计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220698
61(2): 460−469, 2024

# 混合目标与搜索区域令牌的视觉目标跟踪

薛万利[1] 张智彬[1] 裴生雷[2] 张开华[3] 陈胜勇[1]
[1](天津理工大学计算机科学与工程学院 天津 300384)
[2](青海民族大学物理与电子信息工程学院 西宁 810007)
[3](南京信息工程大学计算机学院 南京 130012)
(xuewanli@email.tjut.edu.cn)

# Mixing Tokens from Target and Search Regions for Visual Object Tracking

Xue Wanli[1], Zhang Zhibin[1], Pei Shenglei[2], Zhang Kaihua[3], and Chen Shengyong[1]
[1] (*School of Computer Science and Engineering*, *Tianjin University of Technology*, *Tianjin* 300384)
[2] (*College of Physics and Electronic Information Engineering*, *Qinghai Minzu University*, *Xining* 810007)
[3] (*School of Computer Science*, *Nanjing University of Information Science & Technology*, *Nanjing* 130012)

**Abstract** There are three problems about feature extraction and fusion in the current mainstream tracking framework based on Transformer: 1. The two modules of feature extraction and fusion are used separately, which is easy to produce sub-optimal model training results. 2. Computational complexity of $O\left(N^2\right)$ using self-attention reduces tracking efficiency. 3. The target template selection strategy is simple and is difficult to adapt to the drastic changes in the target appearance during the tracking process. We propose a novel Transformer tracking framework using fast Fourier transform mixing target tokens and search region tokens. For problem 1, an efficient end-to-end approach is proposed to extract and fuse features for unified learning to obtain optimal model; For problem 2, the fast Fourier transform is used to achieve complete information interaction between the target tokens and search region tokens. The computational complexity of this operation is $O\left(N\log(N)\right)$, which greatly improves the tracing efficiency. For problem 3, a template memory storage mechanism based on quality assessment is proposed, which can quickly adapt to the drastic changes in target appearance. Compared with the current state-of-the-art algorithms on three datasets LaSOT, OTB100 and UAV123, our tracker achieves better performance in both efficiency and accuracy.

**Key words** Transformer; fast Fourier transform; feature extraction; feature fusion; object tracking

**摘 要** 当前基于Transformer的主流跟踪框架在特征提取及融合方面存在3个问题:1)分开进行特征提取与融合,易产生次优模型训练结果;2)使用计算复杂度为$O\left(N^2\right)$的自注意力机制会降低跟踪算法效率;3)简单的目标模板选取策略难以自适应跟踪过程中目标表观的剧烈变化.为此,利用快速傅里叶变换对目标与搜索区域的令牌进行有效混合,提出一种新颖的基于Transformer的视觉目标跟踪方案.针对问题1提出一种高效端到端方式将特征提取与融合进行统一学习以获得最优模型.针对问题2采用快速傅里叶变换实现目标与搜索区域令牌之间的完全信息交互,该操作计算复杂度为$O\left(N\log(N)\right)$,有助于提升跟踪效率.针对问题3提出一种基于跟踪质量评估的目标模板记忆存储机制以快速自适应目标表观的剧烈变化.在3个标准数据集LaSOT, OTB100, UAV123上,所提方法与当前最优方法相比在效率和精度上均

**收稿日期**:2022-08-08;**修回日期**:2023-03-13
**基金项目**:国家自然科学基金项目(62376197, 61906135, 61876088, 92048301, 62020106004);江苏省333工程人才项目(BRA2020291)
This work was supported by the National Natural Science Foundation of China (62376197, 61906135, 61876088, 92048301, 62020106004) and the 333 High-level Talents Cultivation of Jiangsu Province (BRA2020291).
**通信作者**:裴生雷(peishenglei@126.com)

取得更好表现.

**关键词** Transformer；快速傅里叶变换；特征提取；特征融合；目标跟踪

**中图法分类号** TP391

视觉目标跟踪是计算机视觉的重要研究方向[1]. 其研究任务是在视频序列首帧中给定任意目标，并在后续视频序列中持续预测目标位置. 目标跟踪被广泛应用于无人驾驶、智能视频监控、人机交互等领域[2]. 如何设计简单、高效的通用视觉目标跟踪方法是一个亟需解决的难题. 尤其在真实复杂场景中，目标表观受光照影响、尺寸变化、严重遮挡等挑战，会产生持续的剧烈变化，从而影响跟踪结果.

近些年，基于卷积神经网络(convolutional neural network, CNN)的目标跟踪获得广泛关注. 然而受限于感受野规模，CNN 更多用于处理目标在时间域或空间域的局部特征，未能有效捕获目标特征之间的长期依赖关系[3]. 当前基于 CNN 的主流跟踪框架主要包括：基于孪生网络(siamese network)[4-7] 的目标跟踪和基于在线学习判别式模型[8-10] 的目标跟踪. 这些方法在获取图像局部特征时表现优异，但在复杂场景中，如目标被频繁遮挡或出现剧烈形变时，则不能很好地建模特征的全局上下文关系.

此外，基于 Transformer 的跟踪方案通过引入全局自注意力(self-attention)机制获取特征间长期依赖关系[11-13]. 在此类方案中，多数跟踪方法采用 CNN 作为主干网络对图像进行特征提取，随后利用 Transformer 设计编码器和解码器. 文献 [14−15] 专注简化跟踪步骤，将 Transformer 作为特征提取器并直接输出预测位置. 然而，此类跟踪方案中起核心作用的自注意力机制，由于其计算复杂度为 $O\left(N^2\right)$，使得其随图像大小增加导致计算量陡增，并直接影响目标跟踪效率.

为了降低计算量，一些研究采用视觉多层感知器(multi-layer perceptron, MLP)来构建主干网络[16-17]. 这些研究使用 MLP 层代替 Transformer 中的自注意力层，并在时域进行令牌(token)间信息交互，所谓令牌即目标和搜索区域对应的样本被切分成若干个不重叠的图像块. 令牌间的信息交互进一步简化时域信息的融合操作. MLP 的引入降低了计算复杂度，但是在训练和测试中随着令牌数量增加，会显著加大 MLP 计算量，同样会影响目标跟踪效率.

受全局滤波网络设计启发[3]，本文提出利用快速傅里叶变换(fast Fourier transform, FFT)对令牌进行高效融合，以降低视觉 MLP 模型在令牌数量增加时产生的计算开销. 首先，利用 FFT 将令牌时域特征转变为频域特征. 随后，在频域空间中捕获目标当前信息及其历史信息与搜索区域信息间的长程依赖关系. 最后，利用快速傅里叶逆变换(inverse FFT, IFFT)将频域特征转换回时域特征. 上述 FFT, IFFT 运算步骤使得所提跟踪方法，能够以较低的对数复杂度在频域空间快速学习目标在时空维度上的交互关系. 此外，为了更好地自适应目标在跟踪过程中的表观变化，提出一种基于质量评估的目标模板记忆存储机制. 该机制根据当前跟踪状态动态更新记忆存储器中稳定的历史目标信息，用于学习适应目标变化的外观模型，有助于在搜索区域内准确匹配目标.

本文的主要贡献有 3 点：

1) 提出一种快速获取令牌间长程依赖关系的跟踪算法. 特征提取与融合以端到端的形式进行学习，同时在频域空间以更小的计算复杂度建模目标令牌与搜索区域令牌间的交互关系.

2) 提出一种基于质量评估的目标模板记忆存储机制，动态自适应地捕捉目标在视频序列中的稳定变化过程，提供高质量的长期历史目标信息.

3) 所提跟踪方法在 3 个公共跟踪数据集 LaSOT[18], OTB100[19], UAV123[20] 上获得优秀评价.

## 1 相关工作

### 1.1 基于 CNN 的目标跟踪

目标跟踪框架通常可以被划分为 3 部分：1)提取图像特征的主干网络；2)目标与搜索区域特征融合模块；3)生成预测位置模块. 多数跟踪方法[5-9] 将 CNN 作为主干网络. 其中，基于孪生网络的跟踪框架以端到端方式进行训练.SiamFC[7] 采用全卷积孪生网络提取目标特征，跟踪过程中不进行目标模板更新.DSiam[21] 基于孪生网络以正则化线性回归模型动态更新目标模板.SiamRPN[6] 利用孪生网络提取目标与搜索区域特征，同时结合目标检测研究中的区域推荐网络对目标位置进行精准定位.SiamRPN++[5] 在训练过程中辅以位置均衡策略缓解 CNN 在训练过程中存在的位置偏见问题. 此外，一些基于在线学习的

判别式目标跟踪模型也取得优异性能.DiMP[8]采用端到端网络模型离线学习目标与背景间的差异，同时在线更新目标模板.PrDiMP[9]基于DiMP[8]将概率回归用于端到端训练，在测试阶段对搜索区域生成关于目标状态的条件概率密度来捕获目标. 基于CNN的目标跟踪在训练时容易造成归纳偏置.

### 1.2 基于Transformer的目标跟踪

当前Transformer网络被广泛用于各项视觉任务中，如目标分类及检测[22-24]. Transformer中的自注意力机制将每个输入元素与其他元素进行相关性计算. 在视觉跟踪研究中，TrDiMP[13]使用Transformer增强目标上下文信息，在编码器中通过自注意力机制增强目标模板特征，利用解码器融合上下文模板进行目标定位. TransT[12]提出一种基于多头注意力机制的特征融合网络，融合后特征分别输入目标分类器及边界回归器. Stark[11]则利用ResNet[25]作为主干网络提取目标特征，以Transformer编码器和解码器进行端到端训练. SwinTrack[14]借鉴Swin Transformer[22]，采用完全基于注意力机制的Transformer进行特征提取及融合. ToMP[26]同样是一种完全基于Transformer的跟踪算法，使用一个并行的2阶段跟踪器来回归目标边界. Mixformer[15]提出一种同时混合注意力模块用于特征提取及融合. 基于Transformer的目标跟踪方法虽然取得出色性能，但是随着搜索区域增大，其计算复杂度也将呈2次方增加，从而影响目标跟踪效率.

### 1.3 基于MLP的混合令牌相关工作

MLP-mixer[16]采用MLP代替Transformer中的自注意力机制进行令牌混合.ResMLP[17]基于MLP-mixer，利用仿射变换代替归一化进行加速处理.gMLP[27]使用空间门控单元在空间维度上给令牌重新加权. 上述MLP混合令牌的研究同样存在计算量增加问题，即随着输入令牌数量增多其时间复杂度会以2次方增加，并且MLP通常存在固定空间权重很难扩展到高分辨率图像的情形.

## 2 方法介绍

图1展示了基于FFT的目标与搜索区域间令牌高效混合的目标跟踪框架，该框架为端到端方式. 首先，初始目标模板大小设置为$H_t\times W_t\times 3$，记忆存储器中存储的历史目标模板的帧数设置为$T$，搜索目标区域的大小为$H_s\times W_s\times 3$. 之后，将记忆存储器内所有目标样本和搜索区域对应图像样本，切分成不重叠的、规格为$\tau\times\tau\times 3$的图像块，这些图像块称为令牌. 将这些令牌拼接起来，组成1维令牌序列. 该序列包含目标信息与搜索区域信息. 下面分2步进行模型的离线训练.

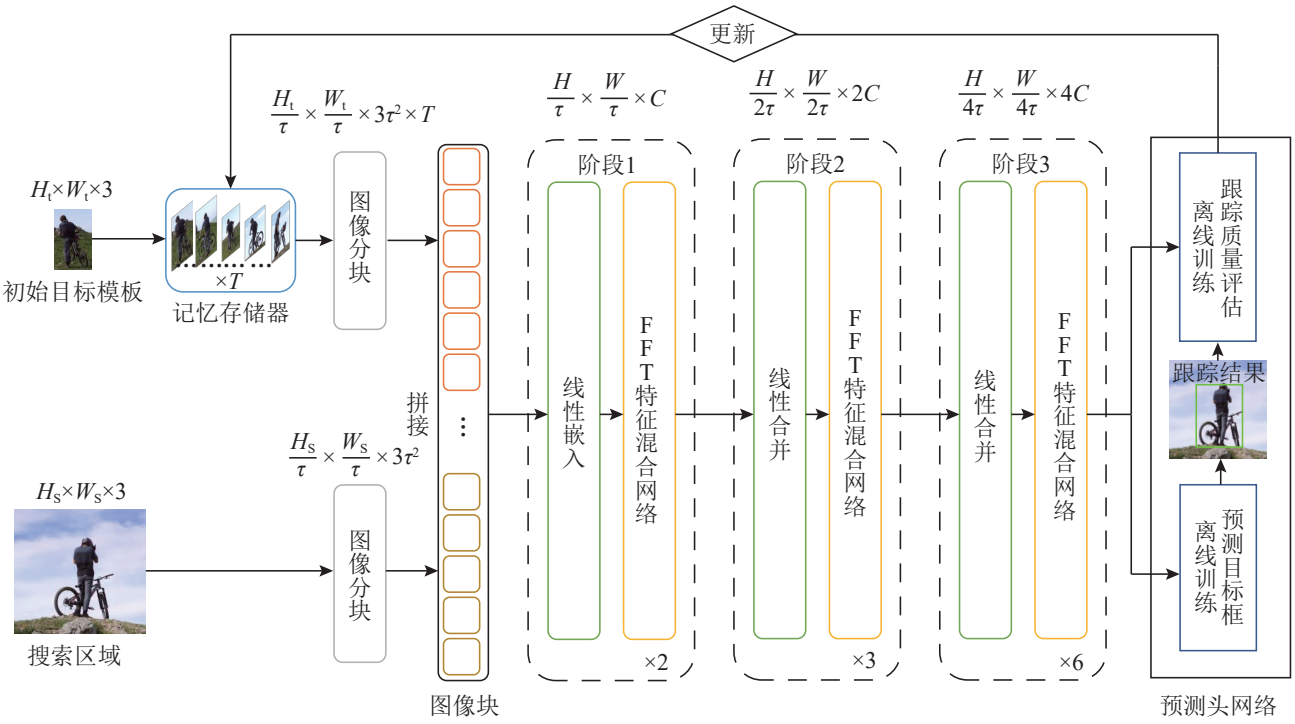



Fig. 1 The tracking algorithm framework proposed in our paper

图1 本文所提跟踪算法框架

1）针对预测目标框分支进行训练. 为了高效学习目标与搜索区域令牌间的长程依赖关系，采用 3 阶段网络设计进行令牌间混合. 在阶段 1 中，使用线性嵌入层将原始令牌投影为维度为$C$的令牌特征，再将这些令牌特征输入至线性嵌入层和包含 2 个 FFT 的令牌混合网络层. 在阶段 2 中，为了扩大模型的感受野，通过线性合并层来减少令牌数量，并将其输出特征维度设置为$2C$，这一过程由线性合并层和 3 个 FFT 的令牌混合网络层组成. 在阶段 3 中，继续进行线性合并，同时经过 6 个 FFT 令牌混合网络层，此时输出的特征维度设置为$4C$. 将在频域空间中获得的融合令牌信息进行 IFFT 运算，将频域特征重新转换为时域特征，并输入由 3 个 Conv-BN-ReLU 网络组成的预测头网络估计目标位置.

2）对跟踪质量评估分支进行离线训练，受 Stark[11] 启发，跟踪质量评估分支由一个 3 层 MLP 网络组成，用于评价当前跟踪质量，以决定是否将当前跟踪结果更新到记忆存储器中.

下面将详细介绍基于 FFT 的令牌混合网络和基于跟踪质量评估的目标模板动态记忆存储机制.

### 2.1 基于 FFT 的令牌混合网络

如图 1 所示，提出的基于 FFT 令牌混合网络层将特征提取与融合进行集成. 具体地，先利用图像分块操作将原始的 2 维目标模板和搜索区域样本转化为$N$个不重叠的$\tau\times\tau\times 3$大小的令牌. 经过裁剪等预处理后，得到一组特征矩阵$\boldsymbol{P}=(\boldsymbol{p}_0,\boldsymbol{p}_1,\cdots,\boldsymbol{p}_{N-1})$，$\boldsymbol{p}_i\in\mathbb{R}^{3\tau^2}$，$i\in[0,N-1]$. 之后，将$\boldsymbol{P}$输入至 FFT 令牌混合网络，在频域空间快速获得目标特征的多尺度交互及搜索区域与目标之间的有效交互. 其中，FFT 令牌融合网络层的结构如图 2 所示，对于第$i$个令牌先将其映射成$C$维向量：

$$\boldsymbol{x}_i=\omega_0\boldsymbol{p}_i+\boldsymbol{b}_0,i\in[0,N-1],\tag{1}$$

其中$\boldsymbol{\omega}_0\in\mathbb{R}^{3\tau^2\times C}$为每个令牌首层可学习权重，$\boldsymbol{b}_0$为首层权重位移参数向量，$N$为输入令牌个数.

FFT 令牌融合网络层的输入特征为$\boldsymbol{X}=(\boldsymbol{x}_0,\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{N-1})\in\mathbb{R}^{C\times N}$，其中$C$为输出通道数. 然后采用式(2)将输入的时域特征转换为频域特征$\boldsymbol{X}'$：

$$\boldsymbol{X}'=F(\boldsymbol{X})\in\mathbb{C}^{H\times W\times N},\tag{2}$$

其中，FFT 函数为$F(\cdot)$用于获得输入特征的频域表达，$W$为输入图像的宽，$H$为输入图像的高.

FFT 令牌混合网络层利用可学习的滤波器$\boldsymbol{K}\in\mathbb{C}^{H\times W\times N}$学习$\boldsymbol{X}'$的频域特征$\boldsymbol{X}''$：

$$\boldsymbol{X}''=\boldsymbol{K}\odot\boldsymbol{X}',\tag{3}$$

其中$\odot$为$\boldsymbol{K}$中每一个元素与$\boldsymbol{X}'$对应位置元素间相乘[3].

最后，根据式(4)将频域特征$\boldsymbol{X}''$转换为时域特征$\boldsymbol{X}^*$，并更新令牌进入下一层特征融合模块.

$$\boldsymbol{X}^*=F^{-1}\left(\boldsymbol{X}''\right),\tag{4}$$

其中$F^{-1}(\cdot)$为 IFFT，用于将频域特征转化为时域特征.

参照 Stark[11]，本文采用一个 3 层 Conv-BN-ReLU 预测头网络来估计目标位置. 具体地，估计过程被建模为预测边界框的左上角和右下角坐标的概率值图，并回归概率值图分布获得预测目标的最终坐标. 不同于 Stark 的预测头网络高度依赖编码器和解码器，本文所提预测头网络由 3 个简单的全卷积网络组成. 离线训练预测头位置分支的损失$L_{\text{loc}}$由$L_1$损失和$L_{\text{giou}}$损失组成，具体定义为：

$$L_{\text{loc}}=\alpha L_1\left(B_i,B_{\text{pred}}\right)+\beta L_{\text{giou}}\left(B_i,B_{\text{pred}}\right),\tag{5}$$

其中$\alpha$为$L_1$损失的权重系数，设置$\alpha=5$；$\beta$为$L_{\text{giou}}$的权重系数，设置$\beta=2$. $B_i$为第$i$帧搜索区域的真实标签，$B_{\text{pred}}$为预测头网络输入预测的目标位置.

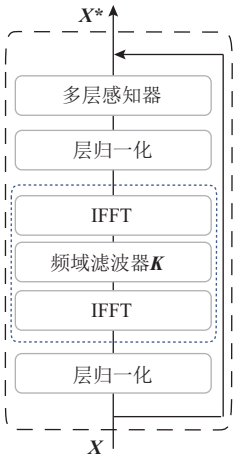



Fig. 2 Structure diagram of FFT tokens fusion network

图 2 FFT 令牌融合网络结构图

### 2.2 基于跟踪质量评估的目标模板记忆存储机制

为了提升跟踪速度的同时规避跟踪过程中引入的累计误差，多数跟踪算法仅采用第 1 帧目标模板进行匹配. 然而在跟踪过程中目标表观通常会出现剧烈变化，此时固定目标模板的跟踪方法容易产生漂移. 部分算法采用跟踪响应图的统计特性来预测当前跟踪质量，如使用峰旁比[28]、平均峰值相关能量[29] 等. 然而基于上述统计数值判断跟踪质量的做法在经历长期不稳定的跟踪后，容易导致不准确的评分结果.

如果跟踪算法可以及时预先获取当前跟踪质量，并将高质量跟踪结果放入记忆存储器中，则能够有效捕获目标在时序上的稳定表观信息变化，为目标与搜索区域的令牌混合提供有效依据.

因此，在预测头网络中添加了一个用于预测当前跟踪质量的分支. 该分支的输入为令牌融合网络层最终输出的令牌时域特征，输出为2个经过softmax函数处理过后的数值$S_{i0}$与$S_{i1}$. 其中$S_{i0}$代表第$i$帧输出的预测目标位置不是目标，$S_{i1}$表示当前预测结果是目标.当$S_{i1}>S_{i0}$时，表示当前跟踪质量良好，可以将当前跟踪结果更新到记忆存储器中，此时设置$\epsilon_i=1$；当$S_{i1}\leqslant S_{i0}$时，表示当前跟踪质量较弱，不适宜将跟踪结果更新至记忆存储器，同时设置$\epsilon_i=0$.$\epsilon_i$表示预测当前跟踪质量评估结果. 离线训练跟踪质量评价分支使用二值交叉熵损失评估，具体定义为：

$$L_{\mathrm{CE}}=l_i\mathrm{lb}(\epsilon_i)+(1-l_i)\mathrm{lb}(1-\epsilon_i), \tag{6}$$

其中$l_i$为第$i$帧样本真实的标签，当$l_i=1$时表示当前搜索区域包含真实目标，当$l_i=0$时表示当前搜索区域不包含搜索目标.

记忆存储器$M$定义为长度$T$的队列，更新间隔设为$T_{\mathrm{INR}}$. 对应的更新策略如算法1所示，当第$i$帧的质量评估为跟踪状态良好时，即$\epsilon_i=1$且符合提取间隔，则将当前跟踪结果加入记忆存储队列$M$. 若记忆存储队列$M$的长度超过$T$，则选择删除$M$队列中首个元素$M_0$. 当跟踪失败或者跟踪质量较低时，所提基于跟踪质量评估的目标记忆存储机制，能够有效缓解目标模板产生误差带来的消极影响.

该机制的可视化展示如图3所示. 第1帧给定初始目标，并将其存入记忆存储器中. 记忆存储器的长度$T$设置为5，根据跟踪质量评价结果，动态地将可靠的目标模板存入$M$中. 第200帧时，目标被完全遮挡，此时质量评估较差，不进行更新存储操作. 至此，$M$中的目标模板分别来自第90帧、第100帧、第110帧、第120帧、第130帧的跟踪结果. 在第260帧时目标重新出现，此时质量评估良好，所以当前$M$存储的目标模板调整为第120帧、第130帧、第240帧、第250帧、第260帧的跟踪结果.

**算法1.** 基于跟踪质量评估的目标模板记忆存储.

输入：$I_0$为第1帧的目标模板区域，

$I_i$为第$i$帧跟踪结果区域，

$\epsilon_i$为第$i$帧跟踪质量评估结果，

$M$为记忆存储器队列，

$T$为记忆存储器长度，

$L_{\mathrm{VID}}$为视频序列长度，

$T_{\mathrm{INR}}$为更新间隔；

输出：$M$为更新后的记忆存储器.

① $M=\varnothing$；

② while $len(M)<T$

③ $M\leftarrow M\cup\{I_0\}$；

④ for $i=1,2,\cdots,L_{\mathrm{VID}}$

⑤ if ($\epsilon_i$==1) and (mod($i,T_{\mathrm{INR}}$)==0)

⑥ $M\leftarrow M\cup\{I_i\}$；

⑦ end if

⑧ if $len(M)<T$

⑨ $M\leftarrow M\mathrm{remove}M_0$；

⑩ end if

⑪ end for

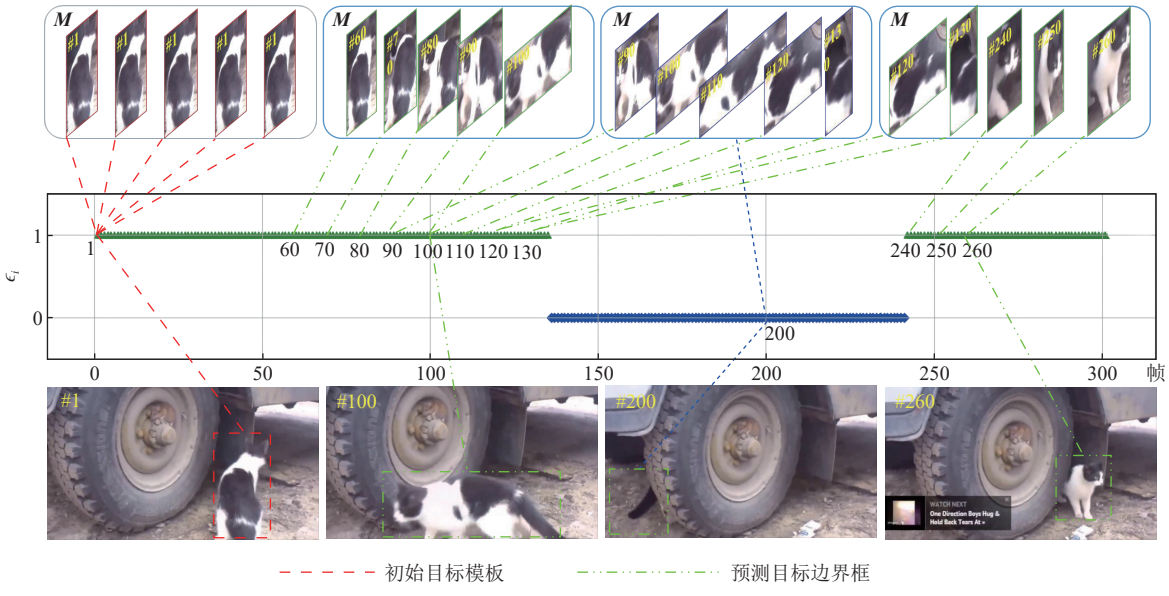



Fig. 3 Visualization of template memory storage algorithm based on quality assessment

图3 基于质量评估的模板记忆存储算法的可视化

## 3 实验结果分析

### 3.1 模型训练设置

目标模板大小为$H_t \times W_t \times 3$，搜索区域的大小为$H_s \times W_s \times 3$. 设置$H_t$，$W_t$两者值均为 128；$H_s$，$W_s$两者值均为 384.记忆存储器长度$T=5$. 记忆器更新模板的间隔$T_{INR}=10$. 图像分块操作中块大小$\tau=4$. 训练数据集为 LaSOT[18]，GOT-10k[30]，TrackingNet[31].

考虑到定位和分类的联合学习可能导致 2 个任务存在次优解[11]. 因此，借鉴 Stark[11] 和 Mixformer[15] 的训练方式，分 2 步训练特征融合模型. 首先，进行 300 批次的预测目标位置分支训练，采用 Adam[32] 损失优化器将学习率设置为 1E−4；其次，进行 100 批次的预测当前跟踪质量的分支训练，学习率设置为 1E−5. 软件环境为 Ubuntu20.04，Python3.6，Torch1.10.3，Cuda11.3.硬件环境为 NVIDIA RTX3090 24 GB.

### 3.2 定量分析

在 LaSOT[18]，OTB100[19]，UAV123[20] 数据集上验证本文方法的有效性. 评价指标为成功率（success ratio）和精度图（precision plot），其中成功率使用成功率曲线下面积（area under curve，AUC）作为排序依据.

LaSOT[18] 数据集包含 1 400 个视频序列，共计 70 类目标. 其中 1 120 个视频用于训练，280 个视频用于测试. 视频序列平均长度 2 400 帧. 数据集包含视野外等 14 种挑战.图 4 显示本文算法与 TrDiMP[13]，TransT[12]，Alpha-Refine[33]，SiamR-CNN[34]，PrDiMP[9]，DiMP[8]，SiamGAT[35]，SiamBAN[36] 8 种优秀算法比较结果. 结果表明本文算法在成功率和精度图中均处于领先水平. 精度图方面比 TransT 高 3.3%，成功率比 Alpha-Refine 高 0.8%.图 5 展示本文算法与 5 种先进算法在不同挑战下的实验结果，可以看出本文算法在多数挑战中均表现优异.

OTB100[19] 数据集包含 100 个视频序列，涉及快速运动等 11 种挑战.图 6 展示本文算法与 TransT[12]，SiamRPN++[5]，SiamBAN[36]，PrDiMP[9]，DiMP[8]，ECO[37]，MDNet[38]，ATOM[10] 的比较结果. 本文方法取得最高的成功率值和精度图值，分别比 SiamRPN++ 提升 0.2% 和 0.5%.

UAV123[20] 数据集由 123 个无人机低空拍摄的视频序列构成. 小目标和频繁遮挡是该数据集的独特挑战. 表 1 显示本文算法与 TrDiMP[13]，TransT[12]，SiamR-CNN[34]，SiamGAT[35]，SiamBAN[36]，PrDiMP[9]，DiMP[8]，SiamRPN++[5] 的比较结果. 本文算法在成功率和精度图评价指标上均排名第一.

### 3.3 定性分析

本节用可视化展示本文算法与 6 种优秀算法在旋转、快速移动、尺寸变换及遮挡等挑战下的表现.

图 7 展示 LaSOT[18] 数据集中 bird-17 视频序列的跟踪结果. 该视频序列具备快速移动、视野外等挑战. 目标在 148~156 帧快速向左移动至视野外，导致 Alpha-Refine[33] 和 TrDiMP[13] 发生跟踪漂移. 在第 184 帧中目标再次回归视野内，只有本文算法可以准确跟踪目标. 由于目标同时发生快速移动、运动模糊、旋转等挑战，其他算法均跟踪失败. 而本文算法拥有记忆存储器中的稳定目标模板，可以增强跟踪器对目标表观的自适应能力，并且在搜索目标时可快速计算目标模板和搜索区域之间的匹配关系，因此可

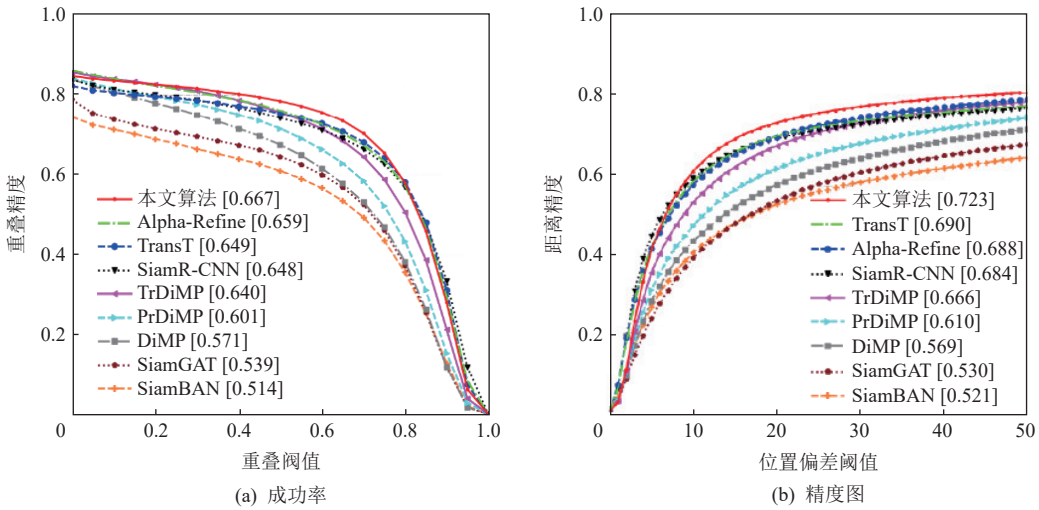



Fig. 4 Comparison of success ratio and precision plot in our algorithm and other state-of-the-art algorithms on LaSOT dataset

图 4 本文算法与其他最先进算法在 LaSOT 数据集上的成功率指标与精度图比较

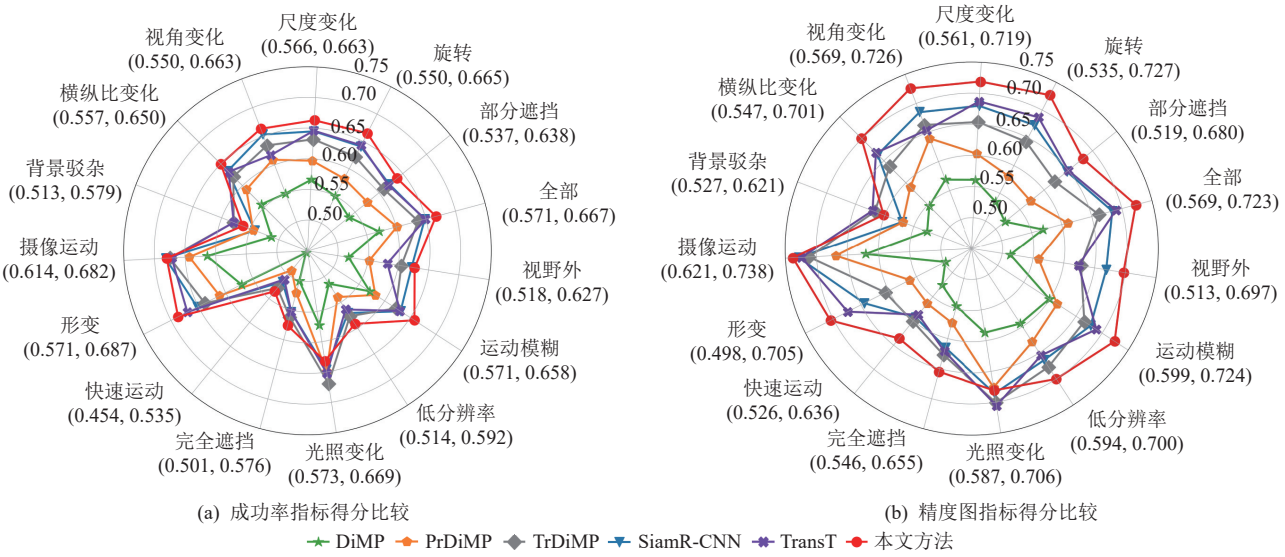



Fig. 5 Score comparison of the indictors in success ratio and precision plot for different challenges on LaSOT dataset

图 5 LaSOT 数据集上不同挑战的成功率指标和精度图指标得分比较

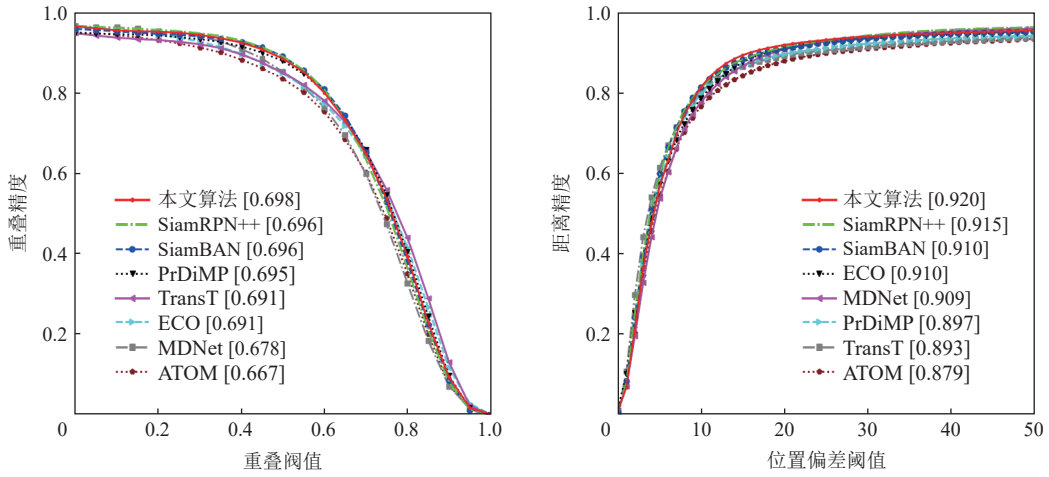



Fig. 6 Comparison of the success ratio and precision plot in our algorithm and other state-of-the-art algorithms on OTB100 dataset

图 6 本文算法与其他最先进算法在 OTB100 数据集上的成功率与精度图比较

**Table 1 Comparison of Our Algorithm and Other State-of-the-art Algorithms on UAV123 Dataset**

**表 1 本文算法与其他先进算法在 UAV123 数据集上的比较**

| 算法 | AUC | 精度图 |
|---|---|---|
| 本文算法 | 0.702 | 0.877 |
| TransT | 0.691 | 0.864 |
| PrDiMP | 0.690 | 0.867 |
| TrDiMP | 0.680 | 0.852 |
| DiMP | 0.662 | 0.838 |
| SiamBAN | 0.650 | 0.820 |
| SiamR-CNN | 0.649 | 0.834 |
| SiamGAT | 0.646 | 0.843 |
| SiamRPN++ | 0.610 | 0.803 |

以高效、稳健地跟踪目标.

图 8 展示 LaSOT[18] 数据集中 bicycle-18 视频序列的跟踪结果. 在此视频中目标受遮挡、旋转等挑战影响. 第 344~400 帧目标被岩石遮挡，导致 TransT[12] 和 SiamGAT[35] 丢失目标. 第 437~517 帧目标发生剧烈旋转，SiamGAT, TransT, PrDiMP[9] 均无法快速应对剧烈旋转引起的外观突变而发生漂移. 本文算法则依托令牌混合方案快速对目标与搜索区域特征进行交互，有效地获取更加稳健的时空特征，最终成功跟踪目标.

### 3.4 消融实验

本节验证本文算法中基于 FFT 的令牌混合网络和基于跟踪质量评估的目标模板动态记忆存储机制

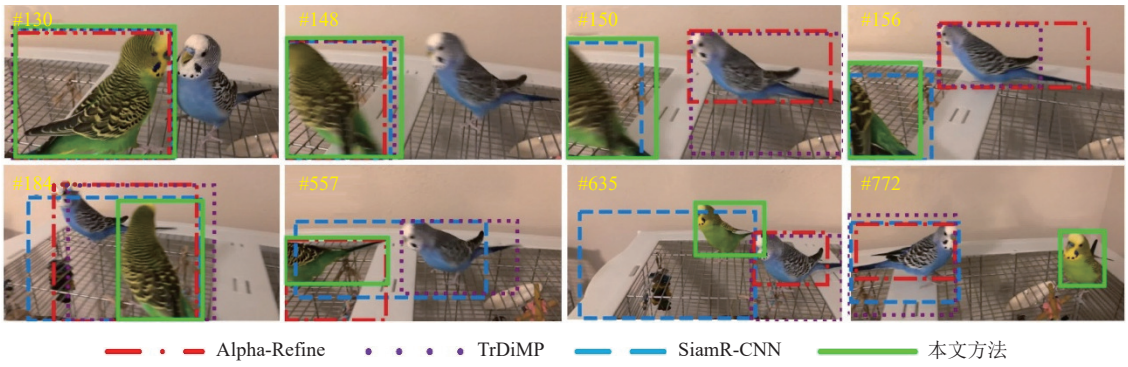



Fig. 7 Tracking results of bird-17 video sequence in LaSOT dataset

图 7 LaSOT 数据集中 bird-17 视频序列中的跟踪结果

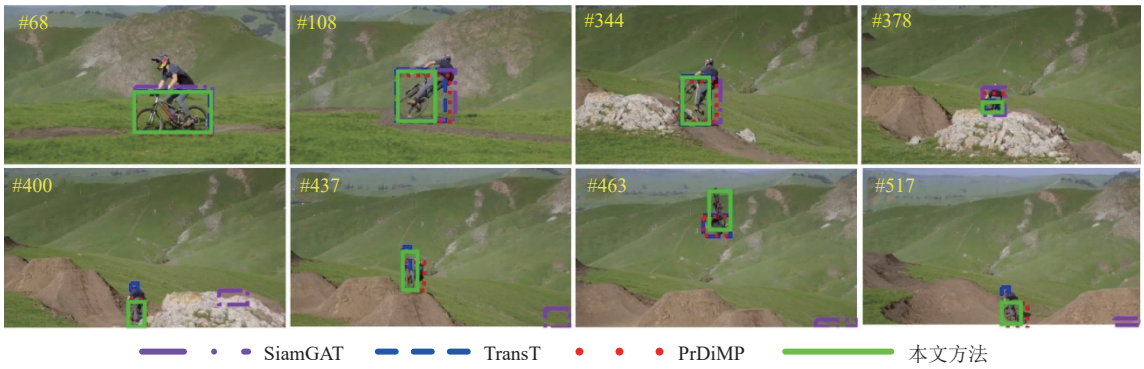



Fig. 8 Tracking results of bicycle-18 video sequence in LaSOT dataset

图 8 LaSOT 数据集中 bicycle-18 视频序列中的跟踪结果

的有效性. 表 2 展示不同变体在 LaSOT[18] 测试集上的成功率和精度图得分.

**Table 2 Results of the Ablation Experiments of Our Proposed algorithm on LaSOT Dataset**

**表 2 在 LaSOT 数据集上本文算法的消融实验结果**

| 模型变体 | 基于 CNN 的融合 | 基于 FFT 的融合 | 记忆存储机制 | 成功率 | 精度图 | 平均跟踪速度/fps |
|---|---|---|---|---|---|---|
| 变体 1 | √ | | | 0.648 | 0.684 | 23 |
| 变体 2 | | √ | | 0.661 | 0.709 | 41 |
| 变体 3 | | √ | √ | 0.667 | 0.723 | 34 |

注：√表示采用的方法.

首先, 探讨基于 FFT 的令牌混合网络的有效性. 表 2 中变体 1 采用基于 CNN 融合目标与搜索区域令牌的方法, 并且仅利用第 1 帧初始目标区域作为目标模板. 变体 2 采用 FFT 融合方法, 同样仅采用第 1 帧初始目标区域作为目标模板进行匹配. 结果显示, 基于 FFT 的融合方法比基于 CNN 的融合方法的成功率和精度图分别高 1.3% 和 2.5%. 基于传统 CNN 的融合方式在训练时只能学习特征间的局部依赖关系, 无法获取全局长程依赖, 且利用 CNN 训练模型存在较大的归纳偏置. 为了更加充分融合目标与搜索区域间的信息同时建立两者间的长程依赖关系, 本文提出利用 FFT 进行令牌间的高效融合. 可以观察到在平均跟踪速度上变体 2 比变体 1 提升近 1 倍, 结果证实基于 FFT 令牌混合网络的有效性.

其次, 变体 3 在变体 2 的基础上增加了基于质量评估的目标模板动态记忆存储机制, 用于获得更新稳定的目标模板信息, 从而自适应目标表观变化. 由于记忆存储机制增加了目标模板数量, 所以对平均跟踪速度上有一定影响. 变体 3 在测试时的平均跟踪速度比变体 2 降低了 7 fps, 但变体 3 在成功率和精度图上, 分别比变体 2 高出 0.6% 和 1.4%. 结果显示基于跟踪质量评估的目标模板动态记忆存储机制有效.

此外, 为了进一步验证本文方法具备高效的特征提取与融合能力. 在 LaSOT 数据集上将本文方法与基于 1 阶段训练的 Mixformer[15] 和基于 2 阶段训练的 TrDiMP[13] 进行对比, 结果如表 3 所示. 与采用 2 阶段训练的 TrDiMP[13] 相比, 本文方法的成功率和精度图分别提升 2.7% 和 5.7%, 同时平均跟踪速度比 TrDiMP[13] 快 8 fps.与基于 1 阶段训练的 Mixformer[15] 相比, 虽然成功率和精度图降低 2.5% 和 2.4%, 但是推理速度比 Mixformer[15] 高 9 fps.实验结果表明本文方法在准确率和推理速度间的平衡能力更好, 同时 34 fps 的平均跟踪速度达到跟踪实时性[11] 要求(>30 fps).

**Table 3 Comparative Experimental Results of Reasoning Speed on LaSOT Dataset**

**表 3 LaSOT 数据集上推理速度的对比实验结果**

| 方法 | 出处 | 成功率 | 精度图 | 平均跟踪速度/fps |
|---|---|---|---|---|
| TrDiMP | CVPR21 | 0.640 | 0.666 | 26 |
| Mixformer | CVPR22 | 0.692 | 0.747 | 25 |
| 本文算法 | | 0.667 | 0.723 | 34 |

## 4 总 结

本文提出了一种端到端的基于傅里叶变换的高效混合目标与搜索区域令牌的视觉目标跟踪方法. 该方法将特征提取与融合相结合, 利用傅里叶变换将令牌的时域特征转换为频域特征, 以便快速学习搜索区域与目标模板之间的长程依赖关系. 为了捕获目标在时序上的外观变化, 提出了一种基于跟踪质量评估的目标模板动态记忆存储机制, 确保更新目标外观模板的合理性. 广泛的实验结果验证了所提方法的有效性.

**作者贡献声明**: 薛万利提出论文整体思路并负责撰写与修改论文; 张智彬负责算法设计与实验并撰写论文; 裴生雷负责算法设计及论文审核; 张开华负责论文修改; 陈胜勇参与了论文思路的讨论及审核.

## 参 考 文 献

[1] Li Xi, Cha Yufei, Zhang Tianzhu, et al. Survey of visual object tracking algorithms based on deep learning[J]. Journal of Image and Graphics, 2019, 24(12): 2057−2080 (in Chinese)
(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)

[2] Liu Peizhong, Wang Hongxiang, Luo Yanmin, et al. Visual tracking algorithm based on adaptive spatial regularization[J]. Journal of Computer Research and Development, 2018, 55(12): 2785−2793 (in Chinese)
(柳培忠, 汪鸿翔, 骆炎民, 等. 一种结合时空上下文的在线卷积网络跟踪算法 [J]. 计算机研究与发展, 2018, 55(12): 2785−2793)

[3] Rao Yongming, Zhao Wenliang, Zhu Zheng, et al. Global filter networks for image classification[C] //Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 980−993

[4] Zhang Zhipeng, Peng Houwen. Deeper and wider Siamese networks for real-time visual tracking[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4591−4600

[5] Li Bo, Wu Wei, Wang Qiang, et al. Evolution of Siamese visual tracking with very deep networks[C] //Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 16−20

[6] Li Bo, Yan Junjie, Wu Wei, et al. High performance visual tracking with Siamese region proposal network[C] //Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8971−8980

[7] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 850−865

[8] Bhat G, Danelljan M, Gool L V, et al. Learning discriminative model prediction for tracking[C] //Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 6182−6191

[9] Danelljan M, Gool L V, Timofte R. Probabilistic regression for visual tracking[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 7183−7192

[10] Danelljan M, Bhat G, Khan F S, et al. Atom: Accurate tracking by overlap maximization[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 4660−4669

[11] Yan Bin, Peng Houwen, Fu Jianlong, et al. Learning spatio-temporal transformer for visual tracking[C] //Proc of the 18th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10448-10457

[12] Chen Xin, Yan Bin, Zhu Jiawen, et al. Transformer tracking[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 8126−8135

[13] Wang Ning, Zhou Wengang, Wang Jie, et al. Transformer meets tracker: Exploiting temporal context for robust visual tracking[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1571−1580

[14] Lin Liting, Fan Heng, Xu Yong, et al. SwinTrack: A simple and strong baseline for transformer tracking[J]. arXiv preprint, arXiv: 2112. 00995, 2021

[15] Cui Yutao, Cheng Jiang, Wang Liming, et al. Mixformer: End-to-end tracking with iterative mixed attention[C] //Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 13608−13618

[16] Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-mixer: An all-MLP architecture for vision[J]. arXiv preprint, arXiv: 2105. 01601, 2021

[17] Touvron H, Bojanowski P, Caron M, et al. ResMLP: Feedforward networks for image classification with data-efficient training[J]. arXiv preprint, arXiv: 2105. 03404, 2021

[18] Fan Heng, Lin Liting, Yang Fan, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 5374−5383

[19] Wu Yi, Lim J, Yang Ming-Hsuan. Online object tracking: A benchmark[C] //Proc of the 26th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2013: 2411−2418

[20] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C] //Proc of the 14th European Conf on Computer Vision. Berlin: Springer, 2016: 445−461

[21] Guo Qing, Feng Wei, Zhou Ce, et al. Learning dynamic siamese

network for visual object tracking[C] //Proc of the 16th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2017: 1763−1771

[22] Liu Ze, Lin Yutong, Cao Yue, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C] // Proc of the 18th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 10012−10022

[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010. 11929, 2020

[24] Wu Haiping, Xiao Bin, Codella N, et al. CVT: Introducing convolutions to vision transformers[C] // Proc of the 18th IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 22-31

[25] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[26] Mayer C, Danelljan M, Bhat G, et al. Transforming model prediction for tracking[J]. arXiv preprint, arXiv: 2203. 11192, 2022

[27] Liu Hanxiao, Dai Zihang, So D, et al. Pay attention to MLPs[C] //Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 9204−9215

[28] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C] //Proc of the 23rd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2010: 2544−2550

[29] Wang Mengmeng, Liu Yong, Huang Zeyi. Large margin object tracking with circulant feature maps[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 4021−4029

[30] Huang Lianghua, Zhao Xin, Huang Kaiqi. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1562−1577

[31] Muller M, Bibi A, Giancola S, et al. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild[C] //Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 300−317

[32] Kingma D P, Jimmy B. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412. 6980, 2014

[33] Yan Bin, Zhang Xinyu, Wang Dong, et al. Alpha-Refine: Boosting tracking performance by precise bounding box estimation[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 5289−5298

[34] Voigtlaender P, Luiten J, Torr P H S, et al. SiamR-CNN: Visual tracking by re-detection[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6578−6588

[35] Guo Dongyan, Shao Yanyan, Cui Ying, et al. Graph attention tracking[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9543−9552

[36] Chen Zedu, Zhong Bineng, Li Guorong, et al. Siamese box adaptive network for visual tracking[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6668−6677

[37] Danelljan M, Bhat G, Shahbaz K F, et al. ECO: Efficient convolution operators for tracking[C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6638−6646

[38] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C] //Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 4293−4302



**Xue Wanli**, born in 1986. PhD, associate professor, master supervisor. Member of CCF and CSIG. His main research interests include visual tracking, sign language recognition, and image stitching.

**薛万利**, 1986 年生. 博士, 副教授, 硕士生导师. CCF 与 CSIG 会员. 主要研究方向为目标跟踪、手语识别和图像拼接.



**Zhang Zhibin**, born in 1996. PhD candidate. His main research interests include visual object tracking and deep learning.

**张智彬**, 1996 年生. 博士研究生. 主要研究方向为视觉目标跟踪、深度学习.



**Pei Shenglei**, born in 1980. PhD, professor, master supervisor. His main research interests include machine learning, data mining, and intelligent decision system.

**裴生雷**, 1980 年生. 博士, 教授, 硕士生导师. 主要研究方向为机器学习、数据挖掘、智能决策系统.



**Zhang Kaihua**, born in 1983. PhD, professor. His main research interests include video object segmentation and visual object tracking.

**张开华**, 1983 年生. 博士, 教授. 主要研究方向为视频目标分割、视觉目标跟踪.



**Chen Shengyong**, born in 1973. PhD, professor. His main research interests include computer vision and machine learning.

**陈胜勇**, 1973 年生. 博士, 教授. 主要研究方向计算机视觉、机器学习.