融合 Swin Transformer 多尺度特征与池化空间特征的目标追踪算法研究

陈苗苗
安徽理工大学 计算机科学与工程学院,安徽 淮南 232001

摘 要:目的 目标追踪在视频监控、汽车自动驾驶、无人机航拍等领域有广泛应用,现有的基于 Transformer 模型的自注意力操作是将 2D 特征转换为 1D 序列,会忽略目标对象的空间先验知识,导致追踪效果不佳,针对这一问题,提出一种名为 Pool-Swin Transformer Tracker (PSTransT)的追踪器。**方法** PSTransT 将 Swin Transformer 的每个阶段与一个池化层相结合,实现了在不同尺度上充分提取特征的能力,同时保留了空间位置信息。具体而言,PSTransT 利用基于跨尺度融合的 Swin Transformer 模型进行上下文建模,该模型在每个阶段都与一个池化层并联,这样可以在保持特征丰富性的同时,有效地捕捉不同尺度上的空间特征。此外,该方法采用基于 Transformer 的特征融合网络,通过 Transformer 的自注意力机制,联合学习模板特征和搜索特征之间的关联性进行特征融合,以更好地捕捉被追踪目标的动态变化和局部上下文信息。**结果** 该方法在多个基准数据集上对 PSTransT 进行了广泛评估,其中在 LaSOT 上达到了 69.2%的成功率,比 SiamPRN++高 19.6%、在 GOT-10k 上达到了 72.2%的平均重叠率(mAO),比 SiamPRN++高 20.5%。**结论** 实验结果表明:在保留上下文信息的同时,保留空间先验信息对目标追踪性能有利,PSTransT 优于其他对比方法。

关键词:目标追踪;Swin Transformer;池化层;上下文建模
中图分类号:TP391.41; TP18 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2025.0003.014

Research on Object Tracking Algorithm Integrating Swin Transformer Multi-scale Features and Pooling Spatial Features

CHEN Miaomiao
School of Computer Science and Engineering, Anhui University of Science and Technology, Anhui Huainan 232001, China

Abstract: Objective Object tracking finds extensive applications in fields such as video surveillance, autonomous driving, and UAV aerial photography. Self-attention operations of existing Transformer-based models transform 2D features into 1D sequences, which ignore spatial priors of target objects, leading to poor tracking performance. To address this limitation, this paper proposed a tracker named Pool-Swin Transformer Tracker (PSTransT). **Methods** PSTransT integrated each stage of the Swin Transformer with a pooling layer, enabling the effective extraction of features across different scales while preserving spatial positional information. Specifically, PSTransT utilized a cross-scale fusion-based Swin Transformer model for contextual modeling, where each stage was parallelized with a pooling layer to effectively capture spatial features at various scales while maintaining feature richness. Furthermore, the method utilized a Transformer-based feature fusion network that leveraged the self-attention mechanism of the Transformer to jointly learn the correlations between template features and search features for feature fusion, aiming to better capture the dynamic changes

收稿日期:2023-10-19 **修回日期:**2023-12-06 **文章编号:**1672-058X(2025)03-0110-08
作者简介:陈苗苗(1999—),女,安徽阜阳人,硕士研究生,从事计算机视觉、单目标跟踪等研究。
引用格式:陈苗苗.融合 Swin Transformer 多尺度特征与池化空间特征的目标追踪算法研究[J].重庆工商大学学报(自然科学版),2025,42(3):110-117.
CHEN Miaomiao. Research on object tracking algorithm integrating swin transformer multi-scale features and pooling spatial features[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2025, 42(3): 110-117.

of the tracked target and local contextual information. **Results** The effectiveness of PSTransT was extensively evaluated on multiple benchmark datasets. This method achieved a success rate of 69.2% on LaSOT, which was 19.6% higher than SiamPRN++, and reached a mean average overlap (mAO) of 72.2% on GOT-10k, surpassing SiamPRN++ by 20.5%. **Conclusion** Experimental results demonstrate that preserving spatial prior information alongside contextual information benefits object tracking performance. PSTransT outperforms other comparative methods.

Keywords: object tracking; Swin Transformer; pool layer; contextual modeling

1 引　言

目标追踪[1]指通过给出初始帧中要跟踪的对象,实现目标对象在时间和空间上的连续跟踪和定位。它的研究始于早期的基于像素级匹配和相关滤波器的方法。目标追踪领域的研究主要集中在两个方向:基于传统算法[2]的方法和基于深度学习[3]的方法。

传统算法主要依赖于手工设计的特征表示和关联模型,比如卡尔曼滤波器和相关滤波器等。卡尔曼滤波器[2]是一种常用的线性动态系统中的滤波方法,通过建立目标状态与观测之间的关系,来预测目标的位置和速度。相关滤波器是一种基于模板匹配的方法,通过计算目标模板与搜索区域的相关性来确定目标的位置。这些方法具有一定的鲁棒性和计算效率,但对目标外观和动态变化敏感度低,并且难以处理目标尺度的变化和形变。

为了解决上述问题,深度学习算法试图设计强大的神经网络来进行鲁棒的特征表示学习,通过端到端的学习有效地捕获目标的特征表示和关系。许多端到端视觉跟踪器被提出,如 Siamese Fully-Convolutional (SiamFC)[3]和 Siamese Region-Proposal Network (SiamRPN)[4],为基于 Siamese 匹配的跟踪铺平了新的道路。它的关键步骤是计算模板与搜索分支之间的相似度,通常采用 ResNet 作为特征提取和相关计算的主干。在此框架下提出了越来越多的 Siamese 跟踪器,尽管这些跟踪器在现有的基准数据集上实现了高性能,但在面对剧烈变形、严重或完全遮挡、快速运动等具有挑战性的因素时,它们的性能仍然有限。

最近,Transformer 在自然语言处理领域取得了显著成功,它的核心模块是自我关注,它很好地模拟了长距离关系。一些研究者已经简单地利用 Transformer 与卷积神经网络(Convolutional Neural Network, CNN)的结合,如 Transformer Tracking (TransT)[5]、Transformer Discriminative Model Prediction (TrDiMP)[6]、Spatio-Temporal Transformer Tracker(STARK)[7]等,这些工作主要使用 Transformer 融合模板和搜索区域,这种结合能够使得模型更好地处理模板和搜索图像上下文之间的关系,提高对搜索区域的表示能力,取得比完全基于 CNN 的追踪器更好的效果。但是它们只是简单地将二维特征映射成一维特征序列,尽管引入了位置编码,但模型并没有直接的机制来处理不同位置之间的相对距离和空间关系,仅仅是计算每个补丁之间的注意力得分。因此,如何在使用 Transformer 学习远程关系时保留目标对象的空间先验性是一个迫切需要解决的问题。

为了保留目标对象的空间先验,本文采用基于 Swin Transformer 和池化层相结合的主干网络进行特征提取,利用 Swin Transformer 对图像的全局特征进行多尺度建模,并利用池化层保留不同位置的空间信息;接着采用基于 Transformer 的特征融合网络促进模板特征和搜索特征的交互;然后采用双头分支进行分类和回归,其中全连接头用于分类任务,卷积头用于回归任务。

2 相关工作

2.1 基于 Siamese 的追踪

随着 SiamFC[3]的发布,视觉追踪被表述为模板与搜索区域之间的相似度匹配问题,随后推出了许多追踪器以提高性能。SiamRPN[8]引入了 RPN 结构来预测目标尺度的长宽比变化;SiamRPN++[9]通过使用更深层次的 ResNet-50 网络作为参数共享主干,将追踪框架的表示能力提升到一个新的水平;Siamese Re-Detection-CNN(SiamR-CNN)[10]结合了 Siamese 框架和 Faster Re-Detection-CNN(Faster R-CNN)[11],实现了在尺度变化和潜在对象上的鲁棒性。这些 Siamese 追踪器在处理运动模糊和变形等属性方面取得了显著的性能,但仍然容易受到相似干扰物和严重遮挡的影响。为了解决这些问题,一些追踪器引入了在线学习方法。KeepTrack[12]引入了学习关联网络来区分干扰物,实现了持续的鲁棒性追踪;UpdateNet[13]打破了线性模板更新策略,提出了学习非线性融合模板的更新网络;Triple Attention Net(TANet)[14]通过采用联合局部-全局搜索方案来实现鲁棒的追踪,在长期视频中表现良好。受到这些追踪器的启发,本文提出了 PSTransT,与其他

Siamese 追踪框架类似,PSTransT 也采用了 Siamese 结构。

2.2 基于 Transformer 的追踪

Vision Transformer(ViT)首次将 Transformer 引入到图像分类中,并取得了与 CNN 相媲美的性能,在 ViT 的推动下,一些追踪方法将 CNN 网络与 Transformer 组件相结合,取得了显著的效果。TransT[5] 首先将 Transformer 结构引入追踪领域,并利用 DETR 的自注意力和交叉注意力来增强自我语境信息和特征交叉融合;TrDiMP[6] 将编码器和解码器网络视为并行分支,并在两个分支之间探索时间上下文;STARK[7] 遵循编码器-解码器框架,在特征提取后将模板与搜索分支合并为一个分支进行特征融合。然而,这些基于 Transformer 的追踪器都受到了 CNN 主干网络上下文表示缺陷的限制。随着 Swin Transformer 的发布,SwinTrack[15] 将 CNN 主干转变为 Transformer,学习到了更出色的特征提取和表示能力,但由于没有直接的机制来处理不同位置之间的相对距离和空间关系,可能会获得次优结果。与之相比,本文提出的 PSTransT 解决了这个问题,并在多个基准数据集上取得了更高的性能。

3 本文方法

本文追踪器网络框架如图 1 所示,由特征提取网络、特征融合网络和头网络 3 个部分组成。特征提取网络以共享权值的方式分别提取模板块和搜索区域块的特征;特征融合网络将模板图像和搜索图像的特征进行拼接,并通过注意力机制对拼接后的特征进行逐层增强,生成搜索图像的最终特征图;头网络通过处理最终的特征图得到分类响应图和边界框回归图。

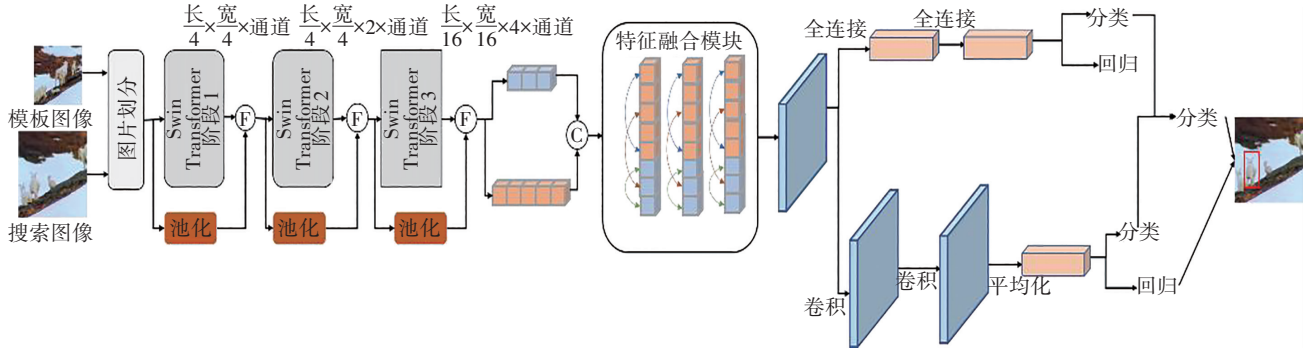

图 1 PSTransT 框架图

Fig. 1 Architecture of PSTransT

3.1 特征提取网络

本文的追踪器遵循经典 Siamese 追踪器的方案,需要将模板图像 $Z \in R^{H_Z \times W_Z \times 3}$ 和搜索区域 $X \in R^{H_x \times W_x \times 3}$ 作为输入。首先,将每个阶段 Swin Transformer 的输出特征定义为 T_i;然后,在每个阶段之前添加池化层,将池化后的输出特征记为 P_i;最后将特征 T_i 和 P_i 进行融合。这样的操作可以在保留空间信息的同时获得丰富的特征表示。

3.1.1 Swin Transformer

Swin Transformer 通过自注意力机制为输入序列中的每个位置分配不同的权重,从而准确地捕捉到不同位置之间特征的相互依赖关系。Swin Transformer 有 4 个阶段,每个阶段都会缩小输入特征图的分辨率,像 CNN 一样逐层扩大感受野。每个阶段由块整合和多个 Swin Transformer 块组成。块整合模块在每个阶段的开头用于降低图像分辨率,它将相邻的图块合并为更大的图块,从而减少整体块的数量,同时保留重要的信息。每个 Swin Transformer 块的具体结构如图 2 所示,它由层归一化(Layer Norman, LN)、多层感知机(Multilayer Perceptron, MLP)、窗口注意力(Window Attention, WA)和滑动窗口注意力(Shift-Window Attention, SWA)组成。

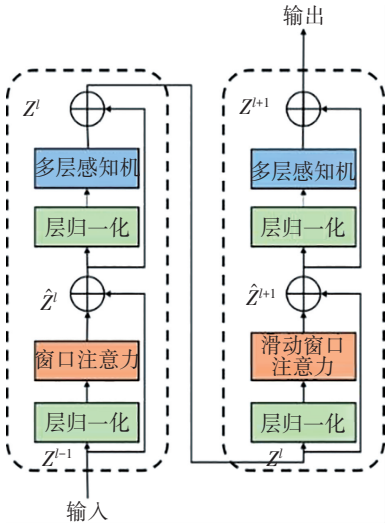

图 2 Swin Transformer 块结构图

Fig. 2 Structure diagram of Swin Transformer block

SWA 和原始的 WA 相结合,共同用于特征的建模和信息交互,WA 和 SWA 的计算可表示为

$$\hat{z}^{i}=\mathrm{WA}(\mathrm{LN}(z^{i-1}))+z^{i-1} \tag{1}$$

$$z^{i}=\mathrm{MLP}(\mathrm{LN}(\hat{z}^{i}))+\hat{z}^{i} \tag{2}$$

$$\hat{z}^{i+1}=\mathrm{SWA}(\mathrm{LN}(z^{i}))+z^{i} \tag{3}$$

$$z^{i}=\mathrm{MLP}(\mathrm{LN}(\hat{z}^{i+1}))+\hat{z}^{i+1} \tag{4}$$

其中,z^{i} 表示输入特征,$\hat{z}^{i}$ 表示输出特征,LN 用于进行层级归一化处理,MLP 用于特征的映射和转换,WA 模块用于计算每个窗口下注意力对局部特征的建模,SWA 模块通过移位用于与其他窗口进行信息交互,进一步加强特征之间的关联性。

SWA 通过移位后,新的窗口包含了原本相邻窗口的元素,导致窗口的数量增多,即由原来的 4 个窗口变成了 9 个窗口。为了解决这个问题,本文对特征图进行移位操作,同时保持原有窗口数量不变,如图 3 所示。通过这种移位操作,能够充分利用窗口内元素的信息,并在计算注意力时保持了相邻窗口的联系。

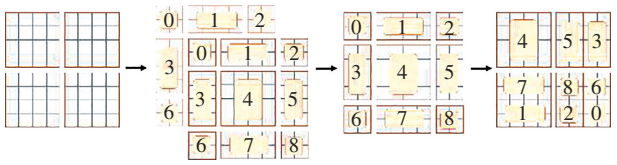

图 3　SWA 原理

Fig. 3　SWA principle

然而,这种序列化的注意力操作可能会忽略目标对象的空间先验导致次优的结果。因此,本文将 Swin Transformer 每个阶段的输出结果与该阶段之前的特征经过池化层进行并联,以获取最终的特征表示。这样设计的优势是可以综合利用空间先验与全局语义信息,因为池化层可以根据空间结构的先验知识对特征进行聚合,保留物体的位置和大小关系,而 Swin Transformer 通过自注意力机制捕捉特征的长程依赖关系,提取全局语义信息,将它们并联使用,可以综合利用空间先验和全局语义信息,使模型具备更好的观测和理解能力。

3.2　特征融合网络

本文没有直接使用原始的 Transformer,而是利用 Transformer 的注意力机制来完成特征融合网络的结构设计,如图 4 所示。在将输入特征融合模块之前,先将模板图像和搜索图像中的特征沿着空间维度连接;接着通过多头自注意力(Multiple Self-attention,MSA)模块计算联合表示上的自关注;然后使用前馈神经网络(Feedforward Netual Network, FFN)对 MSA 生成的特征标记进行细化;最后,执行解连接操作以恢复模板图像特征标记和搜索图像特征标记,生成搜索图像的最终特征映射。整个过程可以表示为

$$H_{z}^{0}=[z_{1}E;z_{2}E;\cdots;z_{i}E]+P_{z} \tag{5}$$

$$H_{x}^{0}=[x_{1}E;x_{2}E;\cdots;x_{3}E]+P_{x} \tag{6}$$

$$A^{1}=\mathrm{Contact}(H_{z}^{0},H_{x}^{0}) \tag{7}$$

$$\mathrm{Att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{8}$$

$$A^{l+1}=(A^{l}+\mathrm{Att}(\mathrm{LN}(A^{l})))+\mathrm{FFN}(\mathrm{LN}(A^{l})) \tag{9}$$

$$Z^{l+1},X^{l+1}=\mathrm{DC}(A^{l+1}) \tag{10}$$

其中,$H_{z}^{0}\in\mathbf{R}^{N_{z}\times D}$ 是模板嵌入,$H_{x}^{0}\in\mathbf{R}^{N_{x}\times D}$ 是搜索区域嵌入,x_{i} 和 z_{i} 分别是搜索图片和模板图片划分的小块,P_{z} 和 P_{x} 是位置编码,E 是一个可训练的线性编码层,SoftMax 是软最大化操作,d 是 $\boldsymbol{Q}$ 的维度,查询(Query,$\boldsymbol{Q}$)是用于度量与键相似性的向量,键(Key, $\boldsymbol{K}$)提供与相应位置或元素相关的信息,值(Value,$\boldsymbol{V}$)包含与相应位置或元素关联的实际特征值,Contact 是连接操作,DC 是解连接操作。

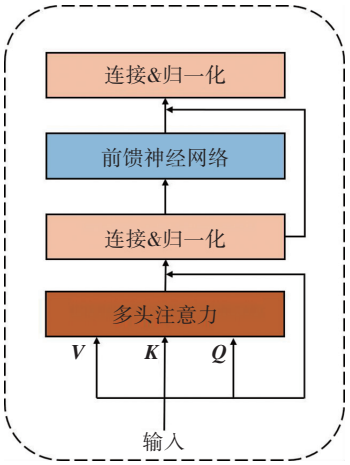

图 4　特征融合网络

Fig. 4　Feature fusion network

3.3　双头网络

双头网络的两个输出头分别对应两个不同的任务。其中一个输出头全连接头(FC-Head)用于对输入图像进行分类预测;而另一个输出头卷积头(Conv-Head)则用于对边界框回归。在训练过程中,FC-Head 不仅要受到本身擅长的分类信息监督还要受到回归信息的监督,同时 Conv-Head 也要受到两个信息的监督。FC-Head 的损失函数如下所示:

$$\vartheta^{\mathrm{fc}}=\lambda^{\mathrm{fc}}L_{\mathrm{cls}}^{\mathrm{fc}}+(1-\lambda^{\mathrm{fc}})L_{\mathrm{reg}}^{\mathrm{fc}} \tag{11}$$

其中,$L_{\mathrm{cls}}^{\mathrm{fc}}$ 和 $L_{\mathrm{reg}}^{\mathrm{fc}}$ 分别是 FC-Head 中的分类和回归损失;λ^{fc} 用来均衡 FC-Head 中的两种损失,根据文献

[27],设 $\lambda^{fc}=0.7$。

Conv-Head 的损失函数计算同 FC-Head:

$$\vartheta^{conv}=(1-\lambda^{conv})L_{cls}^{conv}+\lambda^{conv}L_{reg}^{conv} \tag{12}$$

其中,L_{cls}^{conv} 和 L_{reg}^{conv} 分别是 Conv-Head 中的分类和回归损失;λ^{conv} 用来均衡 Conv-Head 中的两种损失,根据文献[27],设 $\lambda^{conv}=0.8$。

最终的分类结果是两个头融合的结果,而回归结果是只采用卷积得到的结果。对于最终的分类结果,如式(13)所示:

$$L^{cls}=\vartheta^{fc}+\vartheta^{conv}(1-\vartheta^{fc})=\vartheta^{conv}+\vartheta^{fc}(1-\vartheta^{conv}) \tag{13}$$

4 仿真实验与结果分析

4.1 数据集和评估准则

为了训练 PSTransT,本文从 COCO、TrackingNet、LaSOT 和 GOT-10K 数据集中随机抽取训练对。为了评估 PSTransT 的有效性,本文在 4 个追踪数据集上进行测试和验证,包括 LaSOT、GOT-10K、TrackingNet 和 LaSOT-ext。其中,LaSOT 是一个密集标注的大规模数据集,包含 280 个长期视频序列供公众评估;TrackingNet 包含超过 30 k 的训练视频,具有 30 帧间隔标签策略,每个序列平均包含 1 300 帧;LaSOT-ext 是一个最近发布的 LaSOT 数据集,它由来自 15 个对象类的 150 个额外视频组成,是一个长期追踪数据集,每个序列平均包含超过 2 500 帧;GOT-10k 提供 180 个没有公开的地面真相短期视频序列。本文采用多个评估指标,包括成功率(Success)、准确率(Precision)、归一化准确率(Norm Precision)。此外,平均重叠率(mean Average Overlaps,mAO)、$mSR_{0.5}$ 和 $mSR_{0.75}$ 是 GOT-10K 的特定指标,$mSR_{0.5}$ 和 $mSR_{0.75}$ 是为成功率设定不同阈值的 IOU 重叠。

4.2 实现细节

本文模板图像从视频序列的第一帧裁剪,目标物体位于图像的中心,背景面积因子为 2。从当前帧裁剪搜索区域,图像中心为前一帧预测的目标中心位置,搜索区域的背景面积因子为 4。骨干网络的学习率为 5e-5,权值衰减为 1e-4,采用自适应矩估计优化器对模型进行优化,采用梯度裁剪来防止太大的梯度误导优化过程。本文的实验是在 NVIDIA GeForce RTX 3090 GPU 上进行训练的,并在单个 NVIDIA RTX 3060Ti GPU 上测试,在 210 次迭代后,学习率下降了 10 倍。

4.3 实验结果及分析

本文展示在 4 个广泛使用的基准数据集上的追踪性能,并将其与其他最先进的追踪器进行比较,结果如表 1—表 4 所示,最好的结果以粗体显示。

表 1 在 LaSOT 测试集上的比较

Table 1 Comparison on LaSOT dataset

/%

追踪器	成功率	准确率	归一化准确率
SiamPRN++[8]	49.6	—	56.9
DiMP[15]	56.9	53.9	65.0
SiamR-CNN[16]	64.8	—	72.2
TrSiam[6]	62.4	60.0	—
TrDiMP[6]	63.9	61.4	—
STMTrack[17]	60.6	63.3	69.3
TransT[5]	64.9	69.0	73.8
STARK[7]	67.1	—	77.0
KeepTrack[11]	67.1	70.2	77.2
AiATrack[18]	69.0	79.4	73.8
TUC-Net[19]	63.0	67.1	—
PSTransT	**69.2**	**72.8**	**78.0**

表 2 在 GOT-10k 测试集上的比较

Table 2 Comparison on GOT-10k dataset

/%

追踪器	mAO	$mSR_{0.5}$	$mSR_{0.75}$
SiamPRN++[8]	51.7	61.6	32.5
DiMP[15]	61.1	71.7	49.2
SiamR-CNN[16]	64.9	72.8	59.7
TrSiam[6]	66.0	76.6	57.1
TrDiMP[6]	67.1	77.7	58.3
STMTrack[17]	64.2	73.7	57.5
TransT[5]	67.1	76.8	60.9
STARK[7]	68.8	78.1	64.1
AiATrack[18]	69.6	80.4	68.2
TUC-Net[19]	70.0	80.5	—
RANformer[20]	70.6	81.3	66.4
PSTransT	**72.2**	**84.8**	**68.4**

表 3 在 TrackingNet 测试集上的比较

Table 3 Comparison on TrackingNet dataset

/%

追踪器	成功率	准确率	归一化准确率
PrDiMP[21]	75.8	70.4	81.6
SiamFC++[22]	75.4	70.5	80.0
TransT[5]	81.4	80.3	86.7
MAMLTrack[23]	75.7	72.5	82.2
TUC-Net[19]	80.4	72.3	—
PSTransT	**82.1**	**80.4**	**87.2**

表 4　在 LaSOT-ext 测试集上的比较

Table 4　Comparison on LaSOT-ext dataset

/%

追踪器	成功率	准确率	归一化准确率
C-RPN[24]	27.5	32.0	34.4
DiMP[15]	39.2	45.1	47.5
LTMU[25]	41.4	47.3	49.9
SuperDiMP[26]	43.7	—	52.7
PSTransT	**47.1**	**53.0**	**58.2**

从表 1 可以看出,PSTransT 模型获得了 69.2%的成功率得分和 72.8%的准确率得分,超过了另外两个 Transformer 追踪器 TrSiam 和 TransT,成功率得分分别高了 6.8%和 4.3%。这表明 PSTransT 在处理长期视频追踪问题上非常有效。

从表 2 可以看出,PSTransT 在 mAO、$mSR_{0.5}$、$mSR_{0.75}$ 上分别实现了 72.2%、84.8%、68.4%的新性能,比以前的 TransT 高出 5.1%、8.0%、7.5%。这些结果表明:PSTransT 方法在目标追踪中表现出更好的性能,能够更准确地捕捉、追踪目标对象,并更精确地预测目标的位置和形状。

本文在 TrackingNet 上的结果由在线评估服务器报告。从表 3 可以看出,PSTransT 实现了非常好的表现,即成功率、准确率、归一化准确率指标分别为 82.1%、80.4%和 87.2%,与流行的 SiamFC++的结果相比,本文的方法将成功率分数提高了 6.7%,准确率分数提高了 9.9%。

从表 4 可以看出,PSTransT 的追踪器在 LaSOT 数据集上实现了 47.1%的准确率,比之前的追踪器 DiMP 高 7.9%。这个结果表明 PSTransT 在应对长距离追踪问题时表现出卓越的性能。

4.4　消融实验

为了分析所提出方法的有效性,本文对 GOT-10k 进行了详细的消融实验。表 5 展示了每个模块的评估结果。首先同时使用池化层、自注意力模块和交叉注意力模块在 mAO、$mSR_{0.5}$、$mSR_{0.75}$ 上实现了 72.2%、84.8%、68.4%,这表明联合这些模块是有效的。但是当去除交叉注意力模块时,mAO 仅下降了 0.1%,这表明交叉注意力模块在融合模板和搜索区域之间的特征信息方面没有进一步提升结果,为了减少计算量,本文去除交叉注意力模块。当进一步将池化层去除,发现 mAO 得分降低了 2.1%,证明了池化层有助于保留空间信息,从而提高模型对目标区域的感知能力。

图 5 可视化了添加池化层和不添加池化层的分类热力图,第一列是裁剪规则的搜索图片,第二列是添加池化层的结果,第三列是不添加池化层的结果,发现添加池化层的主干网络得到的结果更清晰。

表 5　各个部分的消融

Table 5　Ablation of individual parts

池化层	自注意力模块	交叉注意力模块	结果/%			速度/fps
			mAO	$mSR_{0.5}$	$mSR_{0.75}$	
✓	✓	✓	72.1	84.6	68.0	114
×	✓	×	69.1	81.8	66.3	98
✓	✓	×	**72.2**	**84.8**	**68.4**	**102**

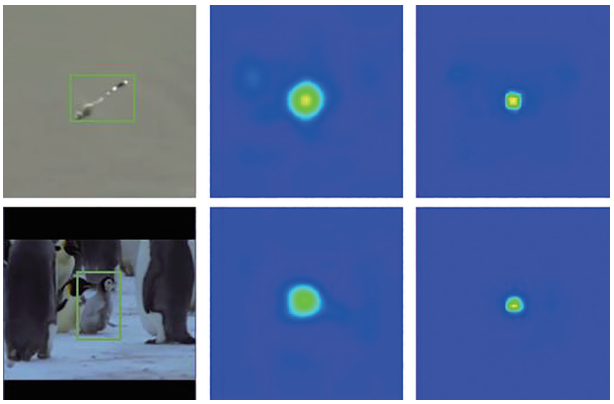

图 5　消融实验的视觉差异

Fig. 5　Visual differences in ablation experiments

4.5　定性分析

本文选择了一些具有代表性的视频序列进行定性分析。图 6 可视化了 TransT 和 PSTransT 的主干网络结构中注意力机制的注意力图。通过注意力图,发现 PSTransT 不仅可以准确定位和标记目标物体在图像中的位置,捕捉目标物体的精确运动轨迹,还可以准确反映目标物体的重要性和注意力,即使在存在相似物体或遮挡的情况下,仍然可以准确区分和突出显示目标热点,这再次充分证明了 PSTransT 是稳健的,能够满足追踪要求。

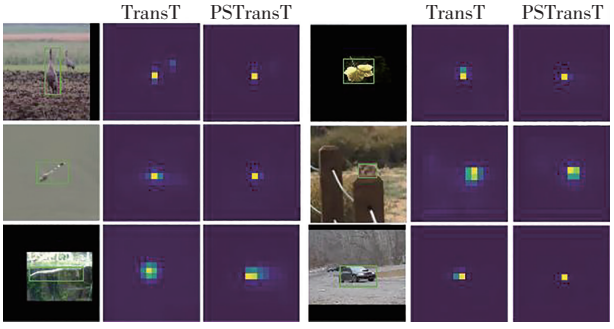

图 6　与 TransT 的定性比较

Fig. 6　Qualitative comparison with TransT

为了将 PSTransT 与最先进的追踪器的追踪结果进行直观比较,本文选择 TransT 和 SiamRPN++两个具有代表性的追踪器作为参考。如图 7 所示,本文选择几个具有代表性的序列进行比较,黄色线代表真实边界框,黑色线代表本文追踪器获取的追踪框,紫色线代表 TransT 追踪器获取的追踪框,红色线代表 SiamRPN++追踪器获取的追踪框。在一些复杂的场景中,PSTransT 能够更加精准和稳定地追踪目标物体,相比之下,TransT 和 SiamRPN++在一些场景下出现了追踪失败或不稳定等问题。而且 PSTransT 能够更好地适应光照变化、遮挡情况,这意味着 PSTransT 更具鲁棒性,能够在复杂环境中保持高质量的追踪。

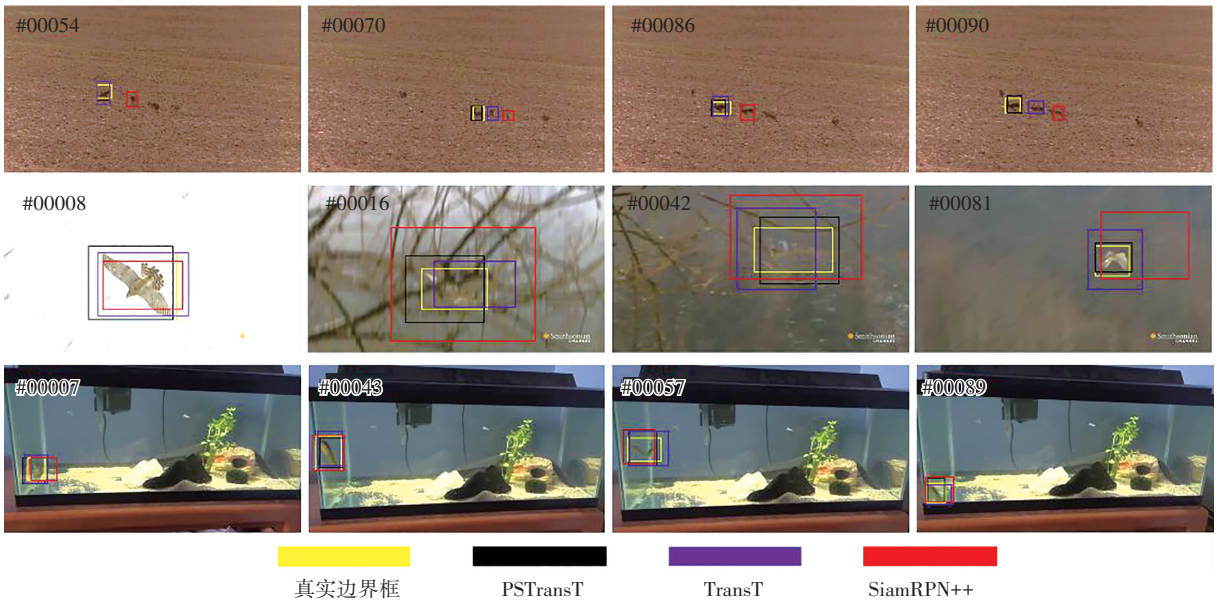

图 7 对 3 个具有挑战性的序列进行定性比较

Fig. 7 Qualitative comparison of three challenging sequences

5 结 论

为了解决 Swin Transformer 模型在目标追踪任务中忽略目标对象空间先验知识的问题,提出一种名为 PSTransT 的追踪器。通过将每个 Swin Transformer 阶段与池化层并联,并结合 Swin Transformer 模型的上下文建模能力,充分提取不同尺度上的特征,并利用池化层保留空间位置信息。采用双头分支进行分类和回归,进一步提高了目标追踪的准确性和稳定性。通过这种融合策略,PSTransT 追踪器能够有效解决 Swin Transformer 模型在目标追踪任务中可能出现的问题,提高了追踪结果的准确性。

在未来的工作中,将设计一个图匹配网络模型,通过判别学习来明确建模密集场景中目标物体的外观,帮助抑制遮挡和干扰的影响,进一步提高目标追踪的性能。期待这种网络的应用能进一步推动目标追踪技术的发展,使其在实际场景中能够达到更高的性能水平。

参考文献:

[1] JAVED S, DANELLJAN M, KHAN F S, et al. Visual object tracking with discriminative filters and siamese networks: A survey and outlook[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(5): 6552-6574.

[2] LEICHTER I, LINDENBAUM M, RIVLIN E. Mean shift tracking with multiple reference color histograms[J]. Computer Vision and Image Understanding, 2010, 114(3): 400-408.

[3] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]// Computer Vision-ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 850-865.

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

[5] CHEN X, YAN B, ZHU J, et al. Transformer tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8126-8135.

[6] 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. IEEE, 2021: 1571-1580.

[7] YAN B, PENG H, FU J, et al. Learning spatio-temporal

transformer for visual tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 10448-10457.

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

[9] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam r-cnn: Visual tracking by re-detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 6578-6588.

[10] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. IEEE, 2015: 1440-1448.

[11] MAYER C, DANELLJAN M, PANI PAUDEL D, et al. Learning target candidate association to keep track of what not to track[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 13444-13454.

[12] ZHANG L, GONZALEZ-GARCIA A, VAN D E WEIJER J, et al. Learning the model update for Siamese trackers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 4010-4019.

[13] LIU Z, ZHAO X, HUANG T, et al. TANET: Robust 3D object detection from point clouds with triple attention[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11677-11684.

[14] LIN L, FAN H, ZHANG Z, et al. Swintrack: A simple and strong baseline for transformer tracking[J]. Advances in Neural Information Processing Systems, 2022, 35: 16743-16754.

[15] BHAT G, DANELLJAN M, VANGOOL L, et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 6182-6191.

[16] VOIGTLAENDER P, LUITEN J, TORR P H S, et al. Siam R-CNN: Visual tracking by re-detection[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 6577-6587.

[17] FU Z, LIU Q, FU Z, et al. STMTrack: Template-free visual tracking with space-time memory networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 13774-13783.

[18] GAO S, ZHOU C, MA C, et al. Aiatrack: Attention in attention for transformer visual tracking[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 146-164.

[19] SONG Z, CHEN Y, LUO P, et al. Transformer Union Convolution Network for visual object tracking[J]. Optics Communications, 2022, 524: 128810

[20] GU F, LU J, CAI C. A robust attention-enhanced network with transformer for visual tracking[J]. Multimedia Tools and Applications, 2023, 82(26): 40761-40782.

[21] DANELLJAN M, VANGOOL L, TIMOFTE R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 7183-7192.

[22] XU Y, WANG Z, LI Z, et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines[J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 12549-12556.

[23] WANG G, LUO C, SUN X, et al. Tracking by instance detection: A meta-learning approach[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 6288-6297.

[24] FAN H, LING H. Siamese cascaded region proposal networks for real-time visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 7952-7961.

[25] DAI K, ZHANG Y, WANG D, et al. High-performance long-term tracking with meta-updater[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 6298-6307.

[26] CHOI S, LEE J, LEE Y, et al. Robust long-term object tracking via improved discriminative model prediction[C]//Computer Vision-ECCV 2020 Workshops. Cham: Springer International Publishing, 2020: 602-617.

[27] WU Y, CHEN Y, YUAN L, et al. Rethinking classification and localization for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 10186-10195.

责任编辑: 李翠薇