Unsupervised Framerate Upsampling from Events

Hiroyuki Okuno *Nagoya University*
Chihiro Tsutake *Nagoya University*
Keita Takahashi *Nagoya University*
Toshiaki Fujii *Nagoya University*

***Abstract*—We propose a method for framerate upsampling from events. We take an unsupervised approach that does not require ground-truth high-framerate videos for pre-training. We also report some promising experimental results.**

***Index Terms*—Framerate upsampling, Event**

I. Introduction

An event camera [1] adopts a bio-inspired sensing mechanism that can record the luminance changes over time. The recorded information, called events, are detected asynchronously at each pixel in the order of microseconds. Events are utilized for framerate upsampling of a video, because the information between the low-framerate video frames (key-frames) can be supplemented from the events. As a seminal work, TimeLens [2] achieved high quality results under the framework of supervised learning; the algorithm was pre-trained on a collection of high-framerate videos as the ground-truth supervisory signal. However, this *generalized* approach may not always optimal for a *specific target scene*. In contrast, we take an unsupervised approach without ground-truth high-framerate videos for pre-training. Our method is implemented as a multi-layer perceptron (MLP) trained *only* on the given key-frames and events of a *specific target scene*. We report some promising experimental results of our method.

II. Proposed method

Let t and (x, y) denote the time and pixel. All we have as input are T key-frames $J_n(x, y)$ $(n \in \{0, 1, ..., T-1\})$ and events $E^{t \to t+\Delta t}(x, y)$ of a target scene. Our goal is to reconstruct video frames $I^t(x, y)$ at arbitrary t.

As shown in Fig. 1, we construct an MLP that takes a coordinate (x, y, t) as input and produces the corresponding pixel value $I_n^t(x, y)$ using a key frame $J_n(x, y)$. We train the MLP so that $I_n^t(x, y)$ are consistent with the given key frames and events. The loss function L is defined as

$$L = L_{\mathrm{MSE}} + \lambda L_{\mathrm{Event}} \tag{1}$$

$$\begin{aligned} L_{\mathrm{MSE}} = \big\{ &\mathrm{D}[I_{n+1}^n(x, y), J_n(x, y)] + \\ &\mathrm{D}[I_{n-1}^n(x, y), J_n(x, y)] + \\ &\mathrm{D}[I_{\lfloor t \rfloor}^t(x, y), I_{\lceil t \rceil}^t(x, y)] \big\} / 3 \end{aligned} \tag{2}$$

$$\begin{aligned} L_{\mathrm{Event}} = \big\{ &\mathrm{D}[\nabla_t I_{\lfloor t \rfloor}^t(x, y), c \cdot E^{t \to t+\Delta t}(x, y)] + \\ &\mathrm{D}[\nabla_t I_{\lceil t \rceil}^t(x, y), c \cdot E^{t \to t+\Delta t}(x, y)] \big\} / 2 \end{aligned} \tag{3}$$

$$\nabla_t I_n^t(x, y) = \log(I_n^{t+\Delta t}(x, y)) - \log(I_n^t(x, y)) \tag{4}$$

where $\lambda > 0$ is a weight and $\mathrm{D}[\alpha, \beta]$ means the average of $|\alpha - \beta|^2$. Equation (2) enforces the consistency between $I_n^t(x, y)$ and given key-frames. Equation (3) enforces the expected relation between $I_n^t(x, y)$ and given events, where c is the temporal contrast threshold of the event sensor. Once the training is completed, $I^t(x, y)$ at arbitrary t can be generated via forward inference on the MLP.

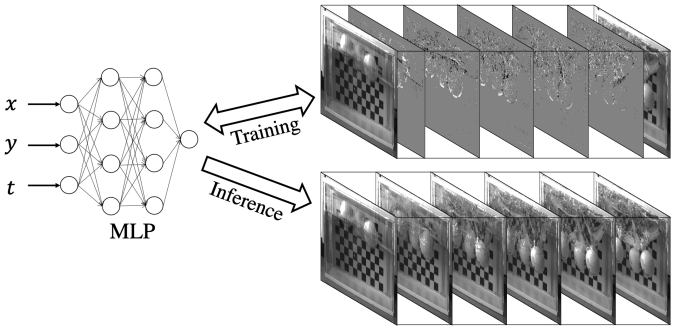

Fig. 1: Overview of our method.

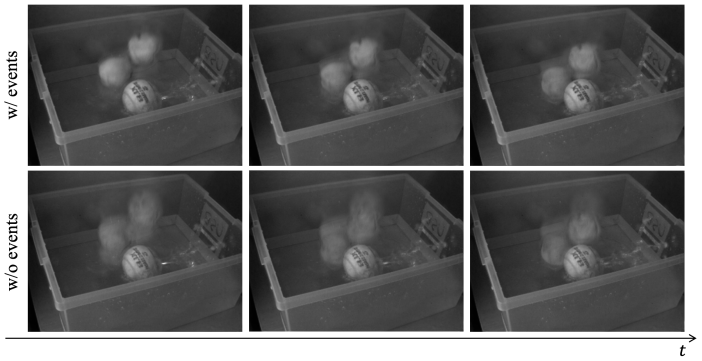

Fig. 2: Visual results of framerate upsampling

III. Results and discussion

We used a DAVIS346 camera that can capture frames and events simultaneously. The recorded videos (40 fps) were upsampled to 400 fps by using our method. Some visual results are presented in Fig. 2. Our method produced visually-plausible results (the contours of fast-moving objects are clearly seen), while disregarding the events (setting $c = 0$) led to poor results.

References

[1] Guillermo Gallego, et al.: "Event-based Vision: A Survey," PAMI, Vol. 44, No. 01, pp. 154-180, 2022.
[2] Stepan Tulyakov, et al.: "Time Lens: Event-based video frame interpolation," CVPR, 2021.