专题 12 / 事件视觉 / 稠密运动估计

E-RAFT：从事件相机估计稠密光流

E-RAFT: Dense Optical Flow from Event Cameras

Mathias Gehrig*，Mario Millhäusler*，Daniel Gehrig，Davide Scaramuzza（*共同第一作者） · 3DV 2021

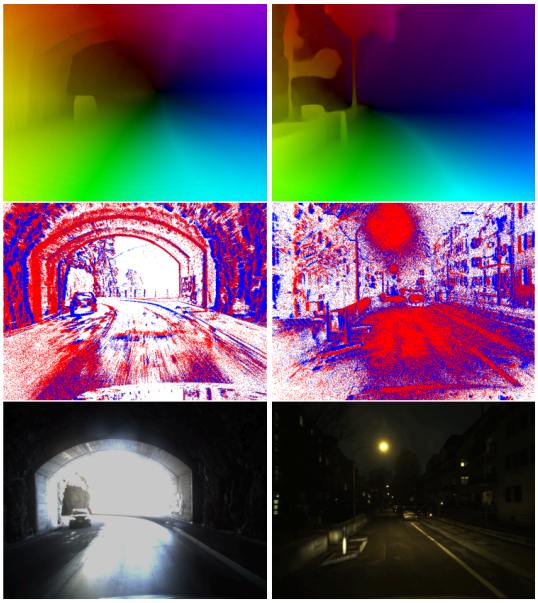
对照 arXiv:2108.10552v3，逐段翻译摘要、正文 §1–7、图表注及公式。读图批注与正文分列。原文十页，页码均从 PDF 首页起算。

摘要 1 引言 2 相关工作：2.1 基于图像的学习方法 3 方法 4 DSEC-Flow 数据集 5 实验 6 结论 7 致谢

摘要

S001 · p.1 我们将特征相关性和序列处理引入事件相机的稠密光流估计。现代帧式方法通常用特征相关性计算匹配代价；相比之下，当时的事件光流方法尚未显式计算这一代价，学习方法多以 U-Net 估计稀疏光流。实验的主要发现是：与只使用卷积层的既有方法相比，引入相关特征能明显提高精度。本方法输出稠密光流，在 MVSEC 上将端点误差降低 23%。我们还发现，以往事件光流方法所用数据集的位移很小，流幅值最大仅约 10 像素。因此，我们构建了一个真实场景数据集，将最大位移扩展至 210 像素，相机像素数约为原来的三倍。在这个数据集上，本方法将端点误差降低了 66%。

1 引言



S002 · p.1 图 1 | 用本文提出的新事件光流数据集训练模型后所得的预测。第二行是输入事件的可视化，第三行是同一场景的图像。

读图 / 三行共享场景坐标。第三行供人辨认街景，而非网络的另一条 RGB 输入；比较暗处的事件边缘与上方稠密颜色分布。

S003 · p.1 光流估计描述运动如何投影到图像平面，是计算机视觉的基本任务。机器人及 AR/VR 中的视觉里程计 [28]、自动驾驶 [16]、动作识别 [31]、高动态范围成像 [18] 和计算摄像 [25] 都以它为基础。传统问题表述是：找出两幅图像之间彼此对应的像素。

S004 · p.1 早期精度最高的方法大多采用能量最小化 [13]；如今，神经网络方法 [34] 已占主导地位。

S005 · p.1 经过数十年发展，基于图像的光流估计已在公开基准 [23] 上取得优异结果。然而，运动模糊 [6] 或传感器动态范围有限造成的过曝区域都会使这些方法失效。事件相机为光流估计提供了另一种有前景的传感器。

S006 · p.1–2 事件相机的采集机制与帧式传感器截然不同：它不按固定间隔拍摄图像，而以异步事件流报告各像素的亮度变化。详细综述见 [9]。其优势在于微秒量级的时间分辨率、超过 120 dB 的动态范围以及没有帧曝光引起的运动模糊。三者对真实场景的高质量光流估计都很重要，同时也意味着必须处理一种完全不同的数据格式。

S007 · p.2 从异步事件流中有效提取运动并不容易。图像可以直接送入卷积网络，事件却在时空上稀疏、不规则且不同步。既有研究包括时空平面拟合、对离散化时空体做变分优化，以及在预处理事件上使用卷积网络。与帧式光流类似，当前领先方法主要由数据驱动。这些方法把事件转换为张量，用 U-Net 提取运动特征，并且只在稀疏位置评估光流。 [3, 2, 40, 41, 38, 29]

S008 · p.2 现代图像光流方法不同于上述事件方法，会借助代价体提取有判别力的特征。网络一般先在较低分辨率上计算图像描述子，再求源、目标描述子之间的相关性。相关值汇集为代价体，可以交给卷积层处理；Teed 等 [34] 则采用迭代更新，从代价体中逐步提取光流。 [35, 32, 36]

S009 · p.2 我们研究的是事件相机的稠密光流，用代价体代替 U-Net。首先，我们表明：把先进的图像光流方法用于事件数据，可以持续取得优于先前方法的性能。接着利用事件数据的顺序性，设计可微的热启动，用上一时刻的预测初始化下一时刻。不同于 [34]，梯度经可微热启动模块沿时间反向传播。序列事件实验表明，训练时引入热启动还能提高精度。该模块不增加参数，因此可直接加入训练好的非循环模型，面向序列数据继续微调。

S010 · p.2 此前，数据驱动的事件光流方法只在 MVSEC [40] 上评测。我们的分析表明，该数据集的难度有限：例如，室外评测序列中，95% 的真值位移小于 3 像素（图 3）。为此，我们为一套真实驾驶数据增加光流真值，新数据集的最大位移可达 210 像素，事件相机像素数约为 MVSEC 的三倍。我们重新训练并评测了已有的先进事件方法。与这类过去只在 MVSEC 上测试的方法相比，我们的方法在 MVSEC 上把 EPE 从 0.35 降至 0.27，降幅约 23%；在新数据集上则从 2.32 降至 0.79，降幅约 66%。 [11]

S011 · p.2 本文的贡献如下：

S012 · p.2 首次针对事件相机的稠密光流估计设计方法，并按稠密预测进行评估。

S013 · p.2 首次在事件光流方法中使用代价体，并引入循环结构以利用时间先验。以往方法主要通过平面拟合、变分优化或纯卷积层估计光流。

S014 · p.2 提出新的真实事件光流数据集。与当时已有数据集相比，它包含更大的像素位移、更高的空间分辨率和质量更好的真值。

S015 · p.2 在 MVSEC 与新数据集上进行了受控对照实验。结果表明，代价体和跨时间的循环对事件光流估计均有效，优于纯卷积结构。

2 相关工作：2.1 基于图像的学习方法

S016 · p.2–3 FlowNet [8] 开创了数据驱动的图像光流估计。随后，结构设计上的进展大幅提高了精度。高性能方法逐渐形成三类共同要素：从粗到细的处理、特征相关体，以及特征变形。Teed 等 [34] 进一步建立多尺度相关体，并迭代提取光流，在 Sintel 和 KITTI 基准上取得了当时领先的结果。 [39, 14, 15, 21, 32, 33, 36, 6, 12]

S017 · p.3 这些方法在常规条件下表现良好，精度却与输入图像质量紧密相连。因此，高动态范围场景和高速运动造成的图像退化，会超出它们的应对能力。

S018 · p.3 为了克服这一内在限制，我们采用事件相机 [20]，利用其不受常规帧式运动模糊影响、动态范围宽的特点。

2.2 事件光流估计

S019 · p.3 传统的事件光流算法可分为三类。第一类是受 Lucas–Kanade 方法启发的迭代异步算法，其中一种还利用图像做光度特征跟踪。第二类是平面拟合，利用事件流在局部时空邻域内近似成平面的几何性质。第三类是变分优化，可引入图像数据或同时估计图像强度。这些方法直接使用事件，通常

还需亮度恒常等强假设。与之不同，我们从数据中学习抽象特征和先验，不以这些假设为前提。 [4, 10, 22, 5, 24, 2, 26]

S020 · p.3 另一条路线采用自监督学习，以图像或纯事件构造监督信号。除 Ye 等的方法外，已有研究大多因为无事件处精度不足，而在评测时掩去这些位置；Ye 等则假定场景静止，由深度和相机自身运动推算光流。上述路线通常以 U-Net 处理事件表征，主要估计和评估稀疏光流。本文则预测并评估稠密光流，显式使用更长的历史，并利用代价体的相关特征在大位移下寻找像素对应。在多个数据集上，这一不同于 U-Net 的设计均优于已有方法。 [41, 19, 37, 27, 40]

2.3 用于事件光流的数据集

Dataset	Train/Test Split	Groundtruth Quality	Resolution (MP)	Translational Motion	Outdoor
DVSFLOW16	✘	**high**	0.05	✘	✘
DVSMOTION20	✘	**high**	0.1	✘	✘
MVSEC	(✘)	medium/low	0.1	✔	✔
Ours	✔	**high**	**0.3**	✔	✔

S021 · p.4 表 1 | 现有事件相机光流数据集对照。

读图 / 从左到右核对训练/测试划分、真值质量、像素数、是否含平移、是否室外。MVSEC 的括号叉记表示划分不统一，而不是没有可用测试数据。

S022 · p.3 据我们所知，当时只有三个事件相机数据集含光流真值。表 1 比较了它们与本文数据集的差异。

S023 · p.3 DVSFLOW16 [30] 和 DVSMOTION20 [1] 是两个小型数据集，用于考查室内场景中相机旋转时的事件光流算法。样本少且相机运动受限，因此不太适合训练和评价神经网络。

S024 · p.3 更接近本文的是 MVSEC [40]，它包含 5 段驾驶序列和在同一环境下采集的 4 段室内序列。室外光流真值来自激光里程计和建图，但有三个主要限制。第一，遮挡和运动物体处理不充分，损害了真值质量。第二，事件相机仅有约 0.1 百万像素，位移偏小（详见第 4 节）。第三，没有明确、固定的训练测试划分；例如，已有工作既在 outdoor day 1 上做评测，又在评测室内序列时用该序列训练。新数据集清除了遮挡含混和运动物体伪影，将像素数提高到约三倍，并设置明确的数据划分和评价指标以支持复现。

3 方法

S025 · p.3 方法受双帧光流 [34] 启发，但输入改为连续事件包，并转换成标准卷积网络能处理的张量。两个共享权重的卷积编码器生成特征嵌入，在原始宽、高各八分之一处建立全对相关体。另一个结构相同、权重独立的编码器提取上下文特征。两种输出共同进入迭代更新，求出逐像素光流。再通过可微热启动，将模型扩展为能够处理连续事件包序列的时间循环结构（第 3.3 节）。

3.1 问题定义

S026 · p.3 通常，稠密光流是寻找图像 I_i 与 I_j 的像素对应，记为 $F_{i\to j}$。在事件相机上，相应问题是：给定两时刻之间的事件，找出时刻 t_i 与 t_j 之间的像素对应 $F_{t_i\to t_j}$。与双图像设定不同，这里严格按时间顺序处理数据。下文把它简写为 $F_{i\to j}$。 [23, 6]

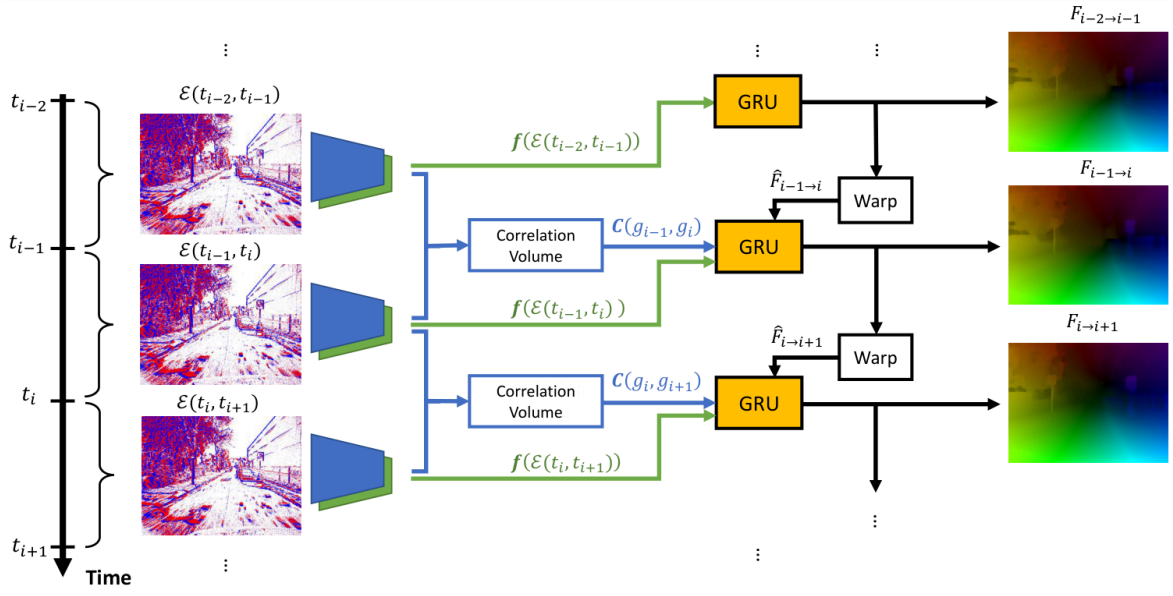
S027 · p.3–4 帧式相机按一定频率成像，事件相机的每个像素独立、异步地响应亮度变化。像素保存对数强度参考值；一旦当前强度相对参考值的偏离足够大，就触发事件并更新参考值。每个事件 $e_k(t)=(x_k,y_k,t,p_k)$ 包含像素坐标、触发时刻及一位极性，后者说明亮度是增加还是降低。将两时刻间的有序事件写作 $E(t_i,t_j)=\langle e(t)\mid t\in[t_i,t_j]\rangle$。

3.2 从事件包构建 RAFT

S028 · p.4 以往神经网络通常只用 t_i 到 $t_{i+1}=t_i+\Delta t$ 的事件构造输入，直接回归 $F_{i\to i+1}$。我们使用两个连续短时事件包：$E(t_{i-1},t_i)$ 和 $E(t_i,t_{i+1})$，来回归同一目标位移。 [40, 41, 27]

S029 · p.4 基线采用 RAFT [34]，因为它容易扩展到第 3.3 节所述的序列设定（图 2）。RAFT 计算三份嵌入：共享编码器为两份输入分别提取特征，用于构建代价体（蓝色）；不同权重的上下文编码器再给出第三份特征（绿色），直接送入 GRU 更新模块。 [7]

S030 · p.4 为了使用卷积层，首先需要合适的事件表征。取前一包从 $t_i-\Delta t$ 至 t_i、后一包从 t_i 至 $t_i+\Delta t$。对每包分别构建体素网格 [41]，把时间轴离散化，并沿时间进行插值，因而能保留大部分时间信息。相关特征编码器分别处理前、后两个事件表征；上下文编码器只看后一包。原因在于，与 t_i 之前的事件相比，从 t_i 到 t_{i+1} 的事件更直接反映待估计流场所在时段的上下文。



S031 · p.5 图 2 | 事件光流的循环估计。相邻两个事件表征用于构造代价体 C；上下文编码器 f 初始化 GRU 的隐状态，并提供额外语义先验。变形模块连接不同时间戳，使信息能够沿时间传播。

读图 / 横向读一次更新：两包→相关体，后包→上下文。再纵向读相邻时刻：Warp 传递的是已经移到新坐标的流，不是把像素索引原样复制。

3.3 可微的光流热启动

S032 · p.4 本节讨论如何使前一时段的光流先验不仅在评测时生效，也参与模型训练。

S033 · p.4 RAFT 默认把初始光流 $F_{i\to i+1}$ 设为零。数据成序列时，也可用上一次预测来初始化。我们把这一步写成可微运算并纳入训练，使网络既能使用时间先验，又能沿时间反向传播梯度。热启动不引入额外参数，可以加到已经训练好的网络上继续微调。整体结构见图 2。 [34]

S034 · p.4 我们想用 $F_{i-1\to i}$ 初始化最终求取 $F_{i\to i+1}$ 的迭代模块。但若直接照抄旧流场，就忽略了源像素位置已随光流移动。必须将旧流场向前变形到下一时间戳。多个源像素可能映射到同一个目标像素，为消除歧义，我们在目标网格上做加权平均散射：将投送的向量按双线性核的权重和归一化。因此，用前一预测初始化新流场可写为：

$$g(x_i) = x_i + F_{i\to i+1}(x_i)$$
$$F_{i\to i+1}(x) = \frac{\sum_{\forall x_{i-1}} k_b(x - g(x_{i-1}))F_{i-1\to i}(x_{i-1})}{\sum_{\forall x_{i-1}} k_b(x - g(x_{i-1}))}$$

S035 · p.4 式（1–2）｜先把源像素移到预测位置，再把到达同一目标邻域的前一时段位移向量按权重平均。

S036 · p.4 其中，权重为

$$k_b(a) = \max\{0, 1 - |a_x|\} \cdot \max\{0, 1 - |a_y|\}$$

S037 · p.4 式（3）｜双线性插值核。

S038 · p.4 该核沿两个坐标轴分别作线性插值，再取乘积。这些运算可取次梯度，能够直接用 PyTorch 等深度学习框架实现。

3.4 监督信号

S039 · p.5 所有方法均用预测光流与真值之间的 L_1 距离监督。本文方法的损失为：

$$L = \sum_{i=1}^{N_i} \sum_{k=1}^{N_k} \gamma^{N_k - k} ||F^{\mathrm{gt}}_{i\to i+1} - F^{k}_{i\to i+1}||_1$$

S040 · p.5 式（4）｜对序列内每一时段、每一次迭代的误差求和，越靠近最终预测的迭代，权重越大。

S041 · p.5 其中 N_i 是序列长度，N_k 是每一时段的迭代次数；$F^{\mathrm{gt}}_{i\to i+1}$ 为第 i 个时段的真值，$F^{k}_{i\to i+1}$ 是同一时段第 k 次迭代的预测。本文取 $\gamma = 0.8$。

4 DSEC-Flow 数据集

S042 · p.5 我们扩展了公开的 DSEC 数据集，以弥补 MVSEC 的主要不足。原 DSEC 已提供由旋转式激光雷达取得的视差真值。其 VGA 事件相机的像素数是 MVSEC 所用 Davis346 的 3.4 倍。DSEC 的视差图还剔除了遮挡，用双目匹配筛选掉运动物体上的激光点。这些条件使进一步构造光流真值成为可能。 [40, 11]

S043 · p.5 对于每一段序列，我们使用相机内参和外参将视差重新投影到三维，再根据真值里程计，在场景静止的假设下，以 10 Hz 计算前向和后向光流。已有过滤虽去除了大多数动态物体上的激光点，却仍有残留。为保证真值精度，我们人工检查全部数据，只截取环境静止的区段。最终数据集含昼夜共 24 段序列，约 7,800 个训练样本和 2,100 个测试样本。图 4 给出了预测示例。 [11]

S044 · p.5 像素位移幅值的分布是判断光流数据集难度的重要依据。若用绝对误差衡量，小位移通常比大位移容易估准；而高速视频若要获得准确、低漂移的光流，又必须建立长距离对应。因此，数据集和算法都要同时兼顾小位移与大位移。 [6, 17]

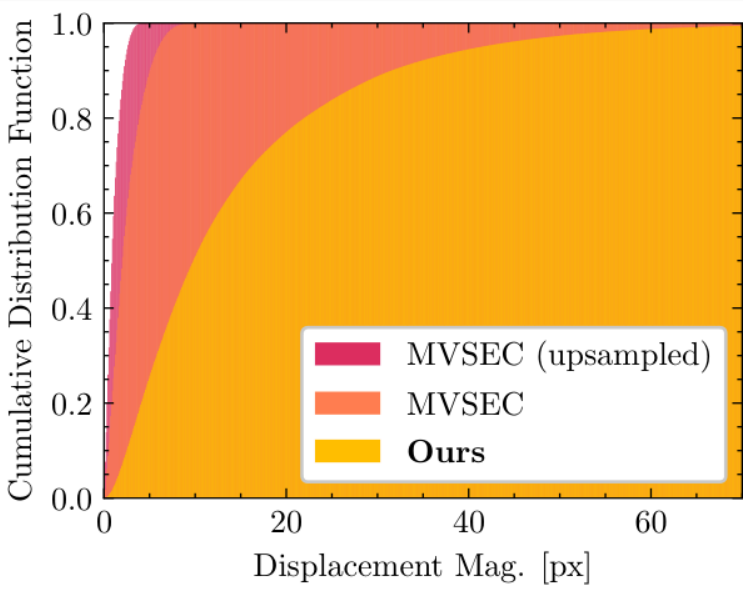
S045 · p.5–6 图 3 比较 DSEC-Flow 和 MVSEC 的真值位移分布。MVSEC 原生光流真值为 20 Hz；沿用 Zhu 等最初的评估方式，后续工作也将真值沿时间轴上采样至相机帧率，约为 45 Hz。 [40, 19, 41, 37]

S046 · p.6 新数据集的位移明显更大。MVSEC 上，80% 的位移小于 4 像素；上采样版本的这一阈值降至 1.7 像素。DSEC-Flow 的第 80 百分位则达到 22 像素。

S047 · p.6 再按图像宽度归一化，DSEC-Flow 的第 80 百分位为 3.4%，MVSEC 原版和上采样版分别为 1.3% 与 0.5%。因此，即使剔除相机分辨率不同带来的影响，新数据集仍更难。

4.1 评价方法

S048 · p.6 测试集包含 7 段序列的约 2,100 张真值光流图。评价既使用端点误差 EPE，也报告误差幅值大于 N 像素的比例，即 NPE。分别取阈值 3、2、1 像素，由粗到细分析大误差；默认指标采用 EPE，因为它对光流精度的刻画更细致。



S049 · p.6 图 3 | 光流幅值的累积分布。原图注写作按图像宽度的像素数归一化；新数据集的位移显著大于 MVSEC 的两个版本。“Upsampled”指依 Zhu 等的方法在时间上对真值上采样。横轴实际单位见图旁译注。

读图 / 在纵轴 0.8 处横向读，第 80 百分位为 1.7、约 4、22 px。图内横轴是 px；正文另报除以图宽后的 0.5%、1.3%、3.4%。

5 实验

S050 · p.6 我们在 MVSEC 和新提出的 DSEC-Flow 上评估方法。既有基线均为在 MVSEC 上训练和评估的自监督方法。为比较已有结构能达到的水平，我们用真值监督，从头在两个数据集上重训 EV-FlowNet。采用相同结构的工作在表 2 所列直接估计光流的基线中，稀疏 EPE 最低。脚注：ECN 只借助深度和相机运动间接计算静态场景光流，不宜直接相比。各方法分别调节学习率，均采用水平翻转、随机裁剪增强。两套数据在训练和评估时，每一时段均迭代 12 次。 [40, 41, 37]

5.1 MVSEC

	Loss Type	Dense EPE	Dense 1PE	Dense 3PE	Sparse EPE	Sparse 1PE	Sparse 3PE
Back to Event Basics [27]	E	-	-	-	0.92	-	5.4
Spike-FlowNet [19]	I	-	-	-	0.84	-	-
ECN * [37]	E	*0.35*	-	0.0	*0.30*	-	0.0
EV-FlowNet [40]	I	-	-	-	0.49	-	0.2
Zhu et. al [41]	E	-	-	-	0.32	-	0.0
EV-FlowNet [40]	S	*0.35*	*3.7*	0.0	0.31	**1.6**	0.0
Ours	S	**0.27**	**1.7**	0.0	**0.24**	*1.7*	0.0

S051 · p.7 表 2 | MVSEC 上将光流真值沿时间上采样至约 45 Hz 的结果。监督类型 E 为事件损失，I 为图像光度损失，S 为有真值监督；星号表示由深度和相机自身运动间接求光流。最优值加粗，次优值用斜体。几乎所有方法的 3PE 都为零，因此该列不标优。

读图 / 先选 Dense 或 Sparse 再比较。监督版 EV-FlowNet 与本文方法在稠密掩膜口径分别为 0.35、0.27 px；稀疏列分别是 0.31、0.24 px。

	Dense			Sparse		
	EPE	1PE	3PE	EPE	1PE	3PE
EV-FlowNet [40]	0.61	15.6	0.45	0.53	10.7	0.60
Ours	**0.47**	**9.2**	**0.24**	**0.46**	**7.5**	**0.49**

S052 · p.7 表 3 | MVSEC 原生 20 Hz 真值上的结果，粗体为最优值。

读图 / 用 20 Hz 原生真值，稠密 EPE 由 0.61 降到 0.47 px；3PE 不再处处为零。时间窗改变了位移大小，不能横跨两张表比较百分比而忽略采样条件。

S053 · p.6 在 outdoor day 2 序列上从头训练采用可微热启动的模型，起初序列长度为 2，10 个 epoch 后增至 5。与已有工作一致，先基于上采样到相机帧率（约 45 Hz）的真值训练并评估。另外还用未经修改的原生真值训练和评估 EV-FlowNet 与本方法，以提高任务难度。

S054 · p.6 为与前人结果保持可比性，表 2 也给出 MVSEC 上的 3 像素误差比例，虽然它通常为零。

S055 · p.6 稀疏与稠密。既有研究往往只评估触发事件的像素。由于并非所有方法都开源，我们用上采样真值比较本方法与有监督基线。表 2 表明，加事件掩膜后，各项指标普遍降低：本方法稀疏 EPE 比稠密 EPE 低约 11%。可见在 MVSEC 上，稠密预测更难。本方法的稠密、稀疏 EPE 分别为 0.27 和 0.24，优于稀疏及稠密评估里的基线与文献结果。 [40, 41, 19, 27]

S056 · p.6 有监督与自监督。在该数据集的已有比较中，只有本文模型和重训的 EV-FlowNet 使用真值监督。无论稀疏还是稠密评估，有监督方法总体优于自监督方法。较弱的有监督基线 EV-FlowNet 的稀疏 EPE 为 0.31，而 Zhu 等方法为 0.32。作者推测，若 MVSEC 光流真值质量更高，有监督学习的优势还会更大。 [41]

S057 · p.6–7 时间上采样与原生频率。表 3 给出两种有监督方法在未上采样的 MVSEC 上的表现。本方法相对 EV-FlowNet 的 EPE 降幅，在两种真值频率下均约为 23%。因此，把 MVSEC 改用原生 20 Hz 真值，尚不足以揭示估计大位移时二者有何不同，下一节将专门考察这一点。但两种方法的 1 像素误差比例都明显增大，这是时间窗变长、位移随之增大的预期结果。

5.2 DSEC-Flow

S058 · p.7 在 DSEC-Flow 上，先作水平翻转并随机裁剪为高 288、宽 384 像素，用 Adam 和 10^{-4} 的学习率训练非循环版本 40 个 epoch，直至收敛。随后加入可微热启动，使用长度为 3 的序列，继续随机裁剪并训练 6 个 epoch。最后将学习率降为原来的十分之一，在完整分辨率上微调后评估。为保证公平，非循环模型及基线也使用同样的微调安排。

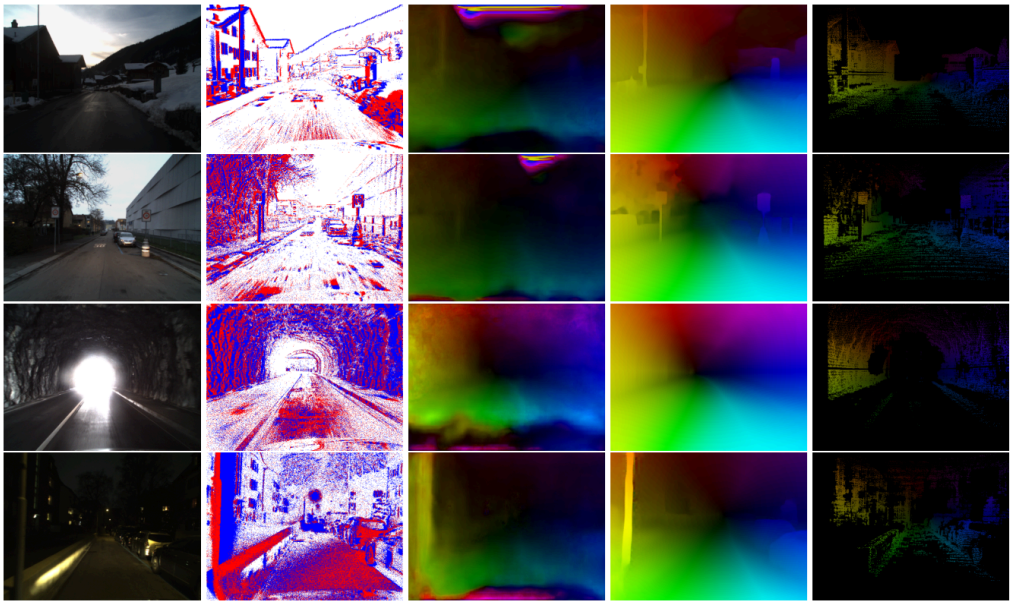
S059 · p.7 表 4 中，本方法的 EPE 为 0.79，EV-FlowNet 为 2.32，后者约为前者的 2.9 倍。在 MVSEC 上，两种方法的误差尚属同一量级；在 DSEC-Flow 上，EV-FlowNet 却明显落后。作者提出的一种可能解释是：EV-FlowNet 是在 MVSEC 的小位移条件下设计的。

	EPE	1PE	2PE	3PE
EV-FlowNet [40]	2.32	55.4	29.8	18.6
Ours	**0.79**	**12.5**	**4.7**	**2.7**

S060 · p.7 表 4 | DSEC-Flow 数据集上的评估结果，最优值加粗。

读图 / DSEC：EPE 以像素计，其余为百分比。本文 / 基线的 3PE 为 2.7% / 18.6%；看平均值之外，还应看长尾。

DSEC 测试集	EPE / px	1PE / %	2PE / %	3PE / %
EV-FlowNet	2.32	55.4	29.8	18.6
E-RAFT	0.79	12.5	4.7	2.7



S061 · p.8 图 4 | DSEC-Flow 测试集的光流预测示例，建议放大 PDF 查看。从左至右依次为仅供观察的图像、事件可视化、EV-FlowNet、本文预测、真值。相比噪声明显的 EV-FlowNet，本文方法能较稳定地给出清晰的运动边界。

读图 / 从左到右比较输入、估计和真值的有效覆盖。估计边缘更锐利并不等于没有激光真值的区域也已被定量验证。

5.3 消融实验

S062 · p.7 接下来用消融实验分析各模块的作用。两套数据集均进行消融；表 5 汇总不同热启动策略及训练、评测组合的结果。

	DSEC-Flow				MVSEC	
	EPE	1PE	2PE	3PE	EPE	1PE
Ours w/o WS	0.82	13.5	5.2	3.0	0.53	12.2
Ours eval WS	0.88	14.1	5.6	3.3	0.51	10.9
Ours w/o WS: 100 iters	0.87	14.9	5.9	3.4	0.60	16.3
Ours train & eval WS: 100 iters	0.88	15.1	6.0	3.5	0.49	10.0
Ours train & eval WS	**0.79**	**12.7**	**4.7**	**2.7**	**0.47**	**9.2**

S063 · p.8 表 5 | DSEC-Flow 与 MVSEC 上的消融结果，最优值加粗。WS 指用前一预测初始化当前光流；eval WS 指训练时未用热启动，仅在推理时启用。

读图 / 在 DSEC，训练不热启动、只在推理热启动使 EPE 从 0.82 恶化到 0.88；训练与推理都启用才到 0.79。表 5 的最终 1PE 为 12.7%，和表 4 的 12.5% 略异。

S064 · p.7 “Ours w/o WS”是处理事件数据的非循环版本，训练和评估都不使用热启动。

S065 · p.7 DSEC-Flow。在训练时没有引入热启动的情况下，仅在评估时启用反而不如基线。不论是否在训练及评估中使用热启动，把迭代数增至每时段 100 次的两种设置也都不如基线。训练和评估均使用热启动的正式方法 EPE 为 0.79，作者报告相对消融方案降低约 4.0%。

S066 · p.7–8 MVSEC。训练与评估都使用热启动同样最好。与 DSEC-Flow 不同，即使训练时不使用，只在评测时打开热启动也比完全不用的误差低。与最接近的消融配置（训练和评估均热启动、迭代 100 次）相比，完整方法的 EPE 又降低了 4.2%。

S067 · p.8 总体来看，在微调时就引入热启动，可在两个数据集上稳定获得几个百分点的改进。若像 [34] 那样仅在推理时加入，实验中则不能稳定复现这种收益。

6 结论

S068 · p.8 本文将帧式光流估计中的设计改造为可处理事件输入的方法。在两套数据集上，新方法都显著优于基线，说明事件视觉文献中常见的 U-Net 架构对这一任务并非最佳选择。另一项主要贡献是新的真实场景光流数据集，它针对已有事件数据集的主要不足作了改进。在此数据集上训练，得到的光流预测质量超过了既有事件数据集能够支持的水平。

7 致谢

S069 · p.8 本工作获瑞士国家科学基金会（SNSF）支持的国家机器人研究能力中心（NCCR Robotics），以及欧洲研究理事会（ERC）资助，项目协议号 864042（AGILEFLIGHT）。

参考文献与复现入口

代码：github.com/uzh-rpg/E-RAFT · 数据：dsec.ifi.uzh.ch

[1] Mohammed Almatrafi, Raymond Baldwin, Kiyoharu Aizawa, and Keigo Hirakawa. Distance surface for event-based optical flow. IEEE transactions on pattern analysis and machine intelligence, 2020.

[2] Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), pages 884–892, 2016.

[3] Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. IEEE Trans. Neural Netw. Learn. Syst., 25(2):407–417, 2014.

[4] Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi, and Mandyam Srinivasan. Asynchronous frameless event-based optical flow. Neural Networks, 27:32–37, 2012.

[5] Tobias Brosch, Stephan Tschechne, and Heiko Neumann. On event-based optical flow detection. Frontiers in neuroscience, 9:137, 2015.

[6] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012.

[7] Kyunghyun Cho, B van Merrienboer, Caglar Gulcehre, F Bougares, H Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In Int. Conf. Comput. Vis. (ICCV), pages 2758–2766, 2015.

[9] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey.

IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[10] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: Asynchronous photometric feature tracking using events and frames. Int. J. Comput. Vis., 2019.

[11] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. IEEE Robot. Autom. Lett., 2021.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. Int. J. Robot. Research, 32(11):1231–1237, 2013.

[13] Berthold K.P. Horn and Brian G. Schunck. Determining optical flow. J. Artificial Intell., 17(1):185 – 203, 1981.

[14] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2018.

[15] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.

[16] Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision, 12(1–3):1–308, 2020.

[17] Joel Janai, Fatma Güney, Jonas Wulff, Michael Black, and Andreas Geiger. Slow flow: Exploiting high-speed cameras for accurate and diverse optical flow reference data. In CVPR, 2017.

[18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph., 36(4):144–1, 2017.

[19] Chankyu Lee, Adarsh Kumar Kosta, Alex Zihao Zhu, Kenneth Chaney, Kostas Daniilidis, and Kaushik Roy. Spike-flownet: event-based optical flow estimation with energy-efficient hybrid neural networks. In Eur. Conf. Comput. Vis. (ECCV), 2020.

[20] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128$128 120 dB 15 μs$s latency asynchronous temporal contrast vision sensor. IEEE J. Solid-State Circuits, 43(2):566–576, 2008.

[21] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.

[22] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In Int. Joint Conf. Artificial Intell. (IJCAI), pages 674–679, 1981.

[23] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2015.

[24] Elias Mueggler, Christian Forster, Nathan Baumli, Guillermo Gallego, and Davide Scaramuzza. Lifetime estimation of events from dynamic vision sensors. In IEEE Int. Conf. Robot. Autom. (ICRA), pages 4874–4881, 2015.

[25] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), June 2020.

[26] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2020.

[27] Federico Paredes-Vallés and Guido CHE de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. arXiv preprint arXiv:2009.08283, 2020.

[28] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics, 34(4):1004–1020, 2018.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, pages 234–241. Springer, 2015.

[30] Bodo Rueckauer and Tobi Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. Frontiers in neuroscience, 2016.

[31] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Conf. Neural Inf. Process. Syst. (NeurIPS), 2014.

[32] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2018.

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. IEEE transactions on pattern analysis and machine intelligence, 2019.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Eur. Conf. Comput. Vis. (ECCV), 2020.

[35] Jia Xu, Rene Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), July 2017.

[36] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. NeurIPS, 5:12, 2019.

[37] C. Ye, A. Mitrokhin, C. Fermüller, J. A. Yorke, and Y. Aloimonos. Unsupervised learning of dense optical flow, depth and egomotion with event-based sensors. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[38] Chengxi Ye, Anton Mitrokhin, Chethan Parameshwara, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow and depth from sparse event data. arXiv e-prints, 2019.

[39] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2020.

[40] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In Robotics: Science and Systems (RSS), 2018.

[41] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.

原作者保留署名及原稿 IEEE 出版标注。按 arXiv 本版本所列 CC BY-SA 4.0（https://creativecommons.org/licenses/by-sa/4.0/）制作并以相同许可分享此中文翻译、标注及排版改编。 arXiv 原文