控制与决策
Control and Decision

基于稀疏注意力的孪生网络目标跟踪算法

陈志旺, 杨天宇, 曹索航, 吕昌昊, 彭勇

引用本文:
陈志旺, 杨天宇, 曹索航, 等. 基于稀疏注意力的孪生网络目标跟踪算法[J]. 控制与决策, 2024, 39(12): 4017-4026.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2023.1352

基于稀疏注意力的孪生网络目标跟踪算法

陈志旺[1,2†], 杨天宇[1,2], 曹索航[1,2], 吕昌昊[3], 彭 勇[4]

(1. 燕山大学 智能控制系统与智能装备教育部工程研究中心，河北 秦皇岛 066004；2. 燕山大学 工业计算机控制工程河北省重点实验室，河北 秦皇岛 066004；3. 燕山大学 电力电子节能与传动控制河北省重点实验室，河北 秦皇岛 066004；4. 燕山大学 电气工程学院，河北 秦皇岛 066004)

摘 要: 利用改进的Inception-Resnet-V2 (IRV2) 网络和局部-全局-局部 (local-global-local, LGL) 模块设计一种结合CNN和Transformer编码结构的孪生网络SiamLGL (siamese local-global-local network) 用于目标跟踪. 首先, 算法特征提取部分采用改进后的IRV2网络, 由于网络的层数更深, 图片经过IRV2网络提取的特征较浅层网络提取的特征效果更优, 特征融合部分采用深度互相关将特征图上的信息进行融合; 其次, 融合后的特征图利用LGL模块获取目标的全局和局部信息, 模块内部采用两个编码器串联, 第1个编码器利用深度可分离卷积获取目标的局部信息, 第2个编码器利用自注意力获取图片的全局特征, 为了降低自注意力结构的时间复杂度, 采用稀疏注意力的方式进行计算, 在降低时间复杂度的同时保证网络的精度; 最后将特征图输入至分类回归网络中, 生成对应的目标位置, 其中分类网络采用二元交叉熵损失函数, 回归网络采用 Distance-IoU (DIoU) 作为损失函数. 算法在GOT-10k、LaSOT、TrackingNet、UAV123、OTB100和VOT2019等6个公开数据集上进行实验评估, 结果验证了算法的有效性.

关键词: 目标跟踪；孪生网络；Inception-Resnet-V2网络；稀疏注意力；Distance-IoU损失

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2023.1352

引用格式: 陈志旺, 杨天宇, 曹索航, 等. 基于稀疏注意力的孪生网络目标跟踪算法[J]. 控制与决策, 2024, 39(12): 4017-4026.

Siamese network object tracking algorithm based on sparse attention

CHEN Zhi-wang[1,2†], *YANG Tian-yu*[1,2], *CAO Suo-hang*[1,2], *LV Chang-hao*[3], *PENG Yong*[4]

(1. Engineering Research Center of the Ministry of Education for Intelligent Control System and Intelligent Equipment, Yanshan University, Qinhuangdao 066004, China; 2. Key Laboratory of Industrial Computer Control Engineering of Hebei Province, Yanshan University, Qinhuangdao 066004, China; 3. Key Lab of Power Electronics for Energy Conservation and Motor Drive of Hebei Province, Yanshan University, Qinhuangdao 066004, China; 4. School of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract: An improved Inception-Resnet-V2 (IRV2) network and local-global-local (LGL) module are used to design a siamese network structure based on CNN and Transformer coding structure for object tracking-SiamLGL (siamese local-global-local network). Firstly, due to the improved (IRV2) network with deep layers, the features extracted by the IRV2 network in the images are better than those extracted by the shallow network. Furthermore, the information on the feature map is fused through deep intercorrelation. Secondly, the fused feature map uses the LGL module to obtain the global and local information of the object, and two encoder layers are used in series inside the module, the first encoder layer with depth-separable convolution obtain the local information of the object, and the second encoder layer with self-attention obtain the global features of the picture. In order to reduce the time complexity of the self-attention structure, the sparse attention approach is used for the computation, which ensures the accuracy of the network while reducing the time complexity. Finally, the feature map is input to the classification and regression network to generate the corresponding object location. The classification network adopts the binary cross entropy loss function, and the regression network adopts Distance-IoU (DIoU) as the loss function. The algorithm is evaluated on six public datasets : GOT-10k, LaSOT, TrackingNet, UAV123, OTB100 and VOT2019. The experimental results verify the effectiveness of the proposed algorithm.

Keywords: object tracking；siamese network；Inception-Resnet-V2 network；sparse attention；Distance-IoU loss

收稿日期: 2023-09-22；**录用日期:** 2024-01-16.

基金项目: 河北省研究生专业学位精品教学案例(库)项目(KCJPZ2023012)；国家自然科学基金项目(61573305)；河北省自然科学基金项目(F2022203038, F2019203511).

†通讯作者. E-mail: czwaaron@ysu.edu.cn.

0 引 言

计算机视觉作为人工智能领域一个重要的分支,其目标跟踪任务一直是研究热点之一. 它根据第1帧中给出的信息预测后续帧中目标的位置等信息,在智能驾驶[1]、人机交互[2]、视频监控[3]等领域具有重要的学术价值和广阔的商业前景. 虽然现有的基于孪生网络的目标跟踪算法在精度和准确率方面都已经比早期的算法有明显提升,但是在一些复杂的场景下仍然存在跟踪不准确的现象[4],包括目标遮挡与消失、外观变化、目标移动和背景干扰等.

基于卷积神经网络的单目标跟踪算法可以分为有锚框网络和无锚框网络两类. 在有锚框跟踪算法中,Bertinetto等[5]提出的SiamFC通过端到端的方式实现实时跟踪,为后续基于孪生网络的目标跟踪算法奠定了强有力的基础. Li等[6]将目标检测中的区域候选网络引入目标跟踪中,使算法可以预测尺度变化的目标. DaSiamRPN[7]通过在SiamRPN中引入干扰物识别模型和扩充训练样本提升算法的泛化能力,获得了较好的长时跟踪效果. 为了避免计算过程中引入大量超参数,研究人员开发了无锚框的跟踪算法. Xu等[8]提出4条跟踪算法的设计准则并据此提出了基于无锚框的SiamFC++算法,通过添加质量评估分支去除锚框所需要的先验知识并获得了较好的效果. STMTrack[9]针对大多数算法只使用第1帧的图片作为模板导致模板历史信息不足的问题提出一种记忆存储分支,通过时空注意力机制解决了跟踪过程中因目标外观变化而引起的跟踪失败问题.

近年来,Transformer[10]在目标跟踪任务中的应用越来越多. DualTFR[11]算法利用多个局部注意力堆叠,并且与全局注意力以及交叉注意力配合,测试结果超越用CNN做特征提取网络的跟踪算法. SwinTrack[12]摒弃了大部分跟踪算法利用CNN作为特征提取网络的方法,设计一种完全基于Transformer注意力的跟踪算法,通过获取更丰富的语义信息收获到更好的特征,从而成为当时整体效果最优的跟踪算法. TransT[13]利用Transformer的编码结构设计出Ego-Context Augment (ECA)和Cross-Feature Augment (CFA)两个模块,采用交叉注意的方式设计出更适合目标跟踪的孪生网络. TMT[14]在模板分支中通过使用多个模板分支并在自注意力部分进行权重共享进一步提升了跟踪的效果. MixFormer[15]将传统孪生网络中特征融合和特征提取网络用混合注意力机制替代,提出了一种基于注意力的端到端跟踪模型.

虽然Transformer结构在目标跟踪任务中应用越来越广泛,但是因为其结构原因会产生较大的时间复杂度,所以制约了Transformer结构在孪生网络中的应用. 为此,本文提出一种基于Transformer编码结构的孪生网络单跟踪算法Siamese Local-Global-Local Network (SiamLGL),采用深度为132层的特征提取网络和Transformer的编码结构,通过使用DIoU[16]作为损失函数,希望提升模型在跟踪任务中的表现. 主要贡献如下:

1) 利用改进的特征提取IRV2[17]网络,充分提取图片中目标的特征,提升跟踪任务的成功率.

2) 提出基于稀疏注意力LGL[18]的孪生目标跟踪算法SiamLGL,利用低时间复杂度的Transformer编码结构使跟踪算法能够关注全局信息.

3) SiamLGL通过在回归分支中使用DIoU作为损失函数进行训练,在6个具有挑战性的数据集上获得具有竞争力的结果.

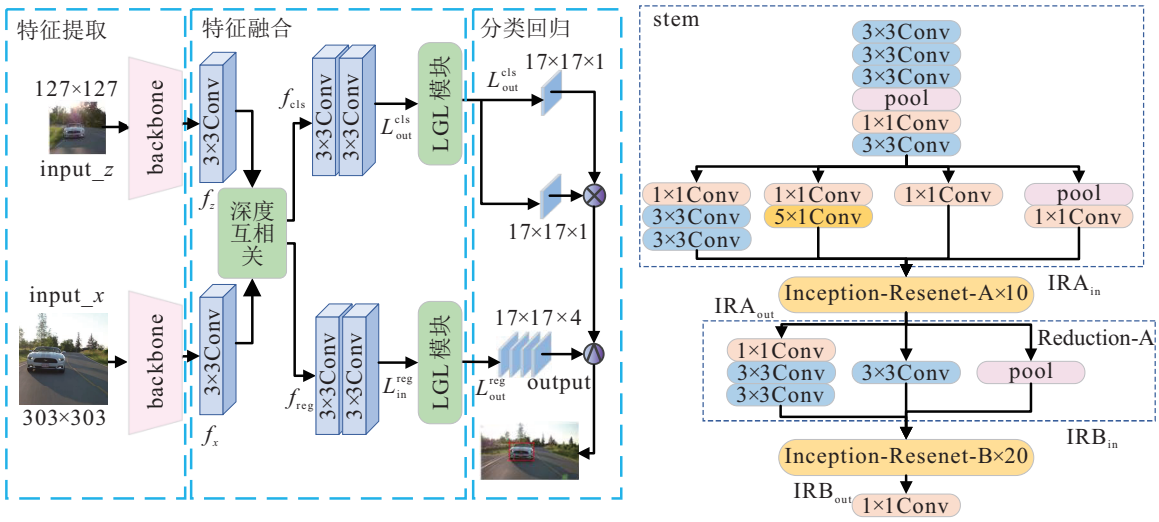
1 SiamLGL算法

算法整体流程如图1所示. 算法由特征提取网络、特征融合网络和分类回归网络组成. 模板帧大小为127×127,检测帧大小为303×303,使用改进后的IRV2网络提取特征. 在特征融合部分,采用深度互相关的方式进行特征融合. 将融合后的特征输入至LGL模块中,进一步突出响应图中特征,最后将特征送入分类回归网络中,在分类网络中采取二元交叉熵损失,回归网络中采取DIoU损失,在分类网络中利用质量评估分支抑制较大的位移,在回归分支中获取当前目标所在区域,最终利用argmax函数获得目标的位置信息.

1.1 特征提取网络

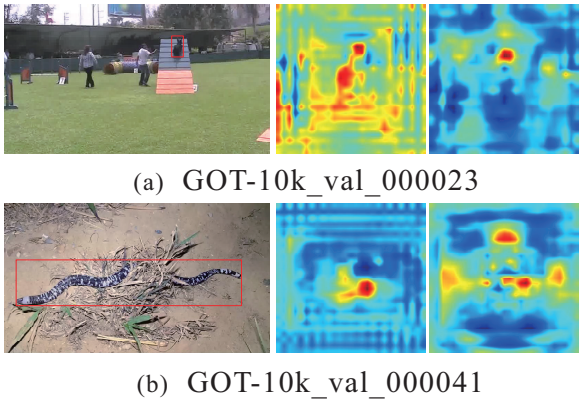
GoogLeNet作为常用的特征提取网络,其网络层数只有22层,无法充分提取目标的深层特征,且单纯叠加Inception结构无法提取到更优的图片特征. 为提升网络提取能力,采用改进后的IRV2网络,其网络层数达到132层. 在特征提取部分引入残差结构,同时增加Inception结构的数量,在加深网络层数的同时提升特征提取能力. 经过提取后的特征热度如图2所示.

图2中,第1列为原图,第2列为使用GoogLeNet网络最后一层提取的热度图,第3列为IRV2网络最后一层提取的热度图. 通过第2列和第3列的对比可以看出,IRV2网络相对于GoogLeNet网络提取到的特征效果更好,在应对各种挑战时效果更优,更适合作为目标跟踪的特征提取网络.



(a) 算法整体架构　　(b) backbone 结构

图 1　SiamLGL 算法整体流程



(a) GOT-10k_val_000023

(b) GOT-10k_val_000041

图 2　经过特征提取后的热度

IRV2 框架可分为 4 个部分: stem 模块、Inception-Resnet-A (IRA) 模块、Reduction-A 模块和 Inception-Resnet-B (IRB) 模块. 模板帧经过特征提取网络后输出的特征图大小为 3×3. stem 模块前半部分采用 Inception-Resnet-V1(IRV1) 的结构, 目的是降低网络的参数量. 为了保证 stem 模块可以获得不同尺度的特征, 在 stem 模块的后半部分增加 4 条支路拼接的结构, 进而增加网络的宽度和对多尺度特征的适应能力, 提取后的特征通过拼接的方式进行连接并作为 stem 模块的输出.

为了充分提取图片特征并保证算法的实时性, IRV2 网络选择使用 10 组 IRA 模块串联. 为了保证网络稳定, 给特征增加一个权重系数, 系数 Scale_A 设为 0.17[10]. IRA 模块如图 3 所示.

将最后一个 IRA 模块的 IRA_out 输入至 Reduction-A 模块进行下采样操作, 图片缩小至 35×35, 不仅保留图片的主要特征, 同时可以将图像中的细节抽象成更高层次的特征, 提升后续网络中特征提取的效果.

为了进一步提取图片的特征, 参考文献 [10], 选

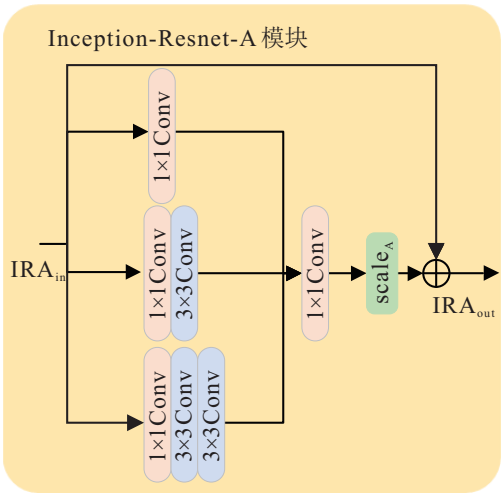


图 3　IRA 模块内部结构

择 20 个 IRB 模块串联. 在 IRB 模块中, 为了保证网络的稳定, 权重系数 Scale_B 选择为 0.1, 同时为了保证 IRB 模块可以串联, 将输出通道数设置为 1088. IRB 模块如图 4 所示.

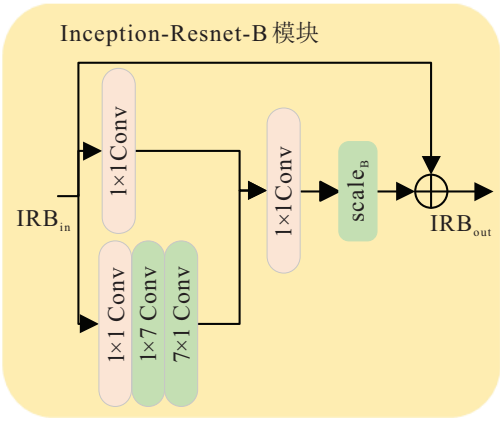


图 4　IRB 模块内部结构

为了充分利用 IRV2 网络提取到的特征, 只取 IRB 模块的输出, 此时输出的特征图大小为 35×35, 为

保证IRV2网络适应跟踪任务,将特征提取网络的输出图片大小调整为27×27,通道数调整为256.

1.2 特征融合

在算法的特征融合部分,首先将从模板帧和检测帧中提取到的特征(f_z和f_x)送至调整层中,对调整后的特征进行融合. 在特征融合部分,采用深度互相关进行相似性映射,有

$$f_j = \text{xorr}_{\text{deepwise}}(f_z, f_x),\ j \in \{\text{cls}, \text{reg}\}. \tag{1}$$

其中: f_{cls}表示深度互相关后输出至分类分支的特征图, f_{reg}表示深度与相关后输出至回归分支的特征图, $\text{xorr}_{\text{deepwise}}(\cdot)$表示深度互相关操作. 在将特征图送入LGL模块前通过两个卷积将特征图调整至目标任务空间中,有

$$L_{\text{in}}^{j} = \text{Conv}(\text{Conv}(f_j)),\ j \in \{\text{cls}, \text{reg}\}. \tag{2}$$

最后将$L_{\text{in}}^{\text{cls}}$、$L_{\text{in}}^{\text{reg}}$输入至LGL模块中. 其中: $L_{\text{in}}^{\text{cls}}$表示调整后送入分类分支的特征图, $L_{\text{in}}^{\text{reg}}$表示调整后送入回归分支的特征图.

LGL模块采用两个编码器串联,第1个编码器利用深度可分离卷积进行注意力的计算,第2个编码器利用多头自注意力获取图片的特征. LGL模块的内部结构如图5所示.

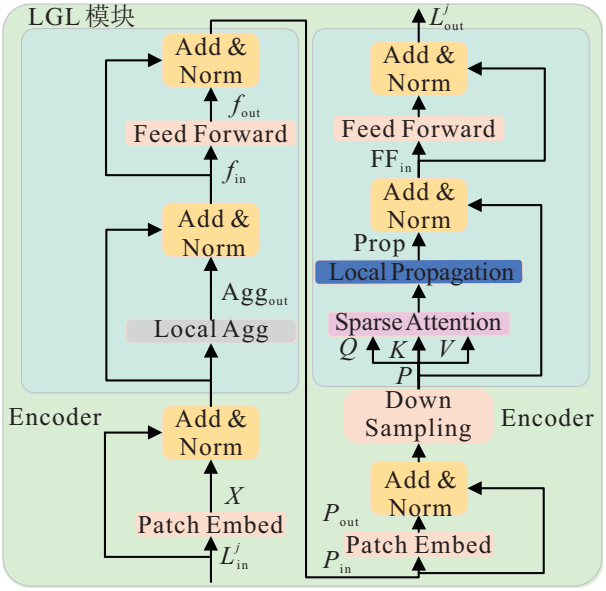


图 5 LGL模块内部结构

1.2.1 Local Agg模块

在进入该模块前,首先将特征图用分组卷积进行编码(Embed(·)),然后进行批标准化操作,有

$$\text{Agg}_{\text{in}} = \text{Norm}(\text{Embed}(L_{\text{in}}^{j}) + L_{\text{in}}^{j}),\ j \in \{\text{cls}, \text{reg}\}. \tag{3}$$

其中: Norm(·)表示利用LayerNorm对数据维度进行批标准化操作; Agg_{in}表示模块的输入,将Agg_{in}送入至模块中,模块采用深度可分离卷积(Deepwise(·))对图片中的每个图像块提取特征,将图像局部之间相近的图像块信息进行聚合,其中聚合信息的窗口大小设定为5×5. 下面利用卷积将空间中位置相近的部分聚集在一个图像块中,有

$$\text{Agg}_{\text{out}} = \text{Conv}(\text{Deepwise}(\text{Conv}(\text{Agg}_{\text{in}}))). \tag{4}$$

在输出位置为了保证能够获取到更丰富的特征,将输入特征与输出特征进行加和,并通过正则化保持网络的稳定,最终的结果f_{in}作为后续前馈神经网络的输入,有

$$f_{\text{in}} = \text{Norm}(\text{Agg}_{\text{in}} + \text{Agg}_{\text{out}}). \tag{5}$$

1.2.2 前馈神经网络(Feed Forward)

前馈神经网络是一个两层的全连接网络,目的是通过非线性变换将模块的输入映射到更高维度的空间中.

首先通过1×1卷积将维度提升至1 024,进而利用非线性变换GELU函数增加特征图的非线性,最后将特征图的维度通过1×1卷积降至256. 前馈神经网络的具体操作如图6所示. f_{out}表示网络的输出,通过加和并进行正则化操作得到第1个编码模块输出的特征图P_{in}为

$$P_{\text{in}} = \text{Norm}(f_{\text{in}} + f_{\text{out}}). \tag{6}$$

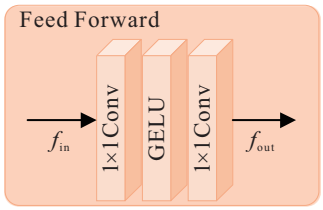


图 6 前馈神经网络内部结构

1.2.3 稀疏注意力模块(Sparse Attention)

在特征图进入第2个编码器前,首先利用分组卷积($\text{Conv}_{\text{group}}(\cdot)$)进行编码,进入稀疏注意力中的特征图$P$可以表示为

$$P = \text{Norm}(P_{\text{in}} + \text{Conv}_{\text{group}}(P_{\text{in}})). \tag{7}$$

为了降低时间复杂度,对编码后的特征利用平均池化(Avg(·))进行下采样,下采样的比率为r,获得第2个编码器的输入QKV,并利用QKV作为第2个编码器的输入特征图,进而将特征图由二维降至一维,获得多头自注意力的Q、K、V,有

$$\text{QKV} = \text{Linear}(\text{Avg}(P)); \tag{8}$$

$$\begin{cases} Q = \text{QKV}[0], \\ K = \text{QKV}[1], \\ V = \text{QKV}[2]. \end{cases} \tag{9}$$

其中Linear(·)表示线性化. 计算注意力

$$\text{Attn} = \text{Linear}\Big(V \cdot \Big(\text{softmax}\Big(\frac{QK^{\text{T}}}{\sqrt{d}}\Big)\Big)\Big). \tag{10}$$

其中: d 为输入特征图的通道, Attn 为注意力的输出. 为适应跟踪任务, 将一维Attn变换为二维 f_{Attn}, 有

$$f_{\text{Attn}} = \text{Reshape}(\text{Attn}). \tag{11}$$

通过Local Propagation模块做反卷积(ConvT(·))将特征图恢复至之前的大小, 最终达到保留全局和局部上下文信息的目的, 有

$$\text{Prop} = \text{ConvT}(f_{\text{Attn}}). \tag{12}$$

在第2个编码器内, 经过多头注意力后模块的输出 FF_{in} 为

$$\text{FF}_{\text{in}} = \text{Norm}(P + \text{Prop}). \tag{13}$$

将输出的特征图再次通过一个前馈神经网络后得到解码器的输出 L_{out}^{j}, 有

$$L_{\text{in}}^{j} = \text{FF}_{\text{in}} + \text{Conv}(\text{GELU}(\text{Conv}(\text{FF}_{\text{in}}))), \tag{14}$$

$$L_{\text{out}}^{j} = \text{Norm}(L_{\text{in}}^{j}). \tag{15}$$

经过解码器后获得的特征图作为LGL模块的输出.

LGL模块时间复杂度较ViT降低是因为ViT中的多头自注意力是对空间中所有图像块进行特征映射, 因此其时间复杂度为 $O(n^2d)$, 此处 $n = HW$. 在Local Agg模块中, 首先将经过编码后的特征图分成 $k\times k$ 个子窗口, 每个子特征图的大小为 $\frac{H}{k}\times\frac{W}{k}$, 对每个子窗口进行自注意力计算, 各子窗口所需的时间复杂度为 $O\Big(\frac{H^2W^2}{k^4}d\Big)$, 整个特征图所需的时间复杂度为 $O\Big(\frac{H^2W^2}{k^2}d\Big)$. 在稀疏自注意力模块中, 由于对图片进行下采样且缩放比例为 r, 用 $\frac{k}{r}\times\frac{k}{r}$ 的特征代表整张特征图中重要的相似度信息, 特征图的大小变为 $\frac{H}{r}\times\frac{W}{r}$, 稀疏注意力的时间复杂度可以表示为

$$O\Big(\frac{k^2}{r^2}\cdot\frac{H}{r}\cdot\frac{W}{r}\cdot d\Big) = O\Big(\frac{k^2HWd}{r^4}\Big). \tag{16}$$

后续的反卷积因为大小恢复至 $H\times W$, 且卷积核的大小为 $r\times r$, 步长为 r, 因此Local Propagation的时间复杂度为 $O(k^2r^2d)$. 这样LGL模块整体的时间复杂度为

$$O\Big(\frac{H^2W^2d}{k^2} + k^2d\Big(\frac{HW}{r^4} + r^2\Big)\Big). \tag{17}$$

利用不等式原理可知, LGL模块的时间复杂度最小为 $2HWd\sqrt{\frac{HW}{r^4}+r^2}$, 为了使整体时间复杂度达到最低, 有 $k^2 = \frac{HW}{\sqrt{\frac{HW}{r^4}+r^2}}$, r 的取值通过2.2节消融实验可知取2时效果最好. 由于输入LGL模块的特征图大小为23×23, 当 $k \approx 9.32$ 时时间复杂度最低, 但是因为需要将窗口大小控制在合适的范围内, 在Local Agg模块中将信息聚合窗口的大小设置为5×5. 所有参数代入到公式中进行计算, LGL模块的时间复杂度相较于ViT下降了约95.66 %, 在既保证算法精度的情况下又降低了模块的时间复杂度.

1.3　分类回归

算法输出部分采用分类分支和回归分支的方式预测目标位置. 分类分支包括分类得分和中心度得分, 分类得分分支将图像块分为正负样本两个分类, 中心度得分分支利用正样本偏移量计算质量评估得分. 回归分支预测正样本距离真实框四边的距离.

分类得分分支使用交叉熵损失函数, 中心度得分分支使用二元交叉熵损失函数, 回归分支使用DIoU作为损失函数. 在回归分支中, 本文算法摒弃IoU损失并采用DIoU损失作为回归分支的损失函数, 这是因为IoU损失缺少中心点距离间的信息. DIoU损失计算如下:

$$L_{\text{DIoU}} = 1 - \frac{|B_{\text{pred}}\bigcap B_{\text{gt}}|}{|B_{\text{pred}}\bigcup B_{\text{gt}}|} + \frac{d^2}{c^2}, \tag{18}$$

$$d = \sqrt{(x_{\text{gt}} - x_{\text{pred}})^2 + (y_{\text{gt}} - y_{\text{pred}})^2}. \tag{19}$$

其中: B_{pred} 为预测框, B_{gt} 为真实框, c 为包围目标框与预测框的最小矩形的对角线长度, d 为两框中心点间的距离, L_{DIoU} 为计算得到的损失, $(x_{\text{gt}}, y_{\text{gt}})$ 为真实框中心点的坐标, $(x_{\text{pred}}, y_{\text{pred}})$ 为预测框中心点的坐标.

算法总体损失计算如下:

$$L_{\text{total}} = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{qua}} + \lambda_3 L_{\text{reg}}. \tag{20}$$

其中: L_{cls} 和 L_{qua} 分别为分类分支和质量评估分支的二元交叉熵损失; L_{reg} 为回归分支的DIoU损失; λ_1、λ_2 和 λ_3 为各损失的权重, 本文分别设置为1、1、3.

2　实验分析

2.1　实验配置

实验操作系统为Ubuntu 20.04, GPU使用2080 Ti. 特征提取网络采用在ImageNet上的预训练模型作为其初始模型. 为了保证模型的泛化性, 采用TrackingNet[19]、COCO[20]、GOT-10k[21]、DET[22]、VID[22]、LaSOT[23]六个数据集进行训练, 最大迭代数30, 批大小设置为28, 前2个迭代进行预热, 学习率从 1×10^{-6} 线性增长至 8×10^{-2}, 后面采用余弦退火的方式将学习率从 8×10^{-2} 降至 1×10^{-6}, 其中特征提取网络部分的学习率为其他部分学习率的0.1倍. 前10个迭代冻结特征提取网络, 后20个迭代放开特征提取网络的参数, 每个迭代训练100万张图片. 采用随机

梯度下降法进行训练,动量设置为0.9,权重衰减系数为10^{-4}.

2.2 消融实验

为了验证IRV2网络、LGL模块和DIoU损失对每个模块的有效性,在GOT-10k数据集上对SiamLGL的各个模块进行分析. 由表1可以看出,3个模块对于跟踪算法的结果产生积极作用.

表1 不同模块有效性分析

DIoU	IRV2	LGL	AO	$SR_{0.5}$	$SR_{0.75}$
			0.601	0.731	0.462
√			0.605	0.716	0.491
√	√		0.619	0.731	0.504
√	√	√	**0.628**	**0.742**	**0.533**

2.2.1 LGL模块位置分析

为了更好地验证LGL模块在算法中的具体位置对最后跟踪结果的影响,分别将此模块放在特征提取网络、特征融合前和分类回归分支前3个位置进行训练,并在GOT-10K测试集上进行测试,实验结果见表2.

表2 LGL不同位置实验结果

位置	AO	$SR_{0.5}$	$SR_{0.75}$
特征提取网络中	0.579	0.691	0.459
特征融合前	0.607	0.725	0.479
分类回归分支前	**0.628**	**0.742**	**0.533**

2.2.2 稀疏注意力效果分析

针对稀疏注意力与多头自注意力,采取实验的方式判断哪种情况下跟踪算法的效果最好. 受网络输入特征图大小的限制,参数r只能取1和2,当$r=1$时等同于使用普通的多头自注意力. 采用LaSOT测试集进行测试,实验结果如表3所示. 可以看出,当注意力部分使用稀疏注意力时,不仅降低了计算量,而且实验效果要好于普通自注意力的实验结果,从而表明了LGL模块的有效性.

表3 稀疏注意力不同参数实验结果

r值	Suc.	Prec.	Norm. Prec.
$r=1$	0.684	0.749	0.719
$r=2$	**0.694**	**0.761**	**0.729**

2.3 与目前领先的跟踪算法比较

将SiamLGL算法与目前较先进的目标跟踪算法在6个数据集上进行比较,其他算法结果均取自对应论文中给出的实验结果.

2.3.1 GOT-10k基准评估

GOT-10k数据集针对野外的通用对象进行跟踪,其测试集包括84个对象类和32个运动类,共180个视频片段. GOT-10k数据集上的测试结果见表4.

表4 GOT-10k测试集上的对比结果

算法	AO	$SR_{0.5}$	$SR_{0.75}$
SiamST[24]	0.621	0.725	0.516
RPT[25]	0.624	0.730	0.504
OSGA[26]	**0.631**	–	0.513
ours	0.628	**0.742**	**0.533**

由表4可知,SiamLGL算法在$SR_{0.5}$和$SR_{0.75}$上取得了最好的分数,但是在AO上较OSGA低0.3%. 这是因为OSGA在进行特征融合时去掉了模板帧中的背景干扰,提升了跟踪的成功率.

表5为不同算法之间运行速度和成功率的对比,可以看出,虽然SiamLGL的跟踪成功率较高,但网络的运行速度只有38 fps,较TCTrack在运行速度上有较大差距.

表5 不同算法运行速度比较

评价指标	SiamLGL	Ocean[27]	TCTrack
运行速度	38	58	**125**
成功率	**0.628**	0.611	0.604

2.3.2 LaSOT基准评估

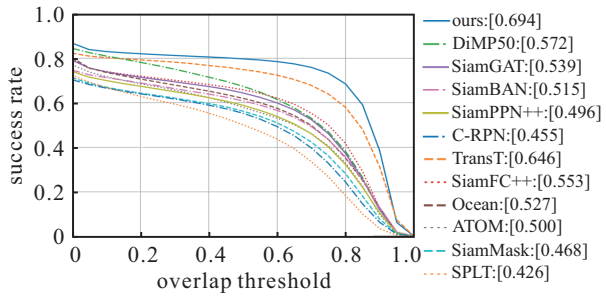
LaSOT数据集是一个大型长时跟踪数据集,图7展现了不同算法在LaSOT数据集上的跟踪结果. 由图7可知,SiamLGL算法在跟踪成功率上相较于使用Transformer进行特征融合的TransT提高了4.8%,表明SiamLGL算法在长时跟踪方面具有很好的效果. 从归一化精度图可以看出,SiamLGL相较于其他算法获取的目标中心更准确.

2.3.3 UAV123基准评估

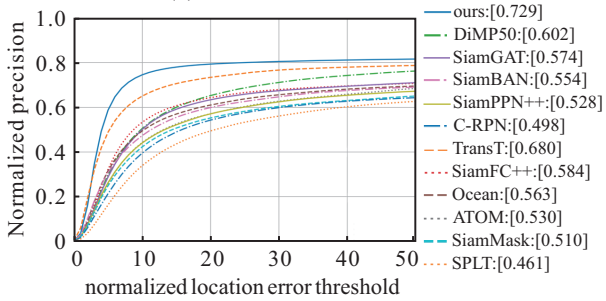
UAV123[28]数据集超过110 k帧,视频在无人机的视角下低空拍摄,在适应形状变化方面对跟踪算法提出了较高的要求. 图8展示了SiamLGL与Ocean等9个较先进的孪生网络目标跟踪算法的比较. 可见,SiamLGL都展现出了最优的跟踪结果. 总体而言,SiamLGL在成功率和精度方面均优于其他9种算法.

2.3.4 TrackingNet基准评估

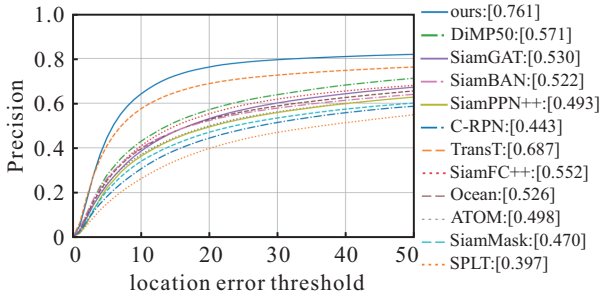
TrackingNet是一个为训练而设计的大型数据集,其庞大的训练集往往可以使跟踪算法获得更佳



(a)　跟踪成功率

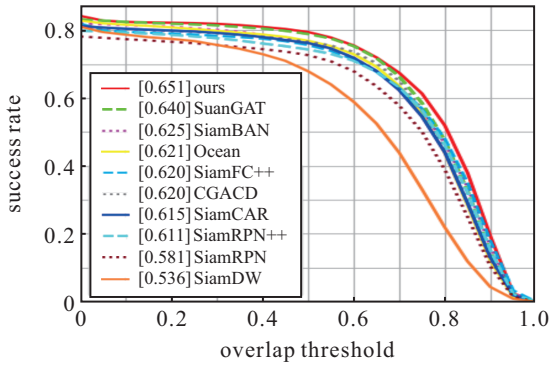


(b)　归一化精度

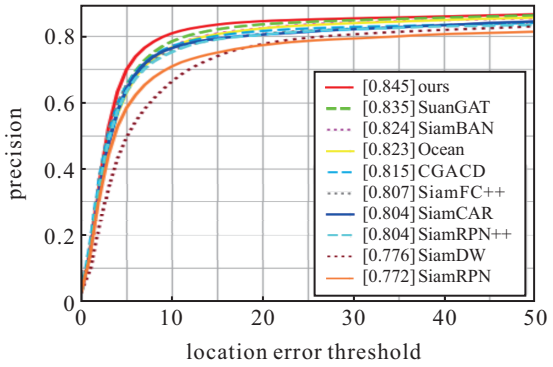


(c)　精确度

图 7　不同算法在 LaSOT 数据集上的比较结果



(a)　跟踪成功率



(b)　跟踪精度

图 8　不同算法在 UAV123 上的比较结果

的性能. 表 6 列出了不同算法在 TrackingNet 测试集上的测试结果. 通过表 6 可以发现, SiamLGL 相比单纯 CNN 架构以及 CNN 与 Transformer 混合架构均具有一定的竞争力.

表 6　不同跟踪算法在 TrackingNet 数据集上的测试结果

算法	Suc.	Prec.	Norm. Prec.
DaSiamRPN-UpdateNet	0.677	0.625	0.752
SiamRPN++	0.733	0.694	0.800
AutoMatch[29]	0.760	0.726	–
ours	**0.794**	**0.755**	**0.835**

2.3.5　OTB100 基准评估

OTB100[30] 数据集包含了其他数据集所缺少的灰度数据, 并且将不同视频按照属性进行分类. 表 7 展示了不同算法在 OTB100 数据集上的测试结果. 由表 7 可知, SiamLGL 算法在 OTB100 上与不同架构的算法比较时仍具有一定的竞争力.

表 7　OTB100 数据集测试结果

算法	Suc.	Prec.
RSABACF[31]	0.631	0.847
SiamDF[32]	0.632	0.856
SiamMTA[33]	0.652	0.870
SimTrack-B/16[34]	0.661	0.857
ours	**0.664**	**0.878**

2.3.6　VOT2019 基准评估

VOT2019[35] 数据集主要针对 RGB 图像中的短期跟踪问题, 表 8 展示了不同算法在 VOT2019 数据集上的测试结果.

表 8　VOT2019 数据集测试结果

算法	Acc.	Rob.	EAO
MVCA[36]	0.527	0.617	0.211
NFS[37]	**0.611**	0.591	0.268
CTT[38]	0.595	0.411	0.299
ours	0.563	**0.366**	**0.302**

由表 8 可见, 本文算法虽然在精度方面没有达到最优, 但是在期望平均覆盖率上较其他算法取得了更优的结果, 由此可见本文算法在 VOT2019 上依然具有一定的竞争力.

2.4　定性分析

本节将 SiamLGL 与 SiamGAT 等其他 9 种目前较先进的孪生网络跟踪算法在 UAV123 数据集上进行

了定性比较. 图 9 详细展示出在面对 UAV123 中的各种挑战时不同跟踪算法的跟踪成功率. 可以看出, SiamLGL 在除小目标外的 11 种挑战中均获得了最优或次优的性能. 从应对的 12 种挑战的成功率曲线可以看出, SiamLGL 算法整体上较其他 9 种算法效果更优.

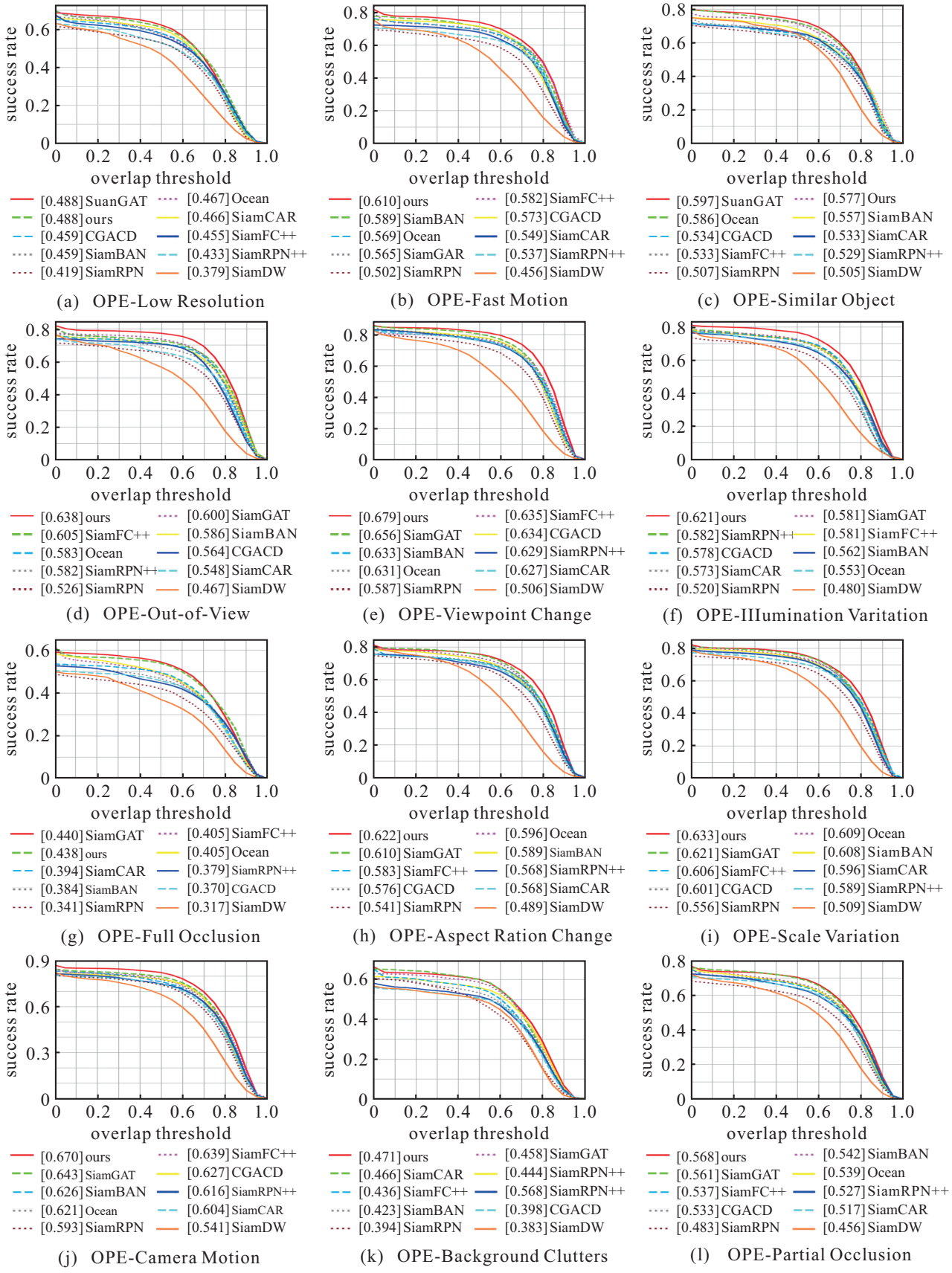


图 9　在 UAV123 的 12 种挑战下 10 种跟踪算法成功率比较结果

3 结 论

本文提出了一种基于全局上下文信息的孪生网络目标跟踪算法SiamLGL,实现了基于LGL模块和IRV2网络相结合的孪生网络目标跟踪算法. 为了充分获取目标特征,设计了改进的IRV2网络,以提高算法的准确性;为了获取图片的全局信息,在进入分类和回归分支前加入LGL模块,利用Transformer结构进一步提升算法对于目标的识别精度;为了使训练过程中能够学习到更详细的回归框的状态信息,采用DIoU损失进行训练. 最终的结果显示,SiamLGL在6个流行的公共数据集上展现出具有一定竞争力的性能,验证了算法的有效性. 另一方面,由于算法的参数量较大,未来的工作将着眼于优化算法的运行速度.

参考文献(References)

[1] 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.
(Hou Z Q, Guo F, Yang X L, et al. Transformer visual object tracking algorithm based on mixed attention[J]. Control and Decision, 2024, 39(3): 739-748.)

[2] 张子烁, 宋勇, 杨昕, 等. 基于动态特征注意模型的三分支网络目标跟踪[J]. 光学学报, 2022, 42(15): 1515001.
(Zhang Z S, Song Y, Yang X, et al. Triplet network based on dynamic feature attention for object tracking[J]. Acta Optica Sinica, 2022, 42(15): 1515001.)

[3] 程旭, 刘丽华, 王莹莹, 等. 基于多帧一致性修正的自监督孪生网络目标跟踪方法[J]. 计算机学报, 2022, 45(12): 2544-2560.
(Cheng X, Liu L H, Wang Y Y, et al. A multi-frame consistency correction based self-supervised Siamese network method for object tracking[J]. Chinese Journal of Computers, 2022, 45(12): 2544-2560.)

[4] 韩瑞泽, 冯伟, 郭青, 等. 视频单目标跟踪研究进展综述[J]. 计算机学报, 2022, 45(9): 1877-1907.
(Han R Z, Feng W, Guo Q, et al. Single object tracking research: A survey[J]. Chinese Journal of Computers, 2022, 45(9): 1877-1907.)

[5] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 850-865.

[6] 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.

[7] Zhu Z, Wang Q, Li B, et al. Distractor-aware siamese networks for visual object tracking[C]. Proceedings of the European Conference on computer vision. Piscataway: IEEE, 2018: 101-117.

[8] Xu Y D, Wang Z Y, Li Z X, 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.

[9] Fu Z H, Liu Q J, Fu Z H, et al. STMTrack: Template-free visual tracking with space-time memory networks[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13774-13783.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 30.

[11] Xie F, Wang C Y, Wang G T, et al. Learning tracking representations via dual-branch fully transformer networks[C]. IEEE/CVF International Conference on Computer Vision Workshops. Montreal, 2021: 2688-2697.

[12] 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.

[13] Chen X, Yan B, Zhu J W, et al. Transformer tracking[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 8126-8135.

[14] 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. Piscataway: IEEE, 2021: 1571-1580.

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

[16] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[17] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-ResNet and the impact of residual connections on learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 4278-4284.

[18] Pan J T, Bulat A, Tan F W, et al. Edgevits: Competing light-weight CNNs on mobile devices with vision transformers[C]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 294-311.

[19] Müller M, Bibi A, Giancola S, et al. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild[C]. Computer Vision — ECCV 2018. Cham: Springer International Publishing, 2018: 310-327.

[20] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]. Computer Vision — ECCV 2014: 13th European Conference. Zurich, 2014: 740-755.

[21] 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.

[22] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[23] Fan H, Lin L T, Yang F, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2019: 5374-5383.

[24] Zhang H, Xing W L, Yang Y F, et al. SiamST: Siamese network with spatio-temporal awareness for object tracking[J]. Information Sciences, 2023, 634: 122-139.

[25] Ma Z A, Wang L Y, Zhang H T, et al. RPT: Learning point set representation for siamese visual tracking[C]. Computer Vision — ECCV 2020 Workshops. Cham: Springer International Publishing, 2020: 653-665.

[26] Zhang J, Miao M, Zhang H, et al. Object semantic-guided graph attention feature fusion network for Siamese visual tracking[J]. Journal of Visual Communication and Image Representation, 2023, 90: 103705.

[27] Zhang Z P, Peng H W, Fu J L, et al. Ocean: object-aware anchor-free tracking[C]. Computer Vision — ECCV 2020. Cham: Springer International Publishing, 2020: 771-787.

[28] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]. Computer Vision — ECCV 2016. Cham: Springer International Publishing, 2016: 445-461.

[29] Zhang Z, Liu Y, Wang X, et al. Learn to match: Automatic matching network design for visual tracking[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021:13339-13348.

[30] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark[C]. IEEE Conference on Computer Vision and Pattern Recognition. Portland, 2013: 2411-2418.

[31] Yu W. Object tracking via background-aware correlation filter with elliptical search area[J]. Engineering Letters, 2023, 31(4): 1747-1758.

[32] Lim S C, Huh J H, Kim J C. Siamese trackers basedon deep features for visual tracking[J]. Electronics, 2023, 12(19): 41-40.

[33] 仲训杲, 范东嘉, 仲训昱, 等. 融合多模板注意力深度网的自适应目标框跟踪算法[J]. 控制与决策, 2024, 39(4): 1123-1132.
(Zhong X G, Fan D J, Zhong X Y, et al. Adaptive target box tracking algorithm byIntegrating multi-template attention deep network[J]. Control and Decision, 2024, 39(4): 1123-1132.)

[34] Chen B Y, Li P X, Bai L, et al. Backbone is all your need: A simplified architecture for visual object tracking[C]. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 375-392.

[35] Kristan M, Pflugfelder R, Leonardis A, et al. The visual object tracking VOT2013 challenge results[C]. Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops. New York: ACM, 2013: 98-111.

[36] Zhang H L, Ma Z H, Zhang J, et al. Multi-view confidence-aware method for adaptive siamese tracking with shrink-enhancement loss[J]. Pattern Analysis and Applications, 2023(3): 26.

[37] Gupta H, Verma O P. Normalization free Siamese network for object tracking[J]. Expert Systems, 2022: e13214.

[38] Yang C, Zhang X M, Song Z X. CTT: CNN meets transformer for tracking[J]. Sensors, 2022, 22(9): 3210.

作者简介

陈志旺 (1978−), 男, 副教授, 博士, 硕士生导师, 主要研究方向为运动物体目标检测与跟踪, E-mail: czwaaron@ysu.edu.cn;

杨天宇 (1996−), 男, 硕士生, 主要研究方向为计算机视觉中目标跟踪, E-mail: 2523644516@qq.com;

曹索航 (1999−), 男, 硕士生, 主要研究方向为计算机视觉中目标跟踪, E-mail: 1127001852@qq.com;

吕昌昊 (1996−), 男, 硕士生, 主要研究方向为智能电网优化运行, E-mail: 316998054@qq.com;

彭勇 (1963−), 男, 教授, 博士生导师, 主要研究方向为生物机器人控制、脑电应用科学, E-mail: PY81@sina.com.