Revisiting Event-Based Video Frame Interpolation (IROS 2023)

AI3610 Brain-Inspired Intelligence Paper Sharing

Yi Ai, Xiangyuan Xue, Shengmin Yang

November 7, 2023

Group 5

Preliminaries & Motivation

Preliminaries

Event Camera

- faster response time and higher dynamic range
- pixel-level output without motion blur
- less power consumption

Problem Definition

- Video Frame Interpolation: Given consecutive video frames $I_0, I_1 \in \mathbb{R}^{W \times H \times 3}$, VFI aims to predict intermediate new frames $\hat{I}_t$ at time $t \in (0, 1)$.
- Event Representation: The events triggered in a given time interval form a sequence $\{e_i = (x_i, y_i, t_i, p_i)\}_{i \in [1,M]}$.

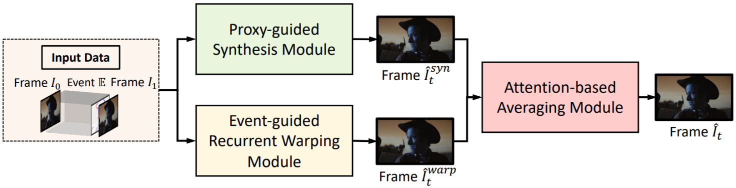
Motivation

- Few of the state-of-the-art methods for VFI fully respect the intrinsic characteristics of events streams.
- Estimating optical flow only with event cameras is difficult.
- Explore to incorporate RGB information in an event-guided optical flow refinement strategy.
- Propose a divide-and-conquer strategy in which event-based intermediate frame synthesis happens incrementally in multiple simplified stages.[1]

Method

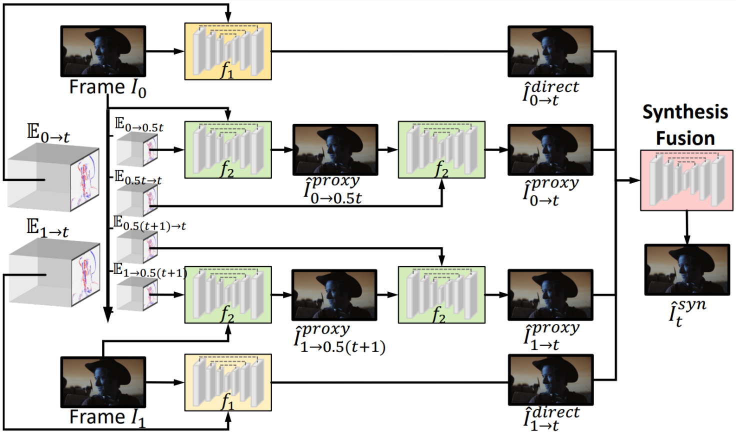
Method

Network Architecture



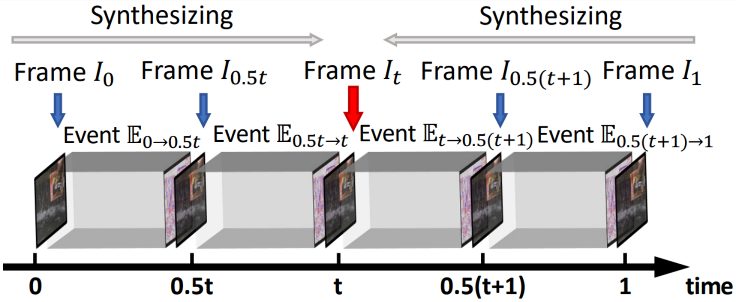
- Propose a VFI framework that relies on two complementary synthesis and warping modules.
- Composed of a proxy-guided synthesis module, an event-guided recurrent warping module, and an attention-based averaging module.

Proxy-Guided Synthesis Module



Method

Proxy-Guided Synthesis Module



- The sequence is divided into subsequences to incrementally generate multiple intermediate frames.

Method

Proxy-Guided Synthesis Module

- Direct Synthesis: Predict within a single step.
- Transitional Synthesis: Depart direct synthesis into multiple steps.

$$\begin{cases} \hat{I}^{\text{direct}}_{0\to t} = f_1(I_0, \mathbb{E}_{0\to t}) \\ \hat{I}^{\text{direct}}_{1\to t} = f_1(I_1, \mathbb{E}_{1\to t}) \end{cases} \quad \begin{cases} \hat{I}^{\text{proxy}}_{0\to \frac{1}{T}t} = f_2(I_0, \mathbb{E}_{0\to \frac{1}{T}t}) \\ \quad\vdots \\ \hat{I}^{\text{proxy}}_{1\to \frac{i}{T}t} = f_2(\hat{I}^{\text{proxy}}_{0\to \frac{i-1}{T}t}, \mathbb{E}_{\frac{i-1}{T}t\to \frac{i}{T}t}) \\ \quad\vdots \\ \hat{I}^{\text{proxy}}_{0\to t} = f_2(\hat{I}^{\text{proxy}}_{0\to \frac{T-1}{T}t}, \mathbb{E}_{\frac{T-1}{T}t\to t}) \end{cases}$$

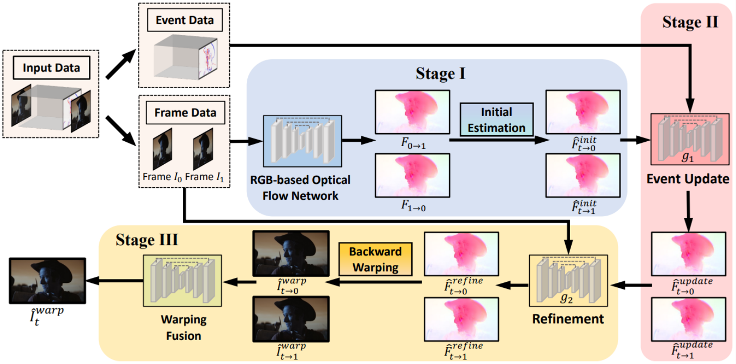
Method

Proxy-Guided Synthesis Module

- Synthesis Fusion: Fuse $(\hat{I}^{\text{direct}}_{0\to t}, \hat{I}^{\text{direct}}_{1\to t}, \hat{I}^{\text{proxy}}_{0\to t}, \hat{I}^{\text{proxy}}_{1\to t})$ through a neural network.
- Loss Function: Composed of two parts, where $L_{\text{perceptual}}$ applies pretrained VGG-16 model.[2]

$$
\begin{aligned}
L_{\text{reconstruct}} &= \|\hat{I}^{\text{syn}}_{t} - I_t\|_1 + \|\hat{I}^{\text{direct}}_{0\to t} - I_t\|_1 + \|\hat{I}^{\text{direct}}_{1\to t} - I_t\|_1 \\
&+ \|\hat{I}^{\text{proxy}}_{0\to t} - I_t\|_1 + \|\hat{I}^{\text{proxy}}_{1\to t} - I_t\|_1 \\
L_{\text{perceptual}} &= \|\psi(\hat{I}^{\text{syn}}_{t}) - \psi(I_t)\|_2^2 \\
L_{\text{synthesis}} &= L_{\text{reconstruct}} + \lambda \cdot L_{\text{perceptual}}
\end{aligned}
$$

Event-Guided Recurrent Warping Module



- Adopt a three-stage architecture to generate a warping-based intermediate video frame.

Method

Event-Guided Recurrent Warping Module

- Initial RGB-Based Estimation: Apply pretrained GMFlow network to estimate optical flow $F_{0\to1}$ and $F_{1\to0}$.[3]

$$\hat{F}^{\text{init}}_{t\to0} = -(1-t)tF_{0\to1} + t^2 F_{1\to0}$$
$$\hat{F}^{\text{init}}_{t\to1} = (1-t)^2 F_{0\to1} - t(1-t)F_{1\to0}$$

- Event-Based Update: Refine predicted optical flow through residual learning.

$$\Delta\hat{F}^{\text{event}}_{t\to0} = g_1(\hat{F}^{\text{init}}_{t\to0}, \mathbb{E}_{t\to0}) \quad \hat{F}^{\text{update}}_{t\to0} = \hat{F}^{\text{init}}_{t\to0} + \Delta\hat{F}^{\text{event}}_{t\to0}$$
$$\Delta\hat{F}^{\text{event}}_{t\to1} = g_1(\hat{F}^{\text{init}}_{t\to1}, \mathbb{E}_{t\to1}) \quad \hat{F}^{\text{update}}_{t\to1} = \hat{F}^{\text{init}}_{t\to1} + \Delta\hat{F}^{\text{event}}_{t\to1}$$

Event-Guided Recurrent Warping Module

- Refinement and Backward Warping: Refine refined optical flow with I_0 and I_1 through residual learning, then backward warp to generate $\hat{I}^{\text{warp}}_{t\to 0}$ and $\hat{I}^{\text{warp}}_{t\to 1}$.

$$\Delta\hat{F}_{t\to 0}, \Delta\hat{F}_{t\to 1} = g_2(\hat{F}^{\text{update}}_{t\to 0}, \hat{F}^{\text{update}}_{t\to 1}, I_0, I_1)$$
$$\hat{F}^{\text{refine}}_{t\to 0} = \hat{F}^{\text{update}}_{t\to 0} + \Delta\hat{F}_{t\to 0}$$
$$\hat{F}^{\text{refine}}_{t\to 1} = \hat{F}^{\text{update}}_{t\to 1} + \Delta\hat{F}_{t\to 1}$$

- Loss Function: Supervise results before and after fusion.

$$L_{\text{warping}} = \|\hat{I}^{\text{warp}}_t - I_t\|_1 + \|\hat{I}^{\text{warp}}_{t\to 0} - I_t\|_1 + \|\hat{I}^{\text{warp}}_{t\to 1} - I_t\|_1$$

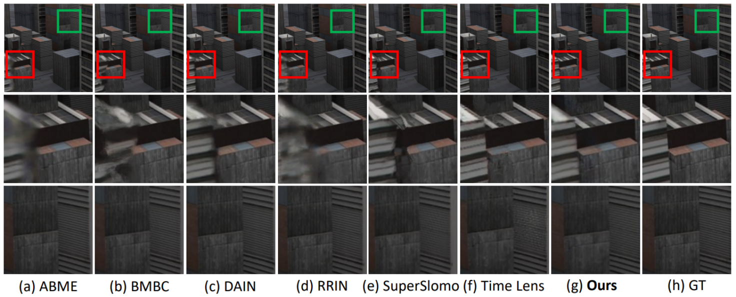
Attention-Based Averaging Module

- Given that the synthesis module predicts directly from events, it has defects along edges caused by noise in the events and insufficient sensitivity in low-texture regions.
- The warping module complements these defects.
- The averaging module learns weights to blend the results in a pixel-wise fashion and yield the final interpolated result $\hat{I}_t$ with reduced distortions and motion blur.

Experiments & Conclusion

Experiment

Result Comparsion



- The method yields interpolated frame without blur and noticeable artifacts in the red boxes, and preserves detailed information in the green boxes.

Experiment

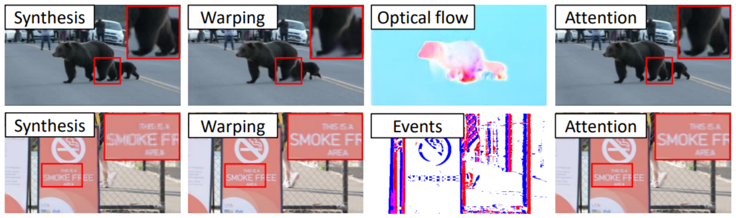
Result Comparsion

	(a) Synthetic Dataset					(b) Real-world Dataset			
	Triplet [34]	Septuplet [34]		Middlebury [33]		HQF [37]		HS-ERGB (close) [6]	
	x2	x2	x4	x2	x4	x2	x4	x6	x8
SuperSlomo [11]	33.44/0.951	33.96/0.943	29.44/0.888	29.68/0.876	26.42/0.819	28.76/0.861	25.54/0.761	28.35/0.788	27.27/0.755
RRIN [12]	34.68/0.962	35.56/0.954	29.41/0.891	31.17/0.894	27.28/0.841	29.76/0.874	26.11/0.778	28.70/0.813	27.44/0.800
BMBC [16]	35.09/0.963	35.48/0.949	30.33/0.897	30.71/0.889	26.45/0.821	30.74/0.875	27.01/0.781	29.32/0.821	27.89/0.808
ABME [17]	36.22/0.969	36.53/0.955	-	31.66/0.900	-	30.58/0.880	-	-	-
DAIN [21]	34.70/0.964	35.29/0.954	29.87/0.900	30.90/0.896	26.65/0.831	29.82/0.875	26.10/0.782	29.03/0.807	28.50/0.801
Time Lens [6]	36.31/0.962	36.87/0.960	35.58/0.949	33.27/0.929	32.13/0.908	30.57/0.903	28.98/0.873	32.19/0.839	31.68/0.835
Ours	36.56/0.965	38.14/0.968	36.34/0.960	32.51/0.909	31.01/0.886	31.75/0.910	28.56/0.850	33.21/0.847	32.95/0.844

- The method outperforms both frame-based and event-based state-of-the-art algorithms on the Vimeo90K dataset in terms of both single-frame interpolation and multi-frame interpolation by a significant margin.[4]
- This demonstrates the advantage of the auxiliary visual information introduced by event cameras and the superiority of synthesis and warping modules.

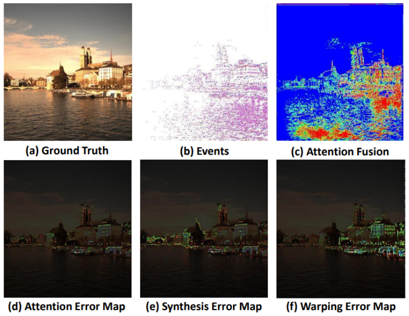
Experiment

Ablation Study



- The synthesis module generates blurry results due to event noise.
- However, the attention module again interpolates clear results by giving preference to the good predictions of the warping module.

Ablation Study


(a) Ground Truth (b) Events (c) Attention Fusion
(d) Attention Error Map (e) Synthesis Error Map (f) Warping Error Map

- visualization of results from different modules
- corresponding interpolation error maps

Experiment

Ablation Study

Module	PSNR↑	SSIM↑
Warping	35.77	0.964
Synthesis	37.71	0.959
Averaging	38.14	0.968

(a)

# Proxies	PSNR↑	SSIM↑
1	35.87	0.954
2	35.95	0.959
4	34.61	0.943

(b)

Case	PSNR↑	SSIM↑
only 1	35.87	0.954
1 & 2	37.71	0.959
1 &2 & 4	37.78	0.956

(c)

Case	PSNR↑	SSIM↑
RGB-guided	33.74	0.938
event-guided	35.77	0.964

(d)

Case	PSNR↑	SSIM↑
w/o event	32.24	0.941
w event	35.77	0.964

(e)

(a) performance of each module

(b) different event segmentation strategies

(c) different fusion strategies

(d) RGB-guided and event-guided warping

(e) importance of event-based updating

Conclusion

Conclusion

- This work shows that a careful consideration of the inherent properties of event cameras in the design of neural architectures can help improve results for VFI.
- Propose an incremental synthesis strategy that breaks down the global prediction into multiple simpler and equivalent short-term prediction steps.
- Extensive experiments show that event-based method demonstrates favorable VFI performance.

References

J. Chen, Y. Zhu, *et al.*, "Revisiting event-based video frame interpolation," *arXiv preprint*, 2023.

J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *ECCV*, pp. 694–711, Springer, 2016.

H. Xu, J. Zhang, *et al.*, "Gmflow: Learning optical flow via global matching," in *CVPR*, pp. 8121–8130, 2022.

T. Xue, B. Chen, *et al.*, "Video enhancement with task-oriented flow," *IJCV*, vol. 127, pp. 1106–1125, 2019.

Thank You!