控制与决策
Control and Decision

基于混合注意力的Transformer视觉目标跟踪算法

侯志强, 郭凡, 杨晓麟, 马素刚, 范九伦

引用本文:
侯志强,郭凡,杨晓麟,马素刚,范九伦. 基于混合注意力的Transformer视觉目标跟踪算法[J]. 控制与决策, 2024, 39(3): 739−748.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1340

基于混合注意力的Transformer视觉目标跟踪算法

侯志强[1†], 郭 凡[1], 杨晓麟[1], 马素刚[1], 范九伦[2]

(1. 西安邮电大学 计算机学院，西安 710121；2. 西安邮电大学 通信与信息工程学院，西安 710121)

摘 要: 基于Transformer的视觉目标跟踪算法能够很好地捕获目标的全局信息,但是,在对目标特征的表述上还有进一步提升的空间. 为了更好地提升对目标特征的表达能力,提出一种基于混合注意力的Transformer视觉目标跟踪算法. 首先,引入混合注意力模块捕捉目标在空间和通道维度中的特征,实现对目标特征上下文依赖关系的建模;然后,通过多个不同空洞率的平行空洞卷积对特征图进行采样,以获得图像的多尺度特征,增强局部特征表达能力;最后,在Transformer编码器中加入所构建的卷积位置编码层,为跟踪器提供精确且长度自适应的位置编码,提升跟踪定位的精度. 在OTB 100、VOT 2018和LaSOT等数据集上进行大量实验,实验结果表明,通过基于混合注意力的Transformer网络学习特征间的关系,能够更好地表示目标特征. 与其他主流目标跟踪算法相比,所提出算法具有更好的跟踪性能,且能够达到26帧/s的实时跟踪速度.

关键词: 计算机视觉；目标跟踪；孪生网络；深度学习；注意力机制；Transformer

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1340

引用格式: 侯志强, 郭凡, 杨晓麟, 等. 基于混合注意力的 Transformer 视觉目标跟踪算法 [J]. 控制与决策, 2024, 39(3): 739-748.

Transformer visual object tracking algorithm based on mixed attention

HOU Zhi-qiang[1†], *GUO Fan*[1], *YANG Xiao-lin*[1], *MA Su-gang*[1], *FAN Jiu-lun*[2]

(1. School of Computer，Xi'an University of Posts & Telecommunications，Xi'an 710121，China；2. School of Communication and Information Engineering，Xi'an University of Posts & Telecommunications，Xi'an 710121，China)

Abstract: The Transformer-based visual object tracking algorithm can capture the global information of the target well, but there is a possibility of further improvement in the presentation of the object features. To better improve the expression ability of object features, a Transformer visual object tracking algorithm based on mixed attention is proposed. First, the mixed attention module is introduced to capture the features of the object in the spatial and channel dimensions, so as to model the contextual dependencies of the target features. Second, the feature maps are sampled by multiple parallel dilated convolutions with different dilation rates to obtain the multi-scale features of the images, and enhance the local feature representation. Finally, the convolutional position encoding constructed is added to the Transformer encoder to provide accurate and length-adaptive position coding for the tracker, thereby improving the accuracy of tracking and positioning. The experimental results of the proposed algorithm on OTB 100, VOT 2018 and LaSOT show that by learning the relationship between features through the Transformer network based on mixed attention, the object features can be better represented. Compared with other mainstream object tracking algorithms, the proposed algorithm has better tracking performance and achieves a real-time tracking speed of 26 frames per second.

Keywords: computer vision；object tracking；siamese network；deep learning；attention mechanism；Transformer

0 引 言

视觉目标跟踪是计算机视觉中的一项关键技术,其目的是预测给定目标在每个视频帧中的位置和形状[1],在机器人视觉、视频监控和无人驾驶等领域有广泛的应用. 但是,由于严重变形、相似物干扰和遮挡等外界因素以及尺度和旋转等目标姿态变化的影响,实现高精度和实时的目标跟踪仍然是一项具有挑战性的任务[2].

收稿日期: 2022-07-26；**录用日期:** 2022-11-10.
基金项目: 国家自然科学基金项目(62072370).
责任编委: 李少远.
†通讯作者. E-mail: hzq@xupt.edu.cn.
*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

近年来,基于深度学习的视觉目标跟踪算法逐渐成为主流,特别是基于Siamese网络的视觉跟踪算法得到了广泛关注. 2016年,Bertinetto等[3]提出了全卷积Siamese网络跟踪算法SiamFC,能够在快速跟踪目标的同时具有很高的跟踪精度和成功率. 结合Siamese网络和区域建议网络(region proposal network, RPN),Li等[4]提出的SiamRPN实现了对跟踪目标的尺度估计和目标定位,在确保跟踪速度的情况下得到了更为精确的目标框. Danelljan等[5]提出了基于最大化重叠度的跟踪算法(accurate tracking by overlap maximization, ATOM),由专门的目标估计和分类组件组成,采用调制机制代替Siamese中的互相关,使得跟踪器具有更强大的判别能力. 在ATOM的IoU (intersection over union, IoU)回归基础上,Bhat等[6]和Danelljan等[7]分别提出了DiMP和PrDiMP算法,其中DiMP实现了端到端的学习,且能够在线更新网络;PrDiMP则将KL散度和信息熵引入DiMP算法中实现对跟踪器的损失计算. 刘如浩等[8]在SiamFC方法的基础上,提出了基于可变形卷积的孪生网络算法,能够强化网络特征提取能力. 陈志旺等[9]通过自适应加权融合目标的深、浅层特征响应,为网络高精度预测目标边界框提供了保障.

上述各方法只利用了基本的CNN网络来提取目标特征,没有关注特征信息在不同空间和通道上的不同分布概率,且相关性网络没有充分利用全局上下文信息,很容易陷入局部最优. 针对这些缺点,近两年学者们提出了多种注意力机制对特征进行增强,并将Transformer结构引入视觉目标跟踪算法,实现对全局特征信息的获取,取得了良好的效果. 如Yan等[10]提出的STARK算法以Transformer为关键结构捕获视频中时空信息间的全局特征依赖关系;Chen等[11]提出的TransT使用了一个新颖的基于注意力的特征融合网络,可有效地组合模板和搜索区域特征;Cui等[12]提出的MixFormer通过注意力机制同时进行特征提取与信息交互,加上一个简单的回归头,直接输出跟踪结果,取得了非常有效的效果. 然而,Transformer在捕获长距离依赖关系时虽然表现良好,但是,对局部特征关系的利用不够全面[13],且通常的视觉Transformer利用绝对位置编码来得到输入序列内的位置关系,但是,绝对位置编码需要与输入序列长度一一匹配,限制了视觉Transformer的灵活性和泛化性.

为了解决上述问题,本文主要做了以下工作.

1) 为了使得跟踪器能够直接关注重要的空间区域和通道,同时避免增大计算量,选择在主干网络ResNet的Block-3层后引入混合注意力机制来获取空间域和通道域中的特征,学习目标特征的空间相互依赖性和通道相互依赖性,并将2个注意力模块的输出进行融合,进一步增强特征表示.

2) 为了增强对局部特征关系的利用,利用多个平行的、空洞率较小的空洞卷积对跟踪序列中的特征进行处理,既避免采样稀疏性,也不会引入过多冗余信息,同时也能够将临近信息通过卷积捕获到生成的特征图中,实现对局部信息的利用.

3) 使用Transformer结构来获取全局信息,增强序列中帧与帧间的联系,实现上下文信息的传播. 同时,在Transformer编码器中加入构建的卷积生成位置编码层,在动态生成长度自适应的位置编码的同时,也一定程度上解决一般位置编码计算量和参数量过大的问题.

4) 在OTB 100、VOT 2018和LaSOT三个数据集上对所提出算法进行实验. 实验结果表明,所提出算法通过构建基于混合注意力的Transformer结构,充分利用局部特征和全局特征,使得跟踪器获得良好的跟踪性能.

1 本文算法

本文提出基于混合注意力的Transformer视觉目标跟踪算法,算法整体框架如图1所示. 整体网络模型主要包含5个部分:基于ResNet 50主干网络的特征提取器(backbone)、混合注意力模块(mixed attention module, MAM)、特征增强模块(feature enhancement module, FEM)、使用卷积位置编码和门控线性单元的Transformer模块(convolutionally positional encoding and GLU transformer, CG-Transformer)以及分类回归头(classification and regression head, Cls & Reg).

图1中:首先,模板和搜索区域经过主干网络ResNet 50提取特征后,引入混合注意力模块MAM对视频序列中的通道和空间信息进行建模;然后,利用特征增强模块FEM获取目标的多尺度特征信息,增强局部特征的表达能力;接着,通过CG-Transformer的编码器-解码器结构进行模板分支与搜索分支间的信息传递和利用,加强帧与帧间的联系,以提高跟踪性能;最后,将编码结果输入至由优化器组成的模型预测器(tracking model)中输出权重,将权重与经过增强的搜索区域特征图进行卷积,最终得到当前帧目标的响应图,并进行回归得出目标边界框.

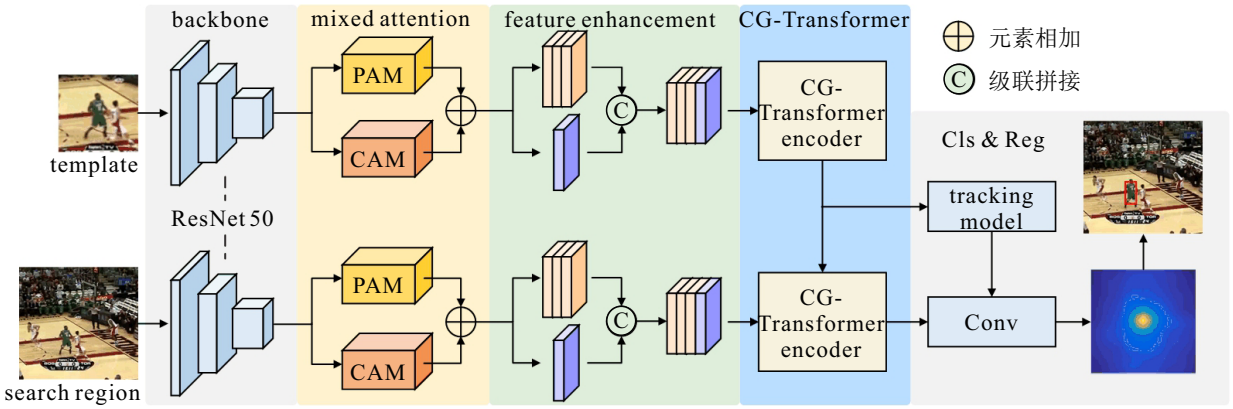


图 1　算法整体框架

下文将详细介绍所提出算法的 3 大模块: 混合注意力模块 MAM、特征增强模块 FEM 以及引入卷积位置编码和门控线性单元的 CG-Transformer 模块. 所提出算法中采用分类回归头与算法 PrDiMP 50[7] 中的结构相同, 此处不再具体介绍.

1.1　混合注意力模块设计

近年来, 视觉注意力机制在多个计算机视觉任务中获得了巨大的成功, 研究者们提出了大量注意力机制对特征进行增强. 受 Non-Local 思想的启发, Fu 等[14] 提出了用于语义分割的 DANet 算法, 从空间角度和通道角度考虑网络架构, 空间注意力可学习到丰富的空间信息; 通道注意力学习并强调目标特征中重要的通道信息.

在跟踪任务中, 感受野的限制使得网络只能捕捉到邻域的信息, 导致跟踪器在目标出视场以及尺度变化时鲁棒性较差, 而空间注意力的引入可使得网络学习如何利用全局上下文信息. 另一方面, 在特征图中每个通道内包含信息的丰富程度不同, 通道注意力的引入可增强通道的特征显著性, 引导网络关注包含信息更多的通道.

浅层特征包含更多的位置、细节信息, 然而, 由于经过的卷积层较少, 其语义性更低, 噪声更多; 深层特征包含语义信息, 但是, 特征图分辨率低且通道数较多, 会导致损失较多信息并增加计算量, 给跟踪任务带来困难. 因此, 所提出算法选择在主干网络 ResNet 50 的 Block-3 层后引入空间与通道混合的注意力模块来提取特征, 并进行元素聚合, 实现了自适应选择重要的区域与通道.

1.1.1　位置注意力模块

位置注意力模块 (position attention module, PAM) 采用自注意力来表征空间信息, 可将更广泛的上下文信息编码至局部特征中, 从而增强其表示能力.

以下是位置注意力模块 PAM 的详细描述. 首先, 给定输入特征 $T_A \in R^{C\times H\times W}$, 利用 1×1 卷积降维后, 分别生成 2 个新的特征映射 $\{T_B, T_C\} \in R^{c\times H\times W}$; 然后, 将 T_B 和 T_C 重组 (reshape) 为 $R^{c\times HW}$; 在 T_B 重组并转置 (transpose) 后得到的 $T_B^{\mathrm{T}} \in R^{HW\times c}$ 与重组后的 T_C 间执行矩阵乘法生成空间注意矩阵, 并应用 Softmax 层计算空间注意力图 $T_S \in R^{HW\times HW}$, 该矩阵对特征的任意 2 个像素间的空间关系进行建模, 有

$$T_{S_{ji}} = \exp(T_{B_i}\cdot T_{C_j}) \Big/ \sum_{i=1}^{N} \exp(T_{B_i}\cdot T_{C_j}). \tag{1}$$

其中: $T_{S_{ji}}$ 衡量第 i 个位置对第 j 个位置的影响, 2 个位置的特征表示越相似, 它们之间的相关性越强; N 为 $H\times W$, 表示特征图的像素数.

同时, 将输入特征 T_A 输入至卷积层, 生成新的特征映射 $T_D \in R^{C\times H\times W}$, 并重组为 $R^{C\times HW}$; 然后, 在 T_D 与空间注意矩阵 T_S 间执行矩阵乘法, 并将结果进行重组得到维度为 $C\times H\times W$ 的张量.

最后, 对上述相乘的结果矩阵乘以比例参数 δ 和原始特征 T_A 进行元素相加运算, 以获得最终输出 $T_P \in R^{C\times H\times W}$, 以此反映远距离背景的最终表征. 整个计算过程可表述为

$$T_{P_j} = \delta \sum_{i=1}^{N} (T_{S_{ji}} T_{D_i}) + T_{A_j}. \tag{2}$$

其中: 比例参数 δ 初始化为 0, 并逐渐学习分配更多权重; T_{P_j} 为 T_P 中的各元素; N 为 $H\times W$, 表示特征图的总像素数.

由式 (2) 可推断, T_P 为空间注意特征图与原始特征的加权和, 可选择性聚合目标的全局上下文信息.

1.1.2　通道注意力模块

通道注意力模块 (channel attention module, CAM) 采用自注意力来建模通道间的相互依赖关系, 通过利用不同通道的特征图间的相互依赖关系, 可改善特定语义的特征表示, 使得网络重点关注某些权重值大的通道.

以下是通道注意力模块CAM的详细描述. 与位置注意力模块不同,本文直接计算得到通道注意力图$T_X \in R^{C\times C}$. 首先,给定输入特征$T_A \in R^{C\times H\times W}$,将$T_A$重组为$\{T_U, T_V\} \in R^{c\times HW}$;然后,在$T_U$的转置与$T_V$间执行矩阵乘法;最后,使用Softmax层来获得通道注意图$T_X \in R^{C\times C}$,即

$$T_{X_{ji}} = \frac{\exp(T_{U_i}\cdot T_{V_j})}{\sum_{i=1}^{C}\exp(T_{U_i}\cdot T_{V_j})}, \tag{3}$$

其中$T_{X_{ji}}$衡量第i个通道对第j个通道的影响.

此外,将T_A再次进行重组得到$T_W \in R^{C\times HW}$,在T_X与T_W间执行矩阵乘法,并将其结果$R^{C\times HW}$重组为$R^{C\times H\times W}$;然后,将结果乘以比例参数ε,并与T_A执行元素求和运算,以获得最终输出$T_Q \in R^{C\times H\times W}$,具体计算如下式所示:

$$T_{Q_j} = \varepsilon\sum_{i=1}^{C}(T_{X_{ji}}T_{W_i}) + T_{A_j}. \tag{4}$$

其中:比例参数ε初始化为0,并逐渐学习分配更多权重;T_{Q_j}为T_Q中的各元素;C为输入特征的通道数.

式(4)表明,每个通道的最终特征是所有通道的特征与原始特征的加权和,从而建立了特征映射间的长期语义依赖关系.

1.1.3 特征聚合模块

为了充分利用空间维度和通道维度中的信息,使得网络直接关注重要的空间区域和通道,使用上文的PAM和CAM模块对ResNet 50中Block-3层的输出特征进行并行处理,且将处理后的特征进行聚合. 整个混合注意力模块最终输出聚合特征T_R,计算过程如下式所示:

$$T_R = \eta(\mu(\mathrm{PAM}(T_P)) + \upsilon(\mathrm{CAM}(T_Q))). \tag{5}$$

其中:T_P和T_Q分别为模块PAM和CAM的输出,μ和υ均为3 × 3卷积、BN和ReLU处理,η为Dropout与1 × 1卷积,PAM(·)为位置注意力计算,CAM(·)为通道注意力计算. 将聚合后的特征T_R应用于后续运算中,使得网络可自适应关注重要的区域和通道.

1.2 特征增强模块设计

近期有学者在工作中提及Transformer的缺点:全局的、长程的注意力机制Transformer能够捕获patch间的长距离依赖关系,但是很容易忽略图像的局部特有性质. 为了缓解Transformer的这个缺点,本文考虑从增强局部特征这一方面解决. 受分割领域的空洞卷积池化金字塔启发,本文引入特征增强模块(feature enhancement module, FEM)来实现图像的多尺度特征提取. 将更广泛的上下文信息通过多个不同空洞率的空洞卷积整合至特征图内,从而增强对目标局部特征的表达能力,为后续引入的Transformer结构提供包含丰富局部信息的特征,一定程度上弥补了Transformer的不足. 特征增强模块的结构如图2所示,其中输入input为经过MAM处理后输出的聚合特征T_R.

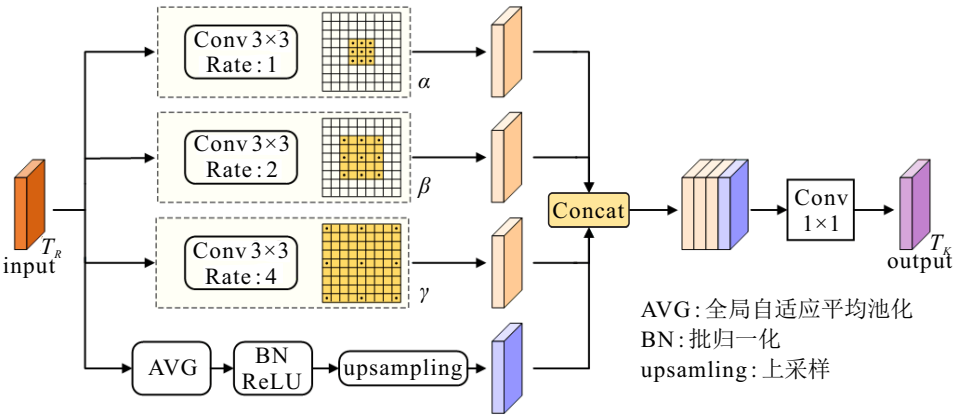


图 2 特征增强模块结构

特征增强模块由3个不同空洞率的空洞卷积$\{\alpha,\beta,\gamma\}$分支和1个含有全局自适应平均池化(global adaptive average pooling, AVG)的分支构成. 与已有使用空洞率的特征增强模块不同,本文采用了多个空洞率较小的空洞卷积,以避免引入过多冗余信息. 设置特征增强模块的参数如下:空洞卷积α空洞率为1,β空洞率为2,γ空洞率为4,3个空洞卷积的卷积核尺寸均为3 × 3,padding值与空洞率选取值相同. 通过不同的填充和膨胀因子,可获得不同尺度的感受野,提取到多尺度的信息;然后,利用全局自适应平均池化(AVG)保留全局的特征信息,并上采样(upsampling)到原来的尺寸;最后,将多个特征图进行Concat拼接,利用1 × 1卷积对拼接后的特征进行整合并降维,得到与输入特征相同维度的特征T_K作为

该模块最终的输出. 利用该特征增强模块, 可提取到输入图像的多尺度信息, 增强局部特征的表达能力.

1.3　CG-Transformer

Transformer 在自然语言处理领域中发挥出的强大作用表明了编码解码结构的有效性. 为了增强视频序列中帧与帧间的联系, 本文引入 Transformer 结构, 将编码器与解码器分别分配至跟踪框架的 2 个并行分支中, 对模板帧和搜索区域进行特征增强. 具体地, 模板分支每次输入多个不同样本, 经过骨干网络和注意力模块处理后, 生成多个模板特征并将其共同输入至编码器, 相互聚合生成编码特征, 然后将获得的编码模板特征输入至解码器, 与当前搜索区域特征进行交叉自注意力计算, 实现了时间信息的传播, 且避免了使用长度固定的绝对位置编码和需要大量计算的相对位置编码, 在 Transformer 编码器中加入构建的卷积位置编码模块 (convolutionally positional encoding, CPE), 动态生成长度自适应的位置编码, 保证了 Transformer 的灵活性.

下面将详细介绍所提出卷积位置编码模块和改进后的 Transformer 结构.

1.3.1　卷积位置编码模块

通常, Transformer 利用绝对位置编码来得到输入序列中元素的位置关系, 这种方式的缺点是需要与输入序列长度一一匹配, 且难以满足计算机视觉任务需要的平移不变性. 为了解决此问题, 有学者提出了可学习的相对位置编码, 使得模型在学习过程中持续考虑位置关系的影响, 但是, 由此带来的问题是会在每层 Transformer 中引入新的参数导致计算量增加.

受文献 [15] 启发, 本文不同于固定长度的绝对位置编码, 实现了利用简单的二维卷积方式自然地处理大小不固定的输入, 且相比于对每个 token 间均需要计算相对位置关系的相对位置编码, 使用带有 0 填充的二维卷积来获取 tokens 间的相对位置关系, 参数量和计算量更小. 卷积能够获取 tokens 间相对位置关系的原因是卷积核尺寸越大, 包含的周边信息越多, 可理解为将原图像顺序排列的特征赋予了空间信息, 使得在有相似物干扰的情况下, 能够很好地进行辨别. 实现过程如图 3 所示.

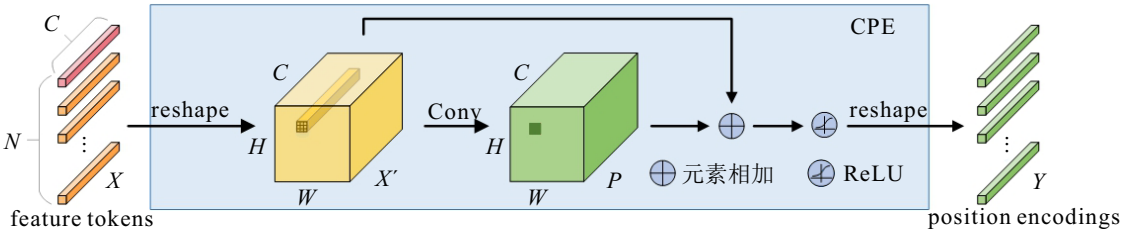


图 3　卷积位置编码模块结构

为了满足卷积处理的条件, 首先, 将扁平化输入序列 $X \in R^{C\times N}$ 重组为三维图像空间中的 $X' \in R^{C\times H\times W}$ 以便于卷积处理, 其中 $N = H \times W$; 然后, 将一个卷积 Conv 重复应用于 X' 中的局部块, 以产生位置编码 $P \in R^{C\times H\times W}$; 接着, 将 P 与 X' 相加构成一个小型残差块并使用 ReLU 函数进行激活; 最后, 再重组为输入序列的形式 $Y \in R^{C\times N}$, 便于 Transformer 后续的处理. 计算过程可由下式表示:

$$Y = \mathrm{CPE}(X) = \omega(\mathrm{ReLU}(\mathrm{Conv}[\varphi(X)] + \varphi(X))). \tag{6}$$

其中: φ 和 ω 均为特征重组操作, 前者将二维序列转换为三维张量, 后者将三维张量转换为二维序列; $\mathrm{CPE}(\cdot)$ 可通过卷积核大小为 k、padding 值为 $\dfrac{k-1}{2}$ 的 0 填充的二维卷积 Conv 来有效实现 ($k \geqslant 3$).

1.3.2　CG-Transformer 整体结构

图 4 为 CG-Transformer 结构, 该部分由 2 层编码器和 1 层解码器组成. 为了减少 Transformer 结构的计算量, 编码器采用单头自注意力模块、门控线性单元 (gate linear unit, GLU) 和第 1.3.1 节描述的卷积位

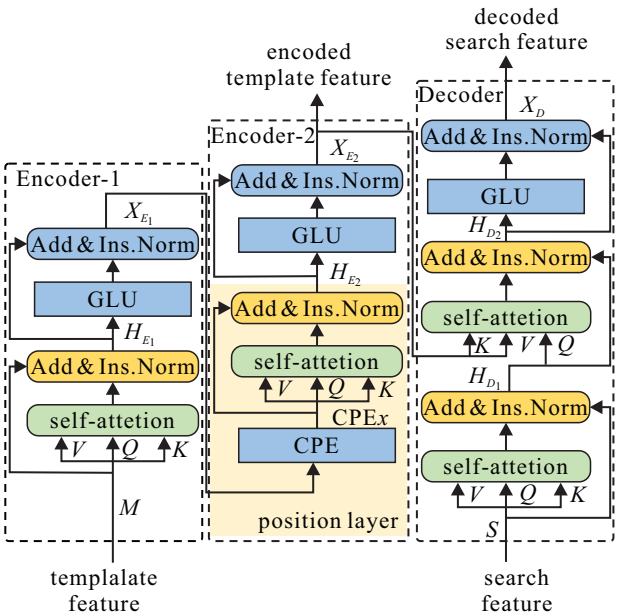


图 4　CG-Transformer 整体结构

置编码模块组合组成,解码器由单头自注意力模块和门控线性单元(GLU) 组成.

在整个Transformer结构中, 自注意力(self-attention)的使用起着关键作用. 首先,编码器和解码器中分别输入模板特征M和搜索区域特征S,将输入特征进行矩阵变换得到查询Q、键K和值V作为自注意力部分的输入;然后,采用点乘来计算Q与K间的相似度,经过Softmax后得到一组注意力权重;最后,将权重与V进行点乘,得到带有注意力权重的向量,具体计算过程如下式所示:

$$\text{Attention}(Q,K,V)=\text{Softmax}\Big(\frac{QK^{\text{T}}}{\sqrt{d_k}}\Big)V, \quad (7)$$

其中d_k为注意矩阵关键维度.

为了节省计算成本,本文采用单头自注意力模块替代以往Transformer的多头自注意力,图4中H_{E_1}可表示为

$$H_{E_1}=\text{Ins.Norm}(\text{Attention}(Q_1,K_1,V_1)+M). \quad (8)$$

其中:M为模板特征,Q_1、K_1和V_1来自模板特征的线性变换,Attention$(\cdot)$计算过程如式(7)所示,Ins.Norm$(\cdot)$表示联合对一个图像块的所有嵌入进行L 2归一化. 则Encoder-1的输出X_{E_1}可表示为

$$X_{E_1}=\text{Ins.Norm}(\text{GLU}(H_{E_1})+H_{E_1}), \quad (9)$$

其中GLU$(\cdot)$为门控单元,在Transformer结构中代替前馈网络,其计算公式如下式所示:

$$\text{GLU}(x,W,V,W_2)=(\text{GELU}(xW)\otimes xV)W_2. \quad (10)$$

其中:x为输入矩阵;W、V和W_2为不同的线性变换单元;$\otimes$为矩阵乘法;GELU$(\cdot)$为GELU激活函数,具体计算如下式所示:

$$\text{GELU}(z)=0.5z\Big(1+\tanh\Big(\sqrt{\frac{2}{\pi}}(z+0.044\,715z^3)\Big)\Big), \quad (11)$$

其中z为输入矩阵.

Encoder-2的输出X_{E_2}可表示为

$$X_{E_2}=\text{Ins.Norm}(\text{GLU}(H_{E_2})+H_{E_2}). \quad (12)$$

其中:GLU$(\cdot)$的计算如式(10)所示;H_{E_2}为Encoder-1的输出X_{E_1}经过位置编码层后的输出,H_{E_2}的具体计算如下式所示:

$$H_{E_2}=\text{Ins.Norm}(\text{Attention}(Q_2,K_2,V_2)+\text{CPE}_X). \quad (13)$$

其中:CPE$_X$为将X_{E_1}输入至上文CPE模块后得到的输出,计算过程如式(6)所示;Q_2、K_2、V_2是由CPE$_X$经过一系列线性变换得到的;Ins.Norm$(\cdot)$为实例归一化;Attention$(\cdot)$计算过程如式(7)所示.

本文使用传统Transformer的解码器,将搜索区域特征作为解码器的输入,与编码器类似,首先,将特征进行线性变换,送入自注意力模块得到H_{D_1},有

$$H_{D_1}=\text{Ins.Norm}(\text{Attention}(Q_3,K_3,V_3)+S). \quad (14)$$

其中:S为搜索区域特征,Q_3、K_3和V_3均由搜索区域特征S线性变换后得到. 然后,将模板编码特征X_{E_2}和经过自注意力后的搜索特征H_{D_1}共同输入至自注意力模块来计算二者的交叉注意力矩阵,计算过程如下式所示:

$$H_{D_2}=\text{Ins.Norm}(\text{Attention}(Q_4,K_4,V_4)+H_{D_1}). \quad (15)$$

其中:Q_4为来自H_{D_1}的线性变换,K_4和V_4为来自Encoder-2的输出X_{E_2}的线性变换. 通过计算二者的交叉注意力矩阵,可建立帧之间的像素级对应关系,将模板区域的特征传递至搜索区域,从而实现上下文信息的传播. 最后,利用GLU[16]作为前馈网络,更好地表示特征间的非线性关系. 解码器Decoder的输出X_D可表示为

$$X_D=\text{Ins.Norm}(\text{GLU}(H_{D_2})+H_{D_2}). \quad (16)$$

2 实验结果与分析

2.1 实验细节

所提出算法均在Ubuntu 16.04系统上使用PyTorch 1.4.0深度学习框架和CUDA 10.1深度学习架构,基于Python编程语言实现. 实验硬件环境为Intel (R) Core (TM) i5-8 400 k 2.80 GHz处理器、内存16 GB的计算机,并使用GPU (NVIDIA TITAN Xp)进行加速.

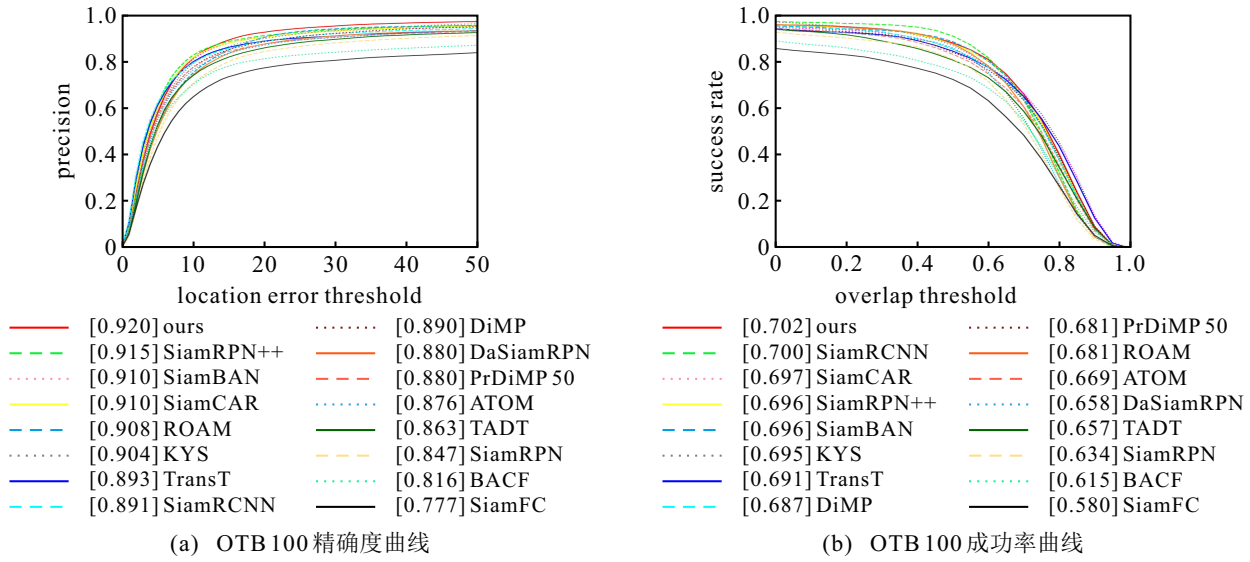
整体网络训练使用的数据集由TrackingNet、LaSOT训练集、GOT-10 k和COCO组成,主干网络使用ImageNet上预训练的ResNet 50. 将Batch Size设置为8, epoch设置为50,每个epoch有30 000个训练样本,使用高斯-牛顿迭代展开最速下降法训练,并使用ADAM优化器进行优化,初始学习率为0.01,每15个epoch进行衰减,衰减系数设为0.2.

2.2 定量分析

为了评估所提出算法的有效性,在OTB 100、VOT 2018和LaSOT测试集上进行了测试评估. 下面对结果进行详细分析.

2.2.1 OTB 100数据集实验结果

OTB 100数据集是比较常见的目标跟踪性能评价的测试数据集,共包含100个带有标注的跟踪序列. 平均每个视频590帧.将所提出方法与近年来一些代表性的跟踪器在OTB 100数据集上进行对比,实



(a) OTB 100 精确度曲线　　(b) OTB 100 成功率曲线

图 5　OTB 100 上不同算法的精确度和成功率曲线

验结果如图5所示. 由图5可见,所提出算法相比于基线算法 PrDiMP 50 在 OTB 100 数据集上的成功率提升了4.0%,精度提升了2.1%.

表1为在OTB 100数据集中,11种不同属性下算法跟踪的成功率,最优结果加粗显示,次优结果由实下划线表示,排名第3的结果由虚下划线表示. 其中:背景杂乱属性(background clutter, BC)下相比于基线算法成功率提高了8.6%,出视野属性(out-of-view, OV)下相比于基线算法成功率提高了6.6%. 这表明利用局部信息与全局信息互补的Transformer,可使得算法有很好的鲁棒性,能够有效避免杂乱背景带来的影响,同时在目标重回视野时能够及时找回目标. 相比于基线算法 PrDiMP 50,所提出算法在各属性上表现均更有优势,其余算法在形变、遮挡、平面外旋转等多个复杂情况下的表现普遍低于所提出算法.

表 1　OTB 100 上不同算法 11 种属性的成功率对比结果

algorithm	BC	DEF	FM	IPR	IV	LR	MB	Occ	OPR	OV	SV
SiamFC	0.525	0.546	0.575	0.567	0.536	0.574	0.587	0.532	0.550	0.473	0.560
SiamRPN	0.598	0.619	0.603	0.632	0.652	0.663	0.624	0.589	0.628	0.548	0.621
BACF	0.605	0.572	0.599	0.582	0.622	0.539	0.575	0.566	0.578	0.548	0.573
PrDiMP50	0.618	0.661	0.671	0.694	0.684	0.670	0.700	0.647	0.667	0.618	0.697
TADT	0.619	0.602	0.655	0.618	0.674	0.664	0.668	0.638	0.644	0.623	0.650
DiMP	0.656	0.663	0.694	0.685	0.691	0.593	0.719	0.667	0.667	0.657	0.687
ATOM	0.631	0.630	0.662	0.652	0.679	0.693	0.658	0.648	0.643	0.612	0.681
ROAM	0.670	0.642	0.685	0.667	0.680	0.677	0.678	**0.677**	0.656	0.631	0.667
DaSiamRPN	0.642	0.645	0.621	0.652	0.655	0.636	0.625	0.611	0.644	0.537	0.637
SiamRPN++	0.691	0.663	0.686	0.694	0.713	0.696	0.703	0.663	0.680	0.646	0.694
KYS	0.662	**0.675**	0.693	0.699	0.720	0.701	0.709	0.657	0.684	0.635	0.693
SiamBAN	0.680	0.662	0.687	**0.717**	**0.724**	**0.719**	0.698	0.648	**0.687**	0.640	0.693
SiamCAR	0.672	0.651	0.703	0.703	0.703	0.712	0.715	0.653	0.679	0.661	0.698
SiamRCNN	0.691	0.645	0.702	0.699	0.716	0.691	**0.735**	0.666	0.684	0.677	0.719
TransT	0.627	0.644	**0.711**	0.687	0.669	0.714	0.731	0.656	0.668	0.684	**0.727**
ours	**0.704**	0.670	0.687	0.699	0.704	0.682	0.713	0.668	0.686	**0.684**	0.711

2.2.2　VOT 2018 数据集实验结果

VOT 2018 由 60 个具有不同挑战因素的视频序列组成,评价指标包括准确性(A)、鲁棒性(R)、丢失数(lost numbers)和期望平均重叠率(EAO). 遵循 VOT 2018 的评估标准,所提出算法在 VOT 2018 数据集上对多个算法进行对比,结果如图6所示. 由图6可见,与 PrDiMP 50 相比,所提出算法在 EAO 值上提高了1.7%. 与近几年的主流跟踪算法相比,所提出算法

表现出较高的EAO值,在VOT 2018上的整体性能表现良好.

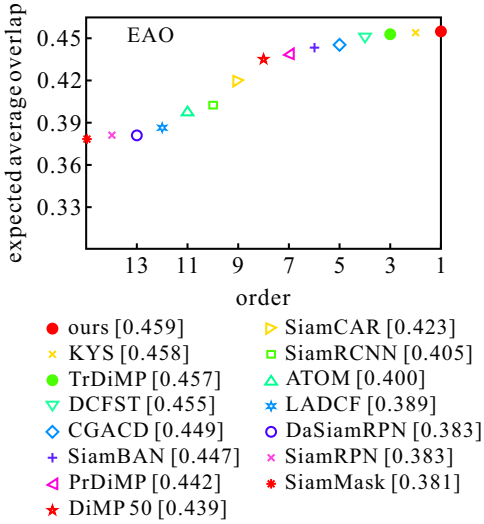


图 6 VOT 2018上不同算法的期望平均重叠率排名

2.2.3 LaSOT数据集实验结果

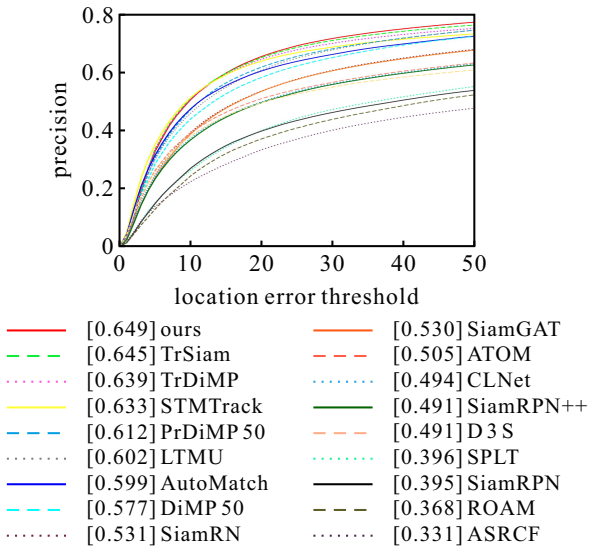
LaSOT是一个具有高质量注释的大规模单目标跟踪数据集,测试集包含280个长视频(平均2 500帧),因此,跟踪器的鲁棒性及其准确性对复杂的场景极为重要. 图7为所提出算法与多个跟踪器的对比结果. 由图7可见,所提出算法(ours)取得了64.9%的精确度,63.1%的成功率和71.6%的归一化精度. 相比于基准算法PrDiMP50分别取得了3.7%、2.8%和2.3%的提升,也超越了TrSiam和TrDiMP等基于Transformer的跟踪算法,反映出所引入的混合注意力机制可为跟踪任务提供优秀的特征表示,使用空洞卷积的特征增强模块和使用CPE的Transformer结构在可以捕获长距离信息的同时也能够增强局部信息的依赖关系,从而提高跟踪器性能.

2.3 定性分析

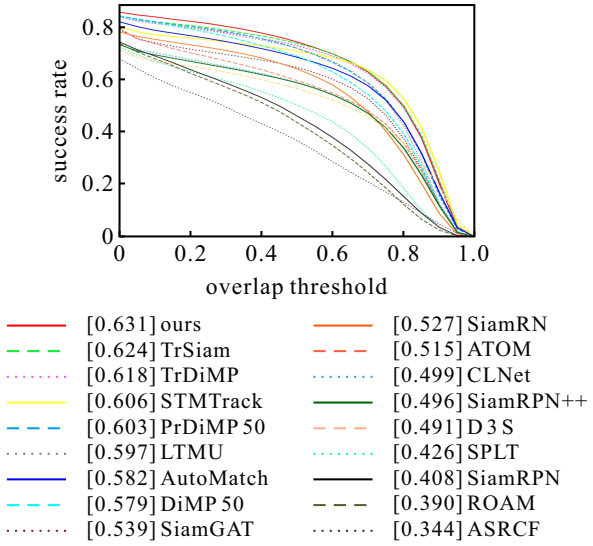
为了定性表明所提出算法的优越性,从OTB 100数据集中选择了4个包含多个场景的视频序列,比较了4个近年流行的跟踪器(SiamRPN[4]、ATOM[5]、DiMP 50[6]和PrDiMP 50[7])与所提出算法(ours)的跟踪结果. 图8为各算法在4个视频序列定性评估的跟踪结果. 所提出算法通过对全局和局部信息的利用,使得跟踪器有较强的稳健性和可靠性.

2.4 消融实验

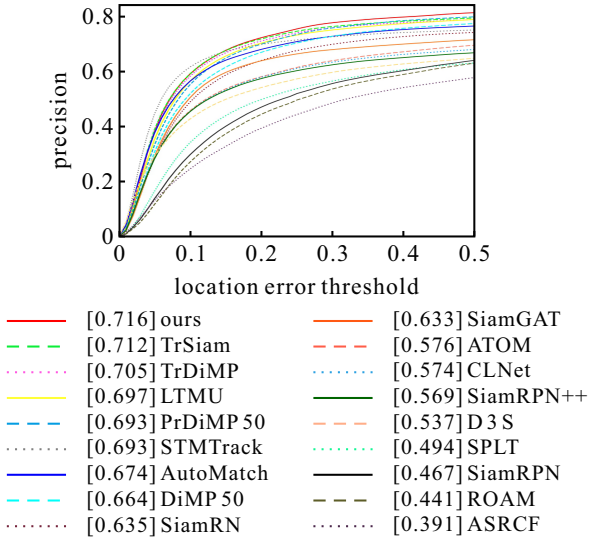
为了测试所提出算法的3个模块对跟踪性能的影响,在LaSOT测试集上进行了消融实验. 实验环境与模型主要超参数均一致,具体如表2所示. 其中第1行为本地复现结果,由于运行环境以及硬件设施不同等原因,本地复现结果与文献[7]提供的结果有所差异(文献[7]中PrDiMP 50算法在LaSOT数据集上的



(a) LaSOT精确度曲线



(b) LaSOT成功率曲线



(c) LaSOT归一化精度曲线

图 7 LaSOT上不同算法的精确度、成功率和归一化精度曲线

成功率为59.8%). 实验结果表明:3个模块能够有效提高跟踪器的成功率,且三者结合后能够进一步提高

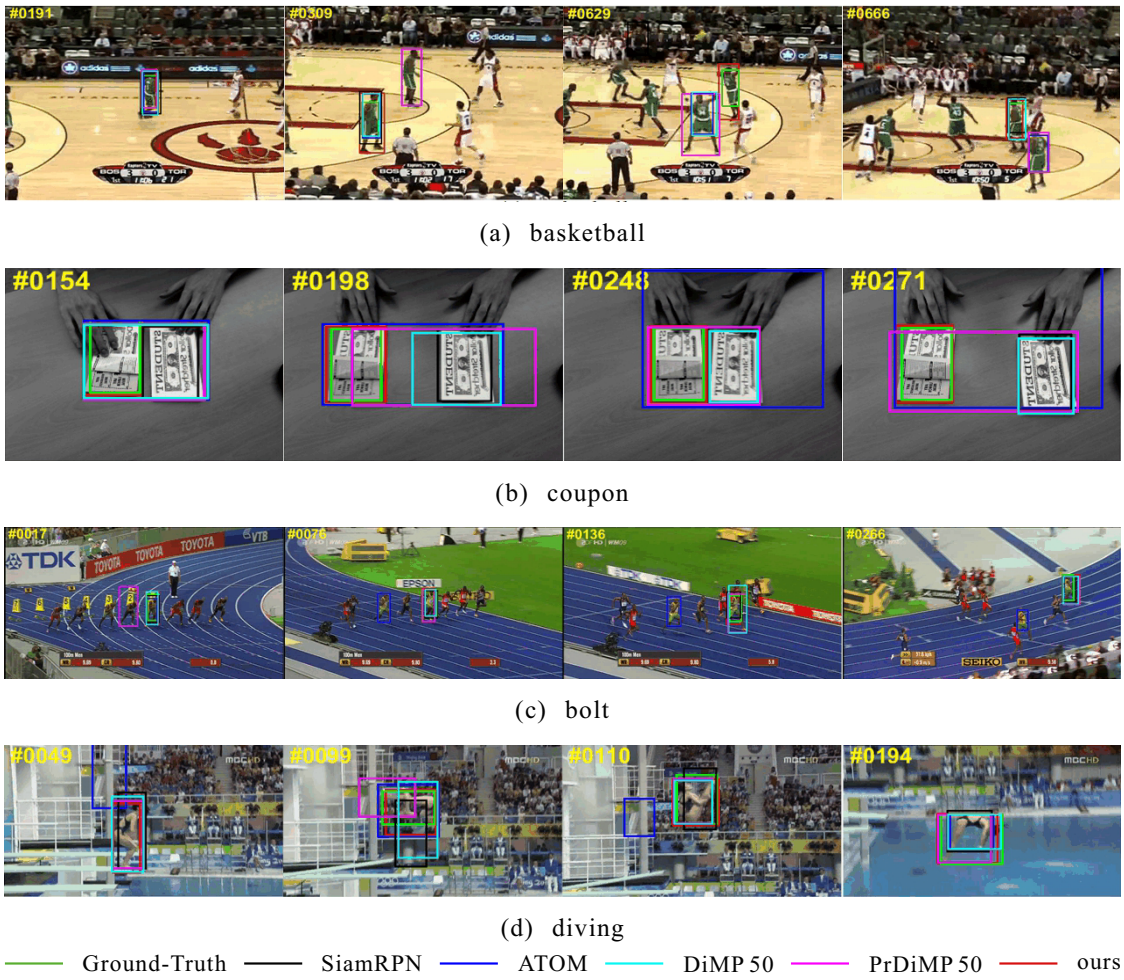


图 8　多个跟踪算法定性分析结果对比

表 2　LaSOT 数据集上消融实验结果

PrDiMP 50	MAM	FEM	CG-Transformer	Succ/%	FPS
✓				60.3	30
✓	✓			61.2	29.47
✓		✓		60.9	29.33
✓			✓	62.4	27.81
✓	✓	✓	✓	63.1	26

跟踪器的性能, 与此同时, 由于均采取了计算量较少的方式来对特征进行处理, 使得跟踪速度也达到了实时要求.

2.5　各算法速度与成功率的比较

为了验证所提出算法的跟踪实时性, 将所提出算法与目前主流的跟踪算法在 OTB 100 数据集上进行了成功率和速度的对比实验. 由图 9 可见, 所提出算法在取得较高成功率的同时, 速度保持在 26 帧/s, 实现了实时跟踪目标. 由于加入了需要大量计算的 Transformer 结构, 使得所提出算法相比于基准算法 PrDiMP 50 速度稍有降低, 但是, 所提出算法摒弃了传统 Transformer 的多头自注意力而采用单头注意力, 保证了算法的实时性.

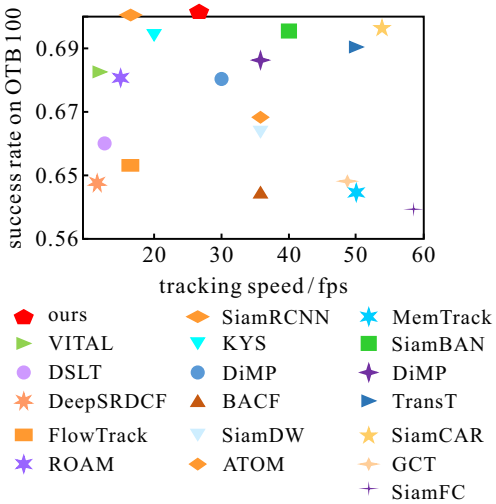


图 9　在 OTB 100 上各算法性能与速度对比

3　结　论

针对现有视觉目标跟踪算法对相似物判别力低、运算量大且无法完全利用目标局部与全局关系等问题, 设计了一种基于混合注意力的 Transformer 视觉目标跟踪算法. 在主干网络中间层嵌入混合注意力提取更为稳健的特征; 使用多个较小空洞率的空洞卷积实现图像的多尺度特征提取, 利用 Transformer

编码解码器在孪生网络的模板分支与搜索分支特征间进行信息传递,提高了目标搜索的准确性;利用0填充的卷积方式实现更灵活且有效的位置编码,提高了网络对于相似物的判别能力,进一步提升了跟踪器的性能.

大量经典数据集上的实验结果表明了所提出算法的有效性,且能够实现实时跟踪效果. 但是,在实验中发现,当目标发生快速运动或图像分辨率过低时,跟踪器不能准确找到目标进行跟踪,该问题是下一步需要继续进行研究的重点.

参考文献(References)

[1] 李玺, 查宇飞, 张天柱, 等. 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.
(Li X, Zha Y F, Zhang T Z, et al. Survey of visual object tracking algorithms based on deep learning[J]. Journal of Image and Graphics, 2019, 24(12): 2057-2080.)

[2] 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.
(Lu H C, Li P X, Wang D. Visual object tracking: A survey[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 61-76.)

[3] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]. European Conference on Computer Vision. Munich, 2016: 850-865.

[4] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8971-8980.

[5] Danelljan M, Bhat G, Khan F S, et al. Atom: Accurate tracking by overlap maximization[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. LongBeach, 2019: 4660−4669.

[6] Bhat G, Danelljan M, van Gool L, et al. Learning discriminative model prediction for tracking[C]. Proceedings of the IEEE International Conference on Computer Vision. Seoul, 2019: 6182-6191.

[7] Danelljan M, van Gool L, Timofte R. Probabilistic regression for visual tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 7183-7192.

[8] 刘如浩, 张家想, 金辰曦, 等. 基于可变形卷积的孪生网络目标跟踪算法[J]. 控制与决策, 2022, 37(8): 2049-2055.
(Liu R H, Zhang J X, Jin C X, et al. Target tracking based on deformable convolution Siamese network[J]. Control and Decision, 2022, 37(8): 2049-2055.)

[9] 陈志旺, 王莹, 宋娟, 等. 特征响应权重自适应的IoU网络跟踪算法改进[J]. 控制与决策, 2022, 37(7): 1752-1762.
(Chen Z W, Wang Y, Song J, et al. Improvement of IoU network tracking with adaptive weighted characteristic responses[J]. Control and Decision, 2022, 37(7): 1752-1762.)

[10] Yan B, Peng H W, Fu J L, et al. Learning spatio-temporal transformer for visual tracking[C]. Proceedings of the IEEE International Conference on Computer Vision. New York, 2021: 10448-10457.

[11] Chen X, Yan B, Zhu J W, et al. Transformer tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York, 2021: 8126-8135.

[12] Cui Y T, Jiang C, Wang L M, et al. MixFormer: End-to-end tracking with iterative mixed attention[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 13608-13618.

[13] Guo J Y, Han K, Wu H, et al. CMT: Convolutional neural networks meet vision transformers[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, 2022: 12165-12175.

[14] Fu J, Liu J, Tian H J, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 3146-3154.

[15] Chu X X, Tian Z, Zhang B, et al. Conditional positional encodings for vision transformers[J/OL]. 2021, arXiv: 2102.10882.

[16] Shazeer N. Glu variants improve transformer[J/OL]. 2020, arXiv: 2002.05202.

作者简介

侯志强(1973−), 男, 教授, 博士生导师, 从事图像处理、计算机视觉等研究, E-mail: hzq@xupt.edu.cn;

郭凡(1999−), 女, 硕士生, 从事视觉目标跟踪算法的研究, E-mail: guofanxs@126.com;

杨晓麟(1997−), 男, 硕士生, 从事视觉目标跟踪算法的研究, E-mail: yxlxupt@126.com;

马素刚(1982−), 男, 博士生, 从事计算机视觉、机器学习等研究, E-mail: msg@xupt.edu.cn;

范九伦(1964−), 男, 教授, 博士生导师, 从事通信与信息技术等研究, E-mail: jiulunf@xupt.edu.cn.