计算机科学
COMPUTER SCIENCE

融合跟踪器:融合图像特征和事件特征的单目标跟踪框架

王琳, 刘哲, 史殿习, 周晨磊, 杨绍武, 张拥军

引用本文

王琳, 刘哲, 史殿习, 周晨磊, 杨绍武, 张拥军. 融合跟踪器:融合图像特征和事件特征的单目标跟踪框架[J]. 计算机科学, 2023, 50(10): 96-103.

WANG Lin, LIU Zhe, SHI Dianxi, ZHOU Chenlei, YANG Shaowu, ZHANG Yongjun. Fusion Tracker:Single-object Tracking Framework Fusing Image Features and Event Features [J]. Computer Science, 2023, 50(10): 96-103.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900075

融合跟踪器:融合图像特征和事件特征的单目标跟踪框架

王 琳[1] 刘 哲[1] 史殿习[1,2,3] 周晨磊[3] 杨绍武[1] 张拥军[2]
1 国防科技大学计算机学院 长沙 410073
2 军事科学院国防科技创新研究院 北京 100166
3 天津(滨海)人工智能创新中心 天津 300457
(wanglin12@nudt.edu.cn)

摘 要 目标跟踪是计算机视觉领域的一项基本研究问题。作为主流目标跟踪方法传感器,传统相机可以提供丰富的场景信息。但是由于受到采样原理的限制,传统相机在极端光照条件下会出现过曝光或欠曝光的问题,且在高速运动场景中存在运动模糊的现象。而事件相机是一种仿生传感器,它能够感知光照强度变化输出事件流,具有高动态范围、高时间分辨率等优点,但难以捕捉静态目标。受传统相机和事件相机的特性启发,提出了一种双模态融合的单目标跟踪方法,称为融合跟踪器(Fusion Tracker)。该方法通过特征增强的方式自适应地融合来自传统相机和事件相机数据中的视觉线索,同时设计一种基于注意力机制的特征匹配网络,将模板帧的目标线索与搜索帧相匹配,建立长期特征关联,使跟踪器关注目标信息。融合跟踪器可以解决特征匹配过程中相关性运算导致的语义丢失问题,提升目标跟踪的性能。在两个公开数据集上的实验展示了所提方法的优越性,并且通过消融实验验证了融合跟踪器中关键部分的有效性。融合跟踪器可以有效提升在复杂场景中目标跟踪任务的鲁棒性,为下游应用提供可靠的跟踪结果。

关键词:目标跟踪;深度学习;事件相机;特征融合;注意力机制
中图法分类号 TP391

Fusion Tracker:Single-object Tracking Framework Fusing Image Features and Event Features

WANG Lin[1],LIU Zhe[1],SHI Dianxi[1,2,3],ZHOU Chenlei[3],YANG Shaowu[1] and ZHANG Yongjun[2]
1 School of Computer Science,National University of Defense Technology,Changsha 410073,China
2 National Innovation Institute of Defense Technology,Academy of Military Sciences,Beijing 100166,China
3 Tianjin Artificial Intelligence Innovation Center,Tianjin 300457,China

Abstract Object tracking is a fundamental research problem in the field of computer vision. As the mainstream object tracking method sensor,conventional cameras can provide rich scene information. However,due to the limitation of sampling principle,conventional cameras suffer from overexposure or underexposure under extreme lighting conditions,and there is motion blur in high-speed motion scenes. In contrast,event camera is a bionic sensor that can sense light intensity changes to output event streams,with the advantages of high dynamic range and high temporal resolution,but it is difficult to capture static targets. Inspired by the characteristics of conventional and event cameras,a dual-modal fusion single-target tracking method,called fusion tracker,is proposed. The method adaptively fuses visual cues from conventional and event camera data by feature enhancement,while designing an attention mechanism-based feature matching network to match object cues of template frames with search frames to establish long-term feature associations and make the tracker focus on object information. The fusion tracker can solve the semantic loss problem caused by correlation operations during feature matching and improve the performance of object tracking. Experiments on two publicly available datasets demonstrate the superiority of our approach and validate the effectiveness of the key parts of the fusion tracker by ablation experiments. The fusion tracker can effectively improve the robustness of object tracking tasks in complex scenarios and provide reliable tracking results for downstream applications.

Keywords Object tracking,Deep learning,Event cameras,Featurefusion,Attention mechanisms

到稿日期:2022-09-08 返修日期:2022-12-09
基金项目:国家自然科学基金(91948303)
This work was supported by the National Natural Science Foundation of China(91948303).
通信作者:史殿习(dxshi@nudt.edu.cn)

1　引言

视觉目标跟踪的目的是预测目标在场景中的位置，其被广泛应用于机器人避障、视频监控、无人驾驶等计算机视觉相关领域。近年来，基于深度学习的方法使用传统相机作为传感设备在目标跟踪任务中表现出良好的性能[1-4]。但是传统相机的帧率和动态范围有限，面对低光照、快速运动、背景杂波等具有挑战性的场景时，传统相机由于过曝光、欠曝光、运动模糊等问题而无法正常工作[5]。

为了应对传统相机所面临的上述挑战，研究者们开发了一种被称为事件相机的视觉传感器[6]。由于事件相机的每个像素由独立电路控制，当光照强度发生变化时，每个像素可以异步且独立地做出响应，因此能够高效准确地感知场景中的动态变化。与传统相机相比，事件相机可以适应不同的照明条件，具有高时间分辨率、低响应延迟和高动态范围等优势[5]。Piatkowska 等通过高斯混合模型对异步事件流进行聚类，实现高遮挡情况下的行人跟踪[7]。Barranco 等提出均值漂移聚类的方法，利用异步事件流实现实时多目标跟踪[8]。

传统相机可以提供丰富的场景信息，但是在高动态范围场景中无法正常工作；而事件相机可以准确感知动态场景的变化，但是不能测量绝对光强度，因此难以捕捉静止的物体。鉴于传统相机和事件相机性能上的互补性，工业界推出了动态和主动像素视觉传感器（Dynamic and Active Pixel Vision Sensor，DAVIS），该传感器由异步事件传感器和同一像素阵列中的传统相机组成，可以同时输出异步事件流和对应的图像帧[9]。Gehrig 等利用 DAVIS 传感器跟踪视觉特征，提取图像帧上的特征，并使用事件异步对其进行跟踪。该方法通过生成事件模型将事件和帧中的像素强度相关联[10]。

本文提出了一种双模态数据融合的视觉目标跟踪方法，称为融合跟踪器。融合跟踪器充分利用传统相机和事件相机的优势，通过一个特征融合网络和一个特征匹配网络提升退化条件（例如具有高动态范围、低光照和快速运动等条件）下的目标跟踪性能。首先，针对复杂场景中图像质量不佳和普通场景中事件数据稀疏的问题，特征融合网络利用来自事件和图像的视觉线索，通过特征增强的方式，获取丰富的场景信息，从而为模板帧和搜索帧的目标匹配提供支持。其次，针对相关性操作容易导致语义信息丢失的问题，特征匹配网络基于注意力机制，通过将模板帧特征和搜索帧特征进行匹配，增强了搜索帧特征的目标信息。最后，本文在公开数据集 VisEvent 和 FE108 中验证了融合跟踪器的性能，并通过消融实验证明了融合跟踪器中关键部分的有效性。

本文的主要创新点和贡献如下：

1）提出了一种特征增强的双模态融合方法，该方法充分利用图像和事件的视觉线索来跟踪场景目标，有效提升了在复杂场景中目标跟踪任务的鲁棒性。

2）针对特征匹配过程中相关性运算导致的语义丢失问题，设计了一种特征匹配网络。该网络对模板帧特征中的目标信息和搜索帧特征进行匹配，从而减少了背景信息的干扰，使模型专注于目标信息。

3）在不同公开数据集中对融合跟踪器进行测试，测试结果表明，相较于 VisEvent-ATOM 方法，本文提出的方法精确率和成功率分别提升了 10.8% 和 8.4%。此外，本文通过消融实验证明了融合跟踪器各部分的有效性。

2　相关工作

2.1　视觉目标跟踪

近年来，基于深度学习的方法在基于传统相机的目标跟踪领域占据主导地位。其中，基于孪生网络的方法在跟踪领域被广泛应用。例如 SiamRPN 将孪生网络与区域建议网络（Region Proposal Network，RPN）相结合，以获得更精确的跟踪结果[11]。

然而，在低光照、快速运动和背景杂波等具有挑战性的场景下，基于传统相机的跟踪器表现并不理想。对此，研究人员提出了众多解决方案，如主动硬样本生成[12-13]和图像去模糊[14]等。但由于成像质量较差，这些方案大多无法很好地解决上述问题。同时，其他传感器也被探索用于目标跟踪任务，如高帧率相机[15]、热敏相机[16]、深度相机[17]等。但高帧率相机对光照敏感，热敏相机价格昂贵，深度相机无法应对高速低光照场景，而事件相机由于具有高动态范围、低延迟、高时间分辨率等特点，可以很好地应对上述挑战。

在基于事件的目标跟踪方面，文献[18]提出时间图像事件表示法，利用事件流的时间信息实现运动补偿。文献[19]将事件表示基于时间图像表示法进一步改进，提出了具有线性时间衰减表示的同步时间表面（Synchronous Time-Surface）来编码事件时空信息。尽管这些工作有着良好的目标跟踪效果，但由于事件相机不能测量绝对光强度，难以捕捉动作缓慢或者静止的物体。因此，这些工作提出的模型在跟踪缓慢移动或静止的物体时性能并不理想，容易出现边界框漂移、目标丢失等情况。

如何充分利用传统相机和事件相机的优势，融合两者的特点进行目标跟踪并提升目标跟踪性能，是目前目标跟踪领域一个亟待研究解决的问题。文献[20]提出了一个大规模目标跟踪数据集，将传统相机和事件相机相结合，用于扩展视觉跟踪器在实际场景中的应用，并且提出了一种跨模态变换器，用于图像和事件之间的特征融合。然而该变换器只是简单地将图像和事件通过注意力机制进行融合，没有平衡图像与事件之间的贡献。与文献[20]相比，本文提出的双模态特征融合方法通过特征增强的方式自适应地融合来自传统相机和事件相机数据中的视觉线索，充分利用了传统相机和事件相机的优势，使得模型可以适应更复杂的场景，从而提升目标跟踪任务的性能。

2.2　注意力机制

Transformer 架构在机器翻译任务中被首次提出[21]，其基本模块为注意力机制，用于聚集输入序列的信息。由于 Transformer 架构具有并行计算和独特的存储机制，其在处理长序列方面比循环神经网络更有优势。Transformer 已经在许多连续任务（自然语言处理[22]、语音处理[23-24]、计算机视觉[25]）中取代了循环神经网络，并逐渐扩展到处理非连续

问题。在文献[26]中，Carion 等将目标检测视为集预测(Set Prediction)问题，采用文献[21]中的编码器-解码器架构作为检测头，提出检测变换器(DEtection Transformer，DETR)方法。在 COCO 数据集[27]上的实验结果表明，基于 Transformer 的 DETR 方法与 Faster R-CNN 基线方法[28]得到的结果相当。由于 DETR 的成功以及检测与跟踪之间的密切关系(如 RPN[28]和 SiamRPN[11])，我们将 Transformer 引入目标跟踪当中，与 DETR 不同的是，我们不直接使用原始 Transformer 中的编码器-解码器架构，而是采用 Transformer 的核心思想，利用注意力机制设计了自注意增强和交叉注意特征增强模块。

3 融合跟踪器

在充分分析传统相机和事件相机的工作原理以及各自的优缺点的基础上，我们利用传统相机和事件相机的优势，提出了一种融合传统相机的图像和事件相机事件数据特征的视觉目标跟踪方法，方法框架如图 1 所示，由特征提取编码器、特征融合模块、特征匹配模块和回归分类预测头等部分构成。方法流程如下：首先，针对事件相机的输出无法适应神经网络输入的问题，本文采用事件累积的方法，离散异步事件的时域，使其可以被神经网络模型处理；其次，将模板分支和搜索分支的图像和事件输入特征提取编码器，得到图像和事件的特征；再次，通过特征融合模块，以特征增强的方式自适应地对图像特征和事件特征进行融合；然后，针对相关性操作容易导致语义信息丢失的问题，通过特征匹配模块对模板帧特征和搜索帧特征进行匹配，以产生更丰富的语义特征映射。最后，分类回归预测头对匹配特征进行二元分类和边界框回归，生成跟踪结果。我们将在下文详细介绍每个部分的细节。

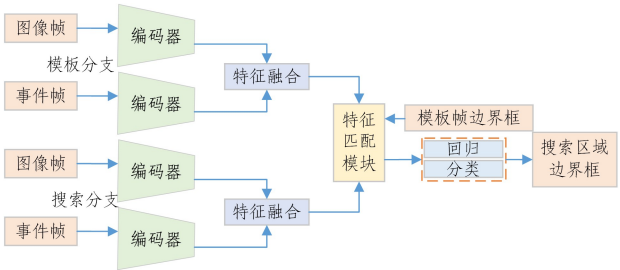


图 1 融合跟踪器总体框架

Fig. 1 General framework of fusion tracker

3.1 输入表示

从感知原理的角度来看，传统相机通过图像同步记录所有像素的光照强度来捕捉全局场景，而事件相机异步测量场景中的光照强度变化。当光强度的变化大于阈值时，像素独立地触发一个事件。事件的极性反映变化的方向，如式(1)所示，事件可以被定义为：

$$\varepsilon=\{e_k\}_{k=1}^{N}=\{[x_k,y_k,t_k,p_k]\}_{k=1}^{N} \tag{1}$$

其中，ε 表示异步事件流；N 表示异步事件流内事件的数量；e_k 表示第 k 个事件；(x_k,y_k) 表示事件 e_k 的像素位置；t_k 表示该事件的时间戳；$p_k\in\{-1,+1\}$ 为事件的极性，极性为正表示光照强度增大，极性为负表示光照强度减弱。

为了适应卷积神经网络的输入，现有方法[29-30]通常在固定的时间间隔内叠加事件，将异步事件流转换为图像帧或网格的表示。在本文工作中我们也采用这样的变换来获得事件图像。

3.2 特征提取编码器

为了充分提取传统相机的图像和事件相机的事件数据的特征，我们采用 ResNet50[31]作为图像和事件的特征编码器。其核心思想是：以视频序列中的第一帧作为模板帧，对其进行裁剪，得到目标及其周围场景的外观信息。搜索帧从前一帧的目标中心坐标扩展到目标边长的 4 倍，通常可以覆盖目标可能的移动范围。然后，将裁剪后的搜索帧和模板帧重塑为正方形，输入特征编码器进行特征提取。为了便于定位跟踪目标，我们删除了 ResNet50[31]的最后一级，保留细粒度的空间信息，将第四级的输出作为最终输出。同时，将第四级下采样单元的卷积步长从 2 改成 1，以获得更大的特征分辨率；并且将第四阶段的 3×3 卷积修改成步长为 2 的扩张卷积，以增大感受野。总的来说，特征提取编码器尽可能地学习模板帧和搜索帧的目标信息，提取高级语义特征，输入到特征融合模块。

3.3 特征融合模块

传统相机可以捕获丰富的纹理和语义线索，而事件相机可以有效捕获边缘信息，并且具有高动态范围。受文献[32]启发，我们设计了一个特征融合模块以有效利用两种模态数据，其结构如图 2 所示。与文献[32]不同的是，我们采用实例归一化，独立地归一化每个训练样本中的每个通道，平均激活倾向于由大均匀区域(通常为背景)中的运动控制。这种归一化与线性整流函数(ReLU)相结合，有助于将背景运动和前景运动分离[33]。特征融合模块旨在融合图像和事件的特征，具体来说，基于事件和图像的特征，我们定义了以下特征增强方法，以生成事件的增强特征 $\hat{\boldsymbol{F}}_e$，如式(4)所示：

$$\hat{\boldsymbol{F}}_{e\to e}=g(\psi_{3\times3}(\boldsymbol{F}_e))\otimes\boldsymbol{F}_e \tag{2}$$

$$\hat{\boldsymbol{F}}_{v\to e}=\left(\psi_{1\times1}\begin{bmatrix}\xi(\psi_{1\times1}(\boldsymbol{F}_v))\\ \xi(\psi_{3\times3}(\boldsymbol{F}_v))\\ \xi(\psi_{5\times5}(\boldsymbol{F}_v))\end{bmatrix}\right)\otimes F_e \tag{3}$$

$$\hat{\boldsymbol{F}}_e=\hat{\boldsymbol{F}}_{e\to e}\oplus\hat{\boldsymbol{F}}_{v\to e}\oplus\boldsymbol{F}_e \tag{4}$$

其中，$\boldsymbol{F}_e$ 表示经过编码器提取的事件特征；$\boldsymbol{F}_v$ 表示经过编码器提取的图像特征；ψ 表示卷积运算；g 表示 Sigmoid 激活函数；$[\cdot]$表示通道方向的拼接；ξ 为实例归一化和 ReLU 线性整流函数；$\hat{\boldsymbol{F}}_{e\to e}$表示基于事件的自我增强特征；$\hat{\boldsymbol{F}}_{v\to e}$表示通过图像对事件的增强特征。同理，可以生成基于传统相机的图像的增强特征 $\hat{\boldsymbol{F}}_v$，此处不再一一赘述。

最后，我们采用一种可学习的正则化参数 W_{F_e} 和 W_{F_v}，通过式(7)融合事件的增强特征与图像的增强特征，自适应地平衡图像和事件的贡献。

$$W_{F_e}=g(\psi_{1\times1}(\xi(\psi_{1\times1}(A(\hat{\boldsymbol{F}}_e))))) \tag{5}$$

$$W_{F_v}=g(\psi_{1\times1}(\xi(\psi_{1\times1}(A(\hat{\boldsymbol{F}}_v))))) \tag{6}$$

$$X=W_{F_e}\hat{\boldsymbol{F}}_e\oplus W_{F_v}\hat{\boldsymbol{F}}_v \tag{7}$$

其中，X 表示融合特征；A 表示自适应平均池化层。

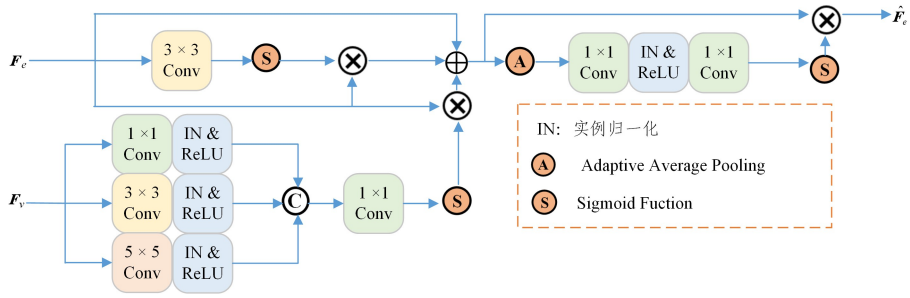


图 2　特征融合模块

Fig. 2　Feature fusion module

3.4　特征匹配模块

为了有效地对搜索帧特征和模板帧特征中的目标信息进行匹配，我们将经过融合后的模板帧特征和搜索帧特征进行展平，分别得到模板帧特征向量$\boldsymbol{X}_T \in \mathbb{R}^{HW\times C}$和搜索帧特征向量$\boldsymbol{X}_S \in \mathbb{R}^{HW\times C}$，并分别将其输入特征匹配模块。特征匹配模块的结构如图 3 所示，我们设计了两个自注意层，通过这两个自注意层分别关注特征向量中有用的语义上下文，如边缘和相似目标。其次，两个交叉注意层同时接收各自分支和另一分支的特征向量，两个自注意层和两个交叉注意层组成一个特征匹配基本块。如图 3 中的虚线框所示，特征匹配基本块重复 N 次（在实现中，$N=4$）。最后通过一个交叉注意层融合这两个特征向量。

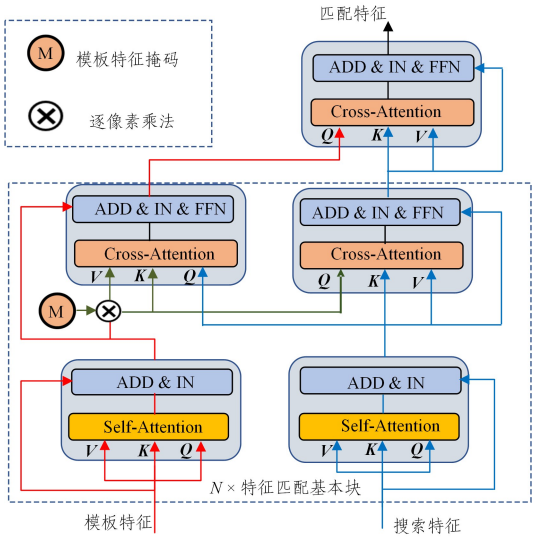


图 3　特征匹配模块

Fig. 3　Feature matching module

注意力机制是设计特征匹配模块的基本要素。给定 $\boldsymbol{Q}\in\mathbb{R}^{HW\times C}$，$\boldsymbol{K}\in\mathbb{R}^{HW\times C}$，$\boldsymbol{V}\in\mathbb{R}^{HW\times C}$，其中 H，W，C 分别表示高度、宽度和通道维度。基于注意的匹配公式如式(8)所示：

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{8}$$

其中，d_k 为 K 的维度。本文采用多头注意机制实现特征匹配模块。如文献[21]所述，将注意力机制扩展到多头注意使得该机制能够考虑多种注意力分布，并关注信息的不同方面。多头注意机制的详细设计与描述可参考文献[21]。

如式(9)所示，自注意层接收图像特征向量，自适应地整合来自特征图不同位置的信息，旨在关注模板帧和搜索帧中的目标信息。

$$\hat{\boldsymbol{S}}_{sa}=\mathrm{IN}\left(Att(\boldsymbol{X}_l^t,\boldsymbol{X}_l^t,\boldsymbol{X}_l^t)+\boldsymbol{X}_l^t\right) \tag{9}$$

其中，IN(·)表示实例归一化；t 表示视频序列帧索引；$l\in\{1,\cdots,N\}$为特征匹配基本块的索引；X 表示模板帧或搜索帧的融合特征。

两个交叉注意层同时接收各自分支和另一分支的特征向量。此外，前馈神经网络(Feed Forward Networks，FFN)用于提升模型的拟合能力。当相机在场景中发生剧烈运动时，传递目标的位置信息是有益的。因此我们通过式(10)构建模板帧的高斯掩码，用于传播目标位置信息。

$$\boldsymbol{m}(y)=\exp\left(-\frac{\|y-c\|^2}{2\sigma^2}\right) \tag{10}$$

其中，c 表示模板帧边界框标签；σ 为标准差；$\boldsymbol{m}(y)$表示生成的高斯掩码。我们将模板帧特征与高斯掩码逐像素相乘来抑制背景区域。模板帧的增强特征$\boldsymbol{F}_T$由式(12)得到：

$$\boldsymbol{X}_T=\mathrm{IN}(Att(\boldsymbol{X}_l^m\otimes\boldsymbol{m},\boldsymbol{X}_l^m\otimes\boldsymbol{m},\boldsymbol{X}_l^t)+\boldsymbol{X}_l^m) \tag{11}$$

$$\boldsymbol{F}_T=\boldsymbol{X}_T+FFN(\boldsymbol{X}_T) \tag{12}$$

其中，$\boldsymbol{X}_l^m$表示在第 l 个特征匹配基本块的模板帧特征；$\boldsymbol{X}_l^t\in\mathbb{R}^{HW\times C}$表示在 t 时刻 l 个特征匹配基本块的搜索帧特征。

搜索帧的增强特征$\boldsymbol{F}_S$由式(14)得到：

$$\boldsymbol{X}_S=\mathrm{IN}(Att(\boldsymbol{X}_l^m,\boldsymbol{X}_l^m,\boldsymbol{X}^t\otimes\boldsymbol{m})+\boldsymbol{X}_l^t) \tag{13}$$

$$\boldsymbol{F}_S=\boldsymbol{X}_S+FFN(\boldsymbol{X}_S) \tag{14}$$

最后，一个交叉注意层融合两个分支的特征向量，输出模板帧与搜索帧的匹配特征 $\boldsymbol{Y}$，如式(16)所示：

$$\boldsymbol{F}_S=\mathrm{IN}(Att(\boldsymbol{F}_S,\boldsymbol{F}_S,\boldsymbol{F}_T)+\boldsymbol{F}_S) \tag{15}$$

$$\boldsymbol{Y}=\boldsymbol{F}_S+FFN(\boldsymbol{F}_S) \tag{16}$$

3.5　分类回归预测头

分类回归预测头由分类分支和回归分支组成，其中每个分支是一个具有隐藏维度和 ReLU 激活函数的三层感知器。对于由特征匹配模块生成的特征向量，预测头对每个向量进行预测，以获得前景/背景分类结果，以及关于搜索帧大小的归一化坐标。

3.6　损失函数

我们采用文献[34]的损失函数，预测头接收特征向量，并输出二元分类和回归结果；选择与标签边界框中的像素对应的特征向量的预测作为正样本，其余作为负样本；所有的样本都参与分类损失的计算，只有正样本参与回归损失的计算。为了减少正负样本之间的不平衡，我们将负样本产生的损失降低为原来的 1/16，使用标准的二进制交叉熵损失进行分类，其定义如式(17)所示：

$$\mathscr{L}_{\mathrm{cls}}=-\sum_j\left[y_j\log(\boldsymbol{p}_j)+(1-y_j)\log(1-\boldsymbol{p}_j)\right] \tag{17}$$

其中，y_j 表示第 j 个样本的真值标签，$y_j=1$ 表示前景，$\boldsymbol{p}_j$ 表示属于由模型预测的前景的概率。对于回归，我们采用 L1 范数损失 $\mathscr{L}_1(.,.)$ 和广义 IoU 损失 $\mathscr{L}_{\mathrm{GIoU}}(.,.)$[35] 的线性组合，其定义如式(18)所示：

$$\mathscr{L}_{\mathrm{reg}}=\sum_j \mathbf{1}_{\{y_j=1\}}\left[\lambda_G \mathscr{L}_{\mathrm{GIoU}}(\boldsymbol{b}_j,\hat{\boldsymbol{b}})+\lambda_1 \mathscr{L}_1(\boldsymbol{b}_j,\hat{\boldsymbol{b}})\right] \tag{18}$$

其中，$y_j=1$ 表示正样本，b_j 表示第 j 个预测边界框，$\hat{b}$ 表示归一化真值边界框。λ_G 和 λ_1 为正则化参数，在实现中分别取 2 和 5。

3.7 算法描述

融合跟踪方法的算法伪代码流程如算法 1 所示。在该算法中，输入为事件流、视频图像帧、图像帧采样时间、模板帧边界框，输出为搜索帧的边界框。第 1 行根据图像帧采样时间累积事件，第 2－16 行描述目标跟踪的流程。其中，第 2－7 行描述对模板帧进行预处理、特征提取和特征融合；第 8－12 行描述对搜索帧进行预处理、特征提取和特征融合；第 13－14 行描述对模板帧特征和搜索帧特征进行匹配；第 15 行描述分类回归预测头根据匹配特征，输出搜索帧边界框。

算法 1 融合跟踪器算法

输入：事件流 E，视频图像帧 F，图像帧采样时间 Δt，模板帧边界框 B_1

输出：搜索帧的目标边界框 B_K

1. 根据图像帧采样时间 Δt，将事件流 **E** 分成 $\mathbf{E}_1,\mathbf{E}_2,\cdots,\mathbf{E}_n$，与图像帧 $\mathbf{F}_1,\mathbf{F}_2,\cdots,\mathbf{F}_n$ 相对应
2. for each E_k and F_k do
3. if k=1 then
4. 使模板帧边界框处于中心位置，将模板事件 $\mathbf{E}_1$ 和模板图像 $\mathbf{F}_1$ 裁剪至 128×128
5. 特征提取网络对 $\mathbf{E}_1$ 和 $\mathbf{F}_1$ 提取特征，得到 $\mathbf{FE}_1$ 和 $\mathbf{FF}_1$
6. 特征融合网络对 $\mathbf{FE}_1$ 和 $\mathbf{FF}_1$ 进行融合，得到 X_1
7. end
8. if k> 1 then
9. 使 k－1 的模板帧边界框处于中心位置，将搜索事件 $\mathbf{E}_k$ 和搜索图像 $\mathbf{F}_k$ 裁剪至 256×256
10. 通过特征提取网络对 $\mathbf{E}_k$ 和 $\mathbf{F}_k$ 进行特征提取，得到 $\mathbf{FE}_k$ 和 $\mathbf{FF}_k$
11. 通过特征融合网络 FE_k 和 FF_k 进行融合，得到 $\mathbf{X}_k$
12. end
13. 通过式(10)，将模板边界框 $\mathbf{B}_1$ 转化为模板掩码 $\mathbf{M}_1$
14. 将 $\mathbf{X}_1$，$\mathbf{M}_1$ 和 $\mathbf{X}_k$ 输入特征匹配网络，得到匹配特征 $\mathbf{Y}_k$
15. 将 $\mathbf{Y}_k$ 输入分类回归预测头，得到搜索帧的边界框 $\mathbf{B}_K$
16. End for

4 实验与分析

为了探索所提出的融合跟踪器对目标跟踪任务的有效性，我们在公开数据集 FE108[32] 和 VisEvent[20] 上进行验证。首先介绍使用的数据集和评价指标并描述实现细节，然后从定量和定性两方面验证所提方法的性能，最后对方法的主要部分进行消融分析。

4.1 数据集和评估指标

我们在单目标跟踪的公开事件相机数据集 FE108[32] 和 VisEvent[20] 上对本文所提方法的有效性进行实验验证。FE108 数据集[32] 由 DAVIS346 事件相机拍摄得到，包含 108 个序列，总时长为 1.5 h。数据序列中包括 21 种不同类型的跟踪目标。运动目标的边界框标签由 Vicon 光学运动捕捉系统提供。数据集涵盖 4 种具有挑战性的场景：低光照、高动态范围、产生运动模糊的快速运动和不产生运动模糊的快速运动，可以充分验证目标跟踪模型在复杂场景中的跟踪鲁棒性。

VisEvent[20] 是面向事件相机的单目标跟踪算法的一个基准数据集，它定义了 17 种场景，涵盖单目标跟踪中的常见挑战，如低光照、相机运动、旋转、缩放变化和遮挡等。该数据集采集了 820 个视频对，其中 709 个为短视频，111 个为长视频。运动目标主要为行人和车辆，其边界框标签由专业的公司进行标注。

为了定量验证跟踪器性能，我们使用两个目标跟踪领域广泛使用的评价指标：精确率(Precision Rate，PR)和成功率(Success Rate，SR)。评价指标表示特定类型帧的百分比。成功率关注真值标签与预测边界框之间的重叠大于阈值的帧。精确率关注在给定阈值内，真值标签和预测边界框之间的中心距离。与其他方法[20,32] 相同，我们使用成功率图的曲线下面积(Aera Under Curve，AUC)代表成功率，使用与 20 像素阈值关联的精确率分数代表精确率。

4.2 训练细节

我们在 Pytorch 中实现所提网络，搜索帧和模板帧的大小分别为 256×256 和 128×128，骨干网络使用 ImageNet 预训练模型[36] 进行初始化。模型的其他参数用 Xavier init[37] 初始化。使用 AdamW[38] 对模型进行训练，将主干网络的学习率设置为 0.000 01，其他参数的学习率设为 0.000 1，batch size 大小设置为 36。在两个 Nvidia RTX3090 GPU 上训练 200 个周期，每个周期迭代 1 000 次，在 100 个周期后，学习率下降为原来的 1/10。

4.3 结果分析

为了验证本文方法的有效性，我们将所提出的方法与多种最先进的方法进行比较。表 1 列出了本文方法在精确率和成功率两个指标上与其他方法的对比结果(跟踪器结果取自 FE108 基准)，可以发现本文方法在精确率和成功率方面都大大优于其他方法。

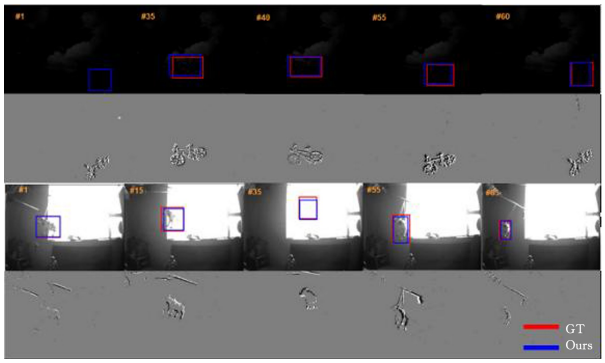
表 1 在 FE108 数据集上的比较结果

Table 1 Comparison results on FE108 dataset

跟踪器	CLNet[1]	SiamBAN[2]	SiamRPN[39]	PrDiMP[40]	ATOM[41]	VisEvent-ATOM[20]	Ours
精确率(↑)	55.5	37.4	33.5	80.5	71.3	79.4	**88.0**
成功率(↑)	34.4	22.5	21.8	53.0	46.5	54.3	**58.9**

具体来说，本文方法的精确率为 88.0%，成功率为 58.9%，与同样融合图像和事件的 VisEvent ATOM[20] 相比分别高出 10.8% 和 8.4%。图 4 展示了本文方法在 FE108 数据集两个序列上的结果。第一行和第二行表示曝光不足场景；第三行和第四行表示过度曝光场景。可以看到，在昏暗或过曝光的环境中，传统相机难以捕捉到目标，而本文方法在

事件和图像的共同帮助下，通过注意机制对模板帧和搜索帧的特征进行匹配，成功地跟踪到了目标。



注：第一和第三行为图像，第二和第四行为事件。

图 4　FE108 数据集上的跟踪结果

Fig. 4　Tracking results on FE108 dataset

为了验证所提方法对挑战性场景的有效性，我们还展示了在 FE108 提供的两种不同挑战条件（高动态范围和低光照）下的跟踪结果，如表 2 所列，本文方法比其他方法表现更好。图 5 给出了所提方法与其他方法在数据集 VisEvent 上的跟踪结果对比。可以发现，SiamRPN[39]，PrDiMP[40]，ATOM[41]等跟踪器在低光照和运动模糊场景下预测的边界框不够精确并且容易被干扰物吸引，导致目标丢失，而本文方法可以准确地跟踪目标。这反映了本文方法通过融合事件和图像的特征，学习序列中的运动信息，获得了更好的目标跟踪能力。

表 2　在 FE108 数据集的高动态范围和低光照条件下的结果

Table 2　Results at high dynamic range and low light conditions on FE108 dataset

方法	高动态范围	低光照
SiamRPN[35]	21.6|15.3	14.5|10.1
SiamBAN[2]	26.6|16.3	26.5|15.5
CLNet[1]	48.3|30.3	23.6|13.7
PrDiMP[36]	66.3|44.3	69.5|44.6
ATOM[37]	56.0|36.6	45.0|28.6
Ours	**93.5**|**64.6**	**94.7**|**64.1**

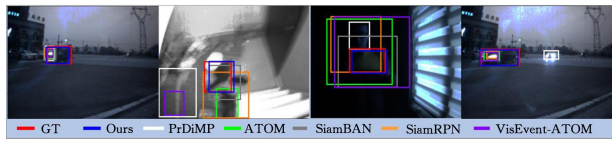
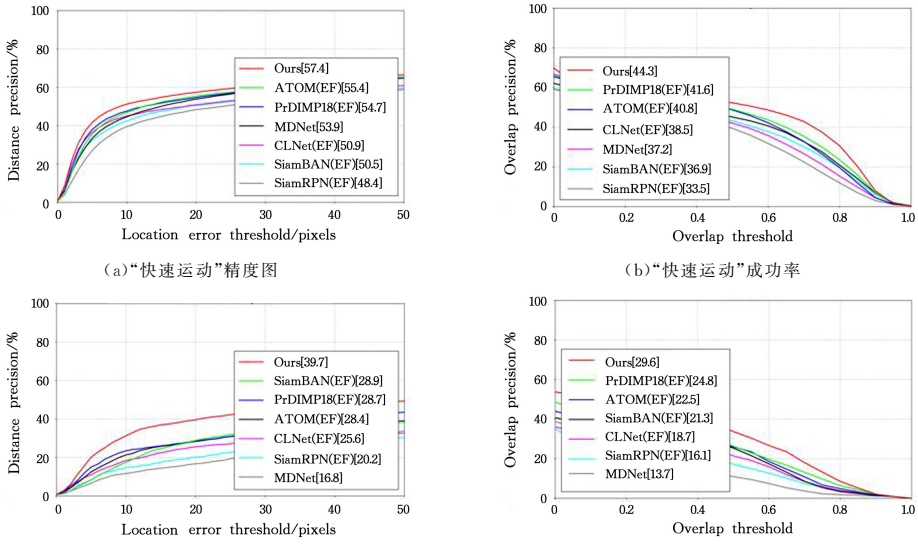


图 5　在 VisEvent 数据集中与多种跟踪器的定性对比

Fig. 5　Qualitative analysis plots on VisEvent dataset compared to multiple trackers

另外，我们将基于传统相机的跟踪器进行扩展，如图 6 所示，图中 EF 表示在将数据输入跟踪模型之前，将事件和图像进行简单融合。在 VisEvent 数据集的两个属性上进行验证，包括“可变形目标（Deformable，DEF）”和“快速运动（FastMotion，FM）”，可以看出，本文方法明显优于其他方法，证明了双模态融合模块可以更有效地融合图像和事件，提升目标跟踪的性能，此外也展示了本文方法对目标的不同形状、大小、运动速度的鲁棒性。



（a）“快速运动”精度图　（b）“快速运动”成功率

（c）“可变形目标”精度图　（d）“可变形目标”成功率

注：第一行和第二行分别为“快速运动”和“可变形目标”属性的精确率图和成功率图。

图 6　跟踪不同挑战场景下的结果

Fig. 6　Results are tracked in different challenge scenarios

4.4　消融实验

本文通过输入不同的数据，验证图像和事件数据融合的有效性。如表 3 所列，当只使用图像（Frame only）时，本文方法在精确率和成功率上分别达到 66.3%和 41.0%；只使用事件（Event only）时，分别可以达到 83.9%和 53.4%。这说明事件相机可以记录更多的细节，从而提升目标跟踪的性能。但因为事件相机只能捕捉场景中发生运动的位置，提供轮廓形状信息，所以仍然需要利用外观和纹理信息来区分目标对象和其他干扰物。当结合这两种方式进行跟踪时，本文方法的整体性能可以提高到 88%和 58.9%，在运动模糊和低光照属性方面取得较好的效果。

为了验证特征匹配模块在所提跟踪器中的贡献，我们删除关键组件并重新训练修改后的模型：（1）没有模板掩码（NoMask）；（2）没有自注意层（No Self-attention）。如表 3

所列，与原始模型相比，去除自注意对性能影响最大，而去除模板掩码对性能也有影响。结果验证了本文方法的确有助于模板帧与搜索帧的匹配。

表 3 在 FE108 数据集上的消融实验结果

Table 3 Results of ablation experiments on FE108 dataset

模型	精确率	成功率
No mask	87.2	58.6
No self- attention	82.5	50.4
Frame only	66.3	41.0
Event only	83.9	53.4
Ours	**88.0**	**58.9**

结束语 本文提出了双模态数据融合的单目标跟踪方法，基于特征增强的特征融合网络自适应地融合异步事件流和传统图像帧中的视觉线索，基于注意力机制的特征匹配网络将模板帧目标特征与搜索帧特征进行匹配，使跟踪器关注目标信息。通过在不同公开数据集中进行测试，验证了所提方法在高动态范围和低光照等方面的优越性，并通过消融实验证明了所提方法各部分的有效性。这表明利用异步事件流和图像帧的互补性可以提高退化条件下目标跟踪的鲁棒性。该模型与其他事件表示兼容，可以使用更轻量级的特征提取网络，具有较强的通用性。但是所提模型比较依赖视频模板帧质量，若模板帧质量不佳，可能会影响后续跟踪结果。未来将研究相邻帧之间的运动线索，利用短期注意，提升目标跟踪性能。

参 考 文 献

[1] DONG X, SHEN J, SHAO L, et al. CLNet: A compact latent network for fast adjusting Siamese trackers[C]// European Conference on Computer Vision. Cham: Springer, 2020: 378-395.

[2] 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. 2020: 7183-7192.

[3] CHENG X, CUI Y P, SONG C, et al. Target tracking algorithm based on spatio-temporal attention mechanism [J]. Computer Science, 2021, 48(4): 123-129.

[4] ZHAO Y, YU Z B, LI Y C. A twin tracking algorithm based on mutual attention guidance[J]. Computer Science, 2022, 49(3): 163-169.

[5] GALLEGO G, DELBRUCK T, ORCHARD G, et al. Event-based vision: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(1): 154-180.

[6] LICHTSEINER P, POSCH C, DELBRUCK T. A 128×128 120dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor[J]. IEEE Journal of Solid-State Circuits, 2008, 43(2): 566-576.

[7] PIATKOWSKA E, BELBACHIR A N, SCHRAML S, et al. Spatiotemporal multiple persons tracking using dynamic vision sensor[C]// 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, 2012: 35-40.

[8] BARRANCO F, FERMULLER C, ROS E. Real-time clustering and multi-target tracking using event-based sensors[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2018: 5764-5769.

[9] MOEYS D P, CORRADI F, LI C, et al. A sensitive dynamic and active pixel vision sensor for color or neural imaging applications [J]. IEEE Transactions on Biomedical Circuits and System, 2017, 12(1): 123-126.

[10] GEHRIG D, REBECQ H, GALLEGO G, et al. EKLT: Asynchronous photometric feature tracking using events and frames [J]. International Journal of Computer Vision, 2020, 128(3): 601-618.

[11] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.

[12] SONG Y, MA C, WU X, et al. Vital: Visual tracking via adversarial learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8990-8999.

[13] WANG X, LI C, LUO B, et al. Sint++: Robust visual tracking via adversarial positive instance generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4864-4873.

[14] GUO Q, FENG W, CHEN Z, et al. Effects of blur and deblurring to visual object tracking[J]. arXiv: 1908.07904, 2019.

[15] GALOOGAHI H K, FANG A, HUANG C, et al. Need for speed: A benchmark for higher frame rate object tracking[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1125-1134.

[16] LI C, LIANG X, LU Y, et al. RGB-T object tracking: Benchmark and baseline[J]. Pattern Recognition, 2019, 96: 106977.

[17] LUKEZIC A, KART U, KAPYLA J, et al. Cdtb: A color and depth visual object tracking dataset and benchmark[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 10013-10022.

[18] MITROKHIM A, FERMULLER C, PARAMESHWARA C, et al. Event-based moving object detection and tracking[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). IEEE, 2018: 1-9.

[19] CHEN H, SUTER D, WU Q, et al. End-to-end learning of object motion estimation from retinal events for event-based object tracking[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(7): 10534-10541.

[20] WANG X, LI J, ZHU L, et al. VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows[J]. arXiv: 2108.05015, 2021.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[J]. arXiv: 1706.03762, 2017.

[22] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[23] LUSCHER C, BECK E, IRIE K, et al. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention--w/o Data Augmentation[J]. arXiv: 1905.03072, 2019.

[24] SYNNAEVE G, XU Q, KAHN J, et al. End-to-end asr: from supervised tosemi-supervised learning with modern architectures

[J]. arXiv:1911.08460,2019.

[25] PARMAR N,VASWANI A,USZKOREIT J,et al. Image transformer[C]// International Conference on Machine Learning. PMLR,2018:4055-4064.

[26] CARION N,MASSA F,SYNNAEVE G,et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision. Cham:Springer,2020:213-229.

[27] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft coco:Common objects in context[C]// European Conference on Computer Vision. Cham:Springer,2014:740-755.

[28] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards Real-Time Object Detection with Region Proposal Networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(6):1137-1149.

[29] CHEN H,SHUTER D,WU Q,et al. End-to-end learning of object motion estimation from retinal events for event-based object tracking[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:10534-10541.

[30] CHEN H,WU Q,LIANG Y,et al. Asynchronous tracking-by-detection on adaptive time surfaces for event-based object tracking[C]// Proceedings of the 27th ACM International Conference on Multimedia. 2019:473-481.

[31] HE K,ZHANG X,REN S,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.

[32] ZHANG J,YANG X,FU Y,et al. Object tracking by jointly exploiting frame and event domain[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:13043-13052.

[33] YANG C,LAMDOUAR H,LU E,et al. Self-supervised video object segmentation by motion grouping[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:7177-7188.

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

[35] UNION G I O. A Metric and a Loss for Bounding Box Regression[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. 2019:658-666.

[36] RUSSAKOVSKY O,DENG J,SU H,et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision,2015,115(3):211-252.

[37] GLOROT X,BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]// Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. 2010:249-256.

[38] LOSHCHILOV I,HUTTER F. Decoupled weight decay regularization[J]. arXiv:1711.05101,2017.

[39] CHEN Z,ZHONG B,LI G,et al. Siamese box adaptive network for visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:6668-6677.

[40] LI B,YAN J,WU W,et al. High performance visual tracking with siamese region proposal network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8971-8980.

[41] DANELLJAN M,BHAT G,KHAN F S,et al. Atom:Accurate tracking byoverlap maximization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4660-4669.

WANG Lin, born in 1998, postgraduate. His main research interests include event camera, deep learning and computer vision.

SHI Dianxi, born in 1966, Ph.D, professor, Ph.D supervisor. His main research interests include distributed object middleware technology, adaptive software technology, artificial intelligence, and robot operating systems.

（责任编辑：何杨）