Time Lens: Event-based Video Frame Interpolation

**研究问题：**Time Lens，利用事件相机和帧插值两者优势的新方法，研究该方法在高度动态场景中的适用性，例如：(1)输入关键帧之间的非线性运动，(2)当照明或运动模糊发生变化时。(3)关键帧之间场景中出现的非刚性运动和新物体。

**想法动机：**最先进的帧插值方法通过从连续的关键帧推断图像中的物体运动来生成中间帧。在没有额外信息的情况下，必须使用一阶近似，即光流，但这种选择限制了可以建模的运动类型，导致在高动态场景中出现误差。事件相机是一种新颖的传感器，通过在帧之间的盲时提供辅助视觉信息来解决这一限制，在高动态场景下发布了一个新的大规模数据集，旨在突破现有方法的限制。

**具体方法：**

1.先假设一个基于事件的 VFI 设置，将左侧 I0 和右侧 I1 关键帧，以及左侧$E_{0\to\tau}$和右侧 $E_{\tau\to1}$时间序列作为输入，目标是在关键帧之间以随机时间 T 插入（一个或多个）新帧$\hat{I}_\tau$ 如图二。

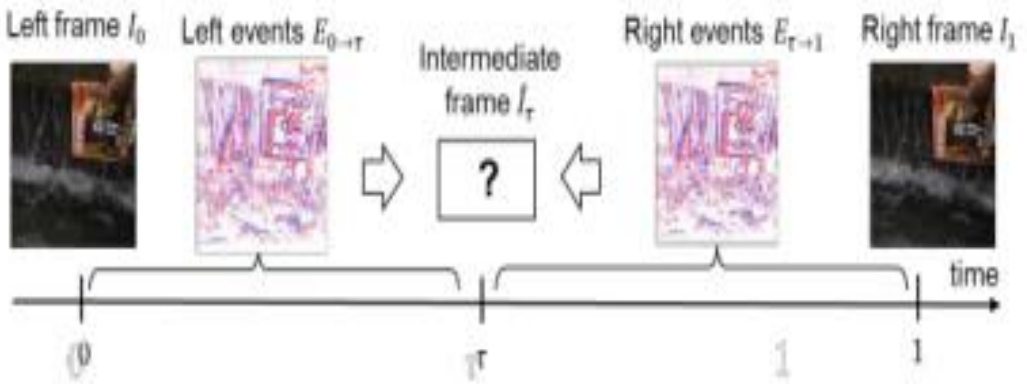



Figure 2: Proposed event-based VFI approach.

提出了一个基于学习的框架，即时间透镜，它由四个专用模块组成，服务于互补的插值方案，即基于扭曲和基于合成的插值。具体而言，(1)基于扭曲的插值模块通过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来改进该估计;(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;最后(4)基于注意力的平均模块旨在将基于扭曲和基于合成的结果进行最佳组合。在这样做的时候，时间镜头结合了扭曲和基于合成的插值技术的优势，使我们能够在处理非线性运动、光线变化和运动模糊的同时，生成具有颜色和高纹理细节的新帧。方法的工作流程如图 3a 所示。

如图 3b 所示，在给定左 I0 和右 I1 RGB 关键帧和事件序列 E0→ τ 和 E τ →1 的情况下，通过合成插值直接回归到一个新的帧 $\hat{I}^{\text{syn}}$ 。该插值方案的优点在于能够处理光照的变化，如图 6 中的水反射和场景中突然出现的新物体，因为与基于扭曲的方法不同，它不依赖于亮度恒定假设。它的主要缺点是当事件信息有噪声或对比度阈值过高时，会导致图像边缘和纹理失真，如图 6 所示。

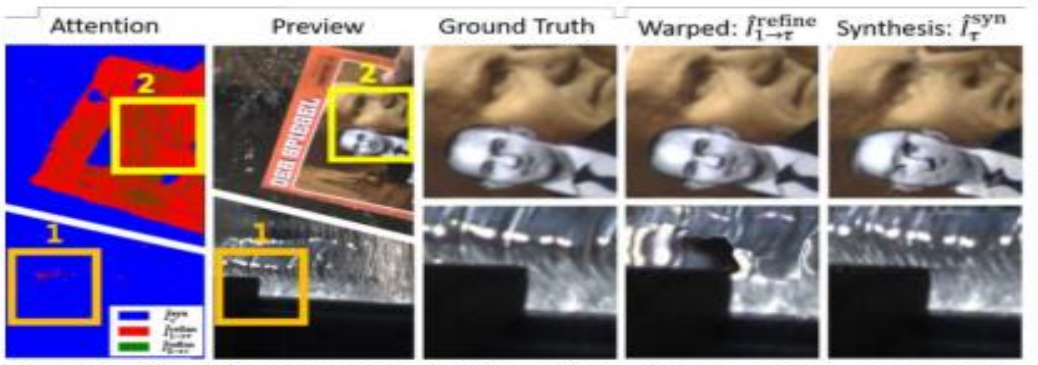
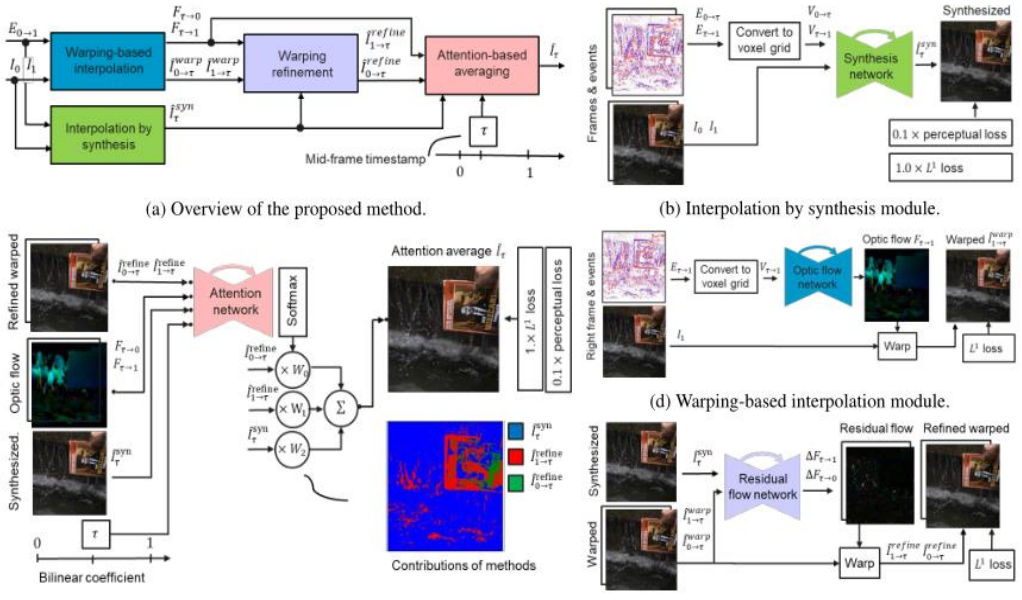



Figure 6: Complementarity of warping- and synthesis-based interpolation.

如图 3d 所示，首先分别使用事件 E τ →0 和 E τ →1 估计潜在新帧$\hat{I}_\tau$和边界关键帧 I0 和 I1 之间的光流 F τ →0 和 F τ →1。通过反转事件序列 E0→ τ 来计算 E τ →0，如图 4 所示。然后，使用计算光流使用可微插值在时间步长 T 中扭曲边界关键帧，从而产生两个新的帧估计，分别是:$\hat{I}^{\text{warp}}_{0\to\tau}$和$\hat{I}^{\text{warp}}_{1\to\tau}$。

扭曲精化模块通过估计基于 warp 的插值结果($\hat{I}^{\text{warp}}_{0\to\tau}$ 和$\hat{I}^{\text{warp}}_{1\to\tau}$)与合成结果$\hat{I}^{\text{syn}}_{\tau}$之间的残馀光流ΔF τ →0 和ΔF τ →1，计算出精化插值帧，即$\hat{I}^{\text{refine}}_{0\to\tau}$和$\hat{I}^{\text{refine}}_{1\to\tau}$。然后，利用估计的剩余光流第二次对$\hat{I}^{\text{warp}}_{0\to\tau}$和$\hat{I}^{\text{warp}}_{1\to\tau}$进行扭曲，如图 3e 所示。

最后，如图 3c 所示的注意力平均模块以逐像素的方式将合成结果$\hat{I}^{\text{syn}}_{\tau}$和基于扭曲的插值结果$\hat{I}^{\text{refine}}_{0\to\tau}$和$\hat{I}^{\text{refine}}_{1\to\tau}$混合在一起，得到最终的插值结果$\hat{I}_\tau$。




(a) Overview of the proposed method.

(b) Interpolation by synthesis module.

(c) Attention-based averaging module.

(d) Warping-based interpolation module.

(e) Warping refinement module.

Figure 3: Structure of the proposed method. The overall workflow of the method is shown in Fig. 3a and individual modules are shown in Fig. 3d, 3b, 3e and 3c. In the figures we also show loss function that we use to train each module. We show similar modules in the same color across the figures.

构建了一个硬件同步混合传感器，该传感器将高分辨率事件相机与高分辨率高速彩色相机相结合。使用这种混合传感器来记录一个新的大规模数据集用于验证视频插值方法。混合摄像机设置如图五。


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

**实现策略：**所有实验都使用 PyTorch 框架完成。对于训练，使用 Adam 优化器和标准设置，批次大小为 4，学习率为 104，每 12 个 epoch 我们减少 10 倍。我们训练每个模块 27 个 epoch。对于训练，我们使用基于的事件模拟器，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、基于扭曲的插值、扭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。我们用感知和 L1 损失来监督网络，如图 3b、3d、3e 和 3c 所示。按照训练的顺序一个模块一个模块地对网络进行微调。为了测量插值图像的质量，使用了结构相似度(SSIM)和峰值信噪比(PSNR)指标。

**实验对比：**进行了消融实验探索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** |

将提出的方法(Time Lens)与四种基于帧的插值方法 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 在流行的视频插值基准数据集(如 Vimeo90k(插值)、Middlebury)上进行了比较。结果如图二。所提出的方法在数据集的平均 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。并且，当跳过并且重建更多帧时，所提出方法的 PSNR 和 SSIM 的下降程度远远低于基于帧的方法，这表明对非线性运动的鲁棒性更好。

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** |

另外还在高质量帧数据集上评估了该方法。在评估中，考虑了方法的两个版本:Time Lens-syn，只在合成数据上进行训练，Time Lens-real，在合成数据上进行训练，并对来自 DAVIS346 相机的真实事件数据进行微调。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** |

另外在高速事件 RGB 数据集上可以看出所提出的方法再次优于基于帧和基于帧+事件的竞争对手。结果如表四。并且在图 7 中，展示了来自 HS-ERGB 测试集的几个例子，这些例子表明，与竞争对手基于帧的方法相比，我们的方法可以在非线性(“Umbrella”序列)和非刚性运动(“Water Bomb”)的情况下插值帧，并且还可以处理光照变化。

Table 4: Benchmarking on the test set of the High Speed Event and RGB camera (HS-ERGB) dataset. We report PSNR and SSIM for all sequences by skipping 5 and 7 frames respectively, and reconstructing the missing frames with each method. By design LEDVDI [15] can interpolate only 5 frames. *Uses frames* and *Uses events* indicate if a method uses frames or events respectively. *Color* indicates whether a method works with color frames. For SSIM and PSNR the scores are averaged over the sequences. Best results are shown in bold and the second best are underlined.

| Method | Uses frames | Uses events | Color | PSNR | SSIM | PSNR | SSIM |
|---|---|---|---|---|---|---|---|
| **Far-away sequences** | | | | **5 frame skip** | | **7 frames skips** | |
| DAIN [3] | ✔ | ✘ | ✔ | 27.92±1.55 | 0.780±0.141 | 27.13±1.75 | 0.748±0.151 |
| SuperSloMo [10] | ✔ | ✘ | ✔ | 25.66±6.24 | 0.727±0.221 | 24.16±5.20 | 0.692±0.199 |
| RRIN [13] | ✔ | ✘ | ✔ | 25.26±5.81 | 0.738±0.196 | 23.73±4.74 | 0.703±0.170 |
| BMBC [28] | ✔ | ✘ | ✔ | 25.62±6.13 | 0.742±0.202 | 24.13±4.99 | 0.710±0.175 |
| LEDVDI [15] | ✔ | ✔ | ✘ | 12.50±1.74 | 0.393±0.174 | n/a | n/a |
| **Time Lens (ours)** | ✔ | ✔ | ✔ | **33.13±2.10** | **0.877±0.092** | **32.31±2.27** | **0.869±0.110** |
| **Close planar sequences** | | | | **5 frame skip** | | **7 frames skips** | |
| DAIN [3] | ✔ | ✘ | ✔ | 29.03±4.47 | 0.807±0.093 | 28.50±4.54 | 0.801 ± 0.096 |
| SuperSloMo [10] | ✔ | ✘ | ✔ | 28.35±4.26 | 0.788±0.098 | 27.27±4.26 | 0.775 ± 0.099 |
| RRIN [13] | ✔ | ✘ | ✔ | 28.69±4.17 | 0.813±0.083 | 27.46±4.24 | 0.800±0.084 |
| BMBC [28] | ✔ | ✘ | ✔ | 29.22±4.45 | 0.820±0.085 | 27.99±4.55 | 0.808±0.084 |
| LEDVDI [15] | ✔ | ✔ | ✘ | 19.46±4.09 | 0.602±0.164 | n/a | n/a |
| **Time Lens (ours)** | ✔ | ✔ | ✔ | **32.19±4.19** | **0.839±0.090** | **31.68±4.18** | **0.835±0.091** |

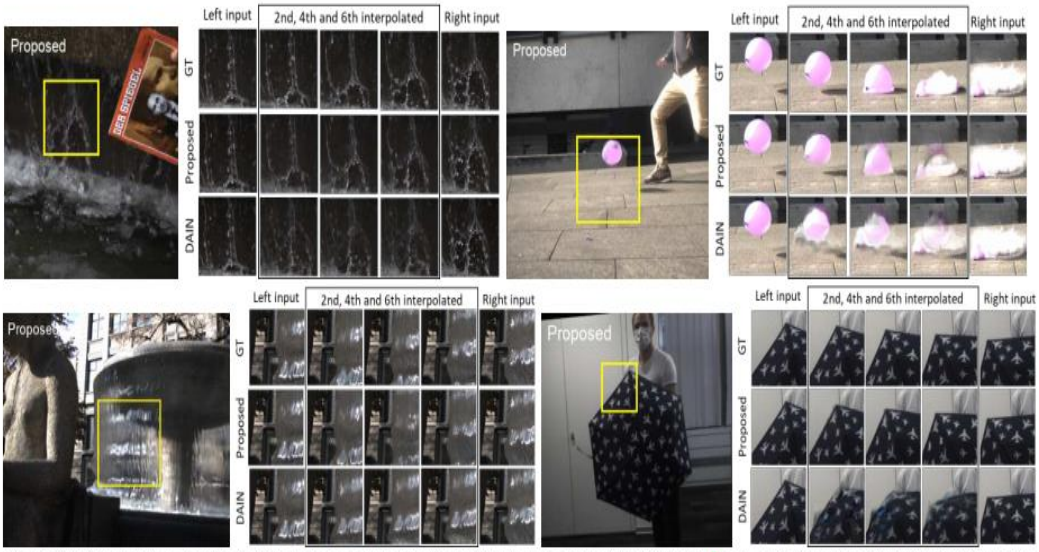



Figure 7: Qualitative results for the proposed method and its closes competitor DAIN [3] on our Dual Event and Color Camera Dataset test sequences: "Fountain Schaffhauserplatz" (top-left), "Fountain Bellevue" (bottom-left) "Water bomb" (top-right) and "Umbrella" (bottom-right). For each sequence, the figure shows interpolation results on the left (the animation can be viewed in Acrobat Reader) and close-up interpolation results on the right. The close-ups, show input left and right frame and intermediate interpolated frames.

**小结：**该方法将事件相机引入视频插值，设计了光学设备收集大量高速帧和事件的配对训练数据，在各方面表现优于先前基于帧及基于事件的方法。缺点在于设备过于昂贵而无法扩展。