E-VFIA : Event-Based Video Frame Interpolation with Attention

Onur Selim Kılıç, Ahmet Akman, A. Aydın Alatan

METU Center for Image Analysis (OGAM) Ankara, TURKEY

ODTÜ GÖRÜNTÜ ANALİZİ UYGULAMA VE ARAŞTIRMA MERKEZİ

ICRA LONDON · 2023

Abstract

Video frame interpolation (VFI) is a fundamental vision task that aims to synthesize several frames between two consecutive original video images. Most algorithms aim to accomplish VFI by using only keyframes, which is an ill-posed problem since the keyframes usually do not yield any accurate precision about the trajectories of the objects in the scene. On the other hand, **event-based cameras provide more precise information between the keyframes of a video.** Some recent state-of-the-art event-based methods approach this problem by utilizing event data for better optical flow estimation to interpolate for video frame by warping. Nonetheless, those methods heavily suffer from the **ghosting effect.** On the other hand, some of kernel-based VFI methods that only use frames as input, have shown that **deformable convolutions**, when backed up with **transformers**, can be a reliable way of **dealing with long-range dependencies.** We propose event-based video frame interpolation with attention (E-VFIA), as a lightweight kernel-based method. E-VFIA fuses event information with standard video frames by deformable convolutions to generate high quality interpolated frames. The proposed method **represents events with high temporal resolution and uses a multi-head self-attention mechanism to better encode event-based information,** while being less vulnerable to blurring and ghosting artifacts; thus, generating crispier frames. The simulation results show that the proposed technique **outperforms current state-of-the-art methods** (both frame and event-based) with a **significantly smaller model size.**

Main Contributions

- E-VFIA, **the first kernel-based algorithm** to utilize deformable convolutions to **fuse event-based information and standard images** for video frame interpolation with events.
- **Significant improvement** (up to $1.04dB$) against the state-of-the-art methods that use only key-frames and events together with key-frames.
- Model has approximately 2.07 **million parameters.**
- Using **voxel grids with higher temporal resolutions** improves performance.
- Utilization of **both temporal and spatial pooling** operations to **associate fast-moving objects** between consecutive images.

Proposed Method

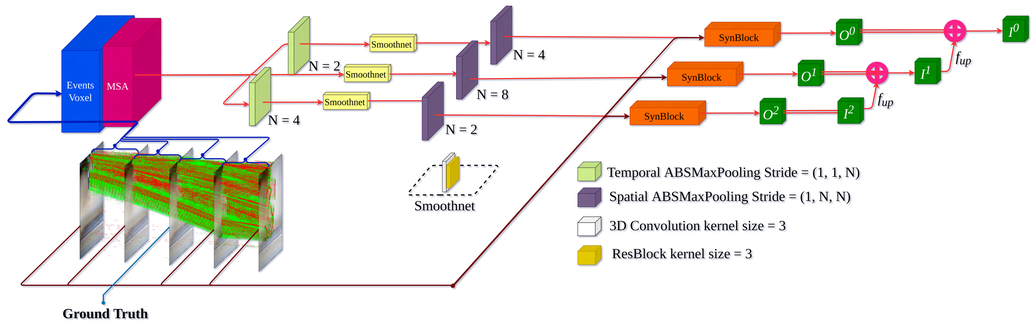


Figure 1. Overview of the proposed method.

Qualitative Comparisons

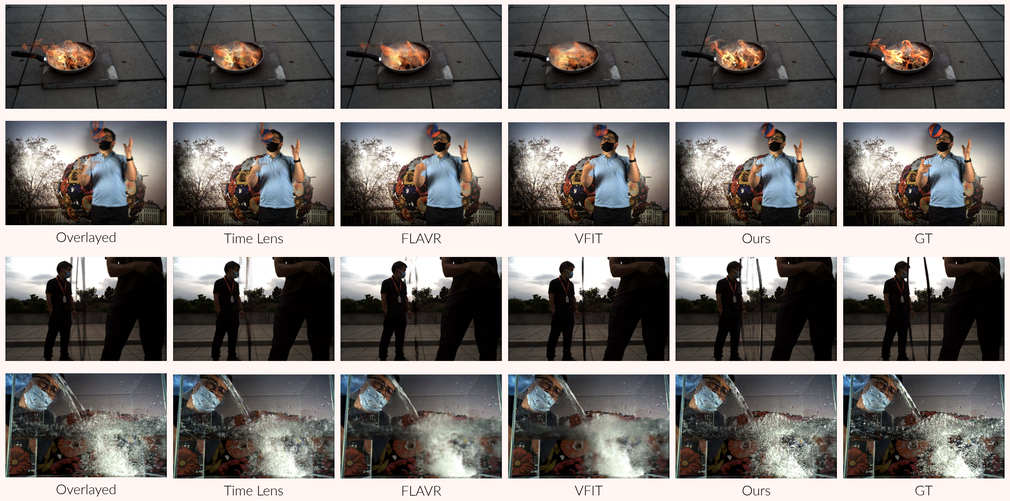
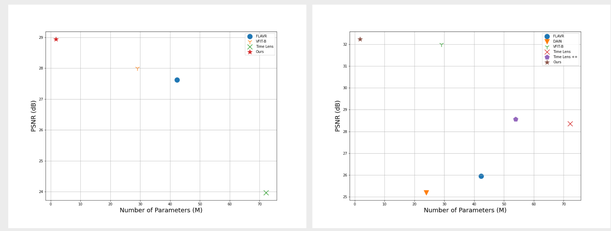


Figure 2. Qualitative comparisons against the state-of-the-art video interpolation algorithm. Our method is **less prone to blur and ghosting effects.**

(a) Time Lens (b) FLAVRFV (c) VFIT (d) Ours

Figure 3. Pixel-wise differences of the fire sample.

Quantitaive comparisons



(a) Full Scale (b) 256x256

Figure 4. Model Sizes vs PSNR

Table 1. Comparison of our method in BS-ERGB in **low resolution**

Method	Input	#Params (M)	PSNR (dB)	SSIM
FLAVRFV [1]	Frames	42.4	31.72	0.9469
VFIT [2]	Frames	29.0	32.08	0.9449
Timelens [3]	Frames Events	72.2	28.36	0.9320
Timelens++ [4]	Frames Events	53.9	28.56	-
Ours	**Frames Events**	**2.07**	**32.23**	**0.9581**

Table 2. Comparison of our method in BS-ERGB dataset in **full scale**

Method	Input	#Params (M)	PSNR (dB)	SSIM
FLAVRFV [1]	Frames	42.4	27.642	0.8729
VFIT [2]	Frames	29.0	28.00	0.8767
Timelens [3]	Frames Events	72.2	23.97	0.7838
Timelens++ [4]	Frames Events	53.9	-	-
Ours	**Frames Events**	**2.07**	**29.04**	**0.8771**

References

[1] T. Kalluri, D. Pathak, M. Chandraker, and D. Tran, "Flavr: Flow-agnostic video representations for fast frame interpolation," *ArXiv*, vol. abs/2012.08512, 2020.

[2] Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang, "Video frame interpolation transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 482–17 491.

[3] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 16 155–16 164.

[4] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. Scaramuzza, "Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17 755–17 764.