智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

聚焦关键信息的目标感知Transformer无人机跟踪

林淑彬, 吴贵山, 杨文元

引用本文:
林淑彬, 吴贵山, 杨文元. 聚焦关键信息的目标感知Transformer无人机跟踪[J]. 智能系统学报, 2025, 20(6): 1483-1492.
LIN Shubin, WU Guishan, YANG Wenyuan. Target-aware Transformer unmanned aerial vehicle tracker: a focus on key information[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(6): 1483-1492.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202506030

DOI: 10.11992/tis.202506030
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20251010.1326.007

聚焦关键信息的目标感知 Transformer 无人机跟踪

林淑彬[1,2], 吴贵山[1,2], 杨文元[3]
(1. 闽南师范大学 计算机学院, 福建 漳州 363000; 2. 闽南师范大学 数据科学与智能应用福建省高校重点实验室, 福建 漳州 363000; 3. 闽南师范大学 福建省粒计算及其应用重点实验室, 福建 漳州 363000)

摘 要: 无人机视觉跟踪是无人机应用的核心技术之一。现有无人机跟踪方法对输入搜索区域进行无差别关注学习, 导致特征判别力下降, 难以应对无人机场景中复杂的背景干扰。本文提出一种聚焦关键信息的目标感知 Transformer 无人机跟踪器。构建一个集成特征学习和目标搜索的单流跟踪框架, 以增强令牌之间的信息交互。提出一种自适应关系建模机制, 通过对目标模板和搜索区域令牌进行关系建模和动态分类, 提前终止对背景令牌的处理, 聚焦关键目标信息。设计了一个特征聚合模块, 保留目标的细节特征, 增强特征表示的判别力, 并引入时序一致性约束以保证特征的稳定性。在 UAV123、DTB70 和 UavDrak135 无人机跟踪基准上的实验表明, 所提出的算法在无人机跟踪方面达到了较优的性能。

关键词: 目标跟踪; Transformer; 自适应令牌终止; 跟踪框架; 特征聚合; 无人机; 背景抑制; 基准
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 1673−4785(2025)06−1483−10

中文引用格式: 林淑彬, 吴贵山, 杨文元. 聚焦关键信息的目标感知 Transformer 无人机跟踪 [J]. 智能系统学报, 2025, 20(6): 1483–1492.
英文引用格式: LIN Shubin, WU Guishan, YANG Wenyuan. Target-aware Transformer unmanned aerial vehicle tracker: a focus on key information[J]. CAAI transactions on intelligent systems, 2025, 20(6): 1483–1492.

Target-aware Transformer unmanned aerial vehicle tracker: a focus on key information

LIN Shubin[1,2], WU Guishan[1,2], YANG Wenyuan[3]
(1. School of Computer Science, Minnan Normal University, Zhangzhou 363000, China; 2. Fujian Province Universities Key Laboratory of Data Science and Intelligence Application, Minnan Normal University, Zhangzhou 363000, China; 3. Fujian Key Laboratory of Granular Computing and Application, Minnan Normal University, Zhangzhou 363000, China)

Abstract: Unmanned aerial vehicle (UAV) visual tracking is a foundational technology in the field of UAV applications. Existing UAV tracking methods focus on the input search area for learning, leading to a decline in feature discrimination and difficulty in dealing with complex background interference in UAV scenarios. This paper proposes a target-aware Transformer UAV tracker that focuses on key information. First, a single-stream tracking framework integrating feature learning and target search is constructed to enhance the information interaction between tokens. Second, an adaptive relationship modeling mechanism is proposed. This mechanism models the relationship between the target template and the search area tokens and dynamically classifies them. As a result, the processing of background tokens is prematurely terminated, and the focus shifts to key target information. A feature aggregation module has been developed to retain the detailed features of the target, enhance the discriminative power of the feature representation, and introduce temporal consistency constraints to ensure the stability of features. Experiments on the UAV123, DTB70, and UavDrak 135 UAV tracking benchmarks demonstrate that the proposed algorithm exhibits superior performance in UAV tracking.
Keywords: target-tracking; Transformer; adaptive token termination; tracking framework; feature aggregation; unmanned aerial vehicle; background suppression; benchmark

收稿日期: 2025−06−25. **网络出版日期**: 2025−10−10.
基金项目: 国家自然科学基金青年科学基金项目(12101289); 福建省自然科学基金项目(2022J01891); 福建省教育厅中青年项目(JAT220202).
通信作者: 杨文元. E-mail: yangwycn@163.com.

无人机(unmanned aerial vehicle, UAV)目标跟踪旨在根据目标的初始状态, 对无人机视频每一帧中目标的位置进行判断, 其应用于森林防火、

灾情检测[1-2]、危难搜救[3]等领域。无人机的高机动性使得目标跟踪会频繁面临极端视角、相似背景干扰和严重遮挡等挑战。此外，有限的负载能力和计算资源对跟踪算法的轻量化提出了更高的要求[4-5]。因此，开发高效、准确且能应对背景干扰的无人机目标跟踪算法，具有重要意义和应用价值。

目前，无人机目标跟踪方法主要分为两大类：基于相关滤波 (correlation filter, CF) 的跟踪器和基于深度卷积神经网络 (convolutional neural network, CNN) 的跟踪器。基于相关滤波的跟踪器[6-7]利用空间域相似性在傅里叶域的快速计算，从而获得较高的跟踪效率。相比之下，基于深度卷积神经网络的跟踪器[8-9]通过学习目标的语义和外观特征，获得更高的跟踪精度，但往往需要大量的计算资源。鉴于无人机平台的资源限制，基于相关滤波的跟踪器因其高效性而在无人机目标跟踪领域占据主导地位。Ma 等[8]将浅层细节特征与深层语义特征相结合，提高跟踪器的鲁棒性。Fu 等[10]在跟踪推理阶段，计算每个背景子区域的置信度，抑制背景区域的模型干扰。He 等[11]利用相邻帧之间的环境残差增强跟踪器的判别能力。此外，许多研究通过抑制背景异常响应[12]、增强夜间光照[13]、构建时空正则化[14]以及增强空间特征学习[15]等策略提高跟踪精度。

尽管这些基于 CF 的无人机跟踪算法在跟踪效率上表现出色，但依赖于手工设计的特征，且有限的端到端训练难以适应无人机跟踪场景的多样性和复杂性。一些研究开始探索将深度学习应用于无人机跟踪领域。Cao 等[16]通过融合浅层空间和深层语义信息应对无人机跟踪场景的复杂性。Cao 等[4]构建自适应时序转换器，增强对时间上下文信息的挖掘能力。而这些基于深度学习的跟踪算法，因计算资源和跟踪效率约束，通常难以满足无人机平台的实时性需求。为兼顾跟踪精度和效率，基于 Siamese 网络[17]的轻量级卷积神经网络被运用于 UAV 跟踪。这些方法通过增强目标特征学习[18]、挖掘时空信息[19]或修剪秩信息[20]等手段，进一步提高 Siamese 网络无人机跟踪的精度和效率。然而，受限于 Siamese 网络仅依赖于局部语义进行相似性匹配，缺乏全局上下文信息，在无人机复杂跟踪场景中，面对光照变化、快速形变等挑战存在不足。

近年来，基于视觉 Transformer（vision Transformer, ViT）[21-22]的网络框架凭借其强大的注意力机制和上下文信息捕捉能力[23]，在目标跟踪领域展现出卓越的性能。一系列基于 ViT 变体[24-27]的跟踪方法[28-29]相继被提出，并取得了显著的成果。Cui 等[30]提出 MixFormer，通过设计双向特征交互模块，将特征提取和特征融合结合，实现了更丰富的目标表征。Chen 等[31]提出焦点窗口技术，利用自适应注意力机制提高目标特征学习的准确性。Ye 等[32]设计的 OSTrack(one-stream tracking framework) 作为代表性的单流 Transformer 跟踪框架，通过将特征提取和目标搜索集成到统一架构中，并引入早期候选淘汰机制，显著提升了效率和性能。Chen 等[33]将 Transformer 中的注意力机制引入目标跟踪，设计一种新的特征融合网络替代传统的相关操作，从而提升跟踪精度和鲁棒性。Li 等[34]探索视角不变特征，以应对无人机跟踪极端视角变化挑战。卢丹等[35]通过在孪生网络目标跟踪方法上融合 Transformer 编解码器模块和自适应加权融合算法，帮助模型更好地建模目标的运动轨迹，提高目标跟踪的准确率和鲁棒性。谌海云等[36]提出轻量级 Transformer 的孪生网络无人机目标跟踪算法，使用 Transformer 对 AlexNet 网络进行改进，在分类回归网络中引入距离交并比，并采用多监督策略训练网络，有效地平衡跟踪精度和跟踪速度。

然而，现有的基于 ViT 的视觉目标跟踪方法仍面临着一些关键挑战。首先，在计算效率方面，大多数 ViT 跟踪器模型参数量大、计算复杂度高，难以满足无人机跟踪等实时要求较高的应用需求。其次，频繁的相似目标和背景干扰，对模型所学习的特征判别力提出了更高要求。因此在无人机目标跟踪过程中存在一些难题需要攻克：一是现有方法普遍采用全局特征交互策略，虽然有利于获取丰富的上下文信息，但同时也引入了大量背景噪声；二是 Transformer 网络将输入图像分割成规则且固定的图像块，并对所有图像块赋予相同的注意力权重进行信息交互，使得跟踪器难以学习到具有足够判别力的目标特征表示，在复杂背景下容易导致目标与背景混淆。

针对上述问题，提出一个聚焦关键信息的目标感知 Transformer 跟踪器（target-aware UAV Transformer tracker focusing on key information, TUTTFKI）。首先，采用单流 Transformer 框架集成特征学习与目标搜索，在统一架构中进行联合优化，以降低计算开销并增强信息交互；然后，构建自适应关系建模机制，通过对目标模板令牌和搜索区域令牌进行相似性度量和动态分类，实现目标和背景

令牌的区分, 并通过提前终止背景令牌的推理, 在减少跟踪过程中计算量的同时显著减少背景干扰, 使跟踪器聚焦于关键信息; 其次, 设计特征聚合模块, 通过逐层聚合来自不同网络层的目标令牌特征, 有效保留目标细节信息, 增强特征表示的判别力, 弥补因固定图像块划分可能导致的目标信息丢失问题; 最后, 引入时序一致性约束, 对目标聚合后的模型学习及更新进行跨帧约束, 保证聚合特征的稳定性, 抑制相似目标干扰, 增强长期跟踪的鲁棒性。

本文主要贡献如下:

1) 基于单流 Transformer 框架对特征提取和融合层面进行联合优化, 并进一步构建自适应关系建模机制。对目标模板令牌和搜索区域令牌进行相似性度量和分类, 实现目标和背景令牌的区分, 并通过提前终止背景令牌的推理, 减少背景干扰, 使跟踪器聚焦于关键信息。

2) 设计特征聚合模块。通过逐层聚合来自不同网络层的目标令牌特征, 有效地保留目标的细节信息, 增强特征表示的判别力, 弥补因固定图像块划分而可能导致的目标信息丢失问题。

3) 引入时序一致性约束。对目标聚合后的模型学习及模型更新进行跨帧约束, 以保证聚合特征的稳定性, 实现抑制相似目标的干扰, 增强算法长期跟踪的鲁棒性。

1　TUTTFKI 跟踪框架

TUTTFKI 利用 Transformer 的主干进行联合特征提取和融合, 简化了跟踪流程, 跟踪框架如图 1 所示。其核心创新点在于: 1) 单流框架设计, 实现特征学习和目标搜索的紧密结合; 2) 自适应关系建模机制, 通过令牌分类和提前终止策略, 实现对关键信息的聚焦; 3) 特征聚合模块, 通过逐层聚合目标令牌特征, 弥补固定图像块划分带来的信息损失。TUTTFKI 通过补丁嵌入层将模板图像 $\boldsymbol{Z}\in\mathbf{R}^{3\times H_Z\times W_Z}$ 和搜索图像 $\boldsymbol{X}\in\mathbf{R}^{3\times H_X\times W_X}$ 分割, 并平展成 $P\times P$ 图像块序列, 图像块数量分别为 $K_Z=H_ZW_Z/P^2$ 和 $K_X=H_XW_X/P^2$。模板图像 $\boldsymbol{Z}$ 以目标对象为中心, 搜索图像 $\boldsymbol{X}$ 代表包含目标的后续帧中的较大区域。

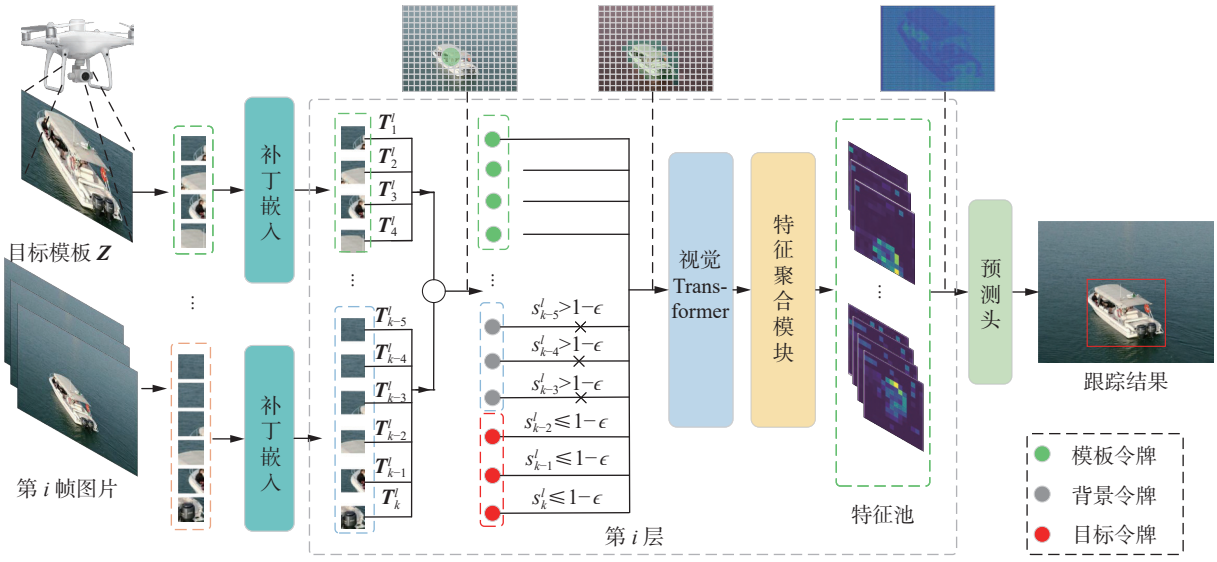


图 1　TUTTFKI 跟踪流程

Fig. 1　TUTTFKI tracking process

图像块序列经过可训练的线性投影层 $E(\cdot)$, 产生 K 个令牌嵌入:

$$\boldsymbol{T}_{1:K}^{0}=E(\boldsymbol{Z},\boldsymbol{X})\in\mathbf{R}^{K\times D}\tag{1}$$

式中: D 是每个令牌嵌入的维数。$\boldsymbol{T}_{1:K}^{L}=[\boldsymbol{T}_{1K_z}^{L};\boldsymbol{T}_{K_z+1:K}^{L}]$, $\boldsymbol{T}_{1:K_z}^{L}$ 表示模板令牌嵌入, $\boldsymbol{T}_{K_z+1:K}^{L}$ 表示搜索区域令牌嵌入。之后, 这些令牌输入到编码器中。每个编码器层通过一个多头注意力模块和一个前馈网络更新输入令牌。设 $\mathfrak{T}^{l}(\cdot)$ 为第 l 层的 Transformer 块, 则令牌嵌入通过第 l 层 Transformer 解码器输出为

$$\begin{aligned}\boldsymbol{T}_{1:K}^{l}=\mathfrak{T}^{l}\left(\boldsymbol{T}_{1:K}^{l-1}\right)=[\boldsymbol{T}_{z}^{l-1};\boldsymbol{T}_{x}^{l-1}]+\\ \mathrm{FFN}\left([\boldsymbol{T}_{z}^{l-1};\boldsymbol{T}_{x}^{l-1}]+\mathrm{MHA}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v})\right)\end{aligned}\tag{2}$$

式中: $[\cdot;\cdot]$ 表示连接操作, $\boldsymbol{q}$、$\boldsymbol{k}$ 和 $\boldsymbol{v}$ 表示传递给多头注意力块的查询、键和值, $\boldsymbol{q}=\boldsymbol{k}=\boldsymbol{v}=[\boldsymbol{T}_{z}^{l};\boldsymbol{T}_{x}^{l}]$。由于模板令牌和搜索令牌是由多头注意力块共同处理, 因此在每个编码器层中, 集成了交叉关系和自适应关系建模。最后一个编码器层的输出被解耦, 并重新按原始空间位置整合为一个二维特征图。

将二维特征图作为目标边界框预测的卷积头部输入，通过一个由多个卷积层–批量归一化层–激活函数层（convolutional-batch normalization-rectified linear unit, Conv-BN-ReLU）组成的完全基于卷积网络的预测头，直接估计目标的边界框，产生目标分类得分 $\boldsymbol{p}\in\widehat{[0,1]}^{\frac{H_X}{P}\times\frac{W_X}{P}}$、局部偏移 $\boldsymbol{o}\in[0,1]^{2\times\frac{H_X}{P}\times\frac{W_X}{P}}$ 以及归一化边界框大小 $\boldsymbol{s}\in[0,1]^{2\times\frac{H_X}{P}\times\frac{W_X}{P}}$。其中，$H$ 和 W 为原图高和宽，P 为下采样率。目标位置由最高分类分数确定：

$$(x_c,y_c)=\arg\max_{(x,y)}\boldsymbol{p}(x,y) \tag{3}$$

通过式 $[(x_t,y_t);(w,h)]=[(x_c,y_c)+\boldsymbol{o}(x_c,y_c);\boldsymbol{s}(x_c,y_c)]$ 估计最终目标边界框。对于跟踪任务，引入加权焦点损失进行目标分类，采用 $\mathcal{L}_1$ 损失和 $\mathcal{L}_{\text{iou}}$ 进行边界框回归。总损失函数为

$$\mathcal{L}_{\text{overall}}=\mathcal{L}_{\text{cls}}+\lambda_{\text{iou}}\mathcal{L}_{\text{iou}}+\lambda_1\mathcal{L}_1+\alpha_{\text{p}}\mathcal{L}^*_{\text{ponder}} \tag{4}$$

具体算法如下所示。

算法 聚焦关键信息的目标感知 Transformer 无人机跟踪

输入 视频序列 $\boldsymbol{I}$，第一帧目标初始化信息；

输出 视频序列中每一帧的目标位置和尺寸。

1）输入第一帧目标初始化模板 $\boldsymbol{Z}\in\mathbf{R}^{3\times H_Z\times W_Z}$ 和搜索图像 $\boldsymbol{X}\in\mathbf{R}^{3\times H_X\times W_X}$；

2）将输入图像和目标分割平展成 K 个图像块，通过式（1）图像嵌入操作为每个图像块生成令牌嵌入；

3）初始化每个令牌的分类 $\boldsymbol{c}_k^0$；

For $l=1$ to L

4）通过式（5）算出每个令牌在 l 层的分类概率；

5）对于类别分类为背景的令牌，即式（6）得分超过 $1-\epsilon$ 的令牌赋值为零；

End For

6）通过特征聚合模块将每层标记为目标的令牌特征进行聚合；

7）将最后一个编码器层的输出解耦，重新按照原始空间位置整合为一个二维特征图；

8）通过预测头式（3）得出目标位置及预测边界框；

9）如果是最后一帧则停止算法，否则重复执行步骤 3）~8）。

10）返回跟踪结果。

2 自适应关系建模机制

TUTTFKI 是基于 ViT 的跟踪框架，通过图像裁剪将目标模板和搜索区域传入跟踪网络。在无人机应用场景中，跟踪目标较小，导致搜索区域令牌数远大于目标模板令牌数。若在跟踪过程中对所有图像块令牌赋予相同的注意力权重，不仅会降低跟踪速度，而且会削弱模型学习到的特征的判别力。

为了使跟踪器更加聚焦于重点区域，提出自适应关系建模机制。通过对搜索区域令牌和目标模板令牌进行交叉建模，标记搜索区域令牌，提前终止对背景概率较高的令牌进行后续推理，将计算资源集中于更可能包含目标的区域。

由于 ViT 网络[22]中每个令牌嵌入的第 e 维度都保留了足够的信息，为避免引入过多计算开销，自适应关系建模机制采用一种高效的实现方式，即在多层感知器层（multilayer perceptron, MLP）中分配一个神经元来完成关系建模。该神经元负责将令牌嵌入的关系信息融合到单流主干跟踪网络中，无需引入任何额外的学习参数。

自适应关系建模机制将模型输入的令牌分为两类，分别为目标令牌 $\boldsymbol{C}_{\text{t}}$ 和背景令牌 $\boldsymbol{C}_{\text{b}}$。对 l 层 k 处的令牌，自适应关系建模机制计算其分类概率得分 c_k^l：

$$c_k^l=B(\boldsymbol{T}_k^l)=\sigma(\gamma\cdot\boldsymbol{T}_{k,e}^l+\beta) \tag{5}$$

式中：$B(\cdot)$ 是通过视觉 Transformer 块 MLP 层的关系建模操作；$\sigma(u)=\dfrac{1}{1+e^{-u}}$ 是逻辑 s 型函数；$\boldsymbol{T}_{k,e}^l$ 表示在维度 e 的 $\boldsymbol{T}_k^l$；β 和 γ 是令牌移位和缩放参数，其针对所有令牌参数跨层共享。通过关系建模操作，每个令牌分类为

$$c_k^n=\sum_{l=1}^{n}c_k^l=\begin{cases}\boldsymbol{C}_{\text{t}}, & c_k^l\leqslant 1-\epsilon\\ \boldsymbol{C}_{\text{b}}, & c_k^l>1-\epsilon\end{cases} \tag{6}$$

式中：ϵ 是一个非常小的常数，且允许 $n=0$。当令牌终止得分不大于 $1-\epsilon$ 时，认为该令牌属于目标令牌 $\boldsymbol{C}_{\text{t}}$，反之则为背景令牌 $\boldsymbol{C}_{\text{b}}$。为避免类别 $\boldsymbol{C}_{\text{b}}$ 的背景令牌干扰模型目标特征的学习，自适应关系建模机制会终止类别 $\boldsymbol{C}_{\text{b}}$ 的背景令牌，对该令牌值清零并阻止其关注其他令牌，同时不会再对该令牌进行更新。

设 N_k 为通过 Transformer 块对令牌 $\boldsymbol{T}_k$ 更新的总数，则

$$N_k=\underset{n\leqslant L}{\operatorname{argmin}}\left\{\sum_{l=1}^{n}c_k^l>1-\epsilon\right\} \tag{7}$$

可以得到所有令牌的终止概率：

$$s_k^l=\begin{cases}0, & l>N_k\\ c_k, & l=N_k\\ 1-\sum_{l=1}^{N_k-1}c_k^l, & l<N_k\end{cases} \tag{8}$$

由于$0 \leqslant s_k^l \leqslant 1$且$\sum_{l=1}^{N_k} s_k^l = 1$，如果对每个令牌的更新次数没有限制，则会倾向于尽可能长时间地等待，以避免出错。为了降低 TUTTFKI 跟踪的计算量，采用等待损失来鼓励令牌提前终止：

$$\mathcal{L}_{\text{ponder}} = \frac{1}{K}\sum_{k=1}^{K}\rho_k = \frac{1}{K}\sum_{k=1}^{K}(N_k + o_k) \tag{9}$$

式中ρ_k表示令牌$\boldsymbol{T}_k$的等待函数。然而，这种损失以$\frac{1}{K}$的权重平等地对待每个令牌，当学习关于令牌的先验知识时，会降低在模型训练阶段对令牌与目标或背景相关性的学习能力。为了进一步提高模型的令牌感知能力，等待损失进一步优化为$\mathcal{L}^*_{\text{ponder}}$：

$$\mathcal{L}^*_{\text{ponder}} = \frac{1}{K}\sum_{k=1}^{K}\rho_k(\mathrm{I}_{\mathrm{t}}(\boldsymbol{T}_k) + \omega_{\mathrm{b}}\mathrm{I}_{\mathrm{b}}(\boldsymbol{T}_k)) \tag{10}$$

式中：$\mathrm{I}_{\mathrm{t}}(\cdot)$和$\mathrm{I}_{\mathrm{b}}(\cdot)$为指示函数，定义为

$$\mathrm{I}_{\{\mathrm{t(b)}\}}(\boldsymbol{T}_k) = \begin{cases} 1, & \boldsymbol{T}_k\text{为目标（背景）令牌} \\ 0, & \text{其他} \end{cases} \tag{11}$$

$\omega_{\mathrm{b}} \geqslant 1$是用于缩放背景令牌的等待损失预定义常数。当$\omega_{\mathrm{b}} = 1$时，$\mathcal{L}^*_{\text{ponder}}$减小到$\mathcal{L}_{\text{ponder}}$。

3　特征聚合模块

基于 ViT 的网络模块输入将图像切割平展为图像块序列。目标边缘图像块通过关系建模机制后，可能由于背景特征更多，被标记为背景，从而提前终止推理。为确保跟踪器能够从浅层网络获得目标的纹理、形状等浅层特征，提出基于 Transformer 的多层特征聚合模块，以保留所有网络层的图像细节。

图 2 给出了特征聚合模块。在令牌关系建模过程中，每个令牌被分配到目标或背景类中。特征聚合模块记录标记为目标的令牌之间的关系。

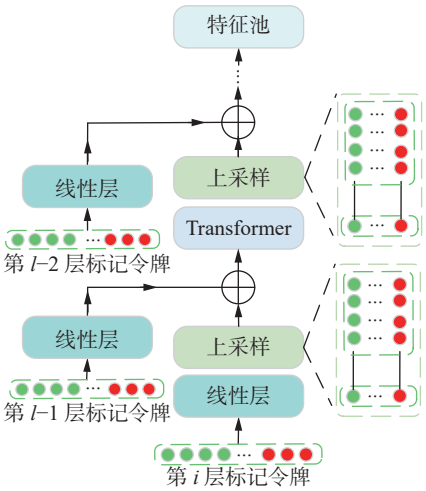


图 2　提出的特征聚合模块

Fig. 2　Proposed feature aggregation module

在标记上采样过程中，根据令牌的标记，将网络保存的令牌特征复制到相应的上采样标记中。将标记令牌上采样后，特征聚合模块继续将前一层网络提取的令牌特征添加到上采样令牌中。然后，这些目标令牌由 Transformer 块进行学习处理。通过逐步执行该操作，直到所有标记为目标类别的令牌被聚合。通过特征聚合模块，可以很容易地将各图像块的局部特征重塑为跟踪目标的全局特征图，以进行后续的目标跟踪。

4　实验结果与分析

所有实验使用 Ubuntu18.04，计算机配置 Intel Core I9-9900K CPU 和 32 GB 内存。TUTTFKI 预测头通过一个轻量级全卷积网络构建，每个输出都由 4 个堆叠的 Conv-BN-ReLU 层组成。根据目标跟踪对象的大小，将模板和搜索区域的尺寸分别设置为 128×128 和 256×256。通过自适应视觉 Transformer 网络[37] 的设计，将实验参数λ_{iou}设为 2，λ_{l}设为 5，α_{d}设为 0.000 1。

4.1　对比数据集

为了验证 TUTTFKI 算法的有效性，以及在多种复杂场景下的跟踪性能，在无人机视频数据集 UAV123、DTB70 及 UAVDark135 上，与 AutoTrack[14]、AbaTrack[38]、AVTrack[34]、DDCTrack[39]、HiFT[16]、SiamAPN++[18]、SiamRPN++[40]、SGLATrack[41]、TCTrack[4] 这 9 种跟踪算法进行对比实验。其中 UAV123 数据集包含共计 123 个无人机拍摄的视频序列，以及超过 11 万帧无人机图像，能够覆盖大部分无人机跟踪场景。DTB70 和 UAVDark135 两个数据集则为小目标跟踪和夜间跟踪图像，代表无人机跟踪的极端挑战环境。此外，在 UAV-123 数据集 12 个挑战属性上进行对比，其中包括长宽比变化（aspect ratio change, ARC）、背景杂波（background clutter, BC）、摄像机运动（camera motion, CM）、快速运动（fast motion, FM）、全遮挡（full occlusion, FOC）、照明变化（illumination variation, IV）、低分辨率（low resolution, LR）、视觉越界（over vision, OV）、部分遮挡（partially occlusion, POC）、尺度变化（scale variation, SV）、类似对象（similar object, SOB）和视点变化（viewpoint change, VC）。

4.2　评价指标

为了公平且有效地评估算法的跟踪精度，采用中心误差E_{CL}（center location error, CLE）和重叠率R_{o}（overlap rate, OR）作为评价指标。中心误差作为跟踪算法精度的评价指标，是指目标真实位置与方法所跟踪到的位置的欧氏距离：

$$E_{\mathrm{CL}}=\sqrt{((x_{\mathrm{t}}-x_{\mathrm{g}}))^2+(y_{\mathrm{t}}-y_{\mathrm{g}})^2} \tag{12}$$

式中：$(x_{\mathrm{t}},y_{\mathrm{t}})$表示算法预测的目标位置中心坐标，$(x_{\mathrm{g}},y_{\mathrm{g}})$表示目标真实位置的中心坐标。设定其距离小于 20 个像素点表示跟踪成功，反之则为失败。

重叠率作为跟踪算法成功率的评价指标，计算的是目标真实的边界框与方法跟踪到的边界框的重叠区域占整个目标区域的比例：

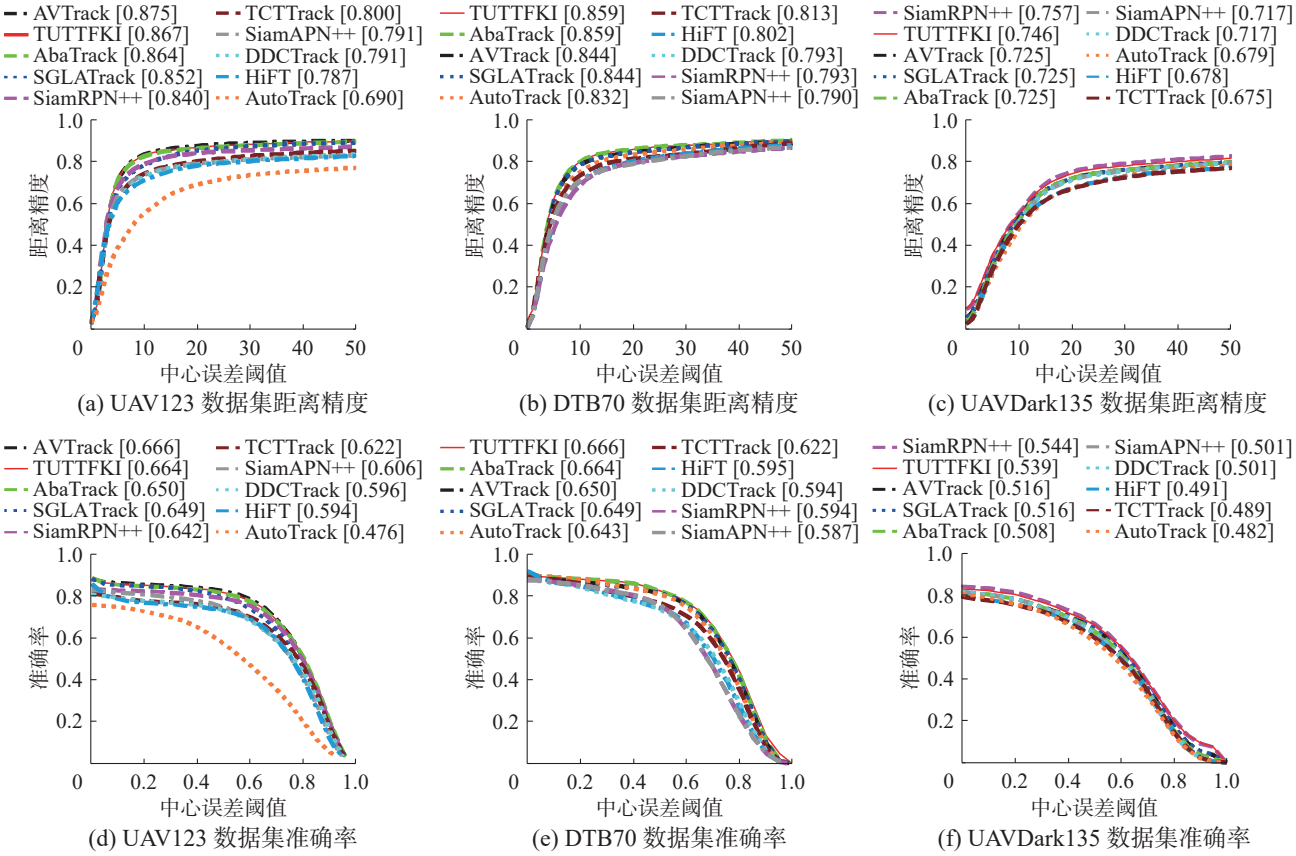
$$R_{\mathrm{o}}=\frac{\mathrm{Area}(B_{\mathrm{t}}\cap B_{\mathrm{g}})}{\mathrm{Area}(B_{\mathrm{t}}\cup B_{\mathrm{g}})} \tag{13}$$

式中：B_{t}表示算法预测的边界框，B_{g}表示目标真实边界框。设置阈值为 0.5，高于 0.5 则判断跟踪成功，反之则为失败。

4.3　结果分析

4.3.1　总体性能评估

图 3 给出了本文方法与对比的 9 种方法的综合评估效果。实验表明：本文提出的方法 TUTTFKI 在 UAV123 和 DTB70 数据集上的距离精度和准确率均取得最佳性能。



(a) UAV123 数据集距离精度　(b) DTB70 数据集距离精度　(c) UAVDark135 数据集距离精度

(d) UAV123 数据集准确率　(e) DTB70 数据集准确率　(f) UAVDark135 数据集准确率

图 3　本文方法与 9 种方法的综合评价结果

Fig. 3　Proposed method is evaluated comprehensively with the results of nine other methods

具体来说，在 UAV123 数据集上，TUTTFKI 的距离精度达到 0.867，准确率达到 0.664，在对比算法中均排名第二。这表明 TUTTFKI 的单流框架设计，集成了目标特征学习和目标搜索，有效增强目标模板和搜索区域的信息交互，提高对目标特征建模的能力。同时，自适应关系建模机制通过对目标令牌和搜索区域令牌进行关系建模，提前终止背景令牌的推理，进一步降低背景干扰，从而学习到更具鲁棒性的目标特征。

在 DTB70 数据集上，TUTTFKI 也取得了令人瞩目的成果，距离精度为 0.859，准确率为 0.666，均获得最优的表现。这说明即使在目标较小、有效信息不足的情况下，TUTTFKI 的自适应关系建模机制和特征聚合模块，仍然能够更专注于目标信息挖掘和提取，有效避免背景信息的影响。

在 UAVDark135 数据集上，TUTTFKI 的性能略逊于 SiamRPN++。TUTTFKI 的距离精度为 0.746，准确率为 0.539，而 SiamRPN++分别达到了 0.757 和 0.544。这可能是由于 UAVDark135 数据集主要包含低光照场景，给 TUTTFKI 的单流框架带来了挑战，使其难以在光线不足的情况下提取足够具有判别力的特征。SiamRPN++作为一种基于深度学习的跟踪器，在学习更鲁棒的特征表示方面可能更具优势，因此在低光照条件下表现更好。尽管如此，TUTTFKI 的自适应关系建模机制仍然发挥了作用，通过终止部分背景令牌的推

理, 尽可能地减少了背景干扰。

为进一步验证 TUTTFKI 在无人机跟踪场景中的时效性, 将其与对比算法中的 4 个基于单流 ViT 跟踪框架的无人机跟踪算法进行跟踪速率对比。通过表 1 给出的结果可见, 5 个算法都能达到实时要求的 25 帧/s, 其中 DDCTrack 通过设计全新的编码器模块, 简化了跟踪框架, 取得 97.2 帧/s 的优异跟踪速率。但是 TUTTFKI 通过构建自适应关系建模机制, 在提高跟踪速率的同时更加关注关键区域信息, 在 69.7 帧/s 跟踪速率下取得优异的跟踪精度, 更好地兼顾了跟踪速度和精度。

表 1　不同算法跟踪速率对比

Table 1　Comparison of tracking speed among different algorithms

基于单流ViT跟踪方法	跟踪速率/ (f/s)
AbaTrack	43.7
AVTrack	59.7
DDCTrack	97.2
SGLATrack	74.8
TUTTFKI	69.7

4.3.2　UAV123 数据集挑战属性分析

为了更加深入地评估所提算法在特定挑战下的跟踪性能, 表 2 和表 3 给出了 TUTTFKI 与其他 9 种先进跟踪器在 UAV123 数据集的 12 种挑战属性下的跟踪结果。结果表明: 在 FOC、POC 和 SOB 3 种属性中, TUTTFKI 取得了最佳性能。在其余的挑战属性中, 除了 ARC 属性外, TUTTFKI 均取得了次优结果, 说明 TUTTFKI 在无人机目标跟踪中表现优异。同时在 UAV123 和 DTB70 数据集上均达到了 69.73 帧/s 的实时跟踪速度。TUTTFKI 之所以能取得优异的性能, 一方面得益于视觉 Transformer 强大的特征提取能力, 这源于其在大规模数据集上的预训练, 使其能够更好地捕获目标的语义信息和特征表示。另一方面, TUTTFKI 将特征学习和目标搜索集成到单流网络中进行联合优化, 并通过自适应关系建模机制, 利用目标和背景的先验信息进行令牌关系建模, 自适应地抑制背景信息的干扰, 从而实现更高效、更鲁棒的无人机目标跟踪。

表 2　不同算法在 UAV123 数据集不同挑战场景的中心位置误差结果对比

Table 2　Comparison of center position errors of different algorithms in various challenging scenarios of the UAV123 dataset

算法名称	ARC	BC	CM	FM	FOC	IV	LR	OV	POC	SV	SOB	VC
AutoTrack	0.628	0.579	0.660	0.542	0.464	0.629	0.595	0.562	0.586	0.654	0.661	0.624
AbaTrack	0.857	0.652	0.898	**0.834**	0.773	0.818	0.739	0.842	0.822	0.850	0.876	0.895
AVTrack	**0.861**	**0.691**	**0.902**	0.828	0.750	**0.845**	**0.752**	**0.852**	0.826	**0.866**	0.877	**0.907**
DDCTrack	0.713	0.639	0.872	0.813	0.721	0.759	0.699	0.801	0.794	0.813	0.793	0.846
HiFT	0.703	0.676	0.728	0.661	0.619	0.709	0.694	0.650	0.704	0.724	0.738	0.690
SiamAPN++	0.741	0.608	0.780	0.743	0.571	0.740	0.636	0.732	0.704	0.765	0.693	0.798
SiamRPN++	0.818	0.635	0.863	0.774	0.661	0.819	0.690	0.816	0.771	0.820	0.800	0.899
SGLATrack	0.849	0.662	0.891	0.827	**0.793**	0.814	0.739	0.822	0.811	0.842	0.851	0.889
TCTrack	0.616	0.454	0.650	0.525	0.473	0.549	0.549	0.555	0.579	0.633	0.668	0.585
TUTTFKI	0.855	0.673	0.898	0.833	0.781	0.826	0.743	0.844	**0.828**	0.854	**0.878**	0.895

注: 加粗数字为最优结果, 带下划线数字为次优结果。

表 3　不同算法在 UAV123 数据集不同挑战场景的重叠率结果对比

Table 3　Comparison of overlap ratio results of different algorithms in various challenging scenarios of the UAV123 dataset

算法名称	ARC	BC	CM	FM	FOC	IV	LR	OV	POC	SV	SOB	VC
AutoTrack	0.415	0.345	0.465	0.363	0.233	0.404	0.338	0.401	0.392	0.444	0.447	0.424
AbaTrack	0.651	0.448	0.690	0.632	0.508	0.611	0.498	0.652	0.606	0.655	0.649	0.708
AVTrack	**0.655**	**0.474**	**0.691**	0.627	0.493	**0.633**	0.503	**0.659**	**0.607**	**0.690**	0.646	**0.712**
DDCTrack	0.624	0.399	0.607	0.582	0.468	0.583	0.481	0.592	0.571	0.603	0.589	0.657
HiFT	0.537	0.392	0.600	0.554	0.358	0.502	0.428	0.522	0.488	0.571	0.514	0.588
SiamAPN++	0.561	0.423	0.600	0.546	0.362	0.556	0.425	0.548	0.513	0.584	0.515	0.631

续表 3

算法名称	ARC	BC	CM	FM	FOC	IV	LR	OV	POC	SV	SOB	VC
SiamRPN++	0.614	0.448	0.658	0.581	0.425	0.607	0.454	0.609	0.563	0.623	0.590	0.682
SGLATrack	0.650	0.443	0.679	0.629	0.507	0.628	0.499	0.639	0.601	0.655	0.647	0.699
TCTrack	0.566	0.397	0.605	0.553	0.391	0.518	0.438	0.575	0.518	0.586	0.531	0.616
TUTTFKI	0.652	0.449	0.690	**0.632**	**0.512**	0.612	**0.508**	0.652	0.605	0.659	**0.650**	0.706

注：加粗数字为最优结果，带下划线数字为次优结果。

4.3.3　消融实验分析

为了验证提出的自适应关系建模机制和特征聚合模块的有效性，在 UAV123 数据集上进行消融实验，结果如表 4 所示。表 4 给出了不同模块组合的跟踪精度和准确率。

表 4　UAV123 数据集上的消融实验结果

Table 4　Ablation study results on the UAV123 dataset

自适应关系建模机制	特征聚合模块	跟踪精度	准确率	跟踪速率/(f/s)
×	×	0.806	0.638	70.47
×	√	0.821	0.642	60.34
√	×	0.846	0.651	83.67
√	√	0.867	0.664	69.73

注："√"表示启用该模块，"×"表示禁用该模块。

从表 4 可以看出，单独使用特征聚合模块，可以将跟踪精度从 0.806 提升到 0.821，准确率从 0.638 提升到 0.642。这表明特征聚合模块能够有效增强特征表示的判别力，从而提高跟踪性能。而单独使用自适应关系建模机制，则可以将跟踪精度提升到 0.846，准确率提升到 0.651，且将跟踪速率提升至 83.67 帧/s。这表明自适应关系建模机制在提高跟踪方法的计算效率的同时，能有效地减少背景干扰，并聚焦于目标区域，显著提高跟踪精度和准确率。

当同时使用自适应关系建模机制和特征聚合模块时，TUTTFKI 取得了最佳性能，精度达到 0.867，准确率达到 0.666。这表明两个模块之间存在协同作用，能够相互补充，共同提升跟踪性能。特征聚合模块通过融合多层特征，增强目标表示的判别力，而自适应关系建模机制则通过减少背景干扰，进一步提高跟踪的鲁棒性。两个模块的结合使得 TUTTFKI 能够更准确地定位目标，并在复杂场景下保持稳定的跟踪性能。

5　结束语

提出一种聚焦关键信息的目标感知 Transformer 无人机跟踪器，将特征学习和目标搜索集成到一个单流主干网络，提升了跟踪精度和效率。尽管 TUTTFKI 在大多数场景下表现出色，但在目标外观剧烈变化、遮挡及恶劣天气等复杂条件下，其跟踪性能仍有提升空间。未来的研究工作将着重从以下几方面展开：1）增强模型对目标外观变化的鲁棒性。将研究如何建模目标的时间上下文信息，例如引入运动估计、光流等技术，或考虑引入循环神经网络等模块以捕捉目标的时序信息，增强模型对目标外观变化的鲁棒性。2）提高模型对遮挡的应对能力。将探索利用多帧信息融合或引入记忆机制等方法，增强模型在目标被遮挡时的跟踪性能。3）模型轻量化。为了更好地将 TUTTFKI 部署到无人机等资源受限的平台，将进一步探索模型压缩和加速方法，例如模型剪枝、知识蒸馏、网络结构搜索等，或采用更轻量级的 Transformer 架构。

参考文献：

[1] WEN Longyin, DU Dawei, ZHU Pengfei, et al. Detection, tracking, and counting meets drones in crowds: a benchmark[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 7808−7817.

[2] 杜江涛, 于家明, 齐辉. 无人机集群不完全信息路径规划方法[J]. 哈尔滨工程大学学报, 2024, 45(11): 2210−2217.
DU Jiangtao, YU Jiaming, QI Hui. Incomplete information path planning method for an UAV cluster[J]. Journal of Harbin Engineering University, 2024, 45(11): 2210−2217.

[3] ALHAFNAWI M, BANY SALAMEH H A, MASADEH A, et al. A survey of indoor and outdoor UAV-based target tracking systems: current status, challenges, technologies, and future directions[J]. IEEE access, 2023, 11: 68324−68339.

[4] CAO Ziang, HUANG Ziyuan, PAN Liang, et al. TCTrack: temporal contexts for aerial tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern

Recognition. New Orleans: IEEE, 2022: 14778−14788.

[5] LI Shuiwang, LIU Yuting, ZHAO Qijun, et al. Learning residue-aware correlation filters and refining scale for real-time UAV tracking[J]. Pattern recognition, 2022, 127: 108614.

[6] BOLME D S, BEVERIDGE J R, DRAPER B A, et al. Visual object tracking using adaptive correlation filters [C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010: 2544−2550.

[7] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(3): 583−596.

[8] MA Chao, HUANG Jiabin, YANG Xiaokang, et al. Robust visual tracking via hierarchical convolutional features[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(11): 2709−2723.

[9] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6931−6939.

[10] FU Changhong, JIN Jin, DING Fangqiang, et al. Spatial reliability enhanced correlation filter: an efficient approach for real-time UAV tracking[J]. IEEE transactions on multimedia, 2021, 26: 4123−4137.

[11] HE Bing, WANG Fasheng, WANG Xing, et al. Temporal context and environment-aware correlation filter for UAV object tracking[J]. IEEE transactions on geoscience and remote sensing, 2024, 62: 5630915.

[12] HUANG Ziyuan, FU Changhong, LI Yiming, et al. Learning aberrance repressed correlation filters for real-time UAV tracking[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2891−2900.

[13] LI Bowen, FU Changhong, DING Fangqiang, et al. ADTrack: target-aware dual filter learning for real-time anti-dark UAV tracking[C]//2021 IEEE International Conference on Robotics and Automation. Xi'an: IEEE, 2021: 496−502.

[14] LI Yiming, FU Changhong, DING Fangqiang, et al. AutoTrack: towards high-performance visual tracking for UAV with automatic spatio-temporal regularization[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11920−11929.

[15] WEN Jiajun, CHU Honglin, LAI Zhihui, et al. Enhanced robust spatial feature selection and correlation filter learning for UAV tracking[J]. Neural networks, 2023, 161: 39−54.

[16] CAO Ziang, FU Changhong, YE Junjie, et al. HiFT: hierarchical feature transformer for aerial tracking[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 15437−15446.

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

[18] CAO Ziang, FU Changhong, YE Junjie, et al. SiamAPN: Siamese attentional aggregation network for real-time UAV tracking[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems. Prague: IEEE, 2021: 3086−3092.

[19] HUANG Bo, CHEN Junjie, XU Tingfa, et al. SiamSTA: spatio-temporal attention based Siamese tracker for tracking UAVs[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal: IEEE, 2021: 1204−1212.

[20] MA Siyu, LIU Yuting, ZENG Dan, et al. Learning disentangled representation in pruning for real-time UAV tracking[C]// Proc of the Asian Conference on Machine Learning. New York: PMLR, 2023: 690−705.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [EB/OL]. (2020−10−22)[2024−06−03]. https://arxiv.org/abs/2010.11929.

[23] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[24] WANG Wenhai, XIE Enze, LI Xiang, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 548−558.

[25] LI Shuiwang, YANG Xiangyang, WANG Xucheng, et al. Learning target-aware vision transformers for real-time UAV tracking[J]. IEEE transactions on geoscience and remote sensing, 2024, 62: 4705718.

[26] 崔阳洁, 屠展, 杨彬淇, 等. 无人机遮挡目标检测与协同跟踪方法[J]. 机器人, 2025, 47(3): 427.
CUI Yangjie, TU Zhan, YANG Binqi, et al. Occluded target detection and multi-UAV cooperative tracking method[J]. Robot, 2025, 47(3): 427.

[27] 刘芳, 卢晨阳, 路言, 等. 基于自适应模板更新的 Transformer 无人机目标跟踪算法[J]. 航空学报, 2025,

46(16): 331687.
LIU Fang, LU Chenyang, LU Yan, et al. Adaptive template update-based Transformer algorithm for UAV target tracking[J]. Acta aeronautica ET astronautica sinica, 2025, 46(16): 331687.

[28] KUGARAJEEVAN J, KOKUL T, RAMANAN A, et al. Transformers in single object tracking: an experimental survey[J]. IEEE access, 2023, 11: 80297−80326.

[29] MA Sugang, ZHAO Bo, HOU Zhiqiang, et al. SOCF: a correlation filter for real-time UAV tracking based on spatial disturbance suppression and object saliency-aware[J]. Expert systems with applications, 2024, 238: 122131.

[30] CUI Yutao, JIANG Cheng, WANG Limin, et al. MixFormer: end-to-end tracking with iterative mixed attention[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 13598−13608.

[31] CHEN Boyu, LI Peixia, BAI Lei, et al. Backbone is all your need: a simplified architecture for visual object tracking[M]//Computer Vision–ECCV 2022. Cham: Springer Nature Switzerland, 2022: 375−392.

[32] YE Botao, CHANG Hong, MA Bingpeng, et al. Joint feature learning and relation modeling for tracking: a one-stream framework[M]//Computer Vision–ECCV 2022. Cham: Springer Nature Switzerland, 2022: 341−357.

[33] CHEN Xin, YAN Bin, ZHU Jiawen, et al. Transformer tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8122−8131.

[34] LI Yongxin, LIU Mengyuan, WU You, et al. Learning adaptive and view-invariant vision transformer for real-time UAV tracking[C]//Proc of the Forty-first International Conference on Machine Learning. Piscataway: ACM, 2024.

[35] 卢丹, 侯娜. 融合 Transformer 的轻量化无人机目标跟踪算法[J]. 计算机工程与设计, 2024, 45(11): 3352−3359.
LU Dan, HOU Na. Lightweight UAV target tracking algorithm with Transformer[J]. Computer engineer and design., 2024, 45(11): 3352−3359.

[36] 谌海云, 王海川, 黄忠义, 等. 引入轻量级 Transformer 的无人机视觉跟踪[J]. 计算机工程与应用, 2024, 60(2): 244−253.
SHEN Haiyun, WANG Haichuan, HUANG Zhongyi, et al. UAV visual tracking with lightweight Transformer [J]. Computer engineering and applications, 2024, 60(2): 244−253.

[37] YIN Hongxu, VAHDAT A, ALVAREZ J M, et al. A-ViT: adaptive tokens for efficient vision transformer[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 10799−10808.

[38] LI Shuiwang, YANG Yangxiang, ZENG Dan, et al. Adaptive and background-aware vision transformer for real-time UAV tracking[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 13943−13954.

[39] DU Guocai, ZHOU Peiyong, YADIKAR N, et al. DDCTrack: dynamic token sampling for efficient UAV transformer tracking[M]//Pattern Recognition. Cham: Springer Nature Switzerland, 2024: 129−144.

[40] LI Bo, WU Wei, WANG Qiang, et al. SiamRPN: evolution of Siamese visual tracking with very deep networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4277−4286.

[41] XUE Chaocan, ZHONG Bineng, LIANG Qihua, et al. Similarity-guided layer-adaptive vision transformer for UAV tracking[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025: 6730−6740.

作者简介：

林淑彬，实验师，主要研究方向为计算机视觉和模式识别。参与福建省自然科学基金项目 1 项。发表学术论文 5 篇。E-mail: greenkure@163.com。

吴贵山，高级实验师，主要研究方向为计算机视觉和机器学习。参与福建省自然科学基金项目 2 项。发表学术论文 7 篇。E-mail: wuabcd@163.com。

杨文元，教授，博士，中国计算机学会（CCF）会员，主要研究方向为计算机视觉、模式识别和机器学习。主持福建省自然科学基金项目 1 项。发表学术论文 30 余篇。E-mail: yangwycn@163.com。