

中图法分类号: TP391.4 文献标识码: A 文章编号: 1006-8961(2025)01-0212-13

论文引用格式: Xu H, Dong S H, Zhang J W and Zheng Y H. 2025. Context-aware attention fused Transformer tracking. Journal of Image and Graphics, 30(01):0212-0224(徐晗, 董仕豪, 张家伟, 郑钰辉. 2025. 融合上下文感知注意力的Transformer目标跟踪方法. 中国图象图形学报, 30(01):0212-0224)[DOI:10.11834/jig.240084]

融合上下文感知注意力的Transformer目标跟踪方法

徐晗, 董仕豪, 张家伟, 郑钰辉*

南京信息工程大学计算机学院, 南京 210044

摘要: 近年来, Transformer跟踪器取得突破性的进展, 其中自注意力机制发挥了重要作用。当前, 自注意力机制中独立关联计算易导致权重不明显现象, 限制了跟踪方法性能。为此, 提出了一种融合上下文感知注意力的Transformer目标跟踪方法。**方法** 首先, 引入SwinTransformer(hierarchical vision Transformer using shifted windows)以提取视觉特征, 利用跨尺度策略整合深层与浅层的特征信息, 提高网络对复杂场景中目标表征能力。其次, 构建了基于上下文感知注意力的编解码器, 充分融合模板特征和搜索特征。上下文感知注意力使用嵌套注意计算, 加入分配权重的目标掩码, 可有效抑制由相关性计算不准确导致的噪声。最后, 使用角点预测头估计目标边界框, 通过相似度分数结果更新模板图像。**结果** 在TrackingNet(large-scale object tracking dataset)、LaSOT(large-scale single object tracking)和GOT-10K(generic object tracking benchmark)等多个公开数据集上开展大量测试, 本文方法均取得了优异性能。在GOT-10K上平均重叠率达到73.9%, 在所有对比方法中排在第1位; 在LaSOT上的AUC(area under curve)得分和精准度为0.687/0.749, 与性能第2的ToMP(transforming model prediction for tracking)相比分别提高了1.1%和2.7%; 在TrackingNet上的AUC得分和精准度为0.831/0.807, 较第2名分别高出0.8%和0.3%。**结论** 所提方法利用上下文感知注意力聚焦特征序列中的目标信息, 提高了向量交互的精确性, 可有效应对快速运动、相似物干扰等问题, 提升了跟踪性能。

关键词: 计算机视觉; 目标跟踪; 上下文感知注意力; Transformer; 特征融合

Context-aware attention fused Transformer tracking

Xu Han, Dong Shihao, Zhang Jiawei, Zheng Yuhui*

School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China

Abstract: Objective Visual target tracking, as one of the key tasks in the field of computer vision, is mainly aimed at predicting the size and position of a target in a given video sequence. In recent years, target tracking has been widely used in the fields of autonomous driving, unmanned aerial vehicles (UAVs), military activities, and intelligent surveillance. Although numerous excellent methods have emerged in the field of target tracking, multifaceted challenges remain, including, but not limited to, shape variations, occlusion, motion blurring, and interference from proximate objects. Currently, target tracking methods are categorized into two main groups: correlation-based filtering and deep learning based. The former approximates the target tracking process as a search image signal domain computation. However, fully utilizing image representation information by using manual features is difficult, which greatly limits the performance of tracking methods. In recent years, deep learning has made significant progress in the field of target tracking by virtue of its powerful visual

收稿日期: 2024-02-06; 修回日期: 2024-04-29; 预印本日期: 2024-05-06

* 通信作者: 郑钰辉 zhengyh@vip.126.com

基金项目: 国家自然科学基金项目(U20B2065); 江苏省自然科学基金项目(BK20211539)

Supported by: National Natural Science Foundation of China (U20B2065); Natural Science Foundation of Jiangsu Province, China (BK20211539)

representation processing capabilities. In recent years, Transformer trackers have made breakthroughs, in which the self-attention mechanism plays an important role. Currently, the independent correlation calculation in the self-attention mechanism is prone to lead to the phenomenon of ambiguous weights, thus hampering the tracking method's overall performance. For this reason, a Transformer target tracking method incorporating context-aware attention is proposed. **Method** First, hierarchical vision Transformer using shifted windows (SwinTransformer) is introduced to extract visual features, and a cross-scale strategy is utilized to integrate deep and shallow feature information to improve the network's ability to characterize targets in complex scenes. The cross-scale fusion strategy is used to obtain key information at different scales, capture templates, and search for image diversity texture features, which helps the tracking network better understand the relationship between the target and the background. Second, a context-aware attention-based codec is constructed to fully fuse template features and search features. For the problem of inaccurate correlation computation that occurs in the attention mechanism, nested computation is used for query key pairs to focus on the target information in the input sequence and incorporates a target mask for assigning weights, which can effectively suppress the noise caused by inaccurate correlation computation, seek the consistency of the feature vectors, and prompt better interaction of feature information. The encoder uses features from the output of the trunk as input and uses global contextual information to reinforce the original features, thus enabling the model to learn discriminative features for object localization. The decoder takes as input the target query and the sequence of enhanced features from the encoder, using a two-branch cross attention design. One of the branches computes the target query and the encoder's inputs to attend to features across the full range of locations and search regions on the template. Finally, a corner prediction header is used to estimate the target bounding box, and the template image is updated by the similarity score results. Specifically, the decoded features are fed into a fully convolutional network that outputs two probability maps for the upper-left and lower-right corners of the target bounding box. The predicted box coordinates are then obtained by calculating the expectation of the probability distribution for the two corners. **Result** Training pairs are randomly selected from the common objects in context (COCO), a large-scale object tracking dataset (TrackingNet), large-scale single-object tracking (LaSOT), and generic object tracking benchmark (GOT-10k) datasets to train the tracking model in this paper. The model's minimum training data unit is a triad consisting of two templates and a search image. It was trained using 500 epochs, using 6×10^4 triples per epoch. The backbone network and the remainder have initial learning rates are 10^{-5} and 10^{-4} . After 400 training sessions, the learning rate decreased by a factor of 10. Extensive testing was conducted on TrackingNet; LaSOT; GOT-10K; online object tracking as a benchmark (OTB100); a benchmark and simulator for UAV tracking (UAV123); and the publicly available need for speed (NfS) dataset, and the results were compared with those of several current state-of-the-art tracking methods, all of which achieved excellent performance. On GOT-10K, the average overlap rate reaches 73.9%, $SR_{0.5}$ and $SR_{0.75}$ reach 84.6% and 69.8%, respectively. $SR_{0.5}$ and $SR_{0.75}$ are the success rate of an overlapping coverage area greater than 0.5 and greater than 0.75. On LaSOT, the area under curve is 68.7%, and the precision rate and normalized precision rate are 78.7% and 74.3%, respectively. On TrackingNet, the success rate is 68.7%, the normalized precision rate is 87.7%, and the accuracy rate is 80.7%. The success rates on the NfS, OTB100, and UAV123 datasets are 68.1%, 69.6%, and 68.3%, respectively. The experimental results prove that the proposed method has good generalization ability. The effectiveness of the proposed method is verified by conducting ablation experiments on the GOT-10K, LaSOT and TrackingNet datasets to validate the effect of different modules on the tracking method. With the use of three feature extraction networks—ResNet-50, SwinTrack-Base, and the cross-scale fusion SwinTransformer—two scenarios were compared between the inclusion of the context-aware attention module and the exclusion of the module. The comparison of the final results shows that the inclusion of the context-aware attention module in SwinTransformer effectively improves the tracking performance. **Conclusion** The proposed method utilizes context-aware attention to focus the target information in the feature sequence, which improves the accuracy of vector interaction. The method effectively copes with the problems of fast motion and similar object interference, and improves the tracking performance. However, the proposed method uses Transformer in both the feature extraction and fusion stages, which leads to a large number of parameters and requires more training time, resulting in low computational efficiency. In the future, the two stages can be merged to integrate feature extraction and fusion.

Key words: computer vision; object tracking; context-aware attention; Transformer; feature fusion

0 引言

目标跟踪作为计算机视觉领域关键任务之一,其主要目的为预测给定视频序列中目标的尺寸和位置。近年来,目标跟踪广泛应用于自动驾驶、无人机、军事活动和智能监控等领域(李玺等,2019)。尽管目标跟踪领域已涌现许多出色的方法,但仍然面临多方面挑战,如形状变化、遮挡、运动模糊以及近似物体干扰等问题。因此,设计一种快速且具备鲁棒性的跟踪方法仍然具有挑战性。

当前,目标跟踪方法分为基于相关滤波和基于深度学习两大类(孟绿和李诚新,2019),前者将目标跟踪过程近似地看做对搜索图像信号域的计算(王皎尧等,2018)。然而,使用手工特征难以充分利用图像表征信息,极大地限制了跟踪方法的性能。近年来,深度学习则凭借其强大的视觉表征处理能力,在目标跟踪领域取得了显著进展。Nam和Han(2016)提出了多域网络跟踪方法(multi-domain convolutional neural networks, MDNet),利用轻量级的卷积网络学习目标特征,然后训练二分类网络实现在线跟踪任务。MDNet虽然取得优异的跟踪性能,但运行速度仅为1帧/s,无法实时跟踪。Bertinetto等人(2016)提出了全卷积孪生网络跟踪方法(fully-convolutional siamese network, SiamFC),利用孪生结构将跟踪问题定义为模板区域与搜索区域之间的匹配重叠度,该方法使用AlexNet(Krizhevsky等,2017)作为特征提取网络,提升了跟踪效率。此后基于全卷积孪生网络的跟踪方法(Li等,2018,2019)在视觉目标跟踪任务中取得显著进展。然而,这些方法通常采用关联函数计算相似度,但线性相互作用存在局限性,限制了跟踪方法进一步发展。

随着自然语言处理领域的Transformer在计算机视觉任务中大放异彩(Dosovitskiy等,2021),Chen等人(2021)率先将Transformer引入目标跟踪领域,提出了基于Transformer的跟踪方法(Transormer tracking, TransT)。TransT设计了一种基于注意力机制的特征融合模块,利用自注意力和交叉注意力增强上下文语义。针对复杂跟踪场景,Yan等人(2021)提出了时空Transformer跟踪方法,使用Transformer对目标对象和搜索区域之间的全局时空特征依赖关系建模,去掉孪生跟踪器中交叉融合,保

持了与目标检测相似的跟踪框架。Song等人(2022)提出了多尺度循环移位窗口Transformer跟踪方法,将提取的特征分割成小窗口并平面化,然后馈送到Transformer以捕获窗口注意力,其在窗口注意模块中使用了循环移位技术,融合了不同尺度的输出。

上述基于Transformer的方法通常使用卷积神经网络(convolutional neural network, CNN)提取特征,利用注意力机制融合目标模板和搜索区域的特征向量,但局部卷积核提取的特征缺乏全局信息。因此Lin等人(2022)提出了一个两分支全注意力机制架构,利用Transformer代替卷积神经网络提取特征信息。进一步地,Cui等人(2022)提出了一段单分支的跟踪框架,利用预训练卷积视觉Transformer网络(convolutions to vision Transformer, CVTTransformer)(Wu等,2021)提取特征并设计了一种混合注意力模块,该模块在同一阶段完成了对特征信息的提取与整合,并通过区域令牌双重注意机制捕获模板区域和搜索区域的目标线索。为提取更具有判别性的特征,Gao等人(2023)提出了自适应令牌划分建模跟踪方法,利用注意力掩蔽策略,缓解了复杂背景中其他物体对目标的干扰问题,使单分支跟踪方法具有更好的灵活性。上述方法充分利用了Transformer的全局特征表示能力,使模型能够捕捉特征信息中的长距离依赖关系。受此启发,本文提出了一种融合上下文感知注意力的Transformer目标跟踪方法。首先使用SwinTransformer(Liu等,2021)提取模板和搜索区域的特征,然后将视觉特征输入到基于上下文感知注意力的编解码器中融合信息和交互学习。最后,使用角点预测头估计目标边界框,并通过相似度分数对比结果更新动态模板。

综上所述,本文主要贡献如下:1)设计了一种上下文感知注意力模块。针对注意力机制相关性计算不准确的问题,对查询键对使用嵌套计算,聚焦输入序列中目标信息,提高特征向量交互的精确性。2)提出一种跨尺度融合的SwinTransformer特征提取网络,通过跨尺度融合策略获取不同尺度的关键信息,捕获模板和搜索图像多样性纹理特征,促使跟踪网络更好地理解目标与背景的关系。3)使用TrackingNet(large-scale object tracking dataset)、LaSOT(large-scale single object tracking)和GOT-10K(generic object tracking benchmark)等多个公开数据集测试了所提方法。在GOT-10K上平均重叠率达到73.9%,

在 LaSOT 和 TrackingNet 上的 AUC 得分为 0.687 和 0.831, 均取得优异性能。此外, 通过消融实验证明了各个模块的有效性。

1 相关工作

Transformer 最初由 Vaswani 等人(2017)提出并应用于自然语言处理任务, 因其优越的全局依赖性和并行计算能力, 被引入图像超分辨率(熊巍等, 2023)、目标检测(Carion 等, 2020)和语义分割(Strudel 等, 2021)等计算机视觉任务中。Transformer 体系结构核心为注意力机制, 主要担负动态信息聚合的角色。该机制通过编码器和解码器组件的堆叠实现, 编解码器层由多个相同的编解码器组成。在 Transformer 结构中, 每个编码器包含两个子层: 自注意力层和全连接前馈层。在解码器中, 除这两个子层外, 中间还包含编码-解码注意力层。输入数据通过位置嵌入算法添加位置信息, 然后经过自注意力层处理捕获的上下文关系。在解码器中, 编解码器注意力子层会聚焦输入信息, 捕捉输入序列中相关特征。

目标跟踪任务中, Wang 等人(2021)将判别相关滤波器应用到编码特征并生成响应映射, 利用该映射在搜索区域中定位目标, 提出了 Transformer 特征

判别跟踪方法。虽然 Transformer 的自注意力机制能够捕获远程依赖关系, 但无法充分捕获特定目标线索, 容易受到杂乱背景的干扰。为此, Fu 等人(2022)提出了基于稀疏 Transformer 的孪生跟踪方法, 通过使用备用注意力模块提供搜索区域中特定目标线索。另外, Xie 等人(2021)使用一组局部注意力块捕获小窗口内注意力, 然后将特征馈送到全局注意力块以捕获长期依赖关系。由于局部注意力在高分辨率特征图上计算标记的小窗口注意力, 而全局注意力在低分辨率特征图上计算同一图像所有标记之间的注意力, 减少了计算复杂度, 在实现高精度的同时保证了实时跟踪, 故本文采用 Transformer 编解码器的结构融合特征信息, 设计了一种上下文感知注意力并应用在编解码器结构中, 提高特征向量权重计算的正确率, 实现鲁棒的跟踪性能。

2 融合上下文感知注意力的 Transformer 跟踪器

2.1 跟踪器框架

所提方法的网络结构如图 1 所示, 主要分为 3 部分: 1) 特征提取网络; 2) 编解码器层; 3) 预测头。首先, 该网络使用模板图像和搜索图像作为输入, 采用基于跨尺度融合的 SwinTransformer 提取视觉特征。

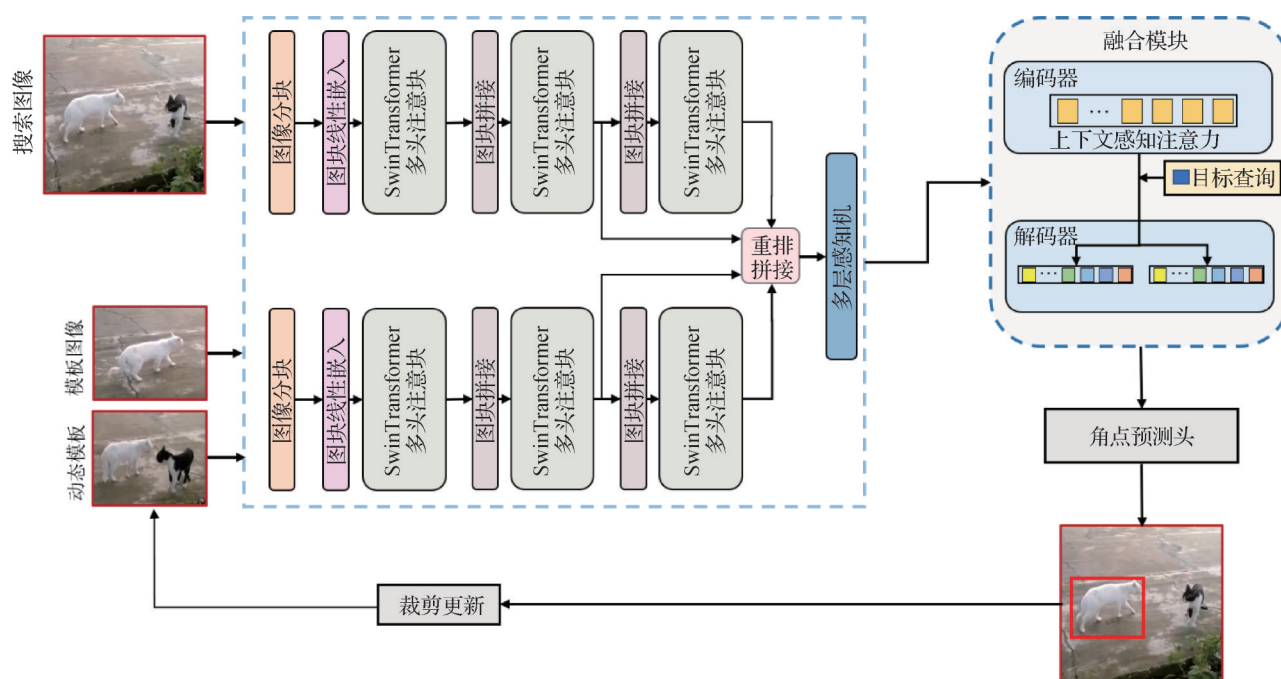


图1 跟踪器总体框架

Fig. 1 Structure of the tracker

接着,将提取的特征输入到带有上下文感知注意力的编解码器中,以促进搜索与模板特征之间的交互学习。其中,Transformer解码器使用目标查询和编码器的输出作为输入,而后经过两分支注意力计算得到融合后的特征向量。最后,预测头通过计算角点概率分布的期望值得到预测框坐标,并绘制出目标位置;同时,通过相似度分数对比结果更新模板图像。

2.2 特征提取网络

给定一个视频帧 T ,输入一对图像:初始目标对象 $z \in \mathbf{R}^{3 \times H_z \times W_z}$ 的模板图像和当前帧 $x \in \mathbf{R}^{3 \times H_x \times W_x}$ 的搜索图像,其中, H 、 W 为高、宽,然后通过共享骨干网络嵌入图像特征。本文选用SwinTransformer-tiny作为基本骨干网络,因为它在性能和效率之间取得了很好的平衡。具体而言,移除SwinTransformer-tiny第4块,获得8倍有效步幅,为区别于SwinTransformer-tiny,本文称之为SwinTrans。SwinTrans主干网络分为3层Trans以及1层多层感知神经网络(multi-layer perceptron, MLP)。首先将图像大小都调整为 $3 \times H \times W$,并相继输入到Trans-1块、Trans-2块、Trans-3块中,通过跨尺度特征融合得到自适应融合纹理特征序列,然后再输入到MLP中。MLP预处理会将通道数从 C 降低到 D ,并将特征映射沿空间维度展开并拼接,从而产生一个长度为 $\frac{H_z}{s} \frac{W_z}{s} + \frac{H_x}{s} \frac{W_x}{s}$ 、维度

为 D 的特征序列, z 为模板区域, x 为搜索区域, s 为通道层数,计算为

$$F_M = \text{MLP}[F_2, \text{Rerrange}(F_3)] \quad (1)$$

式中, $\text{Rerrange}(\cdot)$ 表示提高特征图的分辨率, F_2 、 F_3 是Trans-2块、Trans-3块输出的特征维度序列,表示为 $F_2 \in \mathbf{R}^{2C \times \frac{H}{16} \times \frac{W}{16}}$, $F_3 \in \mathbf{R}^{4C \times \frac{H}{16} \times \frac{W}{16}}$, F_M 为拼接后的输出。

2.3 上下文感知注意力

Transformer的注意力机制使用一个查询和一组键值对作为输入,经过加权计算得到输出。这些值的权重通过查询和相应键之间缩放后点积的归一化指数函数(softmax)计算。Transformer传统注意力机制可以表示为

$$\text{Attn}(Q, K, V) = (\text{softmax}(\frac{\bar{Q}\bar{K}^T}{\sqrt{D}})\bar{V})W \quad (2)$$

式中,分别用 Q 、 K 和 V 表示查询、键和值, $\bar{Q}$ 、 $\bar{K}$ 和 $\bar{V}$ 是通过与权重相乘得到的新向量, W_q 、 W_k 、 W_v 分别为它们的线性转换权重, W 为动态调整权重。在上述Transformer注意力机制中,每个查询、键、值对的相关性计算都是单独开展,这个过程中会忽视其他查询、键、值对的相关性。在视觉特征表达不精确的情形下,这种计算过程会引入错误的相关性,这些错误的相关性会导致后续特征融合以及交叉注意中信息交互不准确,进而导致跟踪失败。故本文提出一个上下文感知注意力,如图2所示。

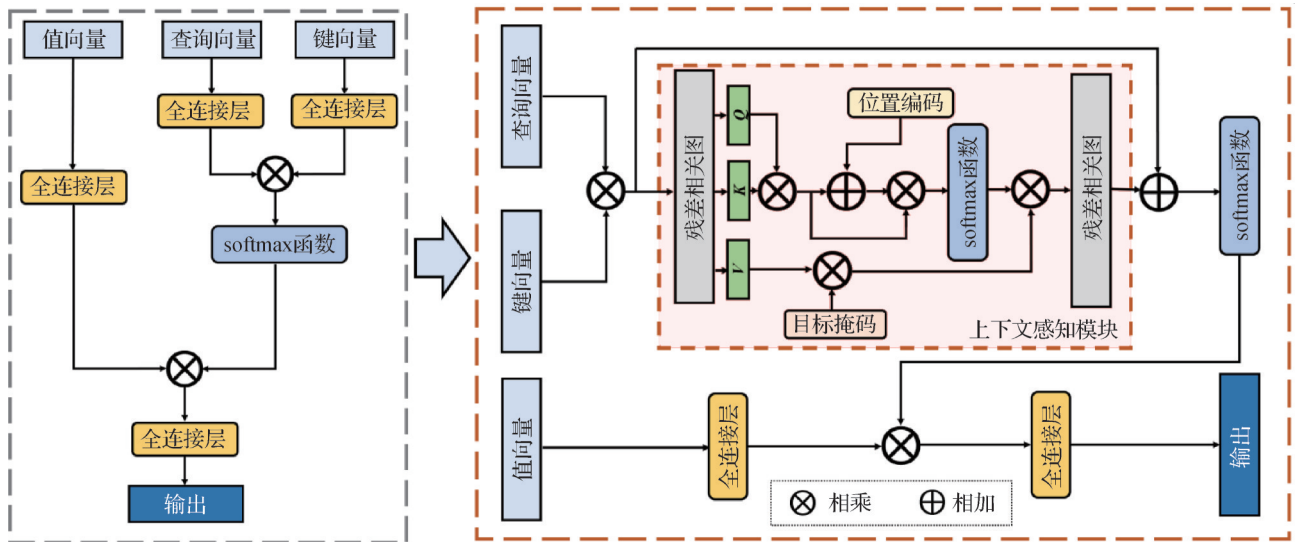


图2 上下文感知注意力

Fig. 2 Context-aware attention

在注意力机制中,若一个键与查询具有高度相关性,则其周围一些键往往也具有相对较高的相关性。基于此,利用相关图中关联之间的信息线索可寻求每个键周围关联的一致性。鉴于模板图像是已知的,为提升模型对目标信息的关注度,使用目标掩码(object mask, OM)为模板特征向量不同区域分配不同的权重,该掩码遵循高斯形状的空间分布,具体为

$$OM(y) = \left(\exp \frac{\|y - e\|^2}{2\sigma^2} \right) \quad (3)$$

式中, e 是目标真实位置, y 为平均值, σ 为超参数,设置为0.05。拉伸后的掩码向量只含模板信息,故只与序列中的模板向量执行点乘操作,搜索区域特征向量不变。通过对查询与键建立新的关联,使用目标掩码增强正确权重,抑制与查询不相关键对的错误关联。接着将 M 中的列作为一系列的相关向量,生成新的查询、键和值,并输出残差相关图,具体为

$$InAttn(M) = \left(\text{softmax} \left(\frac{Q'K'^T}{\sqrt{D}} \right) V' \otimes OM \right) (1 + W_i') \quad (4)$$

式中, M 为残差相关图, Q' 、 K' 和 V' 为新的查询、键和值, K'^T 为其转置矩阵, W_i' 为调整聚合相关性和相同连接的线性变换权重, OM 为目标掩码, $\otimes$ 表示点乘。具体而言,在softmax操作之前引入另一个注意力块细化相关图 M ,并对权重增加目标掩码,新引入的注意力模块称为上下文感知模块,如图2浅色区

域标注所示。

首先,采用线性变换将 Q' 和 K' 的尺寸减小到 $HW \times D(D \ll HW)$ 。在归一化操作后,添加二维正弦编码以提供位置线索。上下文感知注意力会通过新的查询、键、值对生成残差相关图。本质上,对于相关映射 M 中的每个相关向量,都会对原始相关向量聚合生成残差相关向量。通过这种嵌套计算,上下文感知注意力可以表示为

$$EnAttn(M) = (\text{softmax}(M + InAttn(M))V)W \quad (5)$$

上下文感知注意力则是对查询和键先进行一次softmax计算,并在计算时对权重向量加入了目标掩码。相比于传统的Transformer自注意力机制,上下文感知注意力考虑了目标周围环境和序列的整体语境,利用图像中其他区域信息来帮助模型更好地理解当目标的含义,提高了含有正确目标信息的权重。

2.4 编解码器结构

编解码器的结构如图3所示。将主干输出的特征序列 f_i 作为编码器的输入,编码器层由 N 个编码器组成, N 设置为6。每层编码器由1个带有前馈网络的多头自注意模块组成。由于原始编码器的排列不变性,在输入序列中加入了正弦位置嵌入,这使得编码器捕获序列中所有元素之间的特征依赖关系,并使用全局上下文信息强化原始特征,从而使模型能够学习用于对象定位的判别特征。解码器将目标查询和来自编码器的增强特征序列作为输入。解码器层使用单个解码器,并输入一个目标查询预测目标对象的边界框。与原有Transformer解码器不

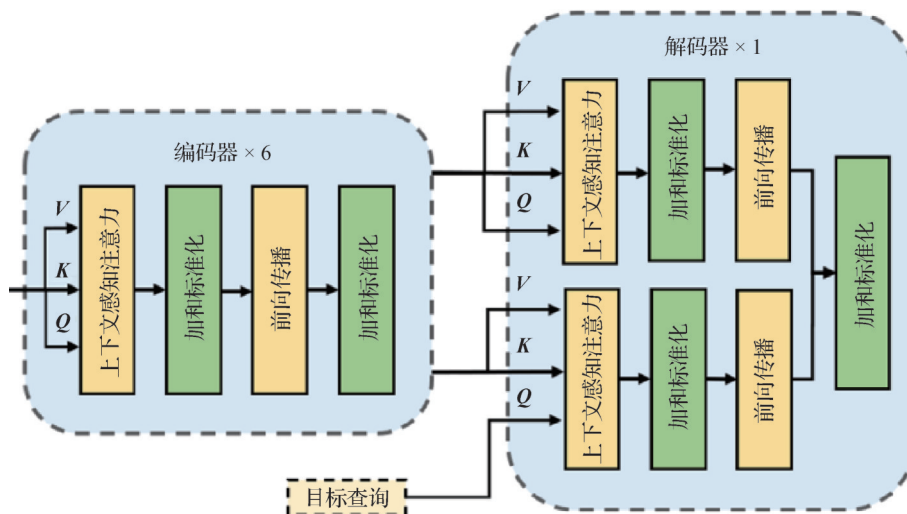


图3 编解码器结构

Fig. 3 Structure of encoder and decoder

同,为了简单起见,删除自注意力块,并引入如图3所示的两分支交叉注意设计。其中一个分支计算目标查询和编码器输出,用以关注模板上全部位置和搜索区域的特征。两个分支的交叉关注块具有相同结构,遵循Transformer中的查询键值设计。最终聚合两个分支,得到更准确的融合效果。此外编解码器都使用了上下文感知注意力模块,可以更好地传播特征语义信息。

2.5 预测头

本文使用一种估计角点概率分布的预测头,如图4所示。其将解码过的特征馈送到一个全卷积网络中,输出概率图并计算出坐标,计算为

$$(\hat{x}_{tl}, \hat{y}_{tl}) = \left(\sum_{x=0}^H \sum_{y=0}^W x \times P_{tl}(x, y), \sum_{x=0}^H \sum_{y=0}^W y \times P_{tl}(x, y) \right) \quad (6)$$

$$(\hat{x}_{br}, \hat{y}_{br}) = \left(\sum_{y=0}^H \sum_{x=0}^W x \times P_{br}(x, y), \sum_{y=0}^H \sum_{x=0}^W y \times P_{br}(x, y) \right) \quad (7)$$

式中, $P_{tl}(x, y)$ 和 $P_{br}(x, y)$ 为左上角和右下角概率图, $(\hat{x}_{tl}, \hat{y}_{tl})$ 和 $(\hat{x}_{br}, \hat{y}_{br})$ 为预测框坐标。

首先从编码器的输出序列中提取搜索区域特征,然后计算搜索区域特征与解码器输出嵌入之间的相似度。接着,通过对相似性分数与搜索区域特征元素相乘,增强重要区域的影响并抑制区分度较低的区域。将新特征序列重构为特征映射,表示为 $f_r \in \mathbf{R}^{D \times \frac{H}{s} \times \frac{W}{s}}$, 然后将其输入到全卷积网络中。预测头显式地模拟坐标估计中的不确定性,为目标跟踪产生了更鲁棒的预测结果。

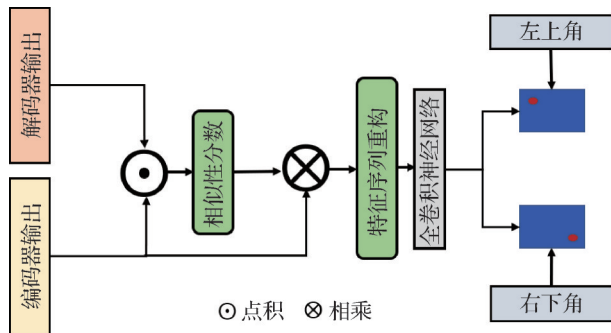


图4 预测头

Fig. 4 Prediction head

2.6 训练以及损失函数

整个网络都是端到端开展训练,结合真实框重叠交并比(intersection-over-union, IoU) L_{IoU} 损失以及 L_1 损失(Rezatofighi等,2019),计算为

$$L = \lambda_{IoU} L_{IoU}(k_n, \hat{k}_n) + \lambda_{L_1} L_1(k_n, \hat{k}_n) \quad (8)$$

式中, k_n 和 $\hat{k}_n$ 分别表示真实值和预测值, λ_{IoU} 和 λ_{L_1} 为训练时需要调整的超参数。

在训练阶段,所有搜索区域图像都包含目标物体,以促使模型能学习定位能力。在训练过程中,首先在第1帧初始化两个模板对应的特征,然后将搜索区域裁剪输入到网络中,生成边界框和置信度评分。跟踪过程中,当目标被完全遮挡、移出视野时或跟踪器漂移时,裁剪出的模板是不可靠的。为判断当前状态的可靠性,采用一个3层感知器的分数预测头,只有搜索区域包含目标时动态模板才可更新。通过训练,模型可学到目标定位和分类能力。

3 实验与分析

3.1 实验环境与参数设置

实验使用的操作系统为Ubuntu20.04.1 64位,系统硬件设施为Intel(R) Xeon(R) Silver 4314 CPU @ 2.40 GHz、NVIDIA RTX 4090 × 2 (49128 MiB)。编译环境为Python3.8和torch-1.7.1。从搜索框中裁剪出目标框面积 5^2 倍的搜索图像,并将其尺寸调整为 320×320 像素,模板图像尺寸为 180×180 像素,是目标框面积的 2^2 倍。Transformer由6个编码器和1个解码器组成,包含注意层和前馈网络。跟踪器中的多头注意力块有4个头,通道宽度为256。上下文感知注意力模块将查询键对的通道维度减少到64。前馈网络块有2048个隐藏单元,使用权重衰减自适应优化器(Loshchilov和Hutter,2019)对网络优化,权重衰减为0.0001。

为训练跟踪模型,从COCO(common objects in context)(Lin等,2014)、TrackingNet(Müller等,2018)、LaSOT(Fan等,2019)和GOT-10K(Huang等,2021)数据集中随机抽取训练对。模型最小训练数据单元是一个三元组,由两个模板和一个搜索图像组成。使用500个批次训练,每个批次使用 6×104 个三元组。初始学习率分别为0.00001和0.0001,在400次训练后,学习率下降了10倍。

实验中采用了多个评估指标:成功率曲线下的面积(area under curve, AUC)、准确率(precision rate, PR)、归一化精度率(normalized precision rate, NPR)、成功率(success rate, SR)和平均重叠率(average

overlaps, AO)。准确率是测量中心在预测点和地面真值中心之间的精确程度。归一化精度率额外考虑了地面真值盒的尺度大小,对精度率归一化处理。平均重叠率计算所有真实边界框和预测边界框之间的平均重叠,SR_{0.5}和SR_{0.75}分别为重叠覆盖面积占比大于0.5和大于0.75的成功率。

3.2 对比实验

为评估所提方法的有效性,在3个大规模数据集 TrackingNet (Müller 等, 2018)、LaSOT (Fan 等, 2019)和 GOT-10K (Huang 等, 2021)以及3个常用小规模数据集 NfS (need for speed) (Galoogahi 等, 2017)、OTB100 (online object tracking as a benchmark) (Wu 等, 2015)和 UAV123 (a benchmark and simulator for UAV tracking) (Mueller 等, 2016)上对比了多种主流跟踪方法。

3.2.1 小规模数据集评估结果

在 NfS、OTB100 和 UAV123 数据集上,将所提方法与 SiamFC++ (fully-convolutional siamese network++) (Xu 等, 2020)、SiamRPN++ (siamese region proposal network++) (Li 等, 2019)、TransT (Chen 等, 2021)、STARK-ST50 (spatial-temporal Transformer) (Yan 等, 2021)、KeepTrack (target candidate association) (Mayer 等, 2021)和 TAT (target-aware Transformer) (Zou 等, 2023)等7种主流跟踪方法对比,结果如表1所示。NfS数据集包含100个物体快速移动的视频,在其常用版本 NfS30 上评估了本文方法。本文方法的 AUC 为 68.1%,比 STARK 提高了 2.9%,在所有对比方法中表现最好。OTB 是评估视觉跟踪方法的开创性基准,然而近年来,这一基准已经高度饱和。尽管如此,实验结果表明,本文方法可以达到与最先进方法相当的性能。UAV123 数据集包括从低空无人机视角捕获的123个视频序列,本文方法的 AUC 达到 68.3%。实验结果表明,本跟踪器适用于无人机视频跟踪场景。

3.2.2 LaSOT数据集评估结果

LaSOT 是一个密集标注的大规模数据集,包含1400个长期视频序列,70个类别。测试集包含280个视频,平均长度2448帧。在 LaSOT 数据集上,将本文方法与 SiamFC、SiamRPN++、STARK-ST50、TransT、TrDiMP (Transformer discriminative model prediction) (Wang 等, 2021)、ToMP (transforming model prediction for tracking) (Mayer 等, 2022)、ATOM

(accurate tracking by overlap maximization) (Danelljan 等, 2019)和 CMAT (convolution mixer and self-attention tracking) (Wang 等, 2024)进行比较,结果如表2所示。

表 1 NfS、OTB100和UAV123数据集的结果比较
Table 1 Comparison of results on NfS, OTB100 and UAV123

方法	AUC		
	NfS	OTB100	UAV123
SiamFC++	49.7	68.3	62.3
SiamRPN++	50.2	69.6	61.3
TransT	65.7	69.4	68.0
STARK-ST50	65.2	68.2	69.1
KeepTrack	66.4	70.9	69.7
TAT	66.7	69.4	67.5
本文	68.1	69.6	68.3

注:加粗字体表示各列最优结果。

表 2 LaSOT数据集结果比较
Table 2 Comparison of results on LaSOT

方法	AUC		
	AUC	NPR	PR
SiamFC	33.6	42.0	33.9
SiamRPN++	49.6	56.9	49.1
ATOM	51.4	57.7	50.5
TrDiMP	59.9	69.0	60.8
TansT	64.9	73.8	69.0
STARK-ST50	66.4	76.3	71.2
ToMP	67.6	78.0	72.2
CMAT	64.1	72.5	65.4
本文	68.7	78.7	74.9

注:加粗字体表示各列最优结果。

从表2可以看出,本文方法在多个评价指标中都取得了最优性能。本文方法的 AUC、NPR 和 PR 分别达到了 68.7%、78.7% 和 74.9%。TransT 和 STARK-ST50 两种方法均使用卷积网络提取特征,并利用Transformer融合交互特征。与二者相比,本文方法的 AUC 指标分别提高了 3.8% 和 2.3%。CMAT 构建了一个混合卷积神经网络和自关注块的

特征提取网络,ToMP是在编码器中嵌入了目标定位和目标扩展,与二者相比,本文方法的AUC指标分别提高了4.6%和1.1%。

3.2.3 TrackingNet数据集评估结果

TrackingNet是一个大规模的短期跟踪数据集,包含超过30 000个训练视频,提供了511个测试视频序列。如表3所示,本文方法在该数据集上实现了优异性能,AUC、NPR和PR分别为83.1%、87.7%和80.7%,充分表明其在长期跟踪场景中具有巨大优势。本文方法优于最近基于Transformer的跟踪器,例如TransT(81.4%、86.7%和80.3%)。TAT使用ResNet50(residual network)(He等,2016)提取特征,本文方法的AUC得分比TAT(Zou等,2023)提高了1.6%。STMP(smaller swin-Transformer model tracking)(Pan等,2023)设计了一个基于SwinTransformer的轻量型的特征提取网络,本文方法的AUC得分比其高0.8%。显然,本文所设计的跨尺度融合SwinTransformer能够捕捉更丰富的特征表示。

表3 TrackingNet数据集结果比较
Table 3 Comparison of results on TrackingNet dataset
/%

方法	AUC	NPR	PR
SiamFC++	75.4	—	80.0
SiamRPN++	73.3	—	80.0
TrDiMP	78.4	83.3	73.1
TransT	81.4	86.7	80.3
STARK-ST50	82.0	86.9	79.1
TAT	81.5	86.7	79.0
STMP	82.3	87.1	80.4
本文	83.1	87.7	80.7

注:加粗字体表示各列最优结果,“—”表示无对应结果。

3.2.4 GOT-10K数据集评估结果

GOT-10K是最近发布的大规模短期视觉跟踪数据集。为了保证训练和测试之间对象类的零重叠,遵循GOT-10K一次性协议,只使用指定子集训练模型。实验结果由官方线上服务器评估,如表4所示,本文方法在所有指标上都有较大提高,AO、SR_{0.5}和SR_{0.75}分别达到73.9%、84.6%和69.8%。与SwinTrack(Swin-Transformer-tracker)(Lin等,2022)、OSTrack(one-stream tracking)(Ye等,2022)和

FFTrack(Transformer tracking via frequency fusion)(Hu等,2024)相比,平均重叠率分别提高了4.5%、0.2%和0.7%。以上结果表明本文方法对不可见类对象也有很好的泛化能力,能够适应不同的环境和场景。

表4 GOT-10K数据集结果比较
Table 4 Comparison of results on GOT-10K dataset
/%

方法	AO	SR _{0.5}	SR _{0.75}
SiamFC++	59.5	69.5	47.9
SiamRPN++	51.7	61.5	32.5
TransT	67.1	76.8	60.9
STARK-ST50	68.8	78.1	64.1
SwinTrack	69.4	78	64.3
OSTrack	73.7	83.2	70.2
FFTrack	73.2	83.1	70.3
本文	73.9	84.6	69.8

注:加粗字体表示各列最优结果。

3.3 消融实验

为验证本文方法的有效性,在GOT-10K、LaSOT和TrackingNet数据集上开展了消融实验,验证不同模块对跟踪方法的影响,如表5所示。消融实验主要从特征提取和特征融合两个方面开展。一方面,本文研究了不同特征提取网络对跟踪方法的影响,选取3种特征提取网络,即ResNet-50、SwinTrack-Base(Lin等,2022)和SwinTrans;另一方面,对比了特征融合阶段的两种融合策略,一种是使用传统注意力,另一种是使用上下文感知注意力。其中,En-

表5 消融实验结果
Table 5 Results of ablation experiments
/%

特征提取网络	特征融合策略	GOT-10K	LaSOT	TrackingNet
		AO	AUC	AUC
ResNet-50	Attn	68.7	67.2	81.8
ResNet-50	En-Attn	69.6	68.2	82.7
SwinTrack-Base	Attn	70.1	67.5	82.1
SwinTrack-Base	En-Attn	72.5	67.3	82.5
SwinTrans	Attn	70.2	67.5	82.3
SwinTrans	En-Attn	73.9	68.7	83.1

注:加粗字体表示各列最优结果。

Attn 代表上下文感知注意力, Attn 代表使用传统Transformer注意力。

从表5可知,使用上下文感知注意力的方法要优于传统注意力的方法。同样地,与ResNet50相比,SwinTrack-Base以及SwinTrans作为特征提取网络可以达到更出色的跟踪结果,这表明相同条件下Transformer比卷积神经网络提取的特征更为优秀。接着对比了SwinTrans和SwinTrack-Base两个特征提取网络,从结果可知使用跨尺度策略的SwinTrans可以表现出更好的性能,充分体现了跨尺度策略的优越性。从表5中全部结果对比可以看出,使用SwinTrans提取特征,通过上下文感知注意力融合信息能有效提高跟踪性能。

3.4 定性分析

为了直观地感受所提跟踪方法的实际跟踪效

果,本文展示了4组具有代表性的视频帧。如图5所示,对比了其他5种主流跟踪方法(ToMP、TransT、STARK-ST50、SiamRPN++和SiamFC)的可视化跟踪结果。如第1组所示,在目标经历遮挡后,本文方法仍可准确跟踪。第3组中目标快速移动过程被遮挡,本文方法丢失了目标,但是在后续视频帧中目标被重新定位。这主要归功于跟踪方法的更新模板准确修正了错误模板图像。从第4组可以看到,当两个类似物体出现复杂运动时,其他跟踪方法均丢失了正确目标,转而跟踪与目标相似的其他物体。相比之下,本文精确地跟踪了目标,这主要因为本文使用了上下文感知注意力,使得跟踪方法对于相似物体干扰具有鲁棒性。简而言之,本文方法在面对相似物干扰、尺度变化、快速运动以及目标遮挡等情况时具有更强的竞争力。

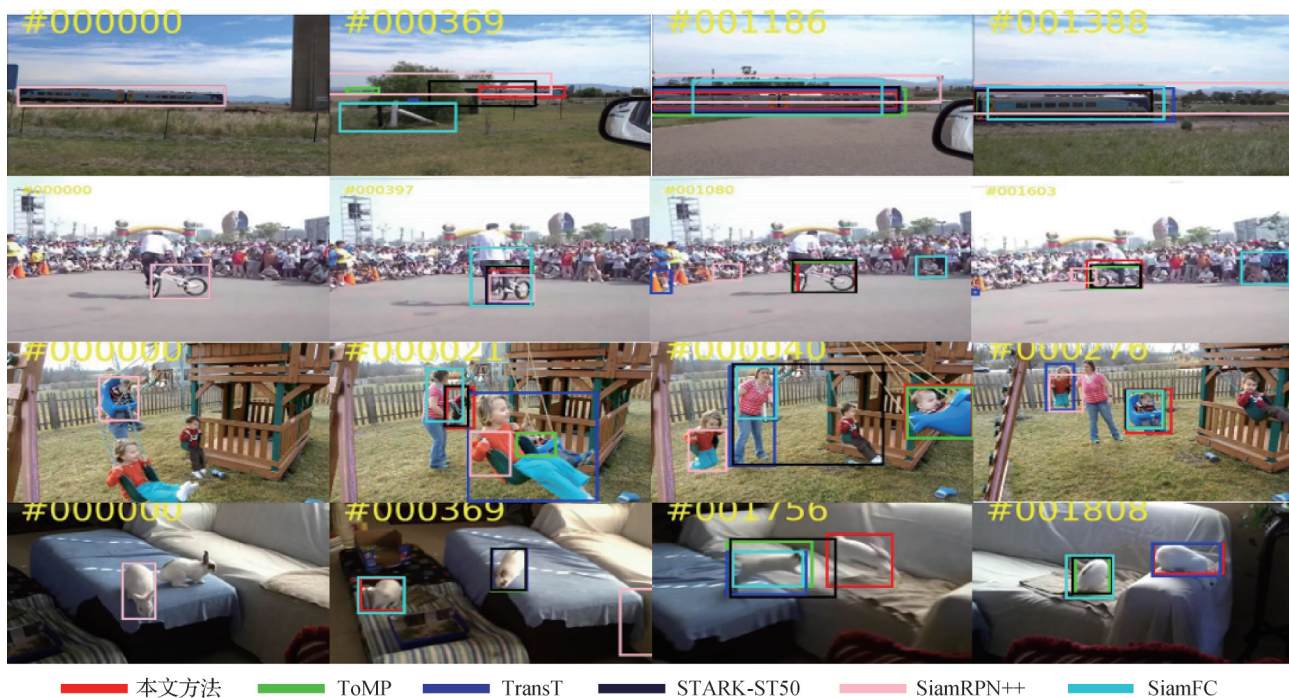


图5 可视化跟踪结果

Fig. 5 Visualization of tracking results

4 结 论

针对目前自注意力机制中相关性计算不精确的问题,本文提出了融合上下文感知注意力的Transformer目标跟踪方法。本文使用SwinTransformer提取特征,通过跨尺度策略融合不同尺度的特征信息,提高了网络对目标与背景的判别能力。在特征交互

阶段,本文设计了一种新颖的上下文感知注意力,通过寻求特征向量的一致性,促使特征信息更好地交互,极大提高了相关性计算的精确率。在LaSOT、TrackingNet、GOT-10K等多个数据集上的实验结果表明,本文方法优于大部分跟踪器,能有效应对快速运动、相似物干扰挑战等问题。并通过消融实验和可视化跟踪结果证明了各模块的有效性。

本文方法虽然取得了更先进的跟踪性能,但在

特征提取与融合两个阶段均采用了Transformer, 导致参数冗余, 无法达到更优的跟踪速度。同时Transformer更关注全局信息, 所以在小目标跟踪场景下, 容易产生目标丢失的现象。后续研究将整合提取和融合两个阶段, 优化模型框架, 降低参数量, 并增加有效的检测机制以应对更复杂的跟踪场景。

参考文献(References)

- Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]
- Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with Transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]
- Chen X, Yan B, Zhu J W, Wang D, Yang X Y and Lu H C. 2021. Transformer tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8122-8131 [DOI: 10.1109/CVPR46437.2021.00803]
- Cui Y T, Jiang C, Wang L M and Wu G S. 2022. MixFormer: end-to-end tracking with iterative mixed attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13598-13608 [DOI: 10.1109/CVPR52688.2022.01324]
- Danelljan M, Bhat G, Khan F S and Felsberg M. 2019. ATOM: accurate tracking by overlap maximization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4655-4664 [DOI: 10.1109/CVPR.2019.00479]
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Housheer N. 2021. An image is worth 16×16 words: Transformers image recognition at scale [EB/OL]. [2024-01-23]. <https://arxiv.org/pdf/2010.11929.pdf>
- Fan H, Lin L T, Yang F, Chu P, Deng G, Yu S J, Bai H X, Xu Y, Liao C Y and Ling H B. 2019. LaSOT: a high-quality benchmark for large-scale single object tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5369-5378 [DOI: 10.1109/CVPR.2019.00552]
- Fu Z H, Fu Z H, Liu Q J, Cai W R and Wang Y H. 2022. SparseTT: visual tracking with sparse Transformers//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI: 905-912 [DOI: 10.24963/ijcai.2022/127]
- Galoogahi H K, Fagg A, Huang C, Ramanan D and Lucey S. 2017. Need for speed: a benchmark for higher frame rate object tracking//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1134-1143 [DOI: 10.1109/ICCV.2017.128]
- Gao S Y, Zhou C L and Zhang J. 2023. Generalized relation modeling for Transformer tracking//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 18686-18695 [DOI: 10.1109/CVPR52729.2023.01792]
- He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]
- Hu X T, Zhong B N, Liang Q H, Zhang S P, Li N, Li X X and Ji R R. 2024. Transformer tracking via frequency fusion. IEEE Transactions on Circuits and Systems for Video Technology, 34(2): 1020-1031 [DOI: 10.1109/TCSVT.2023.3289624]
- Huang L H, Zhao X and Huang K Q. 2021. GOT-10K: a large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5): 1562-1577 [DOI: 10.1109/TPAMI.2019.2957464]
- Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]
- Li B, Wu W, Wang Q, Zhang F Y, Xing J L and Yan J J. 2019. SiamRPN++: evolution of Siamese visual tracking with very deep networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4277-4286 [DOI: 10.1109/CVPR.2019.00441]
- Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018. High performance visual tracking with Siamese region proposal network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8971-8980 [DOI: 10.1109/CVPR.2018.00935]
- Li X, Zha Y F, Zhang T Z, Cui Z, Zuo W M, Hou Z Q, Lu H C and Wang H Z. 2019. Survey of visual object tracking algorithms based on deep learning. Journal of Image and Graphics, 24(12): 2057-2080 (李玺, 查宇飞, 张天柱, 崔振, 左旺孟, 侯志强, 卢湖川, 王茜子. 2019. 深度学习的目标跟踪算法综述. 中国图象图形学报, 24(12): 2057-2080) [DOI: 10.11834/jig.190372]
- Lin L T, Fan H, Zhang Z P, Xu Y and Ling H B. 2022. SwinTrack: a simple and strong baseline for Transformer tracking [EB/OL]. [2024-01-23]. <https://arxiv.org/pdf/2112.00995.pdf>
- Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/

- 978-3-319-10602-1_48]
- Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin Transformer: hierarchical vision Transformer using shifted windows//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9992-10002 [DOI: 10.1109/ICCV48922.2021.00986]
- Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization [EB/OL]. [2024-01-23]. <https://arxiv.org/pdf/1711.05101.pdf>
- Mayer C, Danelljan M, Bhat G, Paul M, Paudel D P, Yu F and Van Gool L. 2022. Transforming model prediction for tracking//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8721-8730 [DOI: 10.1109/CVPR52688.2022.00853]
- Mayer C, Danelljan M, Paudel D P and Van Gool L. 2021. Learning target candidate association to keep track of what not to track//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13424-13434 [DOI: 10.1109/ICCV48922.2021.01319]
- Meng L and Li C X. 2019. Brief review of object tracking algorithms in recent years: correlated filtering and deep learning. *Journal of Image and Graphics*, 24(7): 1011-1016 (孟禄, 李诚新. 2019. 近年目标跟踪算法短评——相关滤波与深度学习. *中国图象图形学报*, 24(7): 1011-1016) [DOI: 10.11834/jig.190111]
- Mueller M, Smith N and Ghanem B. 2016. A benchmark and simulator for UAV tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 445-461 [DOI: 10.1007/978-3-319-46448-0_27]
- Müller M, Bibi A, Giancola S, Alsubaihi S and Ghanem B. 2018. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 310-327 [DOI: 10.1007/978-3-030-01246-5_19]
- Nam H and Han B. 2016. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4293-4302 [DOI: 10.1109/CVPR.2016.465]
- Pan F, Zhao L Y and Wang C L. 2023. Smaller and more accurate swin-Transformer model prediction for tracking//Proceedings of the 35th IEEE International Conference on Tools with Artificial Intelligence. Atlanta, USA: IEEE: 955-961 [DOI: 10.1109/ICTAI59109.2023.00143]
- Rezatofighi H, Tsoi N, Gwak J Y, Sadeghian A, Reid I and Savarese S. 2019. Generalized intersection over union: a metric and a loss for bounding box regression//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 658-666 [DOI: 10.1109/CVPR.2019.00075.]
- Song Z K, Yu J Q, Chen Y P P and Yang W. 2022. Transformer tracking with cyclic shifting window attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 8781-8790 [DOI: 10.1109/CVPR52688.2022.00859]
- Strudel R, Garcia R, Laptev I and Schmid C. 2021. Segmenter: Transformer for semantic segmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7242-7252 [DOI: 10.1109/ICCV48922.2021.00717]
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010
- Wang J, Yin P, Wang Y Y and Yang W H. 2024. CMAT: integrating convolution mixer and self-attention for visual tracking. *IEEE Transactions on Multimedia*, 26: 326-338 [DOI: 10.1109/TMM.2023.3264851]
- Wang J Y, Hou Z Q, Yu W S, Liao X F and Chen C H. 2018. Fast TLD visual tracking algorithm with kernel correlation filter. *Journal of Image and Graphics*, 23(11): 1686-1696 (王皎尧, 侯志强, 余旺盛, 廖秀峰, 陈传华. 2018. 采用核相关滤波的快速TLD视觉目标跟踪. *中国图象图形学报*, 23(11): 1686-1696) [DOI: 10.11834/jig.170655]
- Wang N, Zhou W G, Wang J and Li H Q. 2021. Transformer meets tracker: exploiting temporal context for robust visual tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1571-1580 [DOI: 10.1109/CVPR46437.2021.00162]
- Wu H P, Xiao B, Codella N, Liu M C, Dai X Y, Yuan L and Zhang L. 2021. CvT: introducing convolutions to vision Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 22-31 [DOI: 10.1109/ICCV48922.2021.00009]
- Wu Y, Lim J and Yang M H. 2015. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9): 1834-1848 [DOI: 10.1109/TPAMI.2014.2388226]
- Xie F, Wang C Y, Wang G T, Yang W K and Zeng W J. 2021. Learning tracking representations via dual-branch fully Transformer networks//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 2688-2697 [DOI: 10.1109/ICCVW54120.2021.00303]
- Xiong W, Xiong C Y, Gao Z R, Chen W Q, Zheng R H and Tian J W. 2023. Image super-resolution with channel-attention-embedded Transformer. *Journal of Image and Graphics*, 28(12): 3744-3757 (熊巍, 熊承义, 高志荣, 陈文旗, 郑瑞华, 田金文. 2023. 通道注意力嵌入的Transformer图像超分辨率重构. *中国图象图形学报*, 28(12): 3744-3757) [DOI: 10.11834/jig.221033]
- Xu Y D, Wang Z Y, Li Z X, Yuan Y and Yu G. 2020. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 12549-12556 [DOI: 10.

1609/aaai.v34i07.6944]

Yan B, Peng H W, Fu J L, Wang D and Lu H C. 2021. Learning spatio-temporal Transformer for visual tracking//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10428-10437 [DOI: 10.1109/ICCV48922.2021.01028]

Ye B T, Chang H, Ma B P, Shan S G and Chen X L. 2022. Joint feature learning and relation modeling for tracking: a one-stream framework//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 341-357 [DOI: 10.1007/978-3-031-20047-2_20]

Zou Z J, Liu X X, Zhang Y P, Shu L and Hao J. 2023. Know who you are: learning target-aware Transformer for object tracking//Proceedings of 2023 IEEE International Conference on Multimedia and

Expo. Brisbane, Australia: IEEE: 1427-1432 [DOI: 10.1109/ICME55011.2023.00247]

作者简介

徐晗,男,硕士研究生,主要研究方向为计算机视觉和目标跟踪。E-mail: 202212490766@nuist.edu.cn

郑钰辉,通信作者,男,教授,主要研究方向为图像视频分析、场景理解、视觉跟踪和模式识别。
E-mail: zhengyh@vip.126.com

董仕豪,男,博士研究生,主要研究方向为聚类分析和深度学习。E-mail: dongshihao@nuist.edu.cn

张家伟,男,博士研究生,主要研究方向为多媒体内容安全。
E-mail: zjwei_2020@163.com