Event-based Moving Object Detection and Tracking

作者：Mathias Gehrig, Davide Scaramuzza

期刊：CVPR

年份：2023

Code: https://github.com/uzh-rpg/RVT

Gen1：“Prophesee Gen1 Automotive Detection Dataset License Terms and Conditions”: https://www.prophesee.ai/2020/01/24/prophesee-gen1-automotive-detection-dataset/

1 Mpx：“Prophesee 1MegaPixel Automotive Detection Dataset License Terms and Conditions”: https://www.prophesee.ai/2020/11/24/automotive-megapixel-event-based-dataset/

1、问题背景/研究动机

本文研究的主要问题：**快速（低延迟）、轻量级和高性能的基于事件相机的目标检测**

给定一个固定的带宽，传统相机必须**权衡相机分辨率和帧率**，但在高动态场景中，降低分辨率或帧率可能以**丢失必要的场景细节**为代价 。

现有基于事件相机的目标检测可以分为三个研究方向：**图神经网络**、**脉冲神经网络**和**密集的神经网络**

- **图神经网络**（eg. 10.23-10.29论文）

动态构建时空图，通过子采样事件和寻找现有的在时空上接近的节点来建立新的节点和边。

  - Cons：**计算复杂度较高**，需要消耗大量计算资源和时间；对于**全局特征**的建模能力相对较弱。

- **脉冲神经网络**

每个脉冲神经元都有一个内部状态，神经元只有在达到阈值时才会产生锋电位。

  - Cons：由于非可微性使其**优化困难**；由于需要处理事件的时序信息，导致**计算复杂度较高**；由于数据的稀疏性，导致在目标检测中可能**丢失重要信息**。

- 密集的神经网络

使用**密集的事件表示**，使数据能够进行卷积操作。**早期的方法**直接使用从事件的**短时间窗口**生成的单个事件表示来推断检测，但会在所考虑的时间窗口之外造成信息丢失，使得难以检测移动缓慢的物体。

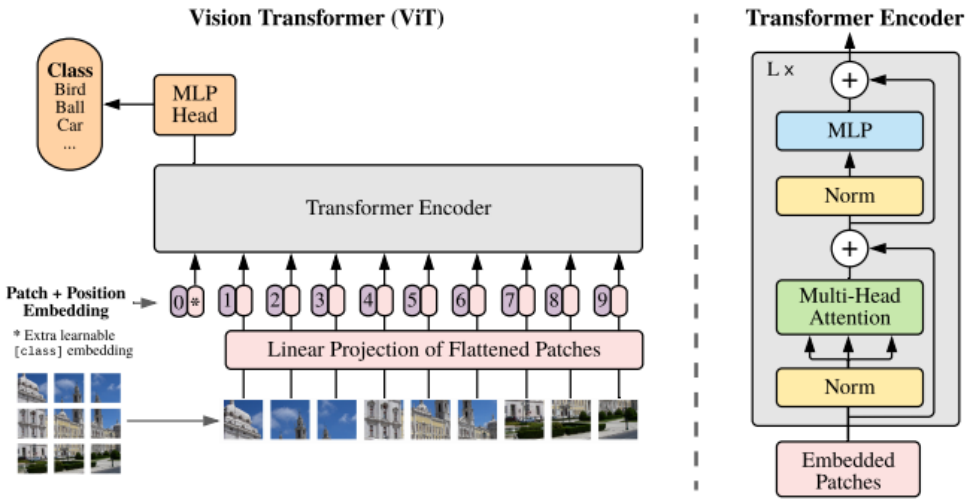
近期的方法通过结合**循环神经网络层**来解决这个问题，大幅提高检测性能。但这些方法的处理延迟仍然**超过40毫秒**，无法充分利用事件相机的**低延迟特性**。

- **关键问题：如何在不需要专门硬件的情况下，同时实现高精度和高效率？**

- Vision Transformer

NLP中**基于注意力**的模型的成功启发了计算机视觉中**基于Transformer架构**的探索。

- **模型架构**



模型由三个模块组成：

1）Linear Projection of Flattened Patches（Embedding层）

2）Transformer Encoder（图右侧）

3）MLP Head（最终用于分类的层结构）

- **核心流程**

ViT主要包括**图像分块处理 (make patches)**、**图像块嵌入 (patch embedding)与位置编码**、**Transformer编码器**和**MLP分类处理**4个部分。

首先，tokenizer将图像拆分为块（patch），并将这些图像块的线性嵌入序列作为Transformer 的输入。接着，Transformer使用注意力方法生成一系列输出块，projector最终将输出块标记重新连接到特征图。**整体流程**大致为：

- 将图像拆分为块（固定大小）
- 展平图像块
- 从这些展平的图像块中创建低维线性嵌入，包括位置嵌入
- 将序列作为输入发送到Transformer编码器
- 使用图像标签预训练 ViT 模型，然后在广泛的数据集上进行训练
- 在图像分类的下游数据集进行微调

2、解决方法

（1）提出了一种**多阶段分层骨干网络**的设计，通过在每个阶段中引入**卷积先验**、**局部和全局自注意力**以及**时间回归**等关键组件，实现**高性能**和**低延迟**的目标检测。

（2）结合传统基于帧的目标检测的神经网络设计，提出了一种**基于Vision Transformers**的目标检测框架：**Recurrent Vision Transformers（RVTs）**。

（3）在**Gen1**和**1 Mpx事件相机数据集**上进行了**消融实验**和评估。

3、具体实现

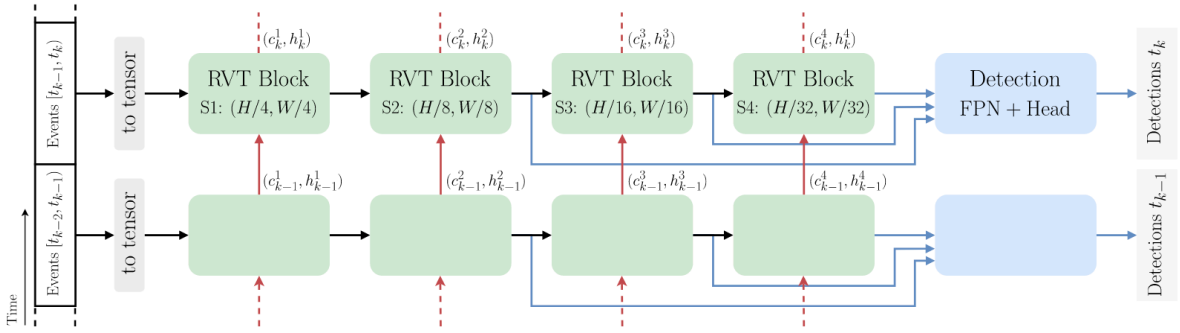


图1：多阶段分层骨干网络架构

1）**事件数据处理**：首先对事件相机的数据进行处理，将输入事件处理成张量（即多维数组），表示事件在空间和时间上的变化。

2）**混合空间和时间特征**：使用Transformer层进行空间特征提取，并使用循环神经网络(LSTM)进行时间特征提取。

3）**多阶段分层网络和参数优化**：在网络设计中，每个阶段由多个重复的模块组成，采用交错的局部和全局自注意力机制以及简化的LSTM单元，以减少参数数量和计算复杂度。

- **事件数据处理**

为了使大量的事件数据**与卷积神经网络层兼容**，对数据进行简单的预处理。

首先，创建一个**四维张量*E***。第一维包含两个分量，表示事件的极性；第二维有T个分量，与时间的T个离散化步骤相关；第三和第四维分别表示事件相机的高度和宽度

接着，对一个时间区间$[t_a, t_b)$内的事件集E进行如下处理：

$$E(p,\tau,x,y) = \sum_{e_k \in \mathcal{E}} \delta(p-p_k)\delta(x-x_k, y-y_k)\delta(\tau-\tau_k),$$
$$\tau_k = \left\lfloor \frac{t_k - t_a}{t_b - t_a} \cdot T \right\rfloor$$

即，创建T个2通道的帧，其中每个像素包含其中一个T时间帧中的正事件或负事件的数量。最后，将**极性和时间维度扁平化**以检索形为$(2T, H, W)$的三维张量，直接实现与卷积的兼容性。

- **混合空间和时间特征**

在目标检测任务中，神经网络需要：

（1）**在像素空间中提取与任务相关的局部和全局特征（空间特征）**，因为物体既可以覆盖非常小的区域，也可以覆盖很大一部分视场；

（2）**从最近的事件以及几秒前的事件中提取特征（时间特征）**，因为一些物体相对于相机是缓慢移动的，以至于它们随着时间的推移产生很少的事件。

由于事件相机数据含有时间维度和空间维度的特性，在本文方法中：

Transformer层用于**空间特征提取**和**循环神经网络（LSTM）**用于**时间特征提取**：

1）**空间特征提取**阶段采用卷积操作，并引入局部和全局自注意力机制，以实现局部和全局特征的混合。

2）**时间特征提取**阶段使用LSTM单元对特征进行聚合，以捕捉事件相机中的时间信息。

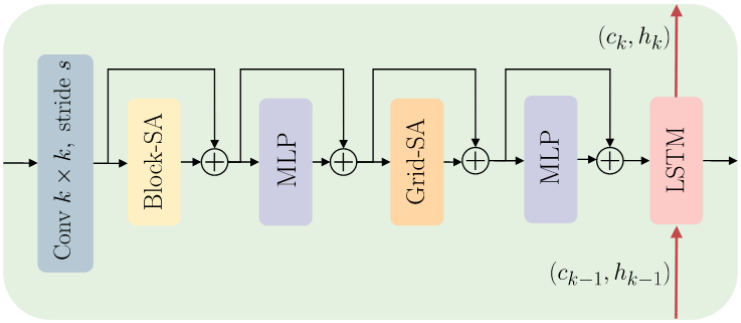


图2：RVT块结构

- **空间特征提取**

首先，使用**具有重叠核的卷积操作**对输入特征进行处理，使其在**提取特征**的同时还可以进行**空间下采样**。且这样的卷积操作引入了**条件位置嵌入**，不需要绝对或相对位置嵌入。

随后，通过**多轴自注意力机制**对特征进行转换，包括**两个阶段**的自注意力操作：**局部特征交互**和**扩张的全局特征混合**。

1）局部特征交互

将特征局部**分组成非重叠的窗口**，并应用多头自注意力机制。图2中的Block-SA块用于建模局部特征之间的交互。

2）全局特征混合

将特征图划分为形为$(G \times G, \frac{H}{G} \times \frac{W}{G}, C)$的网格，得到的窗口大小为$\frac{H}{G} \times \frac{W}{G}$，并对这些窗口应用自注意力机制（图2中的Grid-SA块），实现全局特征的混合，相当于**对特征进行全局、扩张的混合**。

- **时间特征提取**

使用LSTM单元进行时间特征聚合，与之前的工作不同，本文发现**时间和空间特征聚合可以完全分离**。即，可以使用**普通的LSTM单元**，而不需要使用Conv-LSTM单元，使得**LSTM的状态之间互不影响**，从而大大降低计算复杂度和参数数量。

- **LSTM**
  - LSTM单元是一种**递归神经网络（RNN）**，可以在每个时间步骤上独立地处理特征。
  - 在每个时间步骤上，将**当前时间步骤的事件表示**作为**输入**，并将**前一个时间步骤的LSTM状态**作为**上下文信息**传递给当前时间步骤。
  - 通过这种方式，可以在**时间维度**上建立起一种**上下文关联**的机制，从而更好地捕捉到**事件的时间序列信息**，并将其**聚合为一个时间维度上的特征表示**。
  - 对于处理事件相机数据中的时间相关特征非常重要，尤其是在相机视野中移动缓慢、事件较少的对象。

- **RVT块结构**

在RVT块中，每个注意力和MLP模块之前应用**LayerNorm**，之后应用**LayerScale**，并在每个模块之后添加一个**残差连接**。

- **LayerNorm**
  - LayerNorm是一种归一化技术，对每个样本的特征进行**归一化处理**。
  - 通过计算每个特征维度上的**均值**和**方差**，使得每个特征维度的**分布保持一致**。
  - 在RVT块中，LayerNorm可以提高模型的**稳定性**和**收敛性**。
- **LayerScale**
  - LayerScale是一种缩放技术，用于**调整每个模块的学习率范围**。

- 通过在每个模块之后添加一个缩放因子，使得每个模块的**学习率范围更广**。
- 在RVT块中，LayerScale**增加模型的表达能力和学习能力**。

- 残差连接

- 残差连接允许模型直接学习残差（即**输入和输出之间的差异**），不需要通过多个非线性层来逐步逼近，从而更容易**优化模型**。
- 有助于减轻**梯度消失**和**梯度爆炸**问题，提高模型的**训练效果**和**收敛速度**。

4、实验对比

- 数据集

Gen1汽车检测数据集由39个小时的事件相机记录组成，分辨率为304×240。Gen1数据集包含228k个车辆和28k个行人边界框，可用频率为1，2或4Hz。在本实验中，移除边长小于10像素和对角线小于30像素的边界框。

1 MPx数据集也提供了驾驶场景，但提供了更高分辨率的720 × 1280的白天和夜晚记录。它由大约15个小时的事件数据组成，以30或60 Hz的频率标注，共有3个类（小汽车、行人、两轮车）的2500万个边界框标签。在本实验中，移除边长小于20像素和对角线小于60像素的边界框，并将输入分辨率减半至nHD分辨率（640×360）。

- **评价指标**：对于这两个数据集，**平均精度均值(mAP)**是评估的主要指标。

- 消融实验

- 空间交互作用

为了公平比较，实验**保持LSTM和卷积下采样层相同**，只交换注意力和MLP模块。将多轴注意力与ConvNext块和Swin Transformer块进行比较。

ConvNext是一种**卷积神经网络架构**，已经在包括目标检测在内的广泛任务上显示出与基于Transformer的模型具有竞争力的性能。实验使用7 × 7的默认内核大小，并在每个阶段放置三个ConvNext块来近似匹配参考模型的参数数量。

Swin Transformer是一种**基于注意力的模型**，在通过循环移位相互交互的窗口中应用局部自注意力。

	Gen1		1 Mpx		
Block-type	mAP	AP_{50}	mAP	AP_{50}	Params (M)
multi-axis	**47.6**	**70.1**	**46.0**	**72.3**	18.5
Swin	46.7	68.7	44.4	71.7	18.5
ConvNeXt	45.5	65.8	42.3	70.6	18.7

表中可以看出，Swin取得了比ConvNext更好的性能，但是在Gen1和1 Mpx数据集上多轴自注意力都取得了最好的结果。本实验表明，**每一阶段的全局交互(多轴)都有利于纯局部交互(Swin , ConvNext)**。

- **卷积下采样**

原始的Vision Transformer架构没有与卷积层进行局部特征交互。实验比较了**重叠和非重叠卷积核**在**输入层（块嵌入）**和**特征降采样阶段**的情况。

Conv. kernel type	mAP	AP_{50}	AP_{75}	Params (M)
overlapping	**47.6**	**70.1**	**52.6**	18.5
non-overlapping	46.1	68.6	50.5	**17.6**

表中可以看出，**非重叠卷积**在减少参数数量的同时，也**造成了性能的大幅下降；重叠核**的使用以略微增加参数数量为代价换取了**更高的性能**。

因此，本文方法在网络的所有阶段选择**重叠核**。

- **LSTM与卷积**

现有的事件相机目标检测方法**严重依赖于Conv-LSTM单元**。实验对**普通的LSTM单元**和**深度可分离的Conv - LSTM变体**进行比较。

深度可分离Conv-LSTM首先对输入和隐藏状态都使用深度可分离卷积，然后使用逐点(1×1)卷积。

LSTM kernel size	mAP	AP_{50}	AP_{75}	Params (M)
1×1	**47.6**	**70.1**	**52.6**	**18.5**
3×3	46.5	69.0	51.4	40.8
3×3 depth-sep	46.3	67.2	51.2	18.6

表中可以看出，普通的LSTM单元甚至优于两种变体。

基于实验结果，本文方法使用**普通的LSTM单元**。

- **LSTM布局**

在这个消融实验中，研究**只在一个阶段的子集上使用时间递归**的影响，或者根本**不使用时间递归**。对于所有的比较，使模型完全相同，但**在每个时间步的选定阶段重置LSTM的状态**。这样，就可以在比较中**保持参数个数不变**的情况下，**模拟不存在递归层**的情况。

S1	S2	S3	S4	mAP	AP_{50}	AP_{75}
				32.0	54.8	31.4
			✓	39.8	63.5	41.6
		✓	✓	44.2	68.4	47.5
	✓	✓	✓	46.9	70.0	50.8
✓	✓	✓	✓	**47.6**	**70.1**	**52.6**

表中可以看出，完全**不使用递归会导致检测性能的急剧下降**。在每个阶段启用LSTM，从第四个阶段开始均导致**性能增强**。且LSTM单元在**早期阶段**对整体性能也有帮助。

实验表明，**检测框架受益于增加了时间信息的特征**。本文方法将LSTM也保留在第一阶段。

- 基准比较

- **空间交互作用**

在这一部分中，在Gen1和1Mpx数据集上将本文方法与现有方法进行比较。

文中训练了3个模型，一个基模型(RVT-B)，大约有1850万个参数；一个小变体(RVT-S)，大约有990万个参数；一个微小模型(RVT-T)，大约有440万个参数，结构超参数如下表

				Channels		
Stage	Size	Kernel	Stride	RVT-T	RVT-S	RVT-B
S1	1/4	7	4	32	48	64
S2	1/8	3	2	64	96	128
S3	1/16	3	2	128	192	256
S4	1/32	3	2	256	384	512

为了与现有方法进行比较，实验选择了在验证集上表现最好的模型，并在测试集上进行了评估。

			Gen1		1 Mpx		
Method	Backbone	Detection Head	mAP	Time (ms)	mAP	Time (ms)	Params (M)
NVS-S [27]	GNN	YOLOv1 [40]	8.6	-	-	-	0.9
Asynet [34]	Sparse CNN	YOLOv1	14.5	-	-	-	11.4
AEGNN [43]	GNN	YOLOv1	16.3	-	-	-	20.0
Spiking DenseNet [10]	SNN	SSD [30]	18.9	-	-	-	8.2
Inception + SSD [19]	CNN	SSD	30.1	19.4	34.0	45.2	> 60*
RRC-Events [7]	CNN	YOLOv3 [41]	30.7	21.5	34.3	46.4	> 100*
MatrixLSTM [6]	RNN + CNN	YOLOv3	31.0	-	-	-	61.5
YOLOv3 Events [20]	CNN	YOLOv3	31.2	22.3	34.6	49.4	> 60*
RED [38]	CNN + RNN	SSD	40.0	16.7	43.0	39.3	24.1
ASTMNet [26]	(T)CNN + RNN	SSD	46.7	35.6	**48.3**	72.3	> 100*
RVT-B (ours)	Transformer + RNN	YOLOX [15]	**47.2**	**10.2 (3.7)**	47.4	**11.9 (6.1)**	18.5
RVT-S (ours)	Transformer + RNN	YOLOX	46.5	9.5 (3.0)	44.1	10.1 (5.0)	9.9
RVT-T (ours)	Transformer + RNN	YOLOX	44.1	9.4 (2.3)	41.5	9.5 (3.5)	4.4

从表中可以看出，**使用循环层的模型**一致优于其他方法（无论是稀疏的：GNNs、SNNs，还是没有循环层的密集前馈模型：Inception+SSD、RRC-Events、YOLOv3 Events）。本文的基础模型在Gen1数据集上取得了47.2 mAP的性能，在1 Mpx数据集上取得了47.4 mAP的性能。

本文的基础模型在Gen1数据集（304×240分辨率）上达到了10.2ms的推理时间，比RED减少了6 ms的延迟，比ASTMNet减少了3倍以上的推理时间。在1 Mpx数据集（640×360分辨率）上，本文基础模型耗时11.9 ms，比RED快3倍，比ASTMNet快5倍以上。实验表明**在功耗较小的情况下，RVTs具有低延迟推断的潜力**。

5、总结

- 提出了一种**多阶段层次化的主干网络**设计，用于事件相机目标检测。这个设计**结合了传统基于帧的目标检测的思想和事件相机**，实现了**高性能和高效率**的目标检测。
- 引入了**交错的局部和全局自注意力机制**，能够**同时混合局部和全局特征**。这种注意力机制通过**简单的卷积操作**提供了关于**像素数组的网格结构的先验信息**，能够更好地**捕捉目标在图像中的局部和全局上下文信息**，提高检测的准确性。

• 使用LSTM单元进行**时间特征聚合**。本文发现可以**用普通的LSTM单元替代Conv-LSTM单元**，从而**减少参数数量和计算延迟**。

• 本文方法实现了与现有方法相媲美的目标检测性能，在**参数数量和延迟上都有显著的减少**，同时**保持了高性能**。

6、启发

• 文中方法使用了一种非常**简单的事件表示方法**，这种方法**没有充分利用事件数据的特性**。例如，文中方法**直接用全连接层处理时间维度**，只对事件的**先后顺序**有一个弱的先验。**事件数据的低层次高效处理**是未来工作中待研究的问题。

• 本文方法目前**只使用事件流**来检测对象。**帧产生互补信息**，可以考虑**同时利用帧与事件信息**，或许可以显著增强检测性能。因此，未来工作可以考虑，在合适的数据集上对本文方法进行**多模态扩展**。