特征进行增强，构建特征互增强与聚合网络。对特征互增强与聚合网络重复多次获取更具判别性的特征，同时也可借助CAM模块度量2个分支的相似性，获得响应图。使用特征互增强与聚合网络，相对于STARK[11]、TrDiMP[12]重复6次繁琐的编码-解码结构具有更低的计算量，不需要额外地计算各个分支自注意力进行自增强的过程，在不降低性能的同时加快跟踪速度。

2.3　自适应动态调整搜索区域的跟踪策略

经过多次实验，本文发现搜索区域的大小对跟踪效果有较大的影响，现有算法如TransT[10]、STARK[11]等选择一个相对目标尺寸固定放大倍数的搜索区域，但是固定放大倍数的搜索区域无法处理跟踪过程出现的复杂情况。在跟踪任务中目标的运动是不均匀的，而且存在较大的视角变化，搜索区域选择过大，可能包含过多干扰物导致跟踪漂移。搜索区域选择过小，当目标快速运动时，目标可能会离开视野无法跟踪。针对这一问题，本文提出一个通过运动估计动态调整搜索区域的跟踪策略。本文设置初始搜索区域放大倍数为3，进行跟踪并获取连续5帧的目标中心点位置(x_i,y_i)、(x_{i+1},y_{i+1})、(x_{i+2},y_{i+2})、(x_{i+3},y_{i+3})、(x_{i+4},y_{i+4})，相邻2帧的中心点偏差的计算过程如式(9)~式(12)所示：

$$(\Delta x_1,\Delta y_1)=(|x_{i+1}-x_i|,|y_{i+1}-y_i|) \tag{9}$$

$$(\Delta x_2,\Delta y_2)=(|x_{i+2}-x_{i+1}|,|y_{i+2}-y_{i+1}|) \tag{10}$$

$$(\Delta x_3,\Delta y_3)=(|x_{i+3}-x_{i+2}|,|y_{i+3}-y_{i+2}|) \tag{11}$$

$$(\Delta x_4,\Delta y_4)=(|x_{i+4}-x_{i+3}|,|y_{i+4}-y_{i+3}|) \tag{12}$$

并计算相对于x轴和y轴运动距离的最大值，如式(13)~式(16)所示：

$$d_1=\max(\Delta x_1,\Delta y_1) \tag{13}$$

$$d_2=\max(\Delta x_2,\Delta y_2) \tag{14}$$

$$d_3=\max(\Delta x_3,\Delta y_3) \tag{15}$$

$$d_4=\max(\Delta x_4,\Delta y_4) \tag{16}$$

根据4个相邻两帧运动距离的最大值d_1、d_2、d_3、d_4调整搜索区域的放大倍数s。通过多次实验测试，本文初步设置搜索区域放大倍数s和d_1、d_2、d_3、d_4的关系，如式(17)所示：

$$s=\begin{cases}4, d_1,d_2,d_3,d_4\geqslant 25\\ 2.5, d_1,d_2,d_3,d_4\leqslant 18\\ 3, \text{其他}\end{cases} \tag{17}$$

通过后续实验验证，该策略相对于固定搜索区域放大倍数的策略具有更优的性能，而且能够减少大尺寸目标图像不必要的Padding操作，加快推理速度。

3　实验与结果分析

3.1　实验细节

本文实验的所有训练过程软件环境为Ubuntu20.04、PyTorch1.7.1、Python3.8.8，硬件配置为Intel® Xeon® Platinum 8163 CPU和GeForce RTX™ 3090 GPU ×8。推理过程在RTX 3060上进行。

对于骨干网络预训练过程，本文在ImageNet1k上使用PyTorch扩展工具Apex进行实验，结合增强和正则化策略，采用AdamW优化器训练300个周期。本文设置batch size为128，初始学习率为0.001，并采用余弦衰减调整学习率，骨干网络在第280个周期左右性能达到饱和，获得83.1%的Top-1准确率。

对于整个跟踪网络训练过程，本文在GOT-10k[20]、LaSOT[21]、COCO[22]、TrackingNet[23]4个通用的目标跟踪数据集上进行训练，并采用随机采样、变换策略生成训练样本对。裁剪模板图像分支输入图像尺寸为128×128像素，搜索图像分支输入尺寸为256×256像素。骨干网络学习率设置为1e−5，其他参数学习率设置为1e−4，采用分布式数据并行(Distributed Data Paralle，DDP)进行单机多卡训练，每个GPU的batch size设置为50，每个周期训练400 000对图像，共训练120个周期，在第70个周期后学习率衰减0.1。

3.2　结果分析

本文实验将本文所提的算法与近三年表现优异的算法(STARK[11]、TransT[10]、TrDiMP[12]、SiamR-CNN[24]、PrDiMP50[25]、Ocean[26]、DiMP50[27]、SiamRPN++[5]、ATOM[28])进行定性与定量分析，以验证算法的性能。

3.2.1　定性分析

为验证算法的性能，针对OTB2015[29]数据集中目前主流挑战属性进行定性实验，实验结果如图6所示(彩色效果见《计算机工程》官网HTML版)。第1~4行的挑战属性依次为背景干扰、目标遮挡、光照变化、姿态变化。

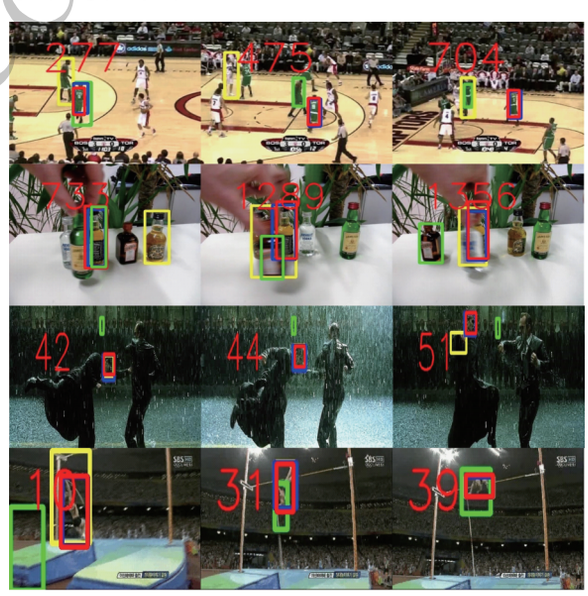


图6　不同算法的预测结果对比

Fig.6　Prediction results comparison among different algorithms

图6中红色标注框为真实标注框，蓝色标注框为本文算法CTFormer的预测框，预测框与真实标注框越贴近，重合度越高代表跟踪效果越好。

1）背景干扰挑战，在第1行序列第277帧中，当背景出现干扰物时，TransT[10]算法跟错目标，在第475帧中当干扰物再次与目标拉近距离时，对比算法均跟踪错误，并导致后续严重的跟踪漂移。这是因为对比算法没有有效调整搜索区域的策略，导致干扰物与目标同时出现在搜索区域中，所提取的特征又不足以分辨干扰物与目标，导致跟踪错误，验证了本文所提动态调整搜索区域策略的有效性。

2）目标遮挡挑战，在第2行序列第733帧中，当目标被部分遮挡时，TransT[10]算法跟踪错误。在第1 289帧中，当目标被轻微遮挡时，对比算法跟踪效果降低，无法准确框选出目标。第1 356帧中，当目标被大部分干扰物遮挡时，ATOM[28]算法跟踪错误，在此过程中本文算法一直能够对目标进行稳定跟踪。其原因为对比算法提取特征的表达能力不足，无法根据目标的部分特征完成整体跟踪，进而说明本文所提的骨干网络和特征聚合增强网络提取特征的表达能力足够强。

3）光照变化挑战，从第3行序列可以看出：当场景中光照发生变化时，对比算法跟踪效果均会不同程度的降低，甚至会发生跟踪漂移，而本文算法能稳定跟踪，说明本文所提网络提取的特征具有更优的鲁棒性。

4）姿态变化挑战，从第4行序列可以看出：当目标发生较大姿态变化时，对比算法的预测框无法根据目标的姿态变化进行有效调整，导致预测框过大且精度降低，甚至当姿态变化剧烈时，导致ATOM[28]算法跟踪错误。

通过以上4个主流挑战属性的对比，充分验证本文算法的有效性，并验证本文算法的主体部分发挥了一定的作用。

3.2.2 定量分析

为更加充分地说明本文算法的有效性，在多个公开数据集上进行大量的定量实验。首先在GOT-10k[20]数据集上进行对比实验。GOT-10k[20]场景丰富挑战难度高，包含10 000多条真实拍摄的视频片段和563个类别，超过150万个手工标注框。GOT-10k[20]是单目标跟踪的一个主流评价基准，以平均重叠度（Average Overlap，AO）和成功率（Success Rate，SR）作为主要的评价指标。遵照其要求，本文仅在GOT-10k[20]上进行训练并与其他算法进行对比，具体对比情况如图7所示。

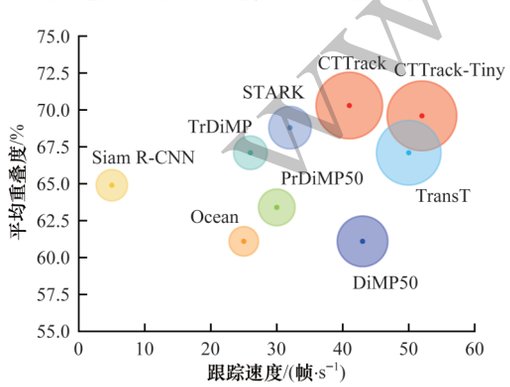


图7 不同算法在GOT-10k数据集上的平均重叠度对比

Fig.7 Average overall comparison among different algorithms on GOT-10k dataset

在图7中，CTTrack为本文的基础版本，CTTrack-Tiny为仅进行2次CAM融合的轻量版本。本文算法基础版本的平均重叠度（AO）达到70.3%，优于当前主流算法，比STARK[11]提高1.5个百分点，比TransT[10]和TrDiMP[12]提高约3.2个百分点，相对于Ocean、PrDiMP50、Siam R-CNN算法普遍提高了5~10个百分点，取得了较优的效果。在跟踪速度方面，本文所提的算法分别以43帧/s和53帧/s的速度超越目前的主流算法，具有较优的实时性。在GOT-10k数据集上不同算法的性能对比如表1所示，加粗表示最优数据。

表1 不同算法在GOT-10k数据集上的性能对比

Table 1 Performance comparison among different algorithms on GOT-10k dataset %

算法	AO	$SR_{0.50}$	$SR_{0.75}$
SiamRPN++	51.7	61.6	32.5
DiMP50	61.1	71.7	49.2
Ocean	61.1	72.1	47.3
PrDiMP50	63.4	73.8	54.3
Siam R-CNN	64.9	72.8	59.7
TrDiMP	67.1	77.7	58.3
TransT	67.1	76.8	60.9
STARK	68.8	78.1	**64.1**
CTTrack-Tiny	69.6	**80.6**	61.9
CTTrack	**70.3**	80.3	63.9

本文在LaSOT[21]数据集上进行测试实验。LaSOT[21]是一个大规模的长时跟踪数据集及评价基准，包含1 400个视频序列，其中，训练集1 120个序列，测试集280个序列，平均每个序列2 500多帧，共有352万个高质量的手工标注框。评价标准一般为曲线下面积（AUC）和归一化精度（P_{norm}）。在LaSOT[21]数据集上不同算法的评价指标对比如表2所示。

表2 不同算法在LaSOT数据集上的评价指标对比

Table 2 Evaluation indicators comparison among different algorithms on LaSOT dataset %

算法	AUC	P_{norm}
SiamRPN++	49.6	56.9
ATOM	51.5	57.6
Ocean	56.0	65.1
DiMP50	56.9	65.0
PrDiMP50	59.8	68.8
TrDiMP	63.9	—
Siam R-CNN	64.8	72.2
TransT	64.9	73.8
STARK	**67.1**	**77.0**
CTTrack-Tiny	65.2	75.0
CTTrack	66.8	76.1

从表2可以看出：本文所提算法的基础版本（CTTrack）的性能指标大幅度领先目前的主流算法，相对于TransT[10]和TrDiMP[12]分别提高1.9和2.9个百分点，比SiamRPN++、ATOM、Ocean、DiMP50、PrDiMP50、Siam R-CNN算法普遍提高了10个百分点。由于本文

算法没有添加任何额外的模板更新策略，因此在长时跟踪上有一定的劣势，导致本文算法以0.3个百分点的差距落后于LaSOT[21]排行榜上的第一名STARK[11]。

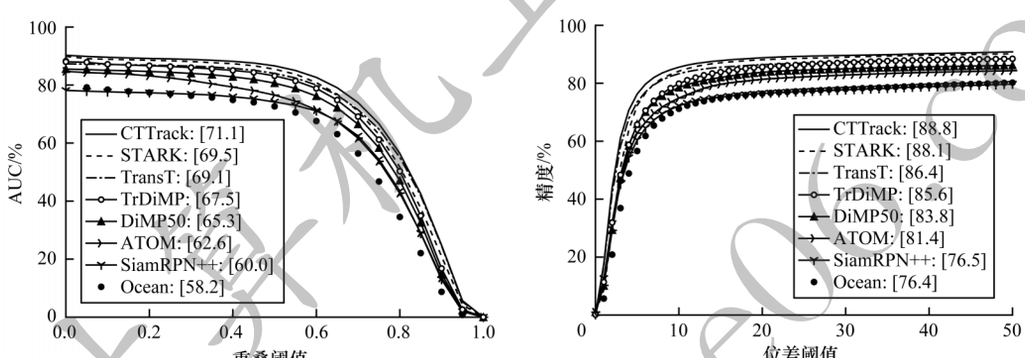
本文在TrackingNet[23]数据集上对不同算法进行对比测试。TrackingNet[23]是一个更大规模的单目标跟踪数据集，超过30 000个视频序列，通过在YouYube视频上采样来表示真实世界的场景，因此涵盖非常丰富的目标类别。评价标准与LaSOT[21]类似一般为曲线下面积(AUC)和归一化精度(P_{norm})。在TrackingNet[23]上不同算法的评价指标对比如表3所示。从表3可以看出：本文算法的基础版本在AUC和P_{norm}均领先于目前的主流算法，甚至超越了STARK[11]，AUC达到82.1%。

本文在UAV123[30]数据集上进行对比测试。UAV123[30]是一个完全由无人机拍摄的数据集，背景干净但视角变化较多，共包含123个视频序列，其中，有20个长视频，相对于OTB2015[29]跟踪难度更高。本文采用AUC和精度(P)作为评价指标。由于不同的测试工具有一定的误差，因此为保证对比实验的公平性，本文将对比算法的原始跟踪结果均在GOT-10k[20]工具包上进行重新测试，具体测试结果如图8所示。从图8可以看出：本文算法无论是AUC还是精度均具有较优的表现。

表3　不同算法在TrackingNet数据集上的评价指标对比

Table 3　Evaluation indicators comparison among different algorithms on TrackingNet dataset　%

算法	AUC	P_{norm}
ATOM	70.3	77.1
SiamRPN++	73.3	80.0
DiMP50	74.0	80.1
PrDiMP50	75.8	81.6
TrDiMP	78.4	83.3
Siam R-CNN	81.2	85.4
TransT	81.4	86.7
STARK	82.0	86.9
CTTrack-Tiny	81.5	**87.3**
CTTrack	**82.1**	87.2



(a)曲线下面积对比　　(b)精度对比

图8　不同算法在UAV123数据集上的评价指标对比

Fig.8　Evaluation indicators comparison among different algorithms on UAV123 dataset

本文在OTB2015[29]和NFS[31]数据集上进行对比测试实验，不同算法的AUC对比如表4所示。

表4　不同算法在OTB2015和NFS数据集上的评价指标对比

Table 4　Evaluation indicators comparison among different algorithms on OTB2015 and NFS dataset　%

算法	AUC	
	OTB2015数据集	NFS数据集
STARK	68.1	66.2
TransT	69.4	65.3
PrDiMP50	69.6	63.5
DiMP50	68.4	61.8
ATOM	66.9	58.4
SiamRPN++	68.7	57.1
本文算法	**70.1**	66.3

OTB2015[29]共有100个视频序列，包含遮挡、光照变化、快速运动等11个挑战属性。NFS[31]数据集包含100个视频序列，共17个物体类别，有2个版本分别为30帧/s和240帧/s，在240帧/s版本上各个算法的指标差距较小。因此，本文仅在30帧/s版本上进行对比测试。从表4可以看出：本文算法的性能相比于主流算法具有最优的性能。

3.3　消融实验

为充分挖掘算法的性能且明确各个策略对网络性能的影响，本文在GOT-10k[20]数据集上进行消融实验，具体消融实验结果如表5所示。

表5　消融实验结果

Table 5　Ablation experiment results

CTFormer	ECN	AAS	ResNet-50	Encoder-Decoder	AO/%	跟踪速度/(帧·s^{-1})
			✓		46.6	66
✓					61.0	63
			✓	✓	66.1	30
		✓	✓	✓	66.6	33
	✓		✓		65.8	45
	✓	✓	✓		66.4	48
✓				✓	69.3	26
✓		✓		✓	69.9	29
✓	✓				69.5	40
✓	✓	✓			70.3	43

在表5中，✓表示使用当前策略，没有✓表示不使用当前策略。从表5可以看出：当使用ResNet-50[7]作为骨干网络时，编码-解码的结构相对于仅利用互注意力的ECN结构具有更优的性能，但是所需的计算量更大；本文设计的CTFormer网络和自适应调整搜索区域的跟踪策略（AAS）有助于提高算法的性能，尤其是在不采用任何特征增强手段时，CTFormer相对于ResNet-50[7]提升近15个百分点；将CTFormer、ECN和AAS策略相结合能有效提高算法的性能，同时具有较优的实时性。

4 结束语

本文提出Transformer[13]的目标跟踪算法CTTrack。设计简单高效的通用目标跟踪骨干网络CTFormer，提取具有判别性的特征以便用于后续的跟踪过程，通过简化现有基于Transformer[13]目标跟踪算法的网络结构，仅使用少量的互注意力进行特征的互增强与聚合，在不降低跟踪精度的前提下减少计算量并加快跟踪速度。针对跟踪过程中搜索区域选择困难的问题，提出根据运动速度自适应调整搜索区域的跟踪策略，在不产生任何计算开销的情况下提升跟踪精度并加快跟踪速度。实验结果表明，相比STARK、TransT、PrDiMP50等算法，CTTrack具有较优的跟踪精度和实时性。后续将结合时空特征与重检测思想，对复杂场景下（如同时出现背景干扰和目标遮挡）目标跟踪问题进行研究，以进一步提高CTTrack算法的稳定性。

参考文献

[1] 李珑，刘凯，李玲. 基于目标检测的时空上下文跟踪算法[J]. 计算机工程，2018，44(9)：263-268，273.
LI L，LIU K，LI L. Spatial-temporal context tracking algorithm based on target detection[J]. Computer Engineering，2018，44(9)：263-268，273. (in Chinese)

[2] 任立成，杨嘉棋，魏宇星，等. 基于特征融合与双模板嵌套更新的孪生网络跟踪算法[J]. 计算机工程，2021，47(7)：239-248.
REN L C，YANG J Q，WEI Y X，et al. Tracking algorithm using siamese network based on feature fusion and dual-template nested update[J]. Computer Engineering，2021，47(7)：239-248. (in Chinese)

[3] BERTINETTO L，VALMADRE J，HENRIQUES J F，et al. Fully-convolutional siamese networks for object tracking[EB/OL]. [2022-02-01]. https://arxiv. org/pdf/1606. 09549. pdf.

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

[5] LI B，WU W，WANG Q，et al. SiamRPN：evolution of siamese visual tracking with very deep networks[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2020：4277-4286.

[6] ZHANG Z P，PENG H W. Deeper and wider siamese networks for real-time visual tracking[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2020：4586-4595.

[7] HE K M，ZHANG X Y，REN S Q，et al. Deep residual learning for image recognition[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2016：770-778.

[8] XU Y D，WANG Z Y，LI Z X，et al. SiamFC++：towards robust and accurate visual tracking with target estimation guidelines[C]//Proceedings of Conference on Artificial Intelligence. [S. l.]：AAAI Press，2020：12549-12556.

[9] GUO D Y，WANG J，CUI Y，et al. SiamCAR：siamese fully convolutional classification and regression for visual tracking[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2020：6268-6276.

[10] CHEN X，YAN B，ZHU J W，et al. Transformer tracking[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2021：8122-8131.

[11] YAN B，PENG H W，FU J L，et al. Learning spatio-temporal Transformer for visual tracking[C]//Proceedings of International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：10428-10437.

[12] WANG N，ZHOU W G，WANG J，et al. Transformer meets tracker：exploiting temporal context for robust visual tracking[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C.，USA：IEEE Press，2021：1571-1580.

[13] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems. Washington D. C.，USA：IEEE Press，2017：5998-6010.

[14] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[15] CARION N，MASSA F，SYNNAEVE G，et al. End-to-end object detection with Transformers[C]//Proceedings of European Conference on Computer Vision. Berlin，Germany：Springer，2020：213-229.

[16] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16×16 words：Transformers for image recognition at scale[C]//Proceedings of International Conference on Learning Representations. Washington D. C.，USA：[s. n.]，2020：1-9.

[17] WANG W H，XIE E Z，LI X，et al. Pyramid vision Transformer：a versatile backbone for dense prediction without convolutions[C]//Proceedings of International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：548-558.

[18] LIU Z，LIN Y T，CAO Y，et al. Swin Transformer：hierarchical vision Transformer using shifted windows[C]//Proceedings of International Conference on Computer Vision. Washington D. C.，USA：IEEE Press，2022：9992-10002.

[19] Tan M X，Le Q V. EfficientNetV2：smaller models and faster training[EB/OL]. [2022-02-01]. https://arxiv. org/abs/2104. 00298.

[20] 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.

（下转第296页）

（上接第288页）

[21] FAN H, LIN L T, YANG F, et al. LaSOT: a high-quality benchmark for large-scale single object tracking [C]// Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 5369-5378.

[22] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2014: 740-755.

[23] MÜLLER M, BIBI A, GIANCOLA S, et al. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 310-327.

[24] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam R-CNN: visual tracking by re-detection [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 6577-6587.

[25] DANELLJAN M, VAN GOOL L, TIMOFTE R. Probabilistic regression for visual tracking [C]//Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 7181-7190.

[26] ZHANG Z P, PENG H W, FU J L, et al. Ocean: object-aware anchor-free tracking [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 771-787.

[27] BHAT G, DANELLJAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking [C]//Proceedings of International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2020: 6181-6190.

[28] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: accurate tracking by overlap maximization [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 4655-4664.

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

[30] WU Y, LIM J, YANG M H. Online object tracking: a benchmark [C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2013: 2411-2418.

[31] GALOOGAHI H K, FAGG A, HUANG C, et al. Need for speed: a benchmark for higher frame rate object tracking [C]// Proceedings of International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 1125-1134.

编辑　薛晋栋