Events-to-Video: Bringing Modern Computer Vision to Event Cameras

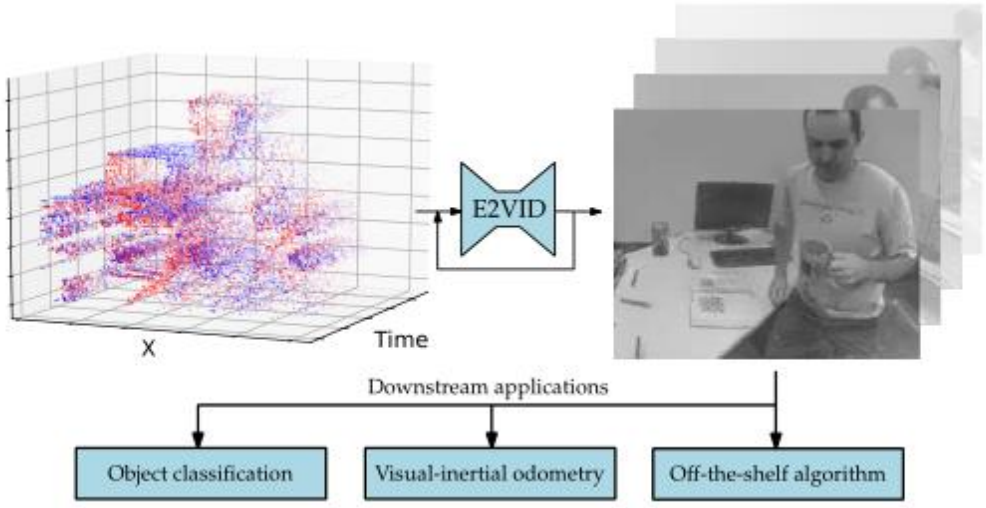
1. 研究问题

构建一种循环网络从事件流中重建视频，并在大量的模拟事件数据上对其进行训练。再进一步将现成的计算机视觉算法应用于从事件数据重构的视频用于对象分类和视觉惯性里程计。

视觉惯性里程计：融合相机和 IMU 数据实现 SLAM 的算法。

2. 方案设计

在事件相机和传统计算机视觉中建立一个桥梁，即学习事件流和图像流之间的映射，从而将现成的计算机视觉技术应用于事件相机。如下图：



- 将左边的事件时空流转换成右边的高质量视频，从而可以直接应用现成的计算机视觉算法。

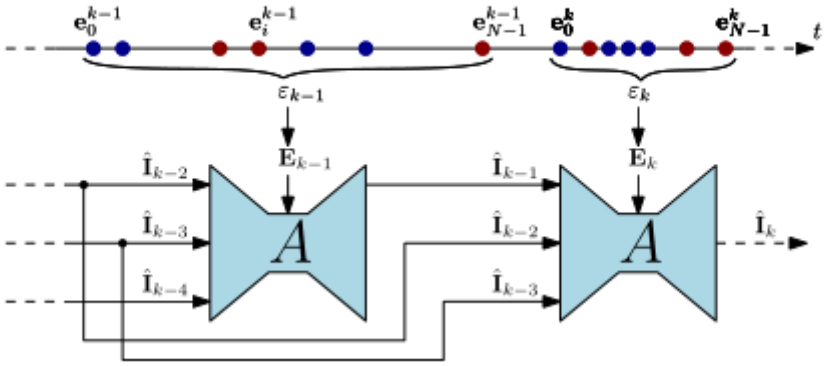
3. 具体方法

与以往将手工制作的平滑先验嵌入到重建框架中不同的是，作者直接使用大量的模拟事件数据从事件中学习视频重建。使用循环网络从长

事件流中重建时间一致的视频，直接学习逐像素的强度估计。

- 视频重构

- 以监督的方式训练网络使用大量的模拟事件序列和相应的真实图像。网络结构如下图：



- 将传入的事件流划分为事件$\varepsilon_k = \{\mathbf{e}_i\}$的连续时空窗口，对于i∈[0,N-1],每个窗口包含固定数量的N个事件。对于每一个新的事件序列ε_k，通过融合之前的K个重构图像$\{\hat{\mathcal{I}}_{k-K}, \ldots, \hat{\mathcal{I}}_{k-1}\}$与新事件$\varepsilon_k$从而将传入的事件流转化为图像序列$\{\hat{\mathcal{I}}_k\}$（$\hat{\mathcal{I}}_k \in [0,1]^{W \times H}$）。

- 事件表示

为了能够使用CNN处理事件流，需要将ε_k转换为固定大小的张量E_k。将事件编码为一个时空体素网格。ε_k中的事件所跨越的持续时间$\Delta T = t_{N-1}^k - t_0^k$离散成B个时间箱。每个事件将其极性$P_i$分配给最接近的两个时空体素。如下：

$$\mathbf{E}(x_l, y_m, t_n) = \sum_{\substack{x_i = x_l \\ y_i = y_m}} p_i \max(0, 1 - |t_n - t_i^*|)$$

- 其中，$t_i^* \triangleq \frac{B-1}{\Delta T}(t_i - t_0)$ 为归一化时间戳，使用 N=25000 个事件和 B=10 个时间箱。

4. 实验对比

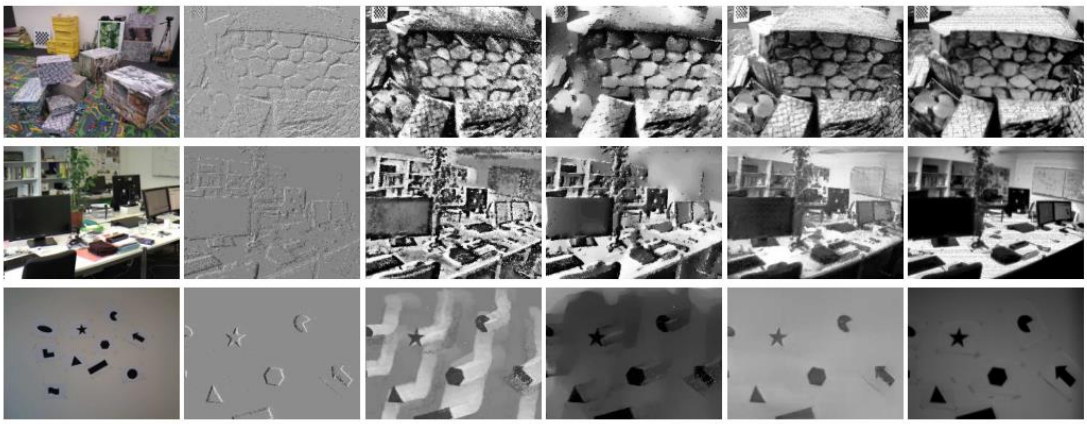
使用 DAVIS240C 传感器记录的事件数据集中的事件序列并去除冗余序列和帧质量较差的序列。并与当前方法进行对比。

- **定量结果：**

Dataset	MSE			SSIM			LPIPS		
	HF	MR	Ours	HF	MR	Ours	HF	MR	Ours
dynamic_6dof	0.10	0.11	**0.08**	0.39	0.44	**0.50**	0.53	0.53	**0.43**
boxes_6dof	0.09	0.07	**0.04**	0.45	0.47	**0.63**	0.51	0.54	**0.36**
poster_6dof	0.06	0.05	**0.04**	0.52	0.55	**0.68**	0.44	0.50	**0.32**
shapes_6dof	0.11	0.14	**0.10**	0.34	0.43	**0.44**	0.63	0.64	**0.53**
office_zigzag	0.09	0.06	**0.05**	0.36	0.43	**0.50**	0.54	0.55	**0.44**
slider_depth	0.08	0.08	**0.06**	0.48	0.51	**0.61**	0.50	0.55	**0.42**
calibration	0.07	0.06	**0.04**	0.41	0.41	**0.52**	0.55	0.57	**0.47**
Mean	0.09	0.08	**0.06**	0.42	0.46	**0.56**	0.53	0.55	**0.42**

- MR 用于流形正则化，HF 用于高通滤波器。
- MSE 为均方误差（越低越好）；SSIM 为结构相似性（越高越好）；LPIPS 为校准感知损失（越低越好）。

- **定性结果**

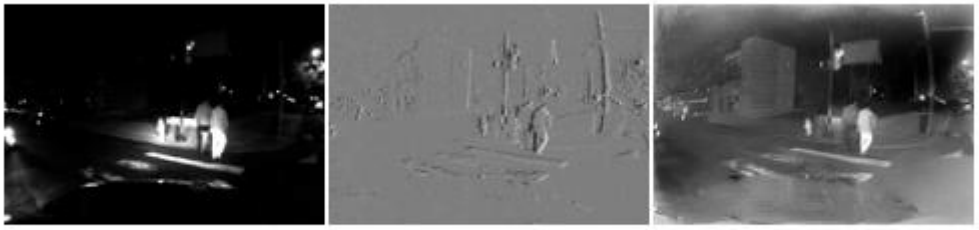


(a) Scene overview (b) Events (c) HF (d) MR (e) Ours (f) Ground truth

- 可以看出该网络可以很好的重建细节且避免了伪影。

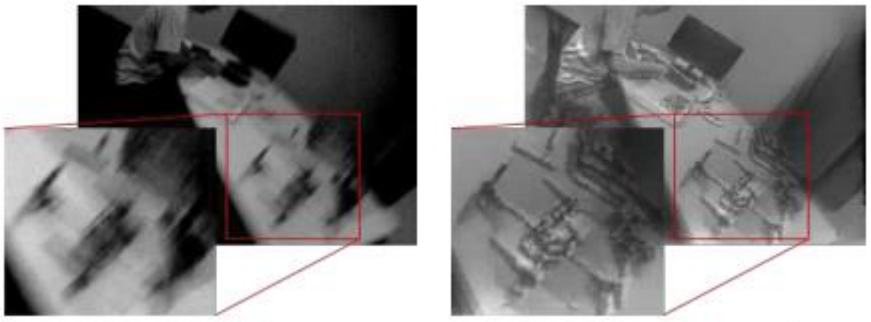
- 亮点

- 该网络能够利用事件的杰出属性来重建低光下的图像，如下：



(a) VI Sensor (b) Events (c) Our reconstruction

- 重建高速运动下的图像，如下：



(a) DAVIS frame (b) Our reconstruction

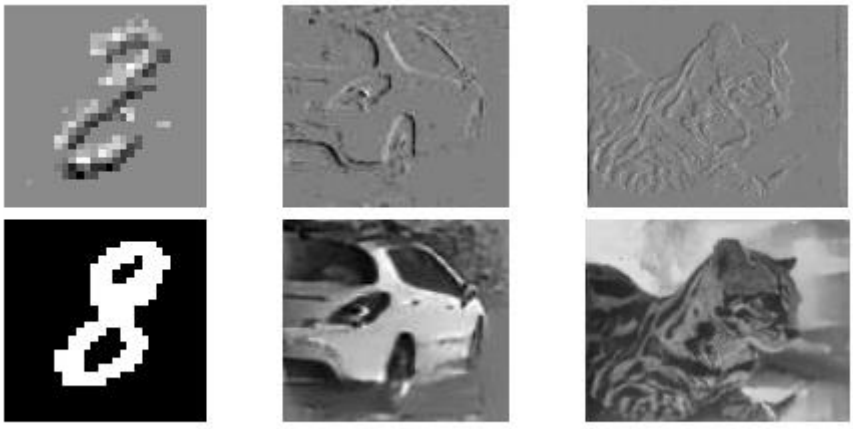
- 限制

引入了延迟，因为在窗口中处理事件而不是 event-per-event，导致可能无法正确重建图像的某些部分，并可能将误差传播到下一次重建。

5. 下游应用程序（Downstream Applications）

- 对象分类（Object Classification）

1. 直接在从事件重建的图像上训练分类网络，对于训练集中的每个事件序列，使用该网络从事件中重建一个图像，如下图，每个数据集的样本用于评估基于事件的对象分类方法.



(a) N-MNIST (b) N-CARS (c) N-Caltech101

2.使用来自训练集的重建图像训练现成的 CNN 进行对象分类,对于 N-MNIST，使用 CNN 从头训练，对于 N-Caltech101 和 N-CARS，使用 ResNet-18。

3.对于基线，直接报告 HATS 中提供的精度。尽管文中方法很简单，但它优于所有基线。结果如下：

	N-MNIST	N-CARS	N-Caltech101
HOTS	0.808	0.624	0.210
HATS/linear SVM	**0.991**	0.902	0.642
HATS/ResNet-18	n.a.	0.904	0.700
Ours (transfer learning)	0.807	n.a.	0.821
Ours (fine-tuned)	0.983	**0.910**	**0.866**

- 值得注意，随着数据集变得更加困难，该方法与其他方法之间差距也在增加。
- 在 N-MNIST 上表现比 HATS 略差，这归因于 N-MNIST 的合成性质。因此与手工制作的特征表示（如 HATS），该方法并没有实质优势。
- 在纯迁移学习设置中(将从事件重构的图像反馈给在真实图像

数据上训练的网络）表现得更好。这是第一次实现图像数据和事件数据之间的直接迁移学习。

- 视觉惯性里程计（Visual-Inertial Odometry）

简介：VIO 是通过一组视觉测量和固定在相机上的惯性测量单元（IMU）的惯性测量来恢复相机的 6 个自由度姿态。

方法：将一个现成的VIO系统应用于该方法从事件中重建的视频，并对 UltimateSLAM 进行评估。结果如下：

SLAM:即时定位与地图构建。

UltimateSLAM (E+I)只使用事件和 IMU,而 UltimateSLAM (E+F+I)使用事件、IMU 和附加帧。

Inputs	Ours E+I	U.SLAM E+I	U.SLAM E+F+I	HF E+I	MR E+I	VINS-Mono F+I
shapes_translation	0.18	0.32	**0.17**	failed	2.00	0.93
poster_translation	**0.05**	0.09	0.06	0.49	0.15	failed
boxes_translation	**0.15**	0.81	0.26	0.70	0.45	0.22
dynamic_translation	**0.08**	0.23	0.09	0.58	0.17	0.13
shapes_6dof	1.09	0.09	**0.06**	failed	3.00	1.99
poster_6dof	**0.12**	0.20	0.22	0.45	0.17	1.99
boxes_6dof	0.62	0.41	**0.34**	1.71	1.17	0.94
dynamic_6dof	0.15	0.27	**0.11**	failed	0.55	0.76
hdr_boxes	0.34	0.44	0.37	0.64	0.66	**0.32**
Mean	0.31	0.32	**0.19**	0.76	0.92	0.91
Median	**0.15**	0.27	0.17	0.61	0.55	0.84

- 总体而言，该方法的中位数误差为 0.15 m，几乎比使用完全相同数据的 UltimateSLAM (E+I) (0.27 m)小两倍。
- 该方法与 UltimateSLAM (E+F+I)的性能相当，而后者需要该方法不需要的额外帧。

6. 总结与展望

总结：提出了一种基于模拟事件数据训练的循环卷积网络的事件到视频重建框架，还展示了该方法作为传统相机和事件相机之间的桥梁，在两种视觉应用上的适用性，即从事件中进行物体分类和视觉惯性里程计。

展望：在多数人认为事件流必须有对应算法才能更好处理时，作者想到了将其重构为视频从而将现成的计算机视觉技术应用于此，是个开创性的方法，但由于在窗口处理事件，引入了一点延迟，这可能导致图像不能很好的重建，并且影响到下一次重建。

个人看法：如果将事件流分为 N 个窗口并联进行，然后进行 N 个窗口的合并，是否会解决这种不好的传播效应？