计算机科学
COMPUTER SCIENCE

基于跨尺度融合特征与轨迹提示的目标跟踪方法

温静，张松松，李旭峰

引用本文

温静，张松松，李旭峰. 基于跨尺度融合特征与轨迹提示的目标跟踪方法[J]. 计算机科学，2025，52(10): 144-150.

WEN Jing, ZHANG Songsong, LI Xufeng. Target Tracking Method Based on Cross Scale Fusion of Features and Trajectory Prompts [J]. Computer Science, 2025, 52(10): 144-150.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800159

基于跨尺度融合特征与轨迹提示的目标跟踪方法

温　静　张松松　李旭峰
山西大学计算机与信息技术学院　太原 030006

摘　要　单纯使用 Transformer 进行目标跟踪的特征提取时，由于没有归纳偏差而无法自适应目标尺度和外观的变化。对此，借助 CNN 引入多尺度特性，提出了一种基于跨尺度融合特征与轨迹提示的目标跟踪方法(Cross Scale Fusion of Features and Trajectory Prompts Tracker，CSFTP-Tracker)。在构建目标跟踪网络输入时，将模板图像与搜索图像同时输入 CNN 与 ViT 网络融合的编码器中，设计了一种多级空间感知金字塔模块(Multi-Level Spatial Awareness Pyramid，MSAP)。首先，对多尺度 CNN 特征通过自注意力机制增强目标位置信息，然后将该多尺度特征与 ViT 中的 F-embeddings 特征相融合，输入 ViT 编码器。这种融合策略不仅增进了 ViT 内部补丁之间的信息交互，还使网络能够同时利用 CNN 的局部特性和 Transformer 的全局依赖能力。其次，将 ViT 提取的融合特征与轨迹提示特征输入解码器中，使用自回归学习目标位置。在 GOT-10k 数据集上的实验结果表明，相较于基线模型，所提出网络的平均重叠率(AO)提升了 1.3%，成功率得分在阈值为 0.5 时($SR_{0.5}$)也提高了 1.4%。

关键词：Transformer；目标跟踪；归纳偏差；编码器；轨迹提示
中图分类号　TP391

Target Tracking Method Based on Cross Scale Fusion of Features and Trajectory Prompts

WEN Jing，ZHANG Songsong and LI Xufeng
School of Computer and Information Technology，Shanxi University，Taiyuan 030006，China

Abstract　When Transformer is used alone for feature extraction in object tracking，the absence of inductive bias makes it difficult to adapt to change in target scale and appearance. To address this，this paper introduces target tracking method based on cross scale fusion of features and trajectory prompts(Cross Scale Fusion of features and Trajectory Prompts Tracker CSFTP-Tracker). In constructing the input for the object tracking network，both the template image and the search image are simultaneously fed into an encoder that fuses CNN and ViT. A key design element is the multi-level spatial-aware pyramid module (Multi-Level Spatial Awareness Pyramid，MSAP). Firstly，the multi-scale CNN features are enhanced with self-attention to strengthen target location information. These multi-scale features are then fused with the F-embeddings features from the ViT and input into the ViT encoder. This fusion strategy not only enhances information interaction between patches within the ViT but also enables the network to leverage both the local features of CNN and the global dependency capabilities of the Transformer. Furthermore，the fused features extracted by the ViT，along with the trajectory prompt features，are fed into the decoder，where autoregressive learning is employed to predict the target's position. Experimental results on the GOT-10k dataset show that，compared to the baseline models，the proposed network improves the average overlap(AO) by 1.3% and increases the success rate score at a 0.5 threshold(SR0.5) by 1.4%.

Keywords　Transformer，Object tracking，Inductive bias，Encoder，Trajectory prompt

1　引言

目标跟踪是计算机视觉[1]和机器人领域的重要任务，是从图像序列或传感器数据中跟踪感兴趣目标对象的位置和运动。在自动驾驶领域，准确的目标跟踪是实现车辆感知和环境理解的关键步骤。

传统跟踪方法侧重利用模板匹配对比图像相似度实现跟踪。在基于深度学习的方法中，如 SiamFC[2]，SiamRPN[3]等算法沿用了模板匹配的思想，采用孪生网络提取 CNN 局部特征计算模板帧和搜索帧的相似度，但在环境变化剧烈时跟踪性能下降。为提升全局建模能力，基于 Transformer[4]的跟踪算法，如 SeqTrack[5]，通过多头注意力增强全局信息，但这也带来了局部空间信息的损失。研究者尝试通过串联 CNN 与 Transformer 整合两者优势，但现有的这类方法仍存

到稿日期：2024-08-29　返修日期：2024-11-29
基金项目：山西省回国留学人员科研资助项目(2022-008)
This work was supported by the Research Project by Shanxi Scholarship Council of China(2022-008).
通信作者：温静(wjing@sxu.edu.cn)

在破坏 CNN 特征局部性的问题。此外，这种模板匹配的思想忽略了时间序列特性，视觉追踪必需结合目标外观、运动状态及时空线索。因此，跟踪的关键在于高效整合 CNN 的局部空间与 Transformer[4] 的全局特征，并利用空间信息增强模型的时空特征处理能力。为此，本文提出了一种混合联结式目标跟踪网络架构。此架构将并行的 CNN 多尺度特征与 patch embedding 融合，串联输入 Transformer 编码器，并结合轨迹提示的 Embedding 交于解码器，实现两者优势互补。具体而言，主要包括 3 个关键部件：1）多级空间感知金字塔模块（MSAP），旨在保留 CNN 的多尺度特性，并整合为与 ViT 的输入维度一致的特征大小；2）CNN To ViT 模块，主要利用注意力机制，将 MSAP 输出的多尺度 CNN 特征与 ViT[6] 的 Patch Embedding 进行融合，以保持 CNN 的局部性和多尺度性，以及 ViT 的全局性；3）在网络中引入轨迹提示，将编码器生成的特征与累积的历史轨迹信息特征一并输入解码器中，通过自回归学习机制来精确预测目标的位置信息。

2　相关工作

在目标跟踪中，主要采用 CNN 或 Transformer 设计跟踪算法，其核心思想主要集中在生成目标模板与搜索区域的特征表示，并通过计算这两个区域之间的相似度或互注意力来确定目标的位置。单独使用 CNN 或 Transformer 会造成缺失局部性或全局建模的问题，因此，融合上述两种网络的跟踪方法成为了研究热点。

常见的融合设计是 CNN-Transformer 串联架构。近年来，Transformer Meets Tracker[7] 算法首次将 Transformer 机制引入基于 CNN 的目标跟踪任务中，随后 TransT[8]，DTT[9]，STARTK[10] 等算法进一步优化了这一融合思路。这些算法的核心流程是：首先，将模板图像与搜索图像送入（如 ResNet 或 VGG 等）骨干网络提取图像特征；随后，将这些图像特征转换为序列化向量，作为 Transformer 的输入。Transformer 利用注意力机制在搜索区域内捕捉与目标模板相似的特征，实现目标定位。然而，这种串联结构为了满足 Transformer 的输入要求，会破坏前序 CNN 特征的局部性和空间结构。

此外，为了提高目标外观变化的适应性，STARK[10]，ODTrack[11] 和 ARTrack[12] 算法均通过引入不同类型的提示信息优化了目标跟踪过程。STARK 通过动态模板适应目标外观变化，ODTrack 通过时空轨迹信息提供丰富的上下文，ARTrack 则通过位置信息 Token 向量实现位置与外观特征的深度融合。这些方法在处理目标外观变化、捕捉动态特征以及融合位置与外观信息方面都为之后的研究提供了新思路。

3　基于跨尺度融合特征与轨迹提示目标跟踪网络

本文提出的混合联结式网络架构借鉴了密集目标检测领域中 ViT-Comer[13] 网络以及融合 Swin Transformer 在多尺度特征与空间池化[14] 上对特征提取不同尺度的思想，如图 1 所示。

本文网络采用并行方式处理图像特征：通过设计的两个关键模块——多级空间感知金字塔模块（MSAP）以及 CNN To ViT 融合模块（CTVF），将 CNN 提取的多级空间感知特征 F-CNN 与 ViT 提取的 F-Embeddings 特征融合，从而构建出兼顾局部性和全局性的语义特征，后串联接入网络编码器。

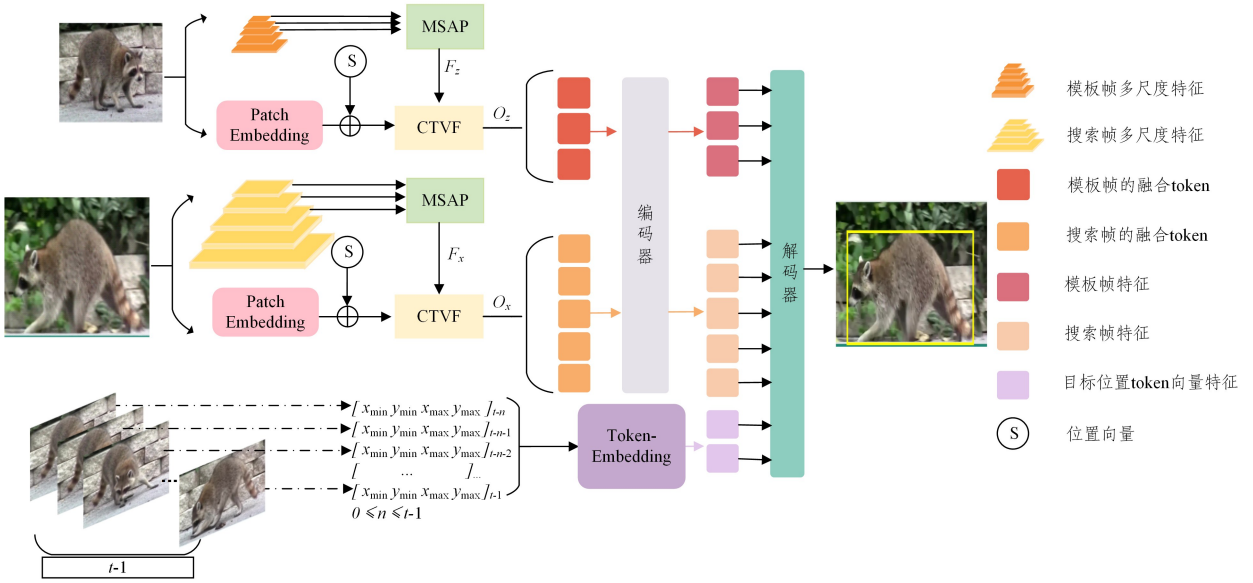

图 1　基于跨尺度融合特征与轨迹提示的目标跟踪网络结构

Fig. 1　Target tracking network structure based on cross scale fusion of features and trajectory prompts

为了提升网络对目标运动轨迹的敏感度与追踪精度，引入了轨迹提示特征机制，将目标的历史边界框坐标转换为一系列离散的 Token 序列，与编码器输出的特征一同送入解码器。通过自回归预测机制，输出目标的位置序列经离散化处理后，被转换回目标的真实坐标作为跟踪的预测结果。

3.1　多级空间感知金字塔模块 MSAP

多级空间感知金字塔模块（MSAP）结构如图 2 所示，由两大核心组件构成：局部注意力[15] 模块与多层感受野卷积模块。前者专注于精确提取感兴趣目标的位置信息，确保模型聚焦于关键区域；后者则通过 Inception 式的卷积核来扩展感

受野，从而捕获并整合丰富的多尺度信息。这两者使得MSAP模块在强化目标位置感知的同时，能全面捕捉图像中的多尺度特征。

具体来说，将模板帧和搜索帧经过CNN后的后三层多尺度特征$\{C_{x5},C_{x4},C_{x3}\}\in C_x$和$\{C_{z5},C_{z4},C_{z3}\}\in C_z$输入局部注意力模块中，如式(3)所示，在局部注意力模块(Local Attnetion)中使用一维卷积和组归一化处理模板和搜索图像的特征，局部注意力模块输入不同尺度的特征，模块对输入特征进行优化，使其保持与原始特征相同的维度，随后通过线性投影层降低特征维度，并按通道维度将其划分为M组($M=2$)。不同的特征组具有不同接受域的卷积层(例如，$k=3\times3,5\times5$)。将不同尺度的特征分为两组，分别使用不同的卷积核，在不同的卷积核中使用padding填充值保持原始特征维度，之后将两组特征进行拼接后输入线性投影层以增加特征维度。

$$C=\{C_i,i=3,4,5\} \tag{1}$$

$$F=\{F_i,i=3,4,5\} \tag{2}$$

$$F_i=FC(DWConv(LA(C_i))) \tag{3}$$

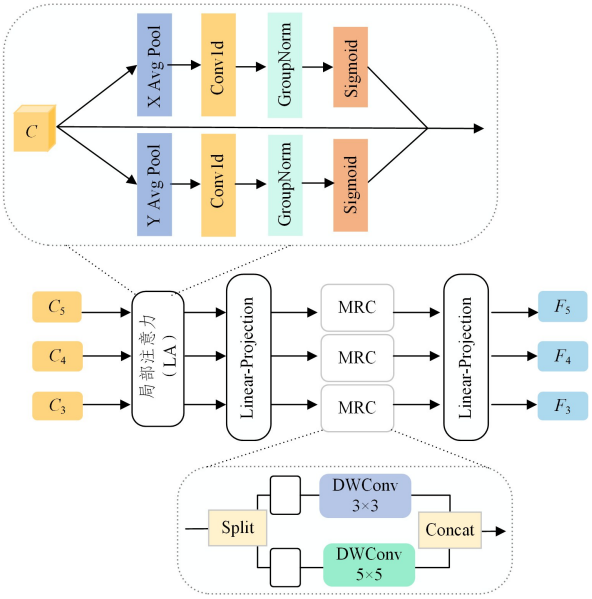

图2 MSAP模块结构

Fig. 2 Structure of MSAP module

3.2 CNNTo ViT Fusion模块

本文提出了一种CTVF(CNN To ViT Fusion)的跨架构特征融合模块。该模块的关键部分包括：多尺度可变自注意力[16]机制以及利用前馈神经网络(FFN)进行非线性变换。如图3所示。

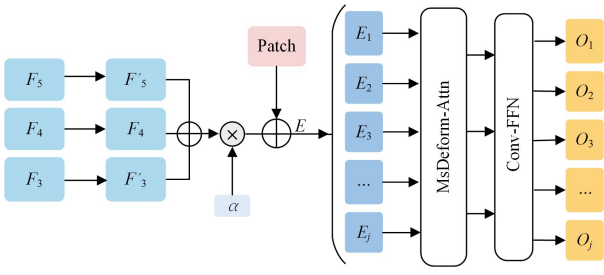

图3 CTVF模块结构

Fig. 3 Structure of CTVF module

CTVF模块在不改变ViT原有结构的基础上，将原输入变为(由MSAP输出的)CNN多尺度特征和位置注意力机制的融合特征，这使得该模块能缓解Transformer中可能面临的归纳偏差不足以及特征尺度单一的问题，从而提升了特征的局部性和全局性表征能力。为了将ViT特征$X\in R^{\frac{H_x}{16}\times\frac{W_x}{16}\times D}$，$Z\in R^{\frac{H_z}{16}\times\frac{W_z}{16}\times D}$与MSAP模块得到的多尺度特征$\{F_{x5},F_{x4},F_{x3}\}\in F_x$，$\{F_{z5},F_{z4},F_{z3}\}\in F_z$融合，首先，由于F4的特征维度与F-Embeddings特征维度相同，因此需要将F5和F3的两个尺度的特征图调整至与F4相同的维度；然后，将3个尺度的特征图叠加在一起；紧接着，引入一组可学习的参数α对特征图加权，并与ViT提取的F-Embeddings特征进行相加操作，这一步骤将并行的两种信息融合为E；接着，将融合后的特征E输入MsDeformAttn模块，捕捉图像中不同尺度下的关键信息，并通过注意力机制强化这些重要特征，同时抑制不相关的背景噪声；最后，再次利用Feed-Forward Network(FFN)对特征进行维度提升，确保处理后的特征维度能满足编码器的输入维度。该过程可以表示为：

$$E=\{E_1,\cdots,E_j,j=patchsize\} \tag{4}$$

$$O=\{O_1,\cdots,O_j,j=patchsize\} \tag{5}$$

$$E=Patch+\alpha F \tag{6}$$

$$O=FFN(MsDeformAttn(Norm(E))) \tag{7}$$

其中，E表示将CNN分支提取的图像特征与ViT产生的F-Embeddings特征进行融合后所得到的新特征；F表示统一后的尺度特征；O则代表这些融合后的特征经过统一的模块处理后生成的最终特征图。

3.3 轨迹提示模块

在解码器阶段专注于将轨迹信息与图像特征相结合作为输入，以推断和预测目标的位置序列。首先，将模板帧的目标坐标(x,y,w,h)作为历史轨迹的初始值赋值给S_i，在后续帧中跟踪到的目标位置也被依次填充至轨迹提示模块中，形成一条完整的目标运动轨迹。接着，使用自然语言处理中对文字处理的Embedding方法，对于跟踪任务来说，就是将对轨迹中的目标位置采用Embedding方法转换为Token[17]向量，将Token向量与编码器产生的图像特征进行融合。在解码器中，向量化的轨迹特征与编码器输出的特征进行自注意力计算，在融合过程中分别将Token向量作为Value值，图像特征作为Query和Key进行计算。这一过程不仅结合了目标的空间位置信息，还融入了其历史运动轨迹的时间信息。最后，解码器采用自回归学习机制，基于当前的融合特征和已预测的序列信息，逐步推断出目标在未来帧中的位置。这一过程通过迭代方式进行，每一次预测都基于前几帧预测的结果和最新的全局信息，从而确保了预测结果的准确性和可靠性。

$$S_i=\{x,y,w,h\} \tag{8}$$

$$Track_i=nn.embedding(S_i) \tag{9}$$

$$History=\{Track_i,i=1,\cdots,historysize\} \tag{10}$$

3.4 损失函数

所提算法使用交叉熵损失[18]来衡量预测边界框与真实框之间的误差。为了更精确地评估预测框与真实框的对齐

程度，引入了 SIOU[19] 损失函数，能更有效地量化预测结果与真实边界框之间的空间对应关系。具体来说，首先从估计的概率分布中提取坐标标记；然后将坐标 Token 映射至预测的边界框，计算它与真实边界框的误差。这种训练和推理使用统一损失函数，消除了分类分支与预测分支结果不匹配的隐患。损失函数为：

$$\mathcal{L}_{ce}=\sum_{j=1}^{L}\log Q(\hat{z}_j\mid s,t,\hat{z}_{<j}) \tag{11}$$

$$\mathcal{L}=\lambda_{ce}\mathcal{L}_{ce}+\lambda_{SIOU}\mathcal{L}_{SIOU} \tag{12}$$

其中，$Q(\cdot)$为 softmax 函数，$\hat{z}_j$为目标真实框，s 为搜索图像，t 为模版图像，$\hat{z}_{<j}$为历史帧的位置信息，$\mathcal{L}_{ce}$和$\mathcal{L}_{SIOU}$分别为交叉熵损失和 SIOU 损失。

4　实验与分析

在 GOT-10k[20]，LaSOT[21] 和 UAV[22] 数据集上训练并测试网络模型，服务器配置一台 NVIDIA A100 GPU，训练的最大迭代次数为 80，使用了 Adam 优化器，其衰减率为 0.05。为了权衡速度与精度，将历史位置信息队列的大小和批处理大小分别设置为 7 和 8。在实验进行中，第 60 次迭代时，学习率将衰减至原来的 1/20。

4.1　结果与分析

如表 1 所列，本文模型与其他单目标跟踪算法在 GOT-10k，LaSOT 和 UAV 数据集上进行了详细比较。结果显示，所提出的算法在所有关键性能指标上均实现了提升。

表 1　在 GOT-10k，LaSOT，UAV123 数据集上的对比结果

Table 1　Comparison results in GOT-10k，LaSOT and UAV123 datasets

（%）

模型	GOT-10k			Lasot			UAV123	
	AO	$SR_{0.5}$	$SR_{0.75}$	*AUC*	P_{Norm}/%	*P*	*AUC*	*P*
SiamFC[2]	34.8	35.3	9.8	33.6	42.0	33.9	46.8	69.3
SiamRPN++[23]	51.7	61.6	32.5	49.6	56.9	49.1	61.3	80.3
DiMP[24]	61.1	71.7	49.2	56.9	65.0	56.7	64.3	62.5
Ocean[25]	61.1	72.1	47.3	56.0	65.1	56.6	57.4	—
PrDiMP[26]	63.4	73.8	54.6	63.9	—	61.4	68.0	62.6
SiamR-CNN[27]	64.9	72.8	59.7	64.8	72.2	—	64.9	83.4
TrDiMP[7]	67.1	77.7	58.3	63.9	—	61.4	67.5	50.1
TransT[8]	64.7	73.5	59.2	64.9	73.8	69.0	69.1	65.8
SparseTT[28]	69.3	79.1	63.8	66.0	70.1	74.8	70.4	—
AutoMatch[29]	65.2	76.6	54.3	58.3	—	59.9	—	—
SwinTrackB[30]	68.6	79.9	62.4	69.3	78.5	76.5	69.8	89.6
SwinTrackL[29]	69.8	78.9	66.0	70.5	79.7	70.8	71.2	91.6
MixFormer-22k[31]	70.7	80.0	67.8	—	—	74.7	—	—
$OSTrack_{256}$[32]	71.0	80.4	68.2	69.1	78.7	75.2	68.3	—
$ARTrack\text{-}B_{256}$[12]	72.6	81.1	70.1	70.4	79.5	76.6	67.7	—
ROMTrack[33]	72.9	82.9	70.2	69.3	78.8	75.6	69.7	—
$OSTrac\ k_{384}$[32]	73.7	83.2	70.8	71.1	81.1	77.6	70.7	—
本文方法	**73.9**	**82.5**	**71.1**	**72.4**	**81.5**	**78.3**	**71.0**	**—**

在 GOT-10k 数据集中，AO（平均重叠率）和 $SR_{0.75}$（重叠阈值为 0.75 时的成功率）是比较难提升的指标，然而，在此模型中，这两个指标均实现了 1%的提升，而 $SR_{0.5}$（重叠阈值为 0.5 时的成功率）的提升则更高。

对比了当前先进的跟踪算法，包括 ARTrack[12]，OSTrack[32] 和 SiamFC++[34]，进行了一系列的视觉对比分析：处理快速运动、遮挡和尺度变化等复杂情况时的表现。

图 4 展示了模型在处理目标尺度变化和遮挡情况时的表现。第一行图片中，目标大小发生显著变化，其他算法在处理尺度变化时未能准确界定目标边界，表现为凸显或远超亦或小于目标范围。本文模型能快速适应这些变化，给出准确目标边界，主要是因为所提出的网络结构中，对 CNN 多尺度特征与 ViT 嵌入特征进行融合，增强了特征表征的多样性和深度。第二行则展示目标存在被遮挡的情况，这使得观测的数据已不在目标外观空间，这时仍然采用相似度获得目标位置是困难的，但是可以倚赖位置状态进行预测，因此本文方法能保证跟踪位置的稳定性，而其他方法则只能在图像可观测的数据上选择最优，易受到遮挡的干扰。

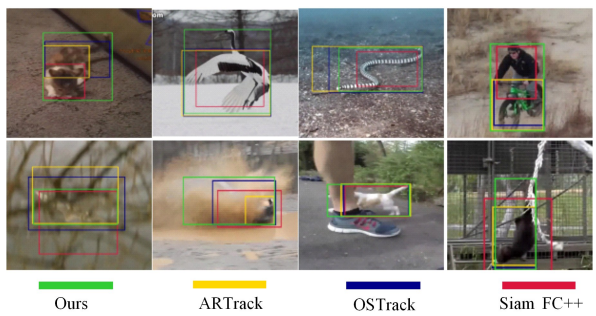

图 4　物体发生遮挡和尺度变化的可视化结果对比

Fig. 4　Comparison of visualization results of object occlusion and scale changes

图 5 对比展示了在复杂环境，特别是存在相似物体干扰条件下，不同算法对目标跟踪的效果。可视结果显示，即便面临相似目标的干扰，本文模型仍能保持跟踪的准确性和连续性，相比之下，其他几种算法则出现了跟踪错误。这是因为本文模型集成了 CNN 多尺度特征、Embedding 特征以及轨迹提示特征，这些特征共同为跟踪过程提供了丰富的空间信息和

目标物体的时序特征，从而增强了模型的稳定性和准确性。

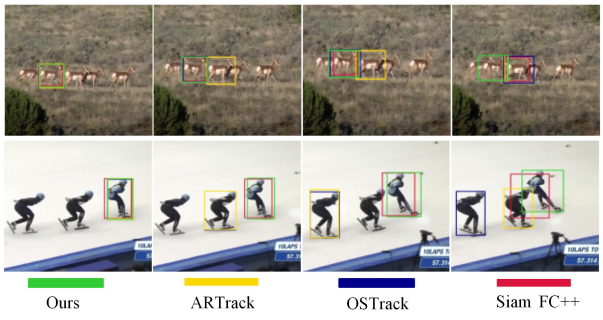

图 5 相似目标出现时可视化结果对比

Fig. 5 Comparison of visualization results for similar targets appearing

4.2 可视化分析

在目标特征图的可视化探索中，采用了解码器输出的特征权重作为注意力图的权重分配依据。这一策略使注意力图能够直观地映射出模型在解码阶段对目标特征重要性的精细评估：权重越高的区域，即代表模型在解码过程中对该部分特征的关注程度越深。

如图 6 所示，由于 ARTrack[12] 的输入特征集中于目标位置的点，其特征图表现出一种明显的特征集中趋势，即主要聚焦在目标的坐标点上。这种策略虽然有效地确保了目标位置的精确捕捉，但也导致忽略了目标整体外观的部分关键特征，从而限制了对目标的全面理解和表征。反观 OSTrack[31] 算法，尽管特征图在展示目标位置时表现出较高的置信度聚焦，表明模型能够定位目标的位置，但在追求位置精度的过程中，未能充分捕捉目标的局部特征。这些局部特征细节有助于区分目标与背景。而该模型仅依赖 ViT 提取目标特征，这可能导致目标特征无法充分关注空间局部性和目标尺度的变化，进而影响对目标的精准识别。特别是在目标外观发生显著变化时，模型可能无法准确地跟踪目标，尤其是在复杂的背景环境中。

另一方面，Siam FC++ 算法的特征图在可视化时，虽然广泛覆盖了目标的整体轮廓，但仅依靠卷积神经网络（CNN）提取目标特征，可能导致目标特征无法有效地关注全局信息，反而可能引入更多的背景信息，进而影响对目标的精准识别，尤其是在背景复杂的场景中。

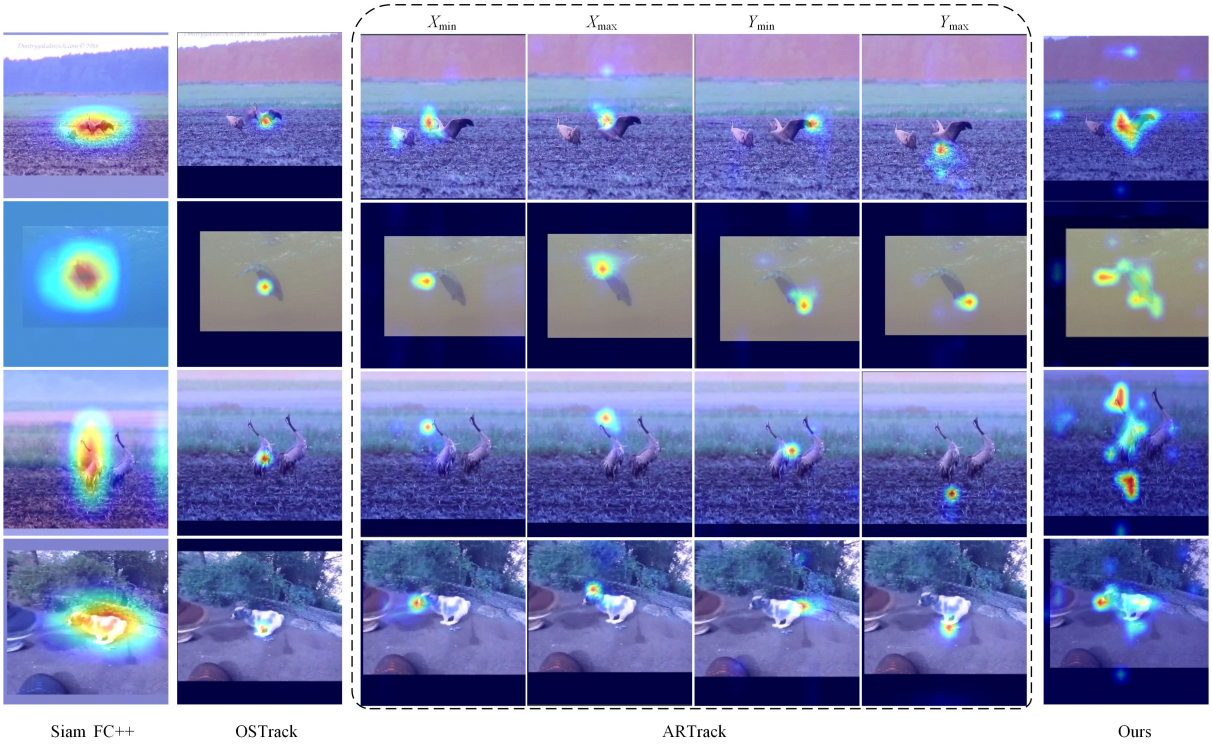

图 6 特征图可视化效果对比

Fig. 6 Visualization comparison of feature maps

本文提出的注意力图设计，通过引入 MSAP 模块，不仅为目标提供了精确的位置信息，还通过多尺度的空间特征捕捉，进一步实现了对目标物体本身的精细聚焦。这种设计使得本文模型能够在锁定目标核心和边界的同时，有效减少不必要的背景干扰，从而提高跟踪过程的效率和准确性。

具体而言，MSAP 模块通过整合多尺度信息，增强了对目标不同尺度特征的捕捉能力，并利用自注意力机制强化了对目标位置的精准识别。与 ARTrack，OSTrack 和 Siam FC++ 相比，本文方法在平衡位置精度与全局特征覆盖方面表现得更加出色，不仅能够有效应对目标尺度变化和外观变化，还能在复杂背景下保持较高的跟踪精度。

4.3 消融实验

为了深入探讨 MSAP 和 CNN To Vit Fusion 模块对整体模型性能的影响，进行了一系列的消融实验，如表 2 所列。对比模型包括：1）基线网络；2）仅保留 MSAP，通过仅保留 MSAP（多级空间感知）机制，直接将 CNN 提取的多级空间感知金字塔特征图与经过 Patch Embedding 处理的模块进行融合相加，这一设计充分利用了多尺度特征的优势，为模型提供

了在多种尺度下对目标特征的全面捕捉；3）仅保留 CTVF，经 CNN 卷积网络的特征图通过线性插值与 F-Embeddings 特征，采用多尺度可变形注意力机制进行深度融合，为模型注入了更多详尽的目标空间特征；4）结合 MSAP 和 CTVF 的方法，性能是最优的。

MSAP 模块与 CTVF 模块各自扮演了不可或缺的角色，它们的结合使得模型在目标检测或跟踪任务上取得了更好的性能。

表 2　消融实验结果的对比分析

Table 2　Comparative analysis of ablation experimental results

（%）

Methods	MSAP	CTFM	GOT-10k		
			AO	$SR_{0.5}$	$SR_{0.75}$
BaseLine			72.6	81.1	70.1
	√		73.2	81.6	70.5
		√	72.9	81.3	70.1
	√	√	**73.9**	**82.5**	**71.1**

结束语　针对目标跟踪模型存在的预测精度低、易受遮挡以及相似物体干扰等问题，本文提出了一种基于跨尺度融合特征与轨迹提示的目标跟踪算法，采用了混合联结式网络架构。该架构包括多级空间感知金字塔模块（MSAP）、CNN To ViT 融合模块（CTVF）以及轨迹提示信息等关键组件，各模块在提升模型的鲁棒性、精度和整体跟踪性能方面起到了至关重要的作用。

通过 MSAP 模块捕获多尺度特征和感兴趣目标的位置信息，增强对目标尺度变化的鲁棒性。而 CTVF 结合了 CNN 在局部性和归纳偏置方面的优势与 Transformer 在全局上下文建模方面的能力，有效提升了其在复杂环境中的跟踪性能。轨迹提示信息模块通过嵌入目标的历史位置信息，与图像信息相互平衡，为目标位置和大小提供准确预测。

然而，本文方法仍存在一些不足之处。首先，由于采用了混合联结式网络架构，模型的计算复杂度较高，导致推理速度相对较慢，在实际应用中可能会面临计算资源和实时性的挑战。其次，尽管轨迹提示信息模块提升了时序跟踪能力，但在处理长时间序列的目标跟踪任务时，模型的稳定性仍有待进一步优化。此外，本文算法在特定场景下（如光照剧烈变化或背景极其复杂的情况下）的表现还有提升空间。未来的工作将继续致力于提高模型的效率和鲁棒性，以满足更加复杂和多样化的应用需求。

参 考 文 献

[1] VOULODIMOS A，DOULAMIS N，DOULAMIS A，et al. Deep learning for computer vision：A brief review[J]. Computational Intelligence and Neuroscience，2018，2018(1)：1-13.

[2] BERTINETTO L，VALMADRE J，HENRIQUES J F，et al. Fully-convolutional siamese networks for object tracking[C]// ECCV 2016 Workshops. Springer，2016：850-865.

[3] 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. New York：IEEE，2018：8971-8980.

[4] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：6000-6010.

[5] CHEN X，PENG H，WANG D，et al. Seqtrack：Sequence to sequence learning for visual object tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2023：14572-14581.

[6] DOSOVITSKIY A. An image is worth 16x16 words：Transformers for image recognition at scale[C]// Proceedings of the International Conference on Learning Representations. 2021.

[7] 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. New York：IEEE，2021：1571-1580.

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

[9] YU B，TANG M，ZHENG L，et al. High-performance discriminative tracking with transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York：IEEE，2021：9856-9865.

[10] 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. New York：IEEE，2021：10448-10457.

[11] ZHENG Y，ZHONG B，LIANG Q，et al. Odtrack：Online dense temporal token learning for visual tracking[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI，2024：7588-7596.

[12] WEI X，BAI Y，ZHENG Y，et al. Autoregressive visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，New York：IEEE，2023：9697-9706.

[13] XIA C，WANG X，LYU F，et al. Vit-comer：Vision transformer with convolutional multi-scale feature interaction for dense predictions[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York：IEEE，2024：5493-5502.

[14] CHEN M M. 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.

[15] XU W，WAN Y. ELA：Efficient Local Attention for Deep Convolutional Neural Networks[J]. arXiv：2403.01123，2024.

[16] ZHU X，SU W，LU L，et al. Deformable DETR：Deformable transformers for end-to-end object detection[C]// Proceedings of the International Conference on Learning Representations. 2021.

[17] CHEN T，SAXENA S，LI L，et al. Pix2seq：A language modeling framework for object detection[C]// Proceedings of the International Conference on Learning Representations. 2022.

[18] DE BOER P T，KROESE D P，MANNOR S，et al. A tutorial on

the cross-entropy method[J]. Annals of Operations Research, 2005, 134(1): 19-67.
[19] GEVORGYAN Z. SIoU loss: More powerful learning for bounding box regression[J]. arXiv: 2303.15067, 2023.
[20] HUANG L, ZHAO X, HUANG K. 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.
[21] FAN H, LIN L, YANG F, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2019: 5374-5383.
[22] MUELLER M, SMITH N, GHANEM B. A Benchmark and Simulator for UAV Tracking[C]// ECCV 2016 Workshops. Springer, 2016: 445-461.
[23] 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. New York: IEEE, 2019: 4282-4291.
[24] BHAY G, DANELLJAN M, GOOL L V, et al. Learning discriminative model prediction for tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Cision. New York: IEEE, 2019: 6182-6191.
[25] ZHANG Z, PENG H, FU J, et al. Ocean: Object-aware anchor-free tracking[C]// Computer Vision ECCV. Berlin: Springer, 2020: 771-787.
[26] DANELLJAN M, GOOL L V, TIMOFTE R. Probabilistic regression for visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2020: 7183-7192.
[27] 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. New York: IEEE, 2020: 6578-6588.
[28] FU Z, FU Z, LIU Q, et al. SparseTT: Visual tracking with sparse transformers[C]// Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. AAAI, 2022: 905-912.
[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. New York: IEEE, 2021: 13339-13348.
[30] LIN L, FAN H, ZHANG Z, et al. Swintrack: A simple and strong baseline for transformer tracking[C]// Proceedings of Advances in Neural Information Processing Systems. 2022: 16743-16754.
[31] CUI Y, JIANG C, WANG L, et al. Mixformer: End-to-end tracking with iterative mixed attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2022: 13608-13618.
[32] YE B, CHANG H, MA B, et al. Joint feature learning and relation modeling for tracking: A one-stream framework[C]// European Conference on Computer Vision. Berlin: Springer, 2022: 341-357.
[33] CAI Y, LIU J, TANG J, et al. Robust object modeling for visual tracking[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. New York: IEEE, 2023: 9589-9600.
[34] XU Y, WANG Z, LI Z, et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020: 12549-12556.

WEN Jing, born in 1982, Ph. D, associated professor, master supervisor, is a member of CCF (No. 22721M). Her main research interests include computer vision and machine learning.

(责任编辑:何杨)