Time Lens: Event-based Video Frame Interpolation

作者：Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, Davide Scaramuzza

期刊：CVPR

年份：2021

1.研究问题

将事件相机引入到视频帧插值中，通过在帧之间的盲时提供辅助视觉信息，结合扭曲和插值的优势来解决帧插值方法在高动态场景下出现误差的情况。

2.想法动机

VFI 通过将中等帧率视频转换成高帧率视频来解决普通用户不易捕获高速瞬间的问题，基于帧的插值方法完全依赖于传统的基于帧的相机的输入以同步和固定的速率记录帧。目前有几种主流方法，如：

- 基于扭曲的方法，将光流估计与图像扭曲相结合，在帧间线性运动和亮度恒定的假设下，计算光流并将输入关键帧扭曲到目标帧，同时利用上下文信息、可见性映射等来改进结果，这些方法大多假设线性运动，也可以处理非线性运动但仍然受到顺序的限制，无法捕获任意运动。
- 基于核的方法，将 VFI 建模为输入关键帧的局部卷积，该方法对运动模糊和光线变化更健壮，但由于卷积核的局部性·，在实践中不能扩展到大型运动。
- 为了克服以上限制，一些作品试图将基于帧的相机的输入与不同

的时空结合，但是这付出了高昂的成本。由此便采用事件相机。

3.该方法的优点

- 引入了名为 Time Lens 的 CNN 框架，结合了扭曲和基于合成的方法。
- 一种新的基于扭曲的插值方法，从事件而不是帧中估计运动，对运动模糊更鲁棒可以估计帧之间的非线性运动。

4.具体方法

假设一个基于事件的 VFI 设置

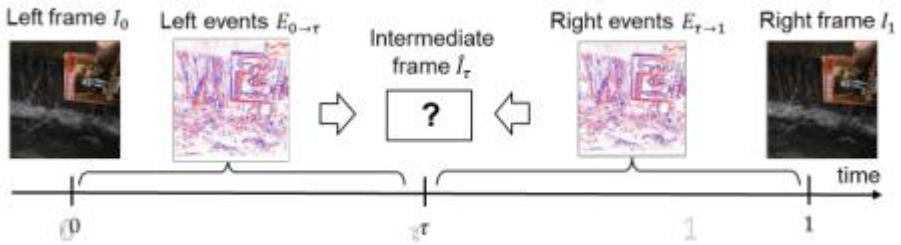
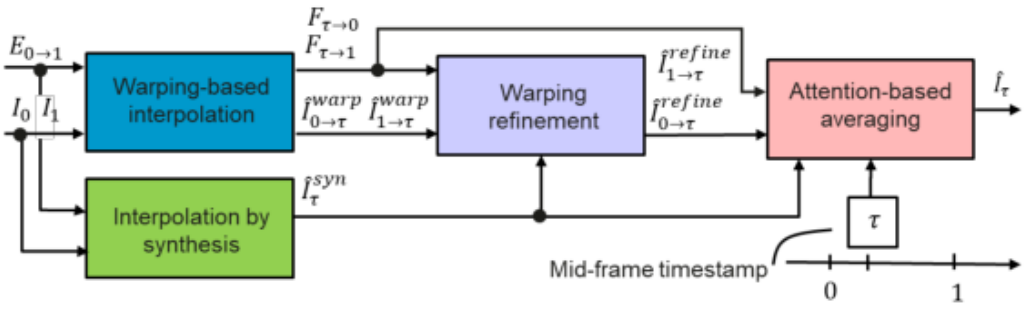


Figure 2: Proposed event-based VFI approach.

- 将左侧 I_0 和右侧 I_1 RGB 关键帧，以及左侧 $E_{0\to\tau}$ 和右侧 $E_{\tau\to 1}$ 事件序列作为输入，目标是在关键帧之间以随机时间步长 τ 插入(一个或多个)新帧 $\hat{I}_\tau$。

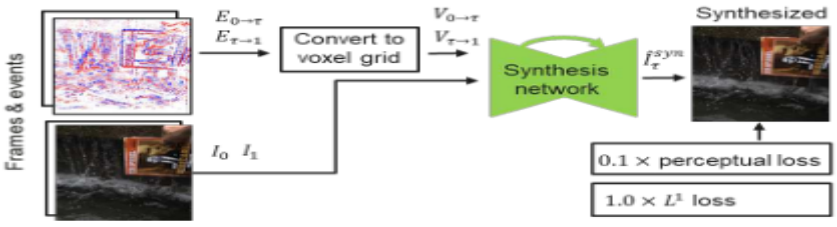
- Time Lens 学习框架



(a) Overview of the proposed method.

- 由四个专用模块组成：（1）基于扭曲的插值模块，利用从各自事件序列估计的光流对边界 RGB 关键帧扭曲来估计新帧；（2）扭曲细化模块，通过计算剩余流量改进该估计；（3）合成插值模块，通过直接融合边界关键帧和事件序列的输入信息估计新帧；（4）基于注意力的平均模块，将基于扭曲和基于合成的结果进行最佳组合。
- 合成插值模块

  给定左 I_0 和右 I_1RGB 关键帧和时间序列 $E_{0\to\tau}$ 和 $E_{\tau\to1}$，通过合成插值直接回归到一个新的帧 $\hat{I}^{syn}$。如图 b。

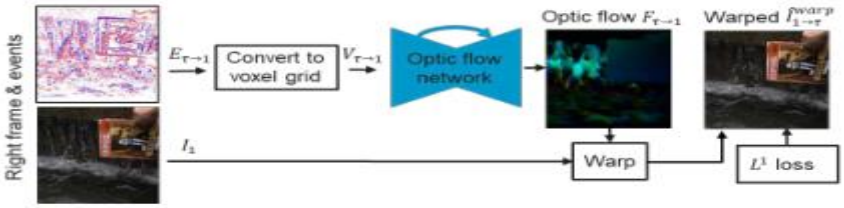


(b) Interpolation by synthesis module.

- 基于扭曲的插值模块

  首先分别使用事件 $E_{\tau\to0}$ 和 $E_{\tau\to1}$ 估计潜在新帧 $\hat{I}_\tau$ 和边界关键帧 I0 和 I1 之间的光流 Fτ→0 和 Fτ→1。我们通过反转事件

序列 $E_{0\to\tau}$ 来计算 $E_{\tau\to0}$，如图 4 所示。然后，计算光流使用可微插值在时间步长 τ 中扭曲边界关键帧，从而产生两个新的帧估计，分别是: $\hat{I}^{warp}_{0\to\tau}$ 和 $\hat{I}^{warp}_{1\to\tau}$。如图 d。



(d) Warping-based interpolation module.

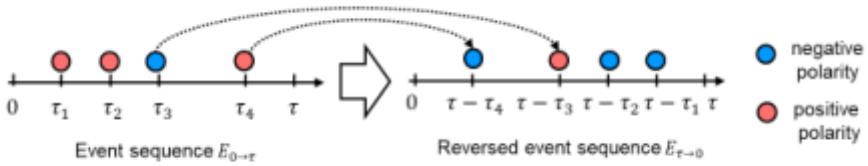
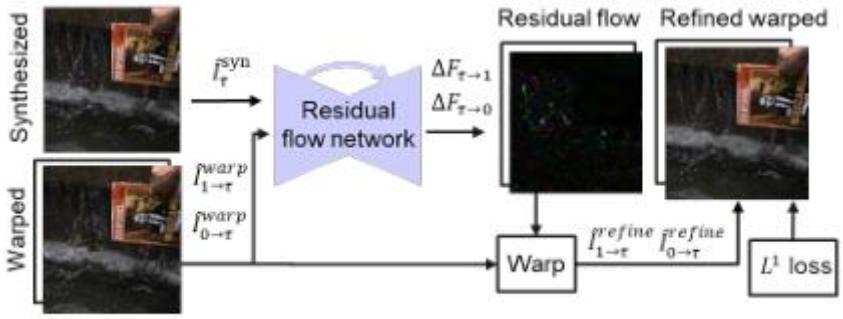


Figure 4: Example of an event sequence reversal.

在这里值得注意的是，与传统的基于扭曲的插值方法**不同的是**，该方法从事件中计算光流，可以自然地处理模糊的非线性运动。

- 扭曲细化模块

  通过估计基于 warp 的插值结果($\hat{I}^{warp}_{0\to\tau}$ 和 $\hat{I}^{warp}_{1\to\tau}$)与合成结果 $\hat{I}^{syn}_{\tau}$ 之间的残馀光流 $\Delta F_{\tau\to0}$ 和 $\Delta F_{\tau\to1}$，计算出细化插值帧，即 $\hat{I}^{refine}_{0\to\tau}$ 和 $\hat{I}^{refine}_{1\to\tau}$。然后，利用估计的剩余光流第二次对 $\hat{I}^{warp}_{0\to\tau}$ 和 $\hat{I}^{warp}_{1\to\tau}$进行扭曲，如图 e。



(e) Warping refinement module.

- 注意力平均模块

  以逐像素的方式将合成结果 $\hat{I}_\tau^{\text{syn}}$ 和基于扭曲的插值结果 $\hat{I}_{0\to\tau}^{\text{refine}}$ 和 $\hat{I}_{1\to\tau}^{\text{refine}}$ 混合在一起，得到最终的插值结果 $\hat{I}_\tau$。该模块利用基于扭曲和基于合成的插值方法的互补性，在结果上优于现在主流方法。

由于缺乏结合同步、高分辨率事件相机和标准 RGB 相机的可用数据集，文中构建了一个硬件同步混合传感器，由此来记录一个新的大规模数据集，即高速事件和 RGB 数据集。如图 5。

Figure 5: Illustration of the dual camera setup. It comprises a Prophesee Gen4 720p monochrome event camera (top) and a FLIR BlackFly S RGB camera (bottom). Both cameras are hardware synchronized with a baseline of 2.5 cm

这里不对此做详细介绍，如有兴趣可参考原文。

5.实验对比

本实验均使用 PyTorch 框架完成使用视频到事件方法从 Vimeo90k 七元数据集生成的具有合成事件的大型数据集。

- 通过逐个添加和训练模块来训练网络，同时冻结所有先前训练模块的权重。按照以下顺序训练模块:基于合成的插值、基于翘曲的插值、翘曲细化和注意力平均模块。之所以采用这种训练方法，是因为从头开始的端到端训练不收敛，预训练后对整个网络的微调只能略微提高结果。
- 为了测量插值图像的质量，使用结构相似度(SSIM)和峰值信噪比(PSNR)指标。

– 消融实验

在图 a 中考察了每个模块对最终插值的贡献，结果如表一。

Table 1: Quality of interpolation after each module on Vimeo90k (denoising) validation set. For SSIM and PSNR we show mean and one standard deviation. The best result is highlighted.

Module	PSNR	SSIM
Warping interpolation	26.68±3.68	0.926±0.041
Interpolation by synthesis	34.10±3.98	0.964±0.029
Warping refinement	33.02±3.76	0.963±0.026
Attention averaging (ours)	**35.83±3.70**	**0.976±0.019**

由表看出

- 在简单的扭曲块之后添加一个扭曲细化块，可以显著改善插值结果。
- 通过基于注意力平均合成和基于扭曲的结果，内插的 PSNR

提高了 1.7 dB。这是因为注意力平均模块结合了这两种方法的优点。

- 在 Vimeo90k(插值)及 Middlebury 数据集上将提出的方法与基于帧的插值方法 DAIN、RRIN、BMBC、SuperSloMo、基于事件的视频重建方法 E2VID 以及两种基于事件和帧的方法 EDI 和 LEDVDI 进行了比较。在评估过程中，取原始视频序列，分别跳过 1 帧或 3 帧，使用插值方法重建它们，并与地面真实跳过的帧进行比较。结果如表二。

Table 2: Results on standard video interpolation benchmarks such as *Middlebury* [2], *Vimeo90k* (interpolation) [43] and *GoPro* [19]. In all cases, we use a test subset of the datasets. To compute SSIM and PSNR, we downsample the original video and reconstruct the skipped frames. For Middlebury and Vimeo90k (interpolation), we skip 1 and 3 frames, and for GoPro we skip 7 and 15 frames due its its high frame rate of 240 FPS. *Uses frames* and *Uses events* indicate if a method uses frames and events for interpolation. For event-based methods we generate events from the skipped frames using the event simulator [6]. *Color* indicates if a method works with color frames. For SSIM and PSNR we show mean and one standard deviation. Note, that we can not produce results with 3 skips on the Vimeo90k dataset, since it consists of frame triplet. We show the best result in each column in bold and the second-best using underscore text.

Method	**Uses frames**	**Uses events**	**Color**	**PSNR**	**SSIM**	**PSNR**	**SSIM**
Middlebury [2]				**1 frame skip**		**3 frames skips**	
DAIN [3]	✔	✘	✔	30.87±5.38	0.899±0.110	26.67±4.53	0.838±0.130
SuperSloMo [10]	✔	✘	✔	29.75±5.35	0.880±0.112	26.43±5.30	0.823±0.141
RRIN [13]	✔	✘	✔	31.08±5.55	0.896±0.112	27.18±5.57	0.837±0.142
BMBC [28]	✔	✘	✔	30.83±6.01	0.897±0.111	26.86±5.82	0.834±0.144
E2VID [31]	✘	✔	✘	11.26±2.82	0.427±0.184	26.86±5.82	0.834±0.144
EDI [25]	✔	✔	✘	19.72±2.95	0.725±0.155	18.44±2.52	0.669±0.173
Time Lens (ours)	✔	✔	✔	**33.27±3.11**	**0.929±0.027**	**32.13±2.81**	**0.908±0.039**
Vimeo90k (interpolation) [43]				**1 frame skip**		**3 frames skips**	
DAIN [3]	✔	✘	✔	34.20±4.43	0.962±0.023	-	-
SuperSloMo [10]	✔	✘	✔	32.93±4.23	0.948±0.035	-	-
RRIN [13]	✔	✘	✔	34.72±4.40	0.962±0.029	-	-
BMBC [28]	✔	✘	✔	34.56±4.40	0.962±0.024	-	-
E2VID [31]	✘	✔	✘	10.08±2.89	0.395±0.141	-	-
EDI [25]	✔	✔	✘	20.74±3.31	0.748±0.140	-	-
Time Lens (ours)	✔	✔	✔	**36.31±3.11**	**0.962±0.024**	-	-
GoPro [19]				**7 frames skip**		**15 frames skips**	
DAIN [3]	✔	✘	✔	28.81±4.20	0.876±0.117	24.39±4.69	0.736±0.173
SuperSloMo [10]	✔	✘	✔	28.98±4.30	0.875±0.118	24.38±4.78	0.747±0.177
RRIN [13]	✔	✘	✔	28.96±4.38	0.876±0.119	24.32±4.80	0.749±0.175
BMBC [28]	✔	✘	✔	29.08±4.58	0.875±0.120	23.68±4.69	0.736±0.174
E2VID [31]	✘	✔	✘	9.74±2.11	0.549±0.094	9.75±2.11	0.549±0.094
EDI [25]	✔	✔	✘	18.79±2.03	0.670±0.144	17.45±2.23	0.603±0.149
Time Lens (ours)	✔	✔	✔	**34.81±1.63**	**0.959±0.012**	**33.21±2.00**	**0.942±0.023**

- 在数据集上的平均 PSNR(高达 8.82 dB 的改进)和 SSIM 分数(高达 0.192 的改进)方面优于其他方法。
- 改进源于在预测阶段使用辅助事件，使得能够执行精确的帧插值。

- 此外，当跳过并尝试重建更多帧时，所提出方法的 PSNR 和 SSIM 分数的下降程度远低于基于帧的方法的分数(高达 1.6 dB vs 高达 5.4 dB)。这表明该方法比基于帧的方法对非线性运动的鲁棒性更好。
  - 还在使用 DA VIS240 事件相机收集的高质量帧(HQF)数据集上评估了该方法，该数据集由无模糊和饱和度的视频序列组成。在这里使用了真实事件。结果如表三。

Table 3: Benchmarking on the High Quality Frames (HQF) DAVIS240 dataset. We do not fine-tune our method and other methods and use models provided by the authors. We evaluate methods on all sequences of the dataset. To compute SSIM and PSNR, we downsample the original video by skip 1 and 3 frames, reconstruct these frames and compare them to the skipped frames. In *Uses frames* and *Uses events* columns we specify if a method uses frames and events for interpolation. In the *Color* column, we indicate if a method works with color frames. In the table, we present two versions of our method: *Time Lens-syn*, which we trained only on synthetic data, and *Time Lens-real*, which we trained on synthetic data and fine-tuned on real event data from our own DAVIS346 camera. For SSIM and PSNR, we show mean and one standard deviation. We show the best result in each column in bold and the second-best using underscore text.

Method	Uses frames	Uses events	Color	PSNR	SSIM	PSNR	SSIM
				1 frame skip		3 frames skips	
DAIN [3]	✔	✘	✔	29.82±6.91	0.875±0.124	26.10±7.52	0.782±0.185
SuperSloMo [10]	✔	✘	✔	28.76±6.13	0.861±0.132	25.54±7.13	0.761±0.204
RRIN [13]	✔	✘	✔	29.76±7.15	0.874±0.132	26.11±7.84	0.778±0.200
BMBC [28]	✔	✘	✔	29.96±7.00	0.875±0.126	26.32±7.78	0.781±0.193
E2VID [31]	✘	✔	✘	6.70±2.19	0.315±0.124	6.70±2.20	0.315±0.124
EDI [25]	✔	✔	✘	18.7±6.53	0.574±0.244	18.8±6.88	0.579±0.274
Time Lens-syn (our)	✔	✔	✔	30.57±5.01	0.903±0.067	28.98±5.09	0.873±0.086
Time Lens-real (ours)	✔	✔	✔	**32.49±4.60**	**0.927±0.048**	**30.57±5.08**	**0.900±0.069**

- 结果与合成数据集上的结果一致，可以得出所提出的方法优于基于帧的方法，且随着跳过帧数的增加，与其他方法性能差距也逐渐拉大。

6. 总结

1. 提出了一种新的视频插值方法 Time Lens。

2. 为了解决现有方法在非线性运动及高速运动下出现误差，结合了扭曲和插值的优势，可以处理不断变化的照明条件

和非线性运动，对运动模糊非线性运动更鲁棒。

7.局限和启示

1.利用了自己构建的硬件传感器，需要收集大量数据，成本较高。

2.在各种真实事件类型上的性能表现是否一定更好不能确定。

启示：可尝试在自监督情况下让模型大量并自主进行真实事件的训练。