基于几何先验引导的可见光-事件目标跟踪方法

赵少川[1,2,3],刘峥玮[1,2,3],徐天阳[4],邵志文[1,2,3],周 勇[1,2,3*],吴小俊[4]

(1. 中国矿业大学计算机科学与技术学院/人工智能学院,江苏徐州 221116;2. 矿山数字化教育部工程研究中心,江苏徐州 221116;
3. 地下空间智能感知与应急物联江苏省产业技术工程化中心,江苏徐州 221116;
4. 江南大学人工智能与计算机学院,江苏无锡 214122)

摘 要: 可见光相机在运动模糊、极端光照等挑战性场景下易出现成像质量退化,导致现有基于可见光模态的目标跟踪方法性能受限。事件相机凭借高时间分辨率、高动态范围和低延迟等优势,可有效弥补可见光传感器的不足,因此可见光-事件目标跟踪逐渐成为研究热点。然而,现有方法高度依赖事件相机的成像质量,当目标处于静止或低速运动状态时,事件信号因缺乏光强变化而难以提供有效的鉴别性信息,制约了跟踪器在复杂场景下的鲁棒性。针对上述问题,本文提出一种基于几何先验引导的可见光-事件目标跟踪方法。考虑到事件信号与角点/边缘信号均对目标边缘区域敏感,且可见光图像在事件信号缺失时可提供稳定的空间几何结构信息,本文利用角点与边缘作为几何先验引导信号,以增强事件模态的表达能力。为实现有效融合,本文设计了两种几何先验引导策略:(1)像素级引导策略,将提取的角点/边缘灰度图与事件图像在通道维度拼接,直接作为跟踪模型输入;(2)特征级引导策略,分别提取事件图像与角点/边缘图像的深度特征,通过信号质量评估模块自适应学习融合权重,实现特征的动态加权融合。两种策略均具备良好的兼容性,可便捷嵌入现有可见光-事件跟踪架构,并以较低计算开销提升模型在事件模态成像质量不佳时的鲁棒性。本文将所提方法集成至TENet(Targetness Entanglement Network)、ViPT(Visual Prompt multi-modal Tracking)与SDSTrack(Self-Distillation Symmetric Tracking)3种先进可见光-事件跟踪模型中,在VisEvent、COESOT和FE240hz三个公开数据集上开展实验验证。实验结果表明,几何先验引导策略显著提升了跟踪精度,尤其在目标静止、形变、部分遮挡等挑战性场景下改善明显。以TENet为例,在FE240hz数据集上的成功率SUC(SUCcess rate)和精确率PRE(PREcision rate)分别提升了1.6%和2.3%;在涵盖多种跟踪场景的COESOT数据集上的两种指标分别提升了0.8%和1.2%;在VisEvent数据集上也分别获得了0.4%和0.5%的增益。综上,本文提出的几何先验引导方法通过引入角点/边缘信息作为事件信号的有效补充,缓解了事件相机在低动态场景下的信息缺失问题,为可见光-事件目标跟踪提供了一种高效且鲁棒的解决方案。

关键词: 计算机视觉;多模态目标跟踪;事件相机;信息融合;特征增强

基金项目: 国家自然科学基金(No.62272461,No.62020106012,No.62332008,No.62336004,No.62472424);中央高校基本科研业务费专项资金资助(No.2025QN1162,No.JUSRP202504007)

中图分类号: TN911.73 **文献标识码:** A **文章编号:** 0372-2112(2026)04-1719-13

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20251160

An RGB-Event Object Tracking Method Based on Geometry Prior Guidance

ZHAO Shaochuan[1,2,3], LIU Zhengwei[1,2,3], XU Tianyang[4], SHAO Zhiwen[1,2,3], ZHOU Yong[1,2,3*], WU Xiaojun[4]

(1. *School of Computer Science and Technology/School of Artificial Intelligence, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*; 2. *Mine Digitization Engineering Research Center of the Ministry of Education, Xuzhou, Jiangsu* 221116, *China*;
3. *Jiangsu Provincial Industrial Technology Engineering Center for Intelligent Sensing and Emergency IoT in Underground Space, Xuzhou, Jiangsu* 221116, *China*; 4. *School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu* 214122, *China*)

Abstract: RGB cameras often suffer from image quality degradation in challenging scenarios such as motion blur and extreme lighting conditions, which limits the performance of existing RGB-based tracker. Event cameras, with their advantages of high temporal resolution, high dynamic range, and low latency, can effectively compensate for the shortcomings of RGB sensors, making RGB-Event object tracking a gradually emerging research focus. However, existing methods heavily rely on the imaging quality of event cameras. When the object is stationary or moving at low speeds, event signals fail to provide effective discriminative information due to the lack of intensity changes, thereby constraining the robustness of trackers in complex scenarios. To address the above issues, this paper proposes an RGB-Event object tracking method guid-

收稿日期:2026-01-26;录用日期:2026-03-19;责任编委:王栋
*通信作者:周勇

ed by geometric priors. Considering that both event signals and corner/edge signals are sensitive to object edge regions, and that RGB images can provide stable spatial geometric structure information in the absence of event signals, this paper employs corners and edges as geometric prior guidance signals to enhance the representational capability of the event modality. To achieve effective fusion, two geometric prior guidance strategies are designed: 1) a pixel-level guidance strategy, which concatenates the extracted corner/edge grayscale maps with event images along the channel dimension as direct input to the tracking model; and 2) a feature-level guidance strategy, which separately extracts high-level features from event images and corner/edge maps, and adaptively learns fusion weights through a signal quality assessment module to achieve dynamic weighted fusion of features. Both strategies offer good compatibility, can be conveniently integrated into existing RGB-Event trackers, enhancing tracking robustness under poor event modality imaging quality with low computational overhead. The proposed method is integrated into three advanced RGB-Event trackers, namely TENet, ViPT, and SDSTrack, and experimental validation is conducted on three public datasets: VisEvent, COESOT, and FE240hz. The results demonstrate that the geometric prior guidance strategy significantly improves tracking accuracy, particularly in challenging scenarios such as object stationary, deformation, and partial occlusion. Taking TENet as an example, the success rate(SUC) and precision rate (PRE) on the FE240hz dataset are improved by 1.6% and 2.3%, respectively; on the COESOT dataset, which covers various tracking scenarios, the two metrics are improved by 0.8% and 1.2%, respectively; and on the VisEvent dataset, gains of 0.4% and 0.5% are achieved, respectively. In summary, the proposed geometric prior guidance method effectively alleviates the information deficiency problem of event cameras in low-dynamic scenarios by introducing corner/edge information as a complement to event signals, providing an efficient and robust solution for RGB-Event object tracking. Our codes are athttps://github.com/LLLzw444/Enhancement-strategy-of-TENet.

Keywords: computer vision; multi-modal object tracking; event camera; information fusion; feature enhancement

Foundation Item(s): National Natural Science Foundation of China (No.62272461, No.62020106012, No.62332008, No.62336004, No.62472424); Fundamental Research Funds for the Central Universities (No.2025QN1162, No.JUSRP202504007)

0 引言

视觉目标跟踪是计算机视觉领域的核心任务之一。给定目标物体在视频序列首帧中的初始状态(包括目标的中心位置和尺寸)后,视觉目标跟踪器负责自动预测目标物体在后续帧中的状态。近年来,该技术已在视觉导航[1]和自动驾驶[2]等多个领域得到广泛应用,成为计算机视觉的重要分支[3-6]。

随着深度学习技术的发展,基于可见光模态的目标跟踪方法已经取得了显著进展。具体而言,基于孪生网络(siamese network)的方法[7-8]凭借其在精度与速度间的良好平衡,成为视觉目标跟踪领域的主流范式。该类方法将来自视频首帧的模板图像和后续帧中的搜索区域映射到同一特征空间,并通过相关操作计算二者的相似度,从而区分目标和背景。然而,模板-搜索区域匹配过程中的线性相关操作在建模全局上下文信息时存在固有局限,成为其跟踪性能进一步提升的瓶颈。为此,基于 Transformer 的目标跟踪方法[9-10]利用注意力机制,将特征提取和模板-搜索区域交互过程统一在一个框架中,能够高效捕捉特征间的长距离依赖关系,从而在多个主流视觉跟踪数据集上取得了显著的性能提升。尽管上述方法从模型建模的角度提升了跟踪器的鉴别能力,但其有效性依赖于可见光相机对场景的高质量成像。一旦遭遇诸如运动模糊、极端光照等挑战性场景,可见光相机在低帧率和低动态范围方面的劣势导致其难以提供有效信息,从而使得在复杂环境下实现稳定、精确的目标跟踪面临较大困难。

为提升跟踪器在全天候应用中的鲁棒性,已有诸多研究通过融合多模态互补信息以克服可见光感知的局限性[11]。其中,事件相机凭借其高时间分辨率、高动态范围和低延迟等优势,能够显著缓解低光照、过曝光和运动模糊等复杂条件下的跟踪性能退化问题,因此近年来受到广泛关注和深入研究[12]。

当前,可见光-事件目标跟踪方法(rgb-event object tracking)主要可分为两类。第一类方法聚焦于设计面向事件模态数据的专用特征提取与跨模态交互模块,并对预训练的可见光特征提取网络进行全面微调,以实现从可见光单模态到可见光-事件模态的迁移学习[13]。尽管该类方法能够针对事件数据特性进行建模,但其可优化参数量较大,而事件模态训练数据相对有限。因此,此类方法不仅增加了训练难度,也限制了模型的泛化性能。第二类方法则采用提示学习(prompt learning)[14]、适配器(adapter)[15]或低秩适应(Low-Rank Adaptation, LoRA)[16]等高效微调策略,在保持预训练可见光跟踪模型主干参数固定的前提下,仅引入少量可训练参数以适应可见光-事件目标跟踪任务,从而显著降低了跨模态学习的计算成本与资源开销。

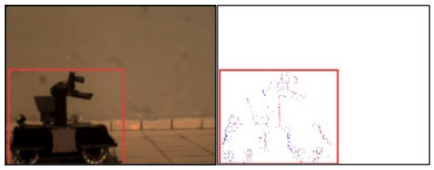
尽管事件模态的引入有效缓解了因可见光成像

质量低而导致的跟踪失败问题,但现有方法的性能高度依赖事件相机的成像质量。例如,在图1(a)和(b)所示场景中,事件数据能够有效增强或作为可见光信息的有机补充;然而,当事件图像成像质量较差时(见图1(c)),其不仅难以准确反映场景的真实状态,还可能作为干扰信息,影响跨模态特征交互过程。

针对上述问题,本文提出一种基于几何先验引导的可见光-事件目标跟踪方法。考虑到事件信号和角点/边缘信号两者本质上对目标的边缘区域较为敏感。然而,事件数据在目标静止时会因缺乏光强变化而中断,此时通过可见光图像提取的角点/边缘作为静态几何先验,实际上是从空间结构的角度模拟了事件数据本应在光强变化时的几何结构信息。因此当动态事件流缺失时,注入的空间几何信息能够提供有效目标鉴别性信息,这符合多模态感知中利用传感器间信息冗余进行互补融合的通用范式。根据信息融合阶段的不同,本文设计了两种引导策略:(1)基于像素级的几何先验引导策略,使用角点/边缘检测算法从可见光图像提取几何特征,并将其以灰度图形式与事件图像在通道维度拼接,作为跟踪模型的输入;(2)基于特征级的几何先验引导策略,将角点/边缘信息输入特征提取模块,通过信号质量评估模块量化其与事件特征的鉴别性,最终以加权方式实现二者的自适应融合。值得注意的是,这两种策略均具备良好的兼容性与可插拔性,能够以较低计算开销嵌入现有的可见光-事件目标跟踪架构,提升模型在事件模态成像质量不佳时的鲁棒性。

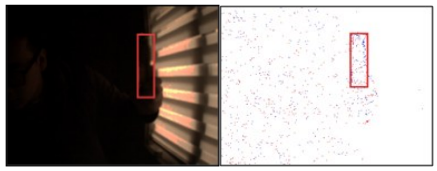
为验证所提出方法的有效性与泛化能力,本文将其集成至3个先进的可见光-事件目标跟踪方法ViPT(VIsual Prompt multi-modal Tracking)[14]、SDSTrack(Self-Distillation Symmetric Tracking)[15]与TENet(Targetness Entanglement Network)[13]中,并在公开数据集VisEvent[11]、COESOT(COlor-Event camera based Single Object Tracking)[17]和FE240hz[18]上进行测试。实验结果表明,所提出的几何先验引导策略显著提升了跟踪精度,尤其在目标处于静止或低速运动场景中改善显著。

综上所述,本文主要贡献如下:(1)利用可见光图像中的角点/边缘信息引导事件信号,以缓解事件相机在成像质量低下时的跟踪性能退化问题;(2)设计了基于像素级和特征级的几何先验引导策略,实现对事件信号及角点/边缘信息的自适应融合;(3)在3个公开数据集和3个基准模型上的实验结果表明,本方法具有良好的兼容性与有效性,能够稳定提升现有跟踪模型的精度。



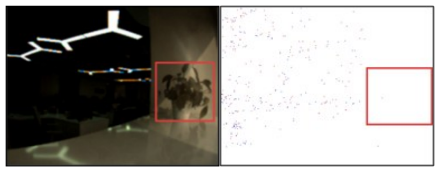
(a) 可见光和事件相机高质量成像的示例

(a) Examples of high quality RGB image and high quality event image



(b) 可见光相机低成像质量、事件相机高成像质量的示例

(b) Examples of low quality RGB image and high quality event image



(c) 可见光相机高成像质量、事件相机低成像质量的示例

(c) Examples of high quality RGB image and low quality event image

图1　本文的研究动机

Figure 1　The motivation of this research

1　相关工作

1.1　基于可见光模态的目标跟踪

近年来,随着深度学习技术的快速发展,基于深度学习的目标跟踪算法获得了广泛关注。其中,SiamFC(Siamese Fully-Convolutional)[7]采用孪生网络结构,将视觉跟踪问题转化为模板与搜索区域间的相似性匹配任务,实现了高效的目标跟踪。然而,SiamFC仅实现了对目标粗糙的定位,缺乏精确的尺度估计机制,无法应对目标大尺度变化的跟踪场景。为此,Li等人[8]提出的SiamRPN(Siamese Region Proposal Network)和SiamRPN++[19]将区域提案网络(Region Proposal Networks,RPN)模块作为边框回归分支,通过对预设锚框(anchor)的精调,高效地完成目标边框的自适应回归。在此基础上,吴非等人[20]引入光流信息作为运动补充,利用全局光流辅助搜索目标,提升了在遮挡和快速运动下的跟踪性能。SiamMask(Siamese Mask)[21]则引入分割分支,通过预测像素级的二值掩码更细粒度地描述目标状态。谢青松等人[22]将目标分割技术与SiamMask结合,并提出了前景优化框架,通过评估前景比例来动态触发尺度优化和角度

优化模块,解决了复杂运动导致的跟踪框包含过多背景的问题。SiamBAN(Siamese Box Adaptive Network)以锚点为中心预测其到目标边界的距离[23],从而舍弃了经由锚框中介的边框回归,避免了一系列超参数的引入,使该方法更加灵活通用。尽管上述方法取得了显著进步,但它们主要依赖线性相关或卷积操作进行信息交互,这种线性建模方式限制了特征表达能力,成为跟踪性能进一步提升的瓶颈。

近年来,基于Transformer的深度神经网络凭借强大的注意力机制,在多个任务上实现性能突破。TransT(Transformer Tracking)[9]将Transformer引入目标跟踪领域,其设计了一种基于注意力机制的特征融合模块,利用自注意力和交叉注意力增强特征的上下文语义。考虑到跟踪过程中有目标形变和跳出视野范围外的情况,Yan等人[24]提出了基于时空建模的Transformer跟踪方法STARK(Spatio-Temporal trAnsfoRmer network for visual tracKing),该方法引入了一种在线更新模板机制,依据对跟踪响应图的质量评估,实现在线模板的自适应选择。然而STARK模型的计算复杂度较高。为实现更简洁高效的目标跟踪,Ye等人[10]则提出一种单流跟踪框架OSTrack(One-Stream Tracking),将特征提取与关系建模任务相结合,良好平衡了跟踪精度与推理速度。因此,该方法被广泛应用于多模态跟踪模型中的可见光模态特征提取分支[25]。

然而,上述跟踪器均以可见光相机作为唯一传感器,其在低光照、运动模糊等恶劣环境下的成像质量无法得到保障。因此,现有基于可见光模态的目标跟踪方法难以从根本上突破由传感器成像质量问题所带来的跟踪失败问题。

1.2 基于可见光-事件模态的目标跟踪

鉴于可见光相机在成像上的局限性,事件相机凭借其在高速运动、极端光照条件下的优异表现,以及低延迟、低功耗的特性,逐渐成为目标跟踪领域的研究热点[26]。该相机可对像素对数亮度变化异步且独立地作出响应:当某一像素自上次事件触发后累积的亮度变化超出设定阈值时,则触发新的事件。

鉴于深度学习需要的足量训练数据,Wang等人[11]提出了大规模可见光-事件模态目标跟踪数据集VisEvent,并将事件流转化为事件图像,以统一模型输入的格式。然而,该操作大幅压缩了时间数据的时间分辨率,增加了信息丢失的风险。为此,Tang等人[17]将事件流信息转换成高时间分辨率的体素格式,并通过交叉注意力机制充分融合多模态信息,大幅提高了模型性能。由于相机间物理位置和时空分辨率的差异,可见光和事件数据并非完美对齐。为此,AFNet(Alignment and Fusion Network)[27]提出了由事件引导的跨模态对齐和交叉相关融合模块,协同实现了可见光与事件数据的对齐与融合。

然而,从可见光模态跟踪模型到可见光-事件模态跟踪模型的迁移学习需消耗大量计算资源,且对于训练数据的数量和质量要求较高。为了解决这些问题,ViPT[14]将事件信号作为视觉提示以实现对基于可见光模态跟踪模型的微调。尽管其简单有效,但不对称的网络架构导致其对可见光模态的过度依赖,限制了多模态数据互补的优势。相反地,SDSTrack[15]针对可见光和事件模态信息,设计了对称的网络架构用于特征提取,并采用基于适配器的模型微调策略以协调多模态信息融合。此外,UnTrack(Unified Tracker)[16]使用低秩适应技术学习跨模态的共享特征表示,将不同模态特征映射到同一空间中,实现跨模态特征对齐,从而大幅减少了模型的参数量和计算量,提升了跟踪模型的训练效率。

上述方法主要在多模态数据的特征提取和融合策略层面进行改进,并没有针对事件模态的特性作专门的设计。考虑到事件数据的空间稀疏性,TENet[13]提出了基于多尺度池化层的事件模态特征提取器,并利用共同引导融合模块,实现了充分的跨模态交互。尽管该方法充分考虑了事件数据的表征形式、特征提取方式以及融合策略,但依然无法解决由事件相机成像质量不佳而引发的跟踪性能退化问题。

1.3 角点检测和边缘检测

在角点检测领域,多种经典算法各具特点。具体来说,Harris算法[28]通过响应图像局部梯度变化,实现了高精度的角点定位。为解决尺度变化问题,SIFT(Scale Invariant Feature Transform)算法[29]通过构建高斯尺度空间,确保了检测的尺度不变性。KAZE算法[30]基于非线性扩散理论构建尺度空间,提升了在模糊和非均匀光照等复杂场景中的检测鲁棒性。ORB(Oriented FAST and Rotated BRIEF)算法[31]融合了FAST特征检测算法[32]和BRIEF(Binary Robust Independent Elementary Features)特征描述符[33]。其核心机制是:对于图像中的某一像素,若在其圆周邻域内存在连续像素灰度值同时高于或低于给定阈值,则判定该点为角点。

在边缘检测领域,经典算法主要基于图像微分运算,通过计算亮度函数的空间梯度来识别强度突变区域,从而定位物体轮廓与边界。其中,Sobel算子[34]通过一阶导数定位边缘,并对噪声具备一定鲁棒性。Laplace算子[35]利用二阶导数信息,能够检测边缘以及孤立点和线端点,但对噪声较为敏感。为获得更优的检测效果,Canny算法[36]通过高斯滤波、梯度计算、

非极大值抑制和双阈值处理等多个阶段,能够提取精确的单像素宽度边缘。贾迪等人[37]通过引入曼哈顿距离,对RGB三通道进行处理,从而能够更好地利用彩色图像的信息进行边缘检测。本文通过角点和边缘检测获取可见光图像中物体的几何结构信息,以引导低质量事件信号,从而进一步提升跟踪的鲁棒性。

2　基于几何先验引导的可见光-事件目标跟踪方法

给定可见光图像与事件流数据作为可见光-事件模态目标跟踪模型的输入。首先,将固定时间窗口内的事件流合成为事件图像,随后将该事件图像与可见光图像共同输入跟踪模型,由模型输出目标边界框的位置预测。该过程可形式化表示为如下公式:

$$\boldsymbol{B}=H\left(\varphi_{\mathrm{R}}\left(\boldsymbol{I}^{\mathrm{R}}\right),\varphi_{\mathrm{E}}\left(\boldsymbol{I}^{\mathrm{E}}\right)\right) \tag{1}$$

其中,$\boldsymbol{I}^{\mathrm{R}}$和$\boldsymbol{I}^{\mathrm{E}}$分别是可见光和事件模态图像(R和E分别代表可见光和事件),$\varphi_{\mathrm{R}}(\cdot)$和$\varphi_{\mathrm{E}}(\cdot)$分别是可见光和事件模态分支的特征提取模块,$H(\cdot)$表示可见光-事件模态特征融合和预测头模块,$\boldsymbol{B}$为模型预测的目标边界框。

然而,在目标静止或低速运动等情况下,事件数据难以提供有效信息。为提升跟踪器在此类场景下的鲁棒性,本文提出一种基于几何先验引导的目标跟踪方法。具体来说,引导信号的提取过程可形式化表示为

$$\boldsymbol{I}^{\hat{\mathrm{R}}}=\mathrm{Guidance}(\boldsymbol{I}^{\mathrm{R}}) \tag{2}$$

本文选用角点或边缘检测算法作为Guidance(·)函数的具体实现。由于角点与边缘信息对光照变化较为敏感,且能有效刻画物体的几何结构特征,本文将其用作事件数据的引导。针对事件模态数据与引导信号的协同表示问题,本文进一步提出了基于像素级和特征级两种几何先验引导策略。

2.1　基于像素级的几何先验引导策略

给定事件图像$\boldsymbol{I}^{\mathrm{E}}$和引导图像$\boldsymbol{I}^{\hat{\mathrm{R}}}$,基于像素级引导策略在通道维度将两者进行级联,从而实现角点/边缘信息对事件数据的引导:

$$\boldsymbol{I}^{\hat{\mathrm{E}}}=\mathrm{Cat}\left(\boldsymbol{I}^{\mathrm{E}},\boldsymbol{I}^{\hat{\mathrm{R}}}\right) \tag{3}$$

其中,Cat(·)表示级联操作,$\boldsymbol{I}^{\hat{\mathrm{E}}}$代表得到引导后的事件数据。与式(1)中直接使用原始事件图像作为输入不同,该策略将引导后的事件信号作为模型输入,进而同可见光特征进行交互和融合:

$$\boldsymbol{B}=H\left(\varphi_{\mathrm{R}}\left(\boldsymbol{I}^{\mathrm{R}}\right),\varphi_{\hat{\mathrm{E}}}\left(\boldsymbol{I}^{\hat{\mathrm{E}}}\right)\right) \tag{4}$$

其中,$\varphi_{\hat{\mathrm{E}}}(\cdot)$为针对引导后的事件数据的特征提取模块。该策略的完整流程如图2所示。值得注意的是,基于像素级的几何先验引导策略对可见光-事件模态跟踪网络架构的改动极小。具体而言,仅需修改网络第一层的输入通道数,以适应从$\boldsymbol{I}^{\mathrm{E}}$到$\boldsymbol{I}^{\hat{\mathrm{E}}}$的维度变化。因此,该策略具有较好的兼容性,能方便地集成到主流可见光-事件跟踪方法中,且几乎不引入额外计算开销。

然而,像素级融合对可见光-事件模态之间的数据对齐要求较高,且在融合策略的设计上存在一定局限性。为此,本文进一步提出了基于特征级的几何先验引导策略。

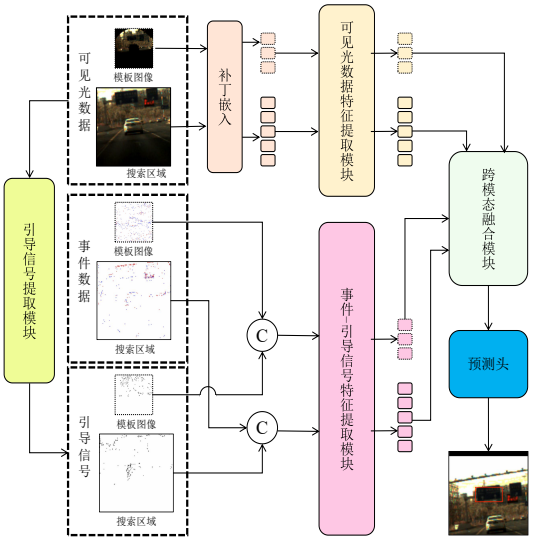


图2　基于像素级的几何先验引导策略

Figure 2　Pixel-level geometric prior guidance strategy

2.2　基于特征级的几何先验引导策略

基于特征级的几何先验引导策略将事件图像$\boldsymbol{I}^{\mathrm{E}}$和引导图像$\boldsymbol{I}^{\hat{\mathrm{R}}}$分别输入各自独立的特征提取网络,生成高层次的特征:

$$\boldsymbol{F}^{\mathrm{E}}=\varphi_{\mathrm{E}}\left(\boldsymbol{I}^{\mathrm{E}}\right),\boldsymbol{F}^{\hat{\mathrm{R}}}=\varphi_{\hat{\mathrm{R}}}\left(\boldsymbol{I}^{\hat{\mathrm{R}}}\right) \tag{5}$$

其中,$\varphi_{\mathrm{E}}(\cdot)$和$\varphi_{\hat{\mathrm{R}}}(\cdot)$分别表示面向事件图像和引导图像的特征提取器,二者结构相同但参数不共享。$\boldsymbol{F}^{\mathrm{E}}$和$\boldsymbol{F}^{\hat{\mathrm{R}}}$分别对应事件信号和引导信号的特征。随后,将二者特征按通道维级联起来,并输入信号质量评估(Signal Quality Assessment,SQA)模块中:

$$\alpha_{\hat{\mathrm{R}}},\alpha_{\mathrm{E}}=\mathrm{Softmax}\left(\mathrm{SQA}\left(\mathrm{Cat}\left(\boldsymbol{F}^{\mathrm{E}},\boldsymbol{F}^{\hat{\mathrm{R}}}\right)\right)\right) \tag{6}$$

如图3所示,该模块依次由3层3×3卷积层(convolutional layer)调制级联特征通道、再由全局池化层(Global Average Pooling,GAP)和一个多层感知机(MultiLayer Perceptron,MLP)组成。其中,卷积层负责对级联后特征作进一步调制,全局池化层压缩特征的空间维度以便后续多层感知机的处理。在经由

Softmax 函数激活之后，输出事件特征和引导特征的融合权重 α_{E} 和 $\alpha_{\hat{\mathrm{R}}}$。事件信号特征和引导信号特征的自适应加权融合过程可被表示为

$$\boldsymbol{F}^{\hat{\mathrm{E}}}=\alpha_{\mathrm{E}}\boldsymbol{F}^{\mathrm{E}}+\alpha_{\hat{\mathrm{R}}}\boldsymbol{F}^{\hat{\mathrm{R}}} \tag{7}$$

其中，$\boldsymbol{F}^{\hat{\mathrm{E}}}$ 表示引导后的事件特征。因此，基于特征级几何先验引导策略的目标跟踪过程可被形式化表示为

$$\boldsymbol{B}=H\left(\varphi_{\mathrm{R}}\left(\boldsymbol{I}^{\mathrm{R}}\right),\boldsymbol{F}^{\hat{\mathrm{E}}}\right) \tag{8}$$

该策略仍保持较强的可插拔性，因其设计与具体的特征提取、关系建模及预测头模块无关。与像素级引导策略相比，基于特征级的引导策略实现了动态信息融合，即所提出的信号质量自适应评估模块可根据具体跟踪场景的事件和引导信号的质量分配特征融合权重，从而突出鉴别性更强的特征来源，抑制不可靠信号。然而，该策略也存在一定局限性。一方面，信号质量自适应评估模块的加入不可避免地增加了模型的参数量和计算量，降低了跟踪效率；另一方面，由于角点/边缘信号反映了物体的结构性信息，在特征提取过程空间维度降维操作可能破坏其完整性，进而潜在地影响跟踪性能。

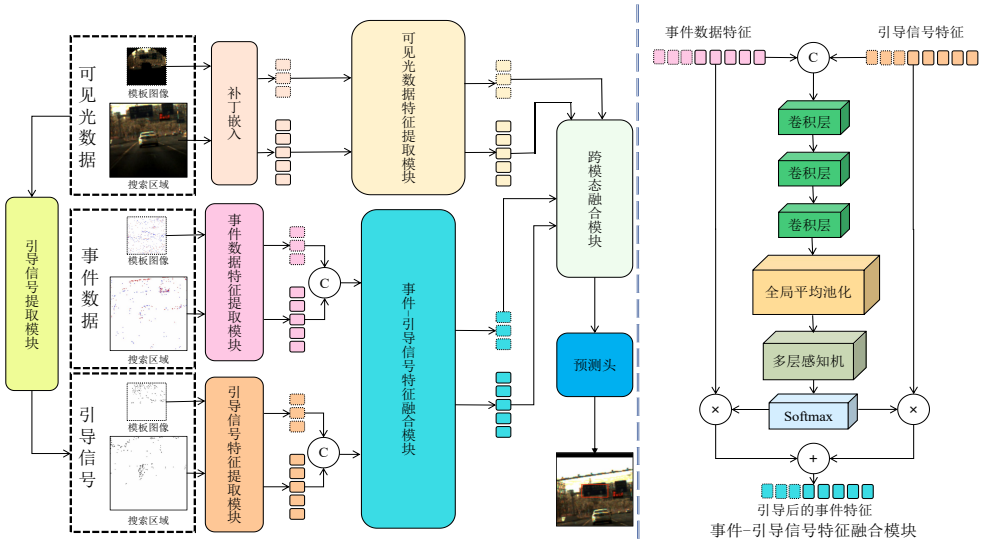


图3 基于特征级的几何先验引导策略

Figure 3 Feature-level geometric prior guidance strategy

2.3 损失函数

本文将真实边界框内对应的特征向量作为正样本，其余则作为负样本。所有样本都会对分类损失产生贡献，而回归损失仅针对正样本。

在分类任务中采用标准二分类交叉熵损失函数，其定义为

$$L_{\mathrm{cls}}=-\sum_{j}\left[y_j\log\left(p_j\right)+(1-y_j)\log(1-p_j)\right] \tag{9}$$

其中：y_j 表示第 j 个样本的标签，$y_j=1$ 代表正样本；p_j 表示跟踪模型预测为前景的概率。在回归任务中，根据预测边界框和归一化后的真实框，采用 L_1 范数损失与广义交并比损失 L_{iou} 的线性组合。L_1 与 L_{iou} 的定义分别为

$$L_1=\frac{\sum_{j}\left|\boldsymbol{B}_j-\hat{\boldsymbol{B}}_j\right|}{n} \tag{10}$$

$$L_{\mathrm{iou}}=-\ln\left(\frac{\mathrm{Intersetion}\left(\boldsymbol{B}_j,\hat{\boldsymbol{B}}_j\right)}{\mathrm{Union}\left(\boldsymbol{B}_j,\hat{\boldsymbol{B}}_j\right)}\right) \tag{11}$$

其中，n 为作用对象元素个数，$\boldsymbol{B}_j$ 为第 j 个预测边界框，$\hat{\boldsymbol{B}}_j$ 表示归一化的真实边界框，$\mathrm{Intersetion}(\cdot)$ 表示交集，$\mathrm{Union}(\cdot)$ 为并集。综上，本文设置正则化参数 $\lambda_{\mathrm{iou}}=2$ 和 $\lambda_1=5$，总体损失函数如下所示：

$$L=L_{\mathrm{cls}}+\lambda_{\mathrm{iou}}L_{\mathrm{iou}}+\lambda_1L_1 \tag{12}$$

3 实验结果及分析

3.1 实验环境与参数设置

实验使用的操作系统为 Ubuntu22.04.5 64 位，系统硬件设施为 Intel（R）Xeon（R）Gold 6330 CPU @ 2.00 GHz、NVIDIA GeForce RTX 3090（24 576 MiB）。编译环境为 Python3.10 和 PyTorch1.11.0。

本文采用 ORB[31] 与 Sobel[34] 算法作为式（2）中的角点和边缘检测算法。由于所提出的几何先验引导策略具有良好的兼容性，可即插即用于主流可见光-事件目标跟踪模型，如无特殊说明，实验均沿用对应基准模型的超参数设置。跟踪模型的训练集为 VisEvent[11]、COESOT[17] 与 FE240hz[18]。

实验部分采用准确率(PREcision rate,PRE)和成功率(SUCcess rate,SUC)作为评估指标。准确率定义为在整个测试序列中,预测边界框中心与真实边界框中心之间距离小于给定阈值的帧所占比例;成功率则指预测边界框与真实边界框的交并比(Intersection of Union,IoU)大于0.5的帧所占比例。此外,每秒帧数(Frames Per Second,FPS)用于评估模型的计算效率,表示每秒可处理的图像帧数,本文方法在计算FPS时,既包含了在CPU上对可见光图像进行角点/边缘提取的时间,也包含了在GPU上运行深度神经网络推理的时间,二者以实时、串行的方式执行。

3.2　与其他先进跟踪方法的性能比较

本文在VisEvent[11]、COESOT[17]和FE240hz[18]测试集上对所提出的事件信号引导后的可见光-事件目标跟踪方法进行了性能验证。实验选取了3种可见光-事件跟踪器——TENet[13]、ViPT[14]和SDSTrack[15]作为基准模型,通过角点图像对事件信号进行引导,并将引导后的模型与多种主流跟踪方法进行综合比较。

3.2.1　在VisEvent数据集上的性能比较

VisEvent[11]是一个大规模可见光-事件双模态数据集,包含820对可见光图像与事件流视频序列,涵盖17个不同类别。该数据集划分为500个训练序列和320个测试序列。在该数据集上的对比方法包括TENet[13]、ViPT[14]、SDSTrack[15]、UnTrack[16]、OSTrack[10]、ProTrack(Prompt Tracker)[38]与PrDIMP50(Probabilistic DIscriminative Model Prediction)[39]。

图4展示了本文方法与对比方法在VisEvent数据集上的评估结果。其中,Ours(TENet)、Ours(ViPT)和Ours(SDSTrack)分别代表了使用几何先验引导策略的TENet、ViPT和SDSTrack跟踪模型。具体而言,本文策略使TENet、ViPT和SDSTrack在SUC指标上分别提升了0.4%、0.4%与0.6%,在PRE指标上分别提升了0.5%、0.5%与0.6%。此外,引导后的TENet模型在两项评价指标上均取得了最优结果,在SUC和PRE指标上分别达到60.4%和74.1%。这表明具有物体几何结构信息的引导信号可作为事件模态信号的有效补充。

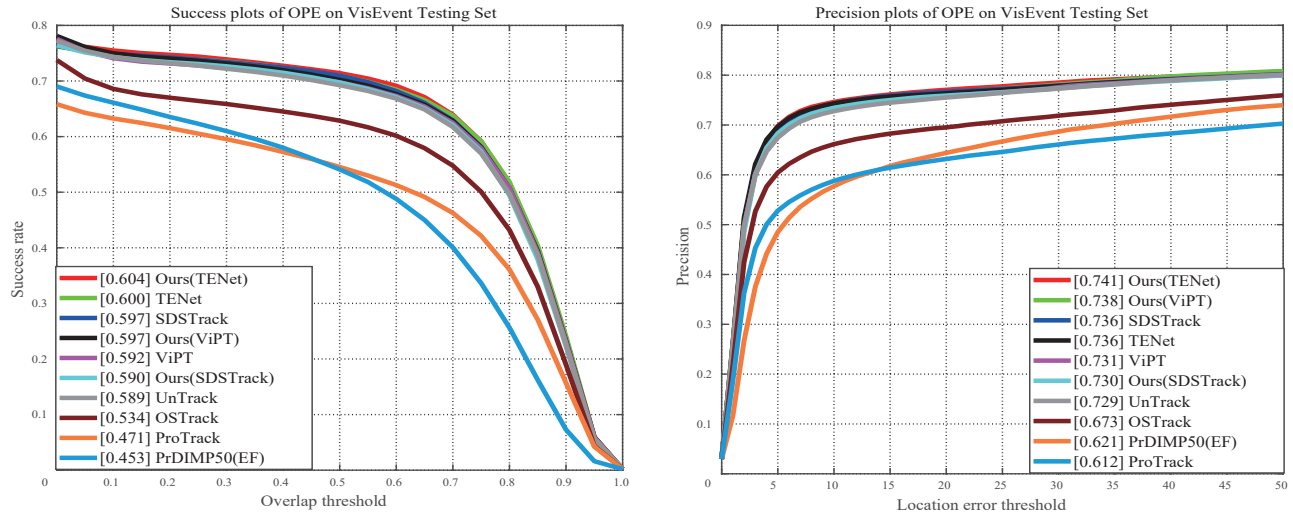


图4　与其他先进跟踪器在VisEvent数据集上的性能比较

Figure 4　Performance comparison with other advanced trackers on the VisEvent dataset

3.2.2　在COESOT数据集上的性能比较

COESOT[17]是一个涵盖90个类别的大规模数据集,包含1 354个视频序列,其中训练集827个,测试集527个。在该数据集上的对比方法包括TENet[13]、ViPT[14]、SDSTrack[15]、AFNet[27]、HDETrack(Hierarchical knowledge Distillation framework for Event stream based Tracking)[40]、CEUTrack(Color-Event Unified Tracking)与CSAM(Cascade Structure Appearance Modeling)[41]。

如表1所示,采用了引导策略的跟踪器相较于其基准模型,性能都取得了不同程度的提升。其中,加粗字体为最优值。具体来说,引导后的TENet模型在SUC和PRE指标上分别达到了0.692和0.777;引导后的ViPT模型的SUC和PRE指标分别为0.684和0.770;引导后的SDSTrack模型的SUC和PRE指标分别为0.672和0.756。值得注意的是,在提升跟踪性能的同时,本文提出的几何先验引导策略对模型跟踪效率的影响较为有限。具体而言,增强后的TENet、ViPT与SDSTrack模型的跟踪速度相较于基准模型仅分别下降了5 fps、4 fps和1 fps。

3.2.3　在FE240hz数据集上的性能比较

FE240hz[18]数据集包含了超过14.3万张可见光和事件数据对。在该数据集上的对比方法包括

TENet[13]、ViPT[14]、SDSTrack[15]、TransT[9]、HDETrack[40]、CEUTrack[17]与AFNet[27]。

如表2所示，引导后的TENet的SUC和PRE指标分别达到了0.651和0.937，与基准模型相比，提升了1.6%和2.3%。然而，引导后的SDSTrack和ViPT在部分指标上与基准模型基本持平或略有下降。这主要是因为FE240hz数据集中包含大量低光照和运动模糊等挑战性属性的视频序列，导致引导信号难以被有效提取。TENet因其模型结构能够较好地保留事件与引导信号中的空间结构信息，从而在一定程度上缓解了引导信号质量较低带来的负面影响。然而，ViPT和SDSTrack模型未对引导信号作专门优化，难以抑制低质量信号的干扰。因此，对引导信号加以适当处理，能够最大限度地发挥几何先验引导策略的作用。

表1 与其他先进跟踪器在COESOT数据集上的性能比较

Table 1 Performance comparison with other advanced trackers on the COESOT dataset

指标	AFNet	HDETrack	CEUTrack	CSAM	TENet	ViPT	SDSTrack	Ours(TENet)	Ours(ViPT)	Ours(SDSTrack)
SUC	0.592	0.531	0.627	0.681	0.684	0.683	0.670	**0.692**	0.684	0.672
PRE	0.678	0.641	0.760	0.767	0.765	0.767	0.757	**0.777**	0.770	0.756
FPS	36	**105**	75	53	55	58	31	50	54	30

注：加黑字体表示最优值。

表2 与其他先进跟踪器在FE240hz数据集上的性能比较

Table 2 Performance comparison with other advanced trackers on the FE240hz dataset

指标	AFNet	HDETrack	CEUTrack	TransT	TENet	ViPT	SDSTrack	Ours(TENet)	Ours(ViPT)	Ours(SDSTrack)
SUC	0.584	0.598	0.556	0.639	0.635	0.645	0.645	**0.651**	0.636	0.647
PRE	0.870	0.922	0.845	0.930	0.914	0.924	0.933	**0.937**	0.910	0.921

注：加黑字体表示最优值。

3.3 消融实验

为进一步探究不同要素对几何先验引导策略的影响，本节通过定量与定性相结合的消融实验，系统分析了不同几何先验引导策略及各类角点/边缘检测算法对跟踪性能的作用。

3.3.1 不同几何先验引导策略对跟踪性能的影响

本节以TENet[13]为基准模型，在VisEvent[11]数据集上评估了不同几何先验引导策略对跟踪性能的影响。为进一步探究特征级引导策略的有效性，如式(13)所示，本文尝试采用基于Transformer的交叉注意力机制进行更加复杂的特征级融合。

$$\begin{aligned}\boldsymbol{F}_{\text{cross}}^{\text{E}}&=\text{Softmax}\left(\frac{\boldsymbol{Q}^{\hat{\text{R}}}\left(\boldsymbol{K}^{\text{E}}\right)^{\text{T}}}{d_k}\right)\boldsymbol{V}^{\text{E}}+\boldsymbol{F}^{\text{E}}\\ \boldsymbol{F}_{\text{cross}}^{\hat{\text{R}}}&=\text{Softmax}\left(\frac{\boldsymbol{Q}^{\text{E}}\left(\boldsymbol{K}^{\hat{\text{R}}}\right)^{\text{T}}}{d_k}\right)\boldsymbol{V}^{\hat{\text{R}}}+\boldsymbol{F}^{\hat{\text{R}}}\\ \boldsymbol{F}^{\hat{\text{E}}}&=\boldsymbol{F}_{\text{cross}}^{\text{E}}+\boldsymbol{F}_{\text{cross}}^{\hat{\text{R}}}\end{aligned}\tag{13}$$

其中，$\boldsymbol{Q}^{\hat{\text{R}}},\boldsymbol{K}^{\hat{\text{R}}},\boldsymbol{V}^{\hat{\text{R}}}$均为角点/边缘特征$\boldsymbol{F}^{\hat{\text{R}}}$按照可学习的权重矩阵映射成的特征；$\boldsymbol{Q}^{\text{E}},\boldsymbol{K}^{\text{E}},\boldsymbol{V}^{\text{E}}$则是由事件数据$\boldsymbol{F}^{\text{E}}$映射的特征；$d_k$是$\boldsymbol{Q}^{\hat{\text{R}}}$和$\boldsymbol{Q}^{\text{E}}$的通道维度。$\boldsymbol{F}_{\text{cross}}^{\hat{\text{R}}}$和$\boldsymbol{F}_{\text{cross}}^{\text{E}}$分别表示使用注意力机制后的角点/边缘特征和事件特征，$\boldsymbol{F}^{\hat{\text{E}}}$是与几何先验引导融合后的事件特征。结果如表3所示，两种特征级融合方法的性能均未能超越基于通道级联的像素级融合。此外，相比于结构更简单的SQA模块，复杂的交叉注意力机制在性能上并未体现出优势，甚至略有不足。这是由于事件信号与引导信号均具有显著的空间结构敏感性，而在特征层级进行融合容易因空间维度的压缩与抽象导致关键细节丢失。同时，各类特征级融合策略之间的最终性能差异并不显著，也表明本文方法对具体融合方式并不敏感。

表3 不同几何先验引导策略对跟踪性能的影响

Table 3 The impact of different geometric prior guidance strategies on tracking performance

引导策略	SUC	PRE
TENet	0.601	0.736
基于像素级的引导策略	0.604	0.741
基于特征级的引导策略(SQA)	0.601	0.740
基于特征级的引导策略(交叉注意力)	0.598	0.735

3.3.2 不同引导信号提取方式对跟踪性能的影响

本节以TENet[13]为基准模型，在VisEvent[11]数据集上评估了不同引导信号提取方式所生成的引导信号对跟踪性能的影响。从表4所示的实验结果可以看出，在多种角点检测算法中，基于ORB算法[31]生成的引导信号对跟踪性能的提升最为显著。这主要得益于ORB算法不仅能够快速定位目标的关键结构点，还具备良好的旋转不变性，能更好地适应复杂多变的动态跟踪场景。相比之下，SIFT[29]与KAZE[30]更适用于信息丰富、纹理清晰的静态图像场景；而Harris[28]角点检测器由于其对尺度变化较为敏感，在处理发生

形变的目标时容易出现角点误检与漏检的情况。

在边缘检测方法中,基于Canny算法[36]生成的引导信号对模型性能的提升效果最好。Canny算法通过高斯滤波、非极大值抑制和双阈值处理等步骤,能够提取噪声抑制良好、连续性强的单像素宽度边缘。相比之下,Sobel[34]和Laplace算子[35]所产生的边缘往往较为粗糙或存在断裂,因此在引导信号质量及跟踪性能方面均不如Canny算法。

本文进一步对比了多种角点与边缘检测算法在几何先验引导任务中的可视化效果。如图5所示,ORB与Canny算法所提取的引导信号在完整性和准确性方面均优于其他方法,且在空间分布上与事件激活区域表现出较高一致性。

此外,为了进一步证明引入传统几何算子的必要性,本文设计了一种可学习的卷积神经网络分支处理可见光图像,使该分支网络自适应提取几何结构信号:

$$\boldsymbol{I}_{\mathrm{CNN}}^{\hat{\mathrm{R}}}=\mathrm{CNN}\left(\boldsymbol{I}^{\mathrm{R}}\right) \tag{14}$$

其中,$\boldsymbol{I}^{\mathrm{R}}$为可见光图像,$\boldsymbol{I}_{\mathrm{CNN}}^{\hat{\mathrm{R}}}$为使用CNN学习先验引导特征。所使用的CNN网络包含3层卷积层,每层后依次接入归一化层与ReLU激活函数,最终输出经Sigmoid函数处理。该分支网络同主干网络一同端到端训练。

如表4结果所示,显式提取的角点/边缘特征在融合效果上优于网络隐式学习得到的特征。这主要归因于角点/边缘等显式方法具有明确的物理意义和良好的可解释性。相比之下,CNN在上采样和下采样的过程中容易造成原有空间结构信息的丢失,因此在跟踪任务中的表现不及显式的方法。

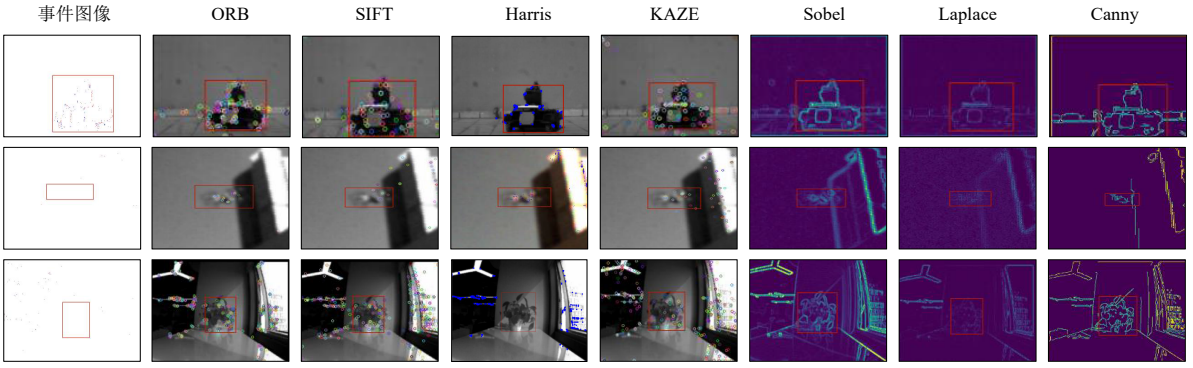


图5　不同角点/边缘算法检测结果的可视化样例

Figure 5　Visualization examples of detection results from different corner/edge algorithms

表4　不同引导信号提取算法对跟踪性能的影响

Table 4　The impact of different guidance signal extraction algorithms on tracking performance

指标	基准模型	基于角点的方法				基于边缘的方法			隐式方法
	TENet	ORB	Harris	SIFT	KAZE	Sobel	Laplace	Canny	CNN
SUC	0.600	0.604	0.602	0.599	0.600	0.600	0.599	0.603	0.597
PRE	0.736	0.741	0.740	0.739	0.738	0.737	0.740	0.741	0.735

3.3.3　跟踪耗时分析

本节以TENet为基准模型,在VisEvent数据集上对比了以下方法在FPS指标上的表现:TENet跟踪器、基于CPU的角点/边缘算子,以及基于CPU和GPU加速角点/边缘算法的本文跟踪方法。

表5中的“角点算子”和“边缘算子”表示基于CPU的ORB和Sobel算子,“Ours(CPU)”和“Ours(GPU)”分别表示使用CPU和GPU进行角点/边缘提取的本文跟踪方法。提取角点与边缘的预处理FPS分别达到274.3和846.9,仅相当于TENet模型运算时耗的约14%和5%。在此基础上,将CPU上的角点/边缘提取与GPU上的深度模型串行执行,整体FPS仅下降3.8和3.5;而若将引导信息提取改用GPU加速,FPS则可提升2.2和1.7,进一步降低了跟踪延时。上述结果表明,在跟踪过程中引入角点/边缘作为几何引导信息,对模型整体运行效率的影响较为有限。

表5　提取引导信号对模型跟踪效率的影响　　单位:帧数/s

Table 5　The impact of extracting the guiding signal on model tracking efficiency　　unit: fps

指标	TENet	角点算子	边缘算子	Ours(CPU)		Ours(GPU)	
				角点引导	边缘引导	角点引导	边缘引导
FPS	38.3	274.3	846.9	34.5	34.8	37.7	36.5

3.3.4　定性分析

本节对提出的方法开展了定性分析。表6展示了采用几何先验引导策略TENet[13]模型在VisEvent[11]数据集的17种不同属性子集上的跟踪结果。实验表明,该方法在10类挑战属性上的性能优于原始TENet模型。例如,在纵横比变化和目标形变属性上,引导策略分别将SUC指标从0.555和0.361提升至

0.571 和 0.381。特别地，在目标静止场景下，仅基于可见光模态的基准跟踪模型 OSTrack，其 SUC 和 PRE 分别为 0.556 和 0.681，均低于 TENet。而基于角点与边缘数据的引导策略在 SUC 指标上相较基准模型分别提升了 3.0% 和 1.7%，在 PRE 指标上分别提升了 3.5% 和 1.3%。这些结果证明，引入角点/边缘数据为事件分支提供了几何先验引导，能够增强模型对目标结构的感知能力。

表 6 在 VisEvent 数据集的 17 个挑战性属性上的跟踪性能对比

Table 6 A comparison of tracking performance across 17 challenging attributes on the VisEvent dataset

属性	TENet		角点引导(ORB)		边缘引导(Sobel)	
	SUC	PRE	SUC	PRE	SUC	PRE
背景杂乱(Background Clutter)	**0.572**	**0.706**	0.570	0.705	0.572	0.703
背景物体运动(Background Object Motion)	**0.564**	0.699	0.561	**0.700**	0.563	0.698
相机运动(Camera Motion)	**0.585**	**0.713**	0.580	0.711	0.576	0.705
快速运动(Fast Motion)	**0.583**	**0.711**	0.580	0.709	0.575	0.698
低光照(Low Illumination)	**0.606**	**0.746**	0.593	0.731	0.593	0.724
运动模糊(Motion Blur)	**0.512**	**0.619**	0.510	0.612	0.487	0.590
出视野(Out-of-View)	**0.451**	0.553	0.442	**0.554**	0.439	0.549
纵横比变化(Aspect Ration Change)	0.555	0.674	**0.571**	**0.693**	0.569	0.690
目标形变(Deformation)	0.361	0.447	**0.381**	**0.463**	0.374	0.458
完全遮挡(Full Occlusion)	0.432	0.564	**0.447**	**0.588**	0.446	0.577
光照变化(Illumination Variation)	0.609	0.743	**0.610**	**0.753**	0.603	0.737
目标静止(No Motion)	0.585	0.699	**0.615**	**0.734**	0.602	0.712
部分遮挡(Partial Occlusion)	0.489	0.623	**0.512**	**0.655**	0.506	0.642
目标旋转(Rotation)	0.524	0.617	**0.541**	**0.634**	0.530	0.617
视角变化(Viewpoint Change)	0.625	0.751	**0.638**	**0.765**	0.626	0.749
过度曝光(Over Exposure)	0.566	**0.736**	0.569	0.734	**0.570**	0.734
尺度变化(Scale Variation)	0.573	0.709	0.577	**0.715**	**0.578**	0.712

注：加黑字体表示最优值。

尽管如此，本文方法在快速运动、运动模糊和低光照等 7 种属性序列中的性能相对于基准模型有所下降。主要源于这些极端场景下可见光相机成像质量显著下降，从而导致角点/边缘信号质量不佳。该结果表明，即便在低质量可见光图像下提取的角点与边缘存在一定偏差，但其负面影响并不显著。

图 6 展示了两个视频序列中模板和搜索区域的可视化示例，包括可见光图像、事件信号、引导信号，以及对应的可见光模态特征图、可见光-事件融合特征图与可见光-事件-引导信号融合特征图。在这两个示例中，事件模态均未能有效表征模板或搜索区域中的目标，可见光-事件融合特征图对目标的表征质量也未优于单独的可见光模态特征图。然而，引导信号则较好地捕捉到了目标的关键几何结构信息，基于其构建的可见光-事件-引导信号融合特征图显著强化了目标区域的响应，最终实现了准确的跟踪结果。

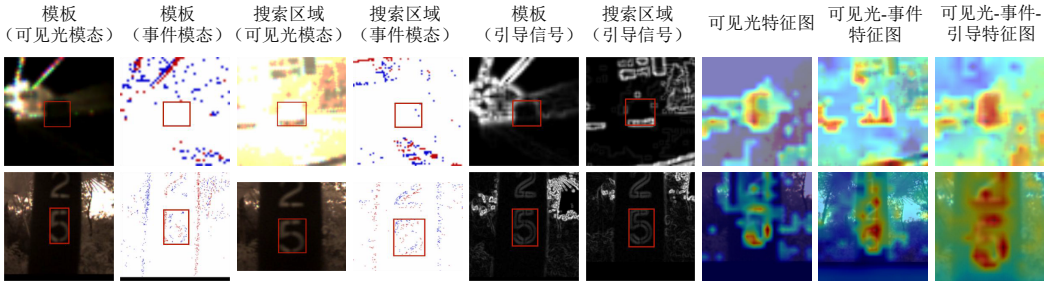


图 6 可见光模态图像、事件模态图像、边缘引导信号与对应跟踪网络特征图的可视化样例

Figure 6 Visualization examples of RGB modal images, event modal images, edge guidance signals, and corresponding tracking network feature maps

4 结论

本文针对事件相机在目标低速运动或静止等挑战性场景下难以提供鉴别性信息的问题,提出了一种基于事件信号引导的可见光-事件目标跟踪方法。该方法通过角点/边缘检测算法从可见光图像中提取目标关键几何结构作为引导信号,并将其与原始事件信号进行融合,从而强化在挑战性条件下的特征表达能力。在 VisEvent、COESOT 和 FE240hz 三个大规模公开数据集上的实验表明,所提方法能够在保持跟踪器实时性能的同时显著提升跟踪精度。此外,消融实验和定性分析进一步验证了方法的有效性。

考虑到本方法在可见光图像质量较低时,角点与边缘特征潜在的退化问题,作者计划在后续工作中设计一种自适应决策机制。该模块能够对当前可见光图像的成像质量进行动态评估,并在可见光图像质量较低时,自适应地判断是否引入几何先验引导信息,以增强模型在可见光图像不同成像质量条件下的跟踪鲁棒性。本文代码已开源,网址:https://github.com/LLLzw444/Enhancement-strategy-of-TENet。

参考文献:

[1] 谷美颖, 李航, 张家伟, 等. 基于视觉的无人机定位与导航方法研究综述[J]. 电子学报, 2025, 53(3): 651-685.
Gu Meiying, Li Hang, Zhang Jiawei, et al. A review of vision-based UAV localization and navigation methods[J]. Acta Electronica Sinica, 2025, 53(3): 651-685. (in Chinese)

[2] 钟钰彬, 杨鹏, 窦磊. 基于纵横比自适应的相关滤波跟踪算法[J]. 电子学报, 2024, 52(6): 2112-2122.
Zhong Yubin, Yang Peng, Dou Lei. Correlation filtering tracking algorithm based on adaptive aspect-ratio[J]. Acta Electronica Sinica, 2024, 52(6): 2112-2122. (in Chinese)

[3] 李玺, 查宇飞, 张天柱, 等. 深度学习的目标跟踪算法综述[J]. 中国图象图形学报, 2019, 24(12): 2057-2080.
Li Xi, Zha 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)

[4] 林慧兰, 赵春蕾, 郝志成, 等. 复杂场景单目标跟踪[J]. 光学 精密工程, 2024, 32(23): 3490-3503.
Lin Huilan, Zhao Chunlei, Hao Zhicheng, et al. Single target tracking in complex scenarios[J]. Optics and Precision Engineering, 2024, 32(23): 3490-3503. (in Chinese).

[5] 陈刚, 张文标, 王洪雁, 等. 三维激光雷达点云级联匹配多目标跟踪方法[J]. 光电工程, 2025, 52(6): 250050.
Chen Gang, Zhang Wenbiao, Wang Hongyan, et al. The cascade matching method for multi-object tracking of 3D LiDAR point clouds[J]. Opto-Electronic Engineering, 2025, 52(6): 250050. (in Chinese)

[6] 刘超军, 段喜萍, 谢宝文. 应用GhostNet卷积特征的ECO目标跟踪算法改进[J]. 激光技术, 2022, 46(2): 239-247.
Liu Chaojun, Duan Xiping, Xie Baowen. Improvement of ECO target tracking algorithm based on GhostNet convolution feature[J]. Laser Technology, 2022, 46(2): 239-247. (in Chinese)

[7] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]//Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2016: 850-865.

[8] Li Bo, Yan Junjie, Wu Wei, et al. High performance visual tracking with Siamese region proposal network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8971-8980.

[9] Chen Xin, Yan Bin, Zhu Jiawen, et al. Transformer tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8122-8131.

[10] Ye Botao, Chang Hong, Ma Bingpeng, et al. Joint feature learning and relation modeling for tracking: A one-stream framework[C]//Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 341-357.

[11] Wang Xiao, Li Jianing, Zhu Lin, et al. VisEvent: Reliable object tracking via collaboration of frame and event flows[J]. IEEE Transactions on Cybernetics, 2024, 54(3): 1997-2010.

[12] Gehrig D, Scaramuzza D. Low-latency automotive vision with event cameras[J]. Nature, 2024, 629(8014): 1034-1040.

[13] Shao Pengcheng, Xu Tianyang, Tang Zhangyong, et al. TENet: Targetness entanglement incorporating with multi-scale pooling and mutually-guided fusion for RGB-E object tracking[J]. Neural Networks, 2025, 183: 106948.

[14] Zhu Jiawen, Lai Simiao, Chen Xin, et al. Visual prompt multi-modal tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 9516-9526.

[15] Hou Xiaojun, Xing Jiazheng, Qian Yijie, et al. SDSTrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 26541-26551.

[16] Wu Zongwei, Zheng Jilai, Ren Xiangxuan, et al. Single-model and any-modality for video object tracking[C]//

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 19156-19166.

[17] Tang Chuanming, Wang Xiao, Huang Ju, et al. Revisiting color-event based tracking: A unified network, dataset, and metric[J]. Pattern Recognition, 2026, 172: 112718.

[18] Zhang Jiqing, Yang Xin, Fu Yingkai, et al. Object tracking by jointly exploiting frame and event domain[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 13023-13032.

[19] Li Bo, Wu Wei, Wang Qiang, 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. Piscataway: IEEE, 2018: 4282-4291.

[20] 吴非, 张建林. 结合全局光流的孪生区域提名网络目标跟踪算法[J]. 半导体光电, 2023, 44(3): 422-428.
Wu Fei, Zhang Jianlin. Siamese region proposal network object tracking algorithm with global optical flow[J]. Semiconductor Optoelectronics, 2023, 44(3): 422-428. (in Chinese)

[21] Wang Qiang, Zhang Li, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 1328-1338.

[22] 谢青松, 刘晓庆, 安志勇, 等. 基于前景优化的视觉目标跟踪算法[J]. 电子学报, 2022, 50(7): 1558-1566.
Xie Qingsong, Liu Xiaoqing, An Zhiyong, et al. Visual object tracking algorithm based on foreground optimization[J]. Acta Electronica Sinica, 2022, 50(7): 1558-1566. (in Chinese)

[23] Chen Zedu, Zhong Bineng, Li Guorong, et al. SiamBAN: Target-aware tracking with Siamese box adaptive network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(4): 5158-5173.

[24] Yan Bin, Peng Houwen, Fu Jianlong, et al. Learning spatio-temporal transformer for visual tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 10428-10437.

[25] Tan Yuedong, Wu Zongwei, Fu Yuqian, et al. XTrack: Multimodal training boosts RGB-X video object trackers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2025: 5734-5744.

[26] Gallego G, Delbrück T, Orchard G, et al. Event-based vision: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 154-180.

[27] Zhang Jiqing, Wang Yuanchen, Liu Wenxi, et al. Frame-event alignment and fusion network for high frame rate tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 9781-9790.

[28] Harris C, Stephens M. A combined corner and edge detector[C]//Proceedings of the Alvey Vision Conference. Manchester: Alvey Vision Club, 1988: 147-152.

[29] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[30] Alcantarilla P F, Bartoli A, Davison A J. KAZE features[C]//Proceedings of the 12th European Conference on Computer Vision. Heidelberg: Springer, 2012: 214-227.

[31] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]//Proceedings of the 2011 International Conference on Computer Vision. Piscataway: IEEE, 2011: 2564-2571.

[32] Rosten E, Drummond T. Machine learning for high-speed corner detection[C]//Proceedings of the 9th European Conference on Computer Vision. Heidelberg: Springer, 2006: 430-443.

[33] Calonder M, Lepetit V, Strecha C, et al. BRIEF: Binary robust independent elementary features[C]//Proceedings of the 11th European Conference on Computer Vision. Heidelberg: Springer, 2010: 778-792.

[34] Kanopoulos N, Vasanthavada N, Baker R L. Design of an image edge detection filter using the Sobel operator[J]. IEEE Journal of Solid-State Circuits, 1988, 23(2): 358-367.

[35] Marr D, Hildreth E. Theory of edge detection[J]. Proceedings of the Royal Society of London. Series B. Biological Sciences, 1980, 207(1167): 187-217.

[36] Canny J. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, PAMI-8(6): 679-698.

[37] 贾迪, 孟祥福, 孟琭, 等. RGB空间下结合高斯曼哈顿距离图的彩色图像边缘检测[J]. 电子学报, 2014, 42(2): 257-263.
Jia Di, Meng Xiangfu, Meng Lu, et al. Color image edge detection combining with Gauss Manhattan distance map in RGB space[J]. Acta Electronica Sinica, 2014, 42(2): 257-263. (in Chinese)

[38] Yang Jinyu, Li Zhe, Zheng Feng, et al. Prompting for multi-modal tracking[C]//Proceedings of the 30th ACM

International Conference on Multimedia. New York: ACM, 2022: 3492-3500.

[39] Danelljan M, Van Gool L, Timofte R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 7181-7190.

[40] Wang Xiao, Wang Shiao, Tang Chuanming, et al. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 19248-19257.

[41] Zhang Tianlu, Debattista K, Zhang Qiang, et al. Revisiting motion information for RGB-event tracking with MOT philosophy[C]//Proceedings of the 38th International Conference on Neural Information Processing System. New York: Curran Associates Inc., 2024: 2836.

作者简介

赵少川　男,1996年8月出生于江苏省徐州市。中国矿业大学计算机科学与技术学院/人工智能学院博士后。主要研究方向为计算机视觉与模式识别,包括视觉目标跟踪、神经网络安全与多模态学习等领域。

E-mail: shaochuan_zhao@cumt.edu.cn

刘峥玮　男,2004年4月出生于湖南省郴州市。中国矿业大学计算机科学与技术学院/人工智能学院硕士研究生。主要研究方向为视觉目标跟踪。

E-mail: TS25170019A31LD@cumt.edu.cn

徐天阳　男,1989年5月出生于江苏省常州市。江南大学人工智能与计算机学院(软件学院)副教授。主要研究方向为模式识别、计算机视觉、人工智能。

E-mail: tianyang_xu@163.com

邵志文　男,1994年12月出生于安徽省马鞍山市。中国矿业大学计算机科学与技术学院/人工智能学院副教授、硕士生导师。主要研究方向为情感智能、细粒度目标检测、视觉内容生成、时间序列预测。

E-mail: zhiwen_shao@cumt.edu.cn

周　勇　男,1974年9月出生于江苏省徐州市。中国矿业大学计算机科学与技术学院/人工智能学院教授、博士生导师。主要研究方向为机器学习、人工智能、数据科学与工程。

E-mail: yzhou@cumt.edu.cn

吴小俊　男,1967年12月出生于江苏省镇江市。江南大学人工智能与计算机学院(软件学院)教授、博士生导师。主要研究方向为计算机视觉与模式识别。

E-mail: wu_xiaojun@jiangnan.edu.cn