Paper Reading

Dachun Kai

USTC

August 6, 2021

Outline

1. Paper 1: EFI-Net: Video Frame Interpolation from Fusion of Events and Frames
2. Paper 2: Time Lens: Event-based Video Frame Interpolation

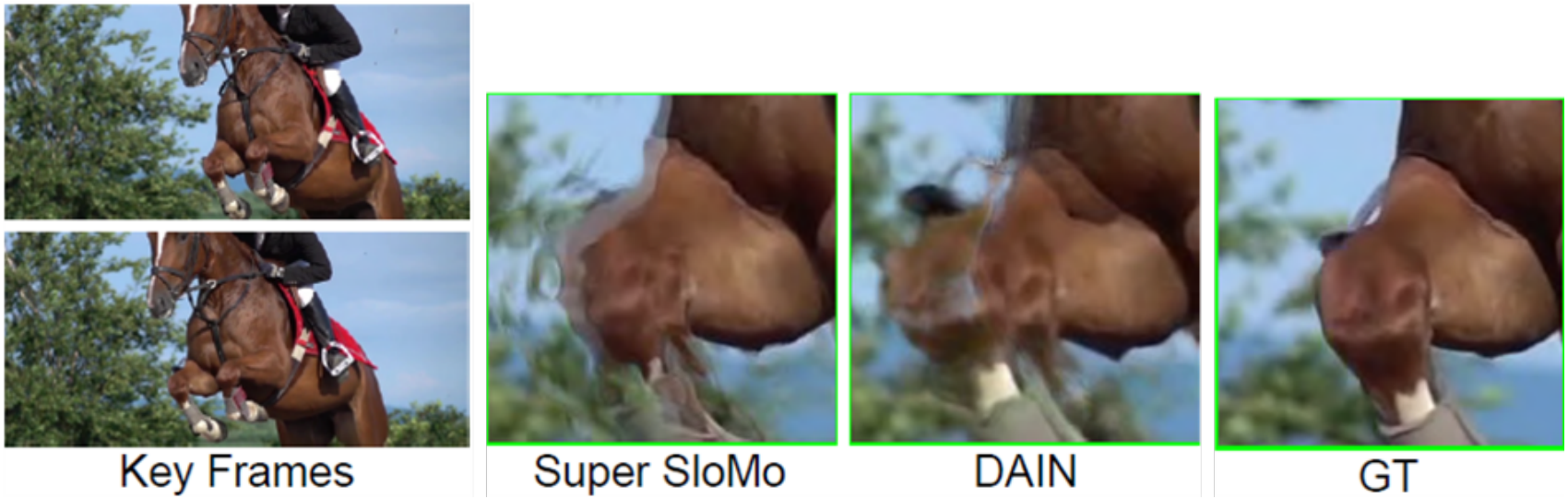
Paper 1: EFI-Net*(CVPR2021W)*

EFI-Net: Video Frame Interpolation from Fusion of Events and Frames

Genady Paikin Yotam Ater Roy Shaul Evgeny Soloveichik
Samsung Israel R&D Center
Tel Aviv
{genady.p, yotam.ater, roy.s, evgeny.s}@samsung.com

Problem Definition

Classical CNN-based VFI(Video Frame Interpolation) methods, DAIN[1] and SSM[2] suffer from Occlusion.

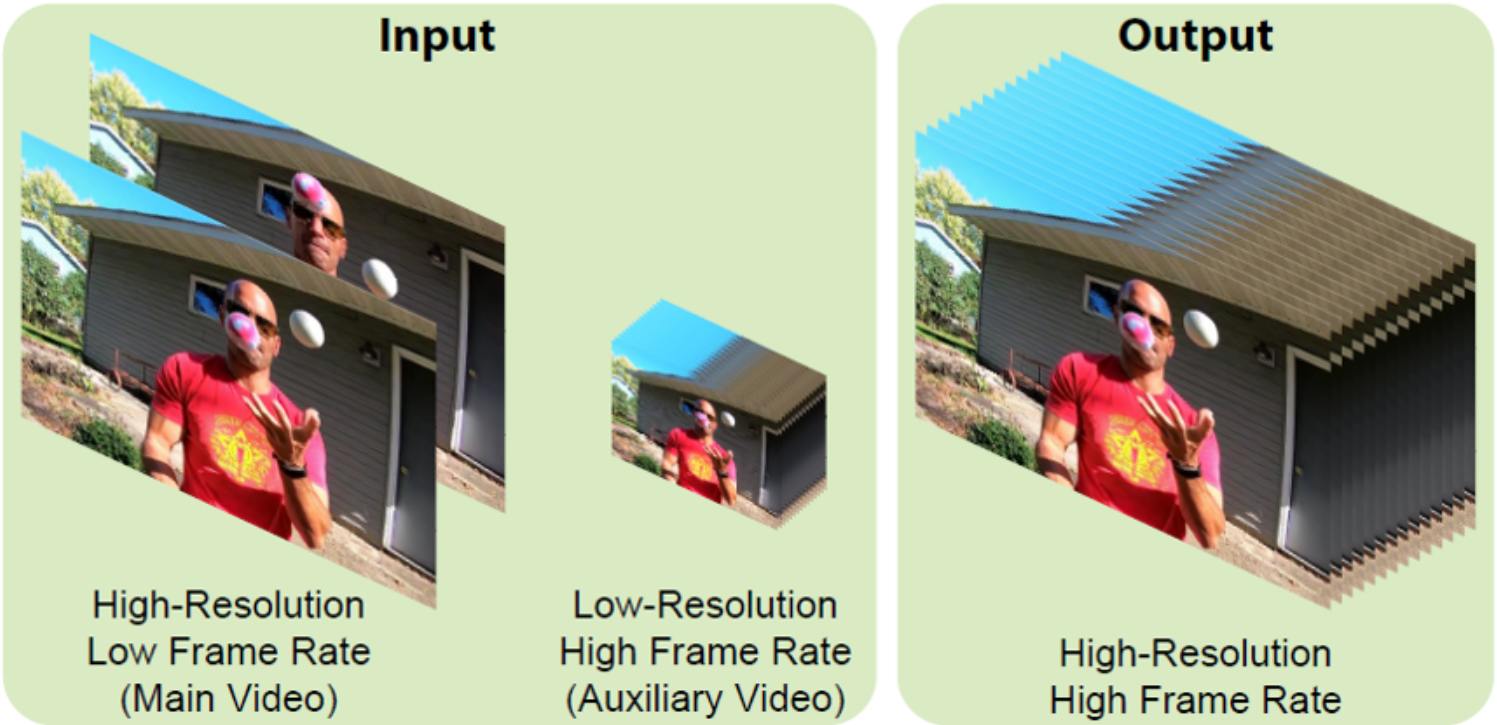


[1] Wenbo Bao et al. "Depth-aware video frame interpolation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3703–3712.

[2] Huaizu Jiang et al. "Super slomo: High quality estimation of multiple intermediate frames for video interpolation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9000–9008.

Problem Definition

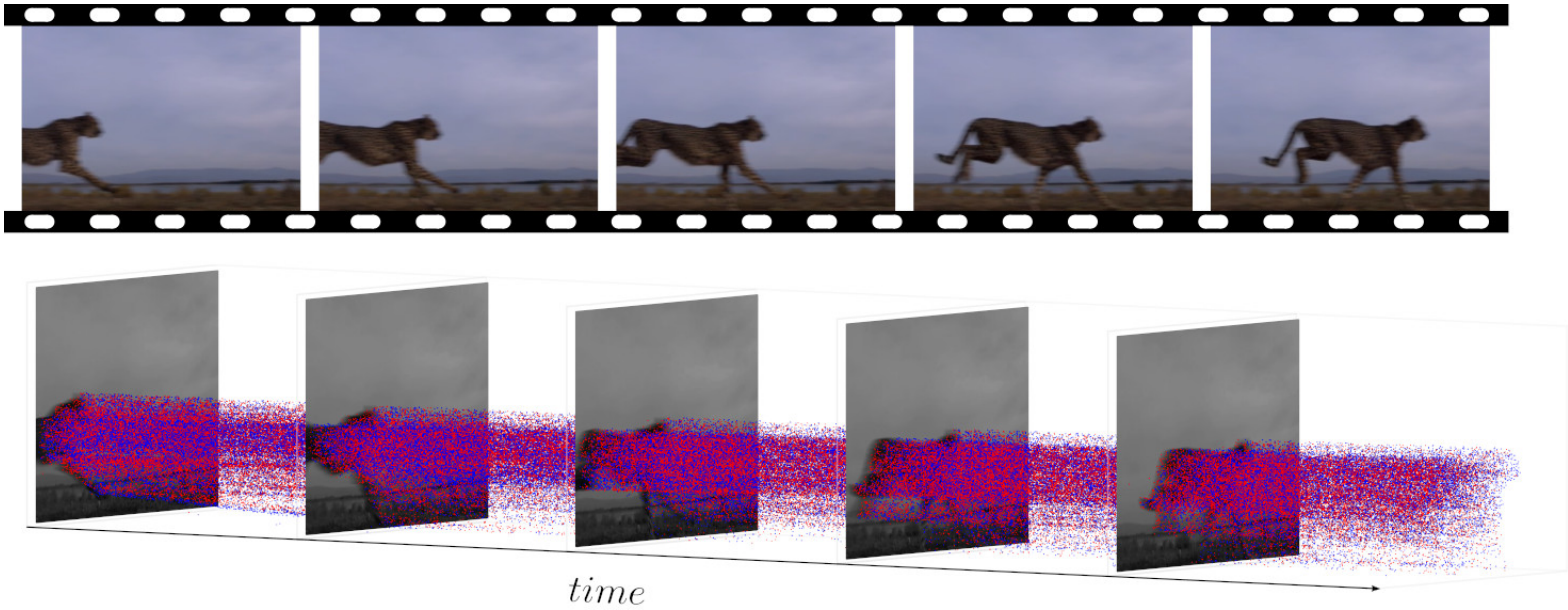
DSM(Deep Slow Motion)[1] fuse High-Resolution, Low Frame Rate(Main Video) with Low-Resolution but High Frame Rate(Auxiliary Video).



[1] Avinash Paliwal and Nima Khademi Kalantari. "Deep slow motion video reconstruction with hybrid imaging system". In: *IEEE transactions on pattern analysis and machine intelligence* 42.7 (2020), pp. 1557–1569.

Problem Definition

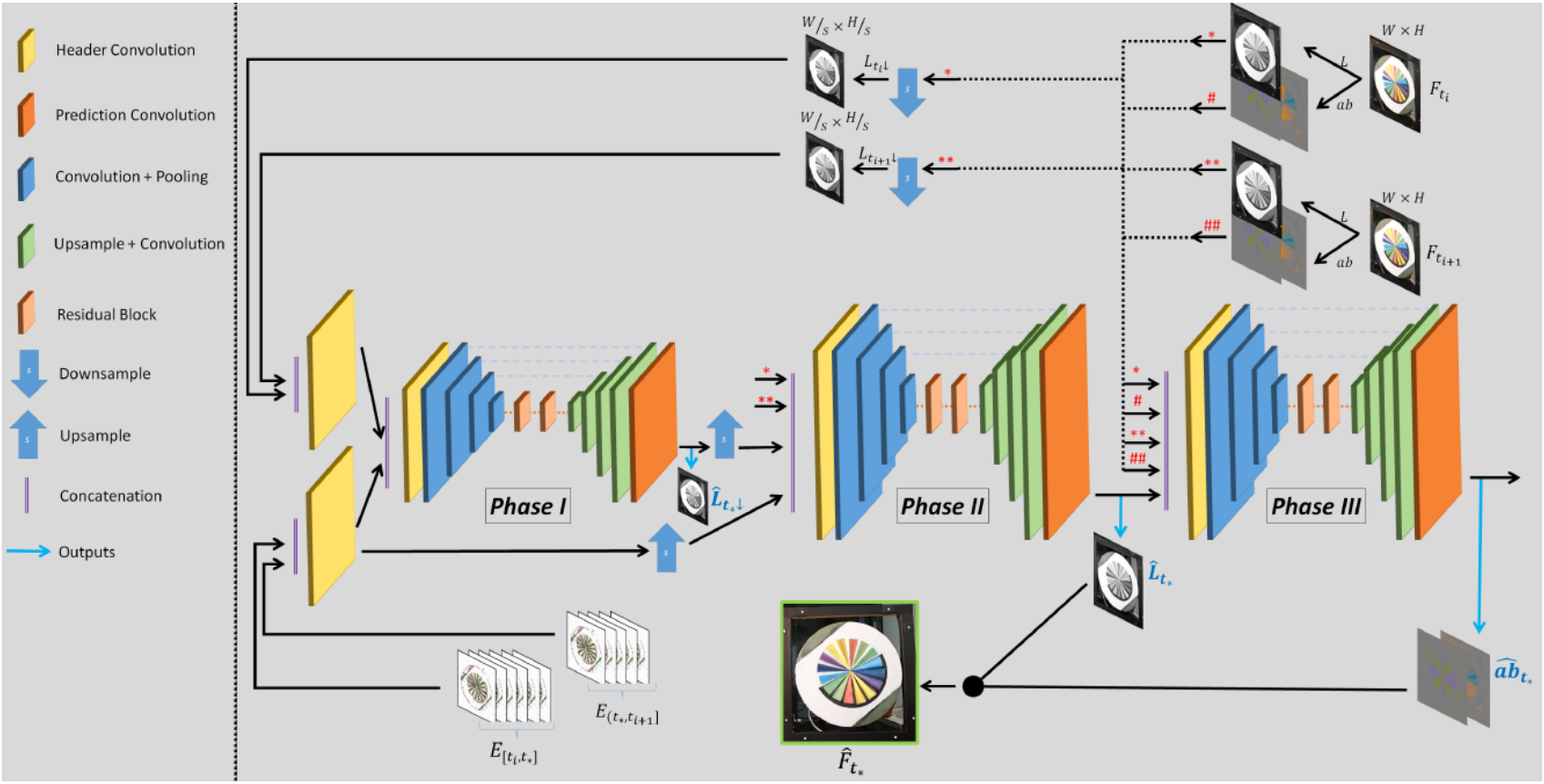
Great performance of Event Camera



- High temporal resolution(microsecs)
- Low Latency
- Low Power(SNN, neuromorphic chip)
- No motion blur
- High dynamic range

Proposed Method

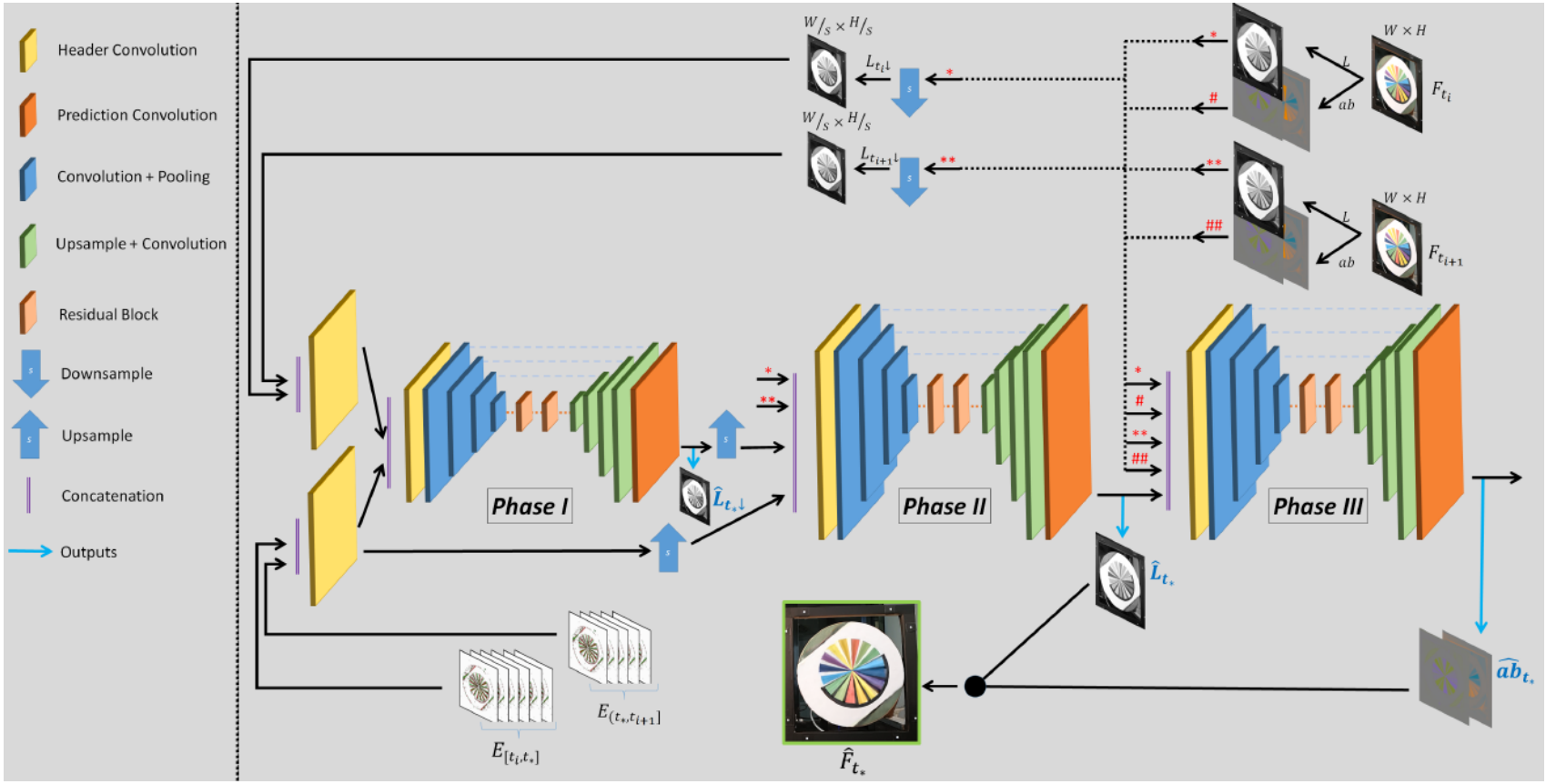
The architecture of proposed EFI-Net



- Phase I: Low resolution intensity interpolation
- Phase II: High resolution intensity interpolation
- Phase III: Re-colorization

Proposed Method

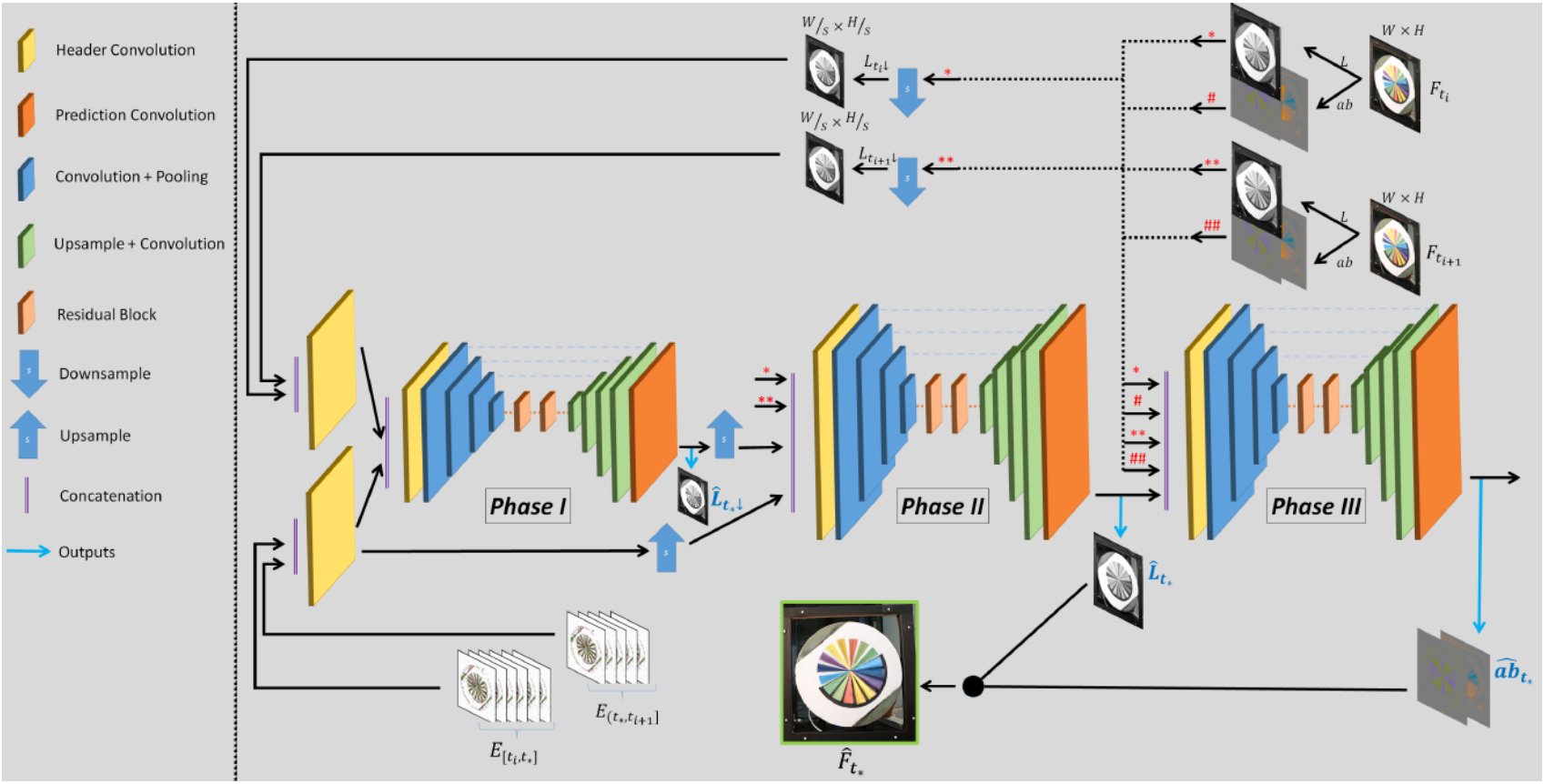
The architecture of proposed EFI-Net



- Phase I: convert RGB to CIELAB space $\rightarrow$ event representation $\rightarrow$ fuse L and E $\xrightarrow[U-Net]{output}$ $\hat{L}_{t^*\downarrow}$ (low resolution intensity frame).

Proposed Method

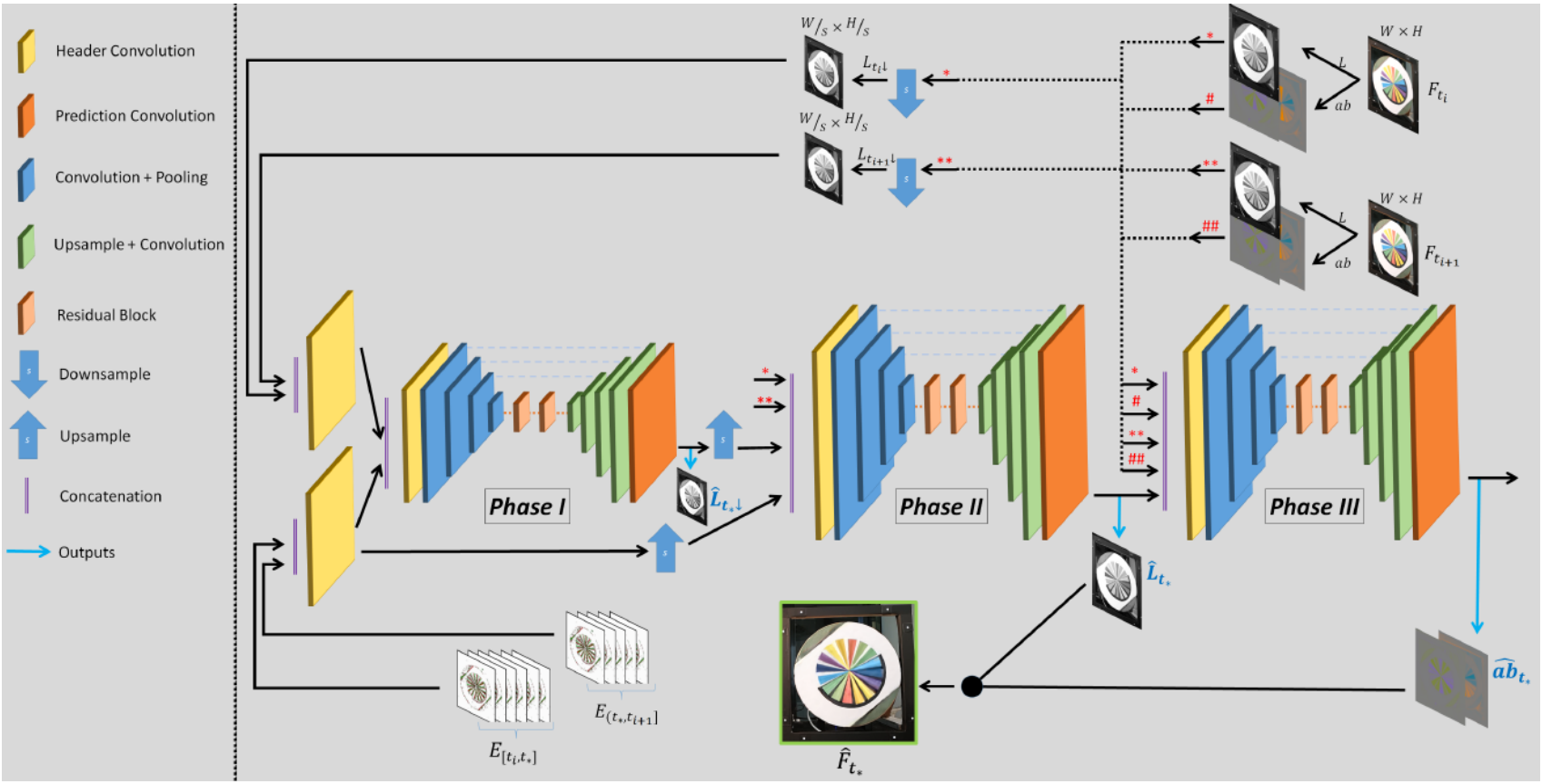
The architecture of proposed EFI-Net



- Phase II: $\hat{L}_{t^*\downarrow}$, $E_{Phase\ I}$ upsample $\rightarrow$ concat with L_{t_i}, $L_{t_{i+1}}$ $\xrightarrow[U-Net]{output}$ $\hat{L}_{t^*}$(high resolution intensity frame).

Proposed Method

The architecture of proposed EFI-Net



- Phase III: estimate color components $\hat{ab}_{t^*}$
- $\hat{L}_{t^*} + \hat{ab}_{t^*} \xrightarrow{output} \hat{F}_{t_*}$

Proposed Method

Loss Functions

$$\mathcal{L}_{\text{total}} = \lambda_P \left(\mathcal{L}_P \downarrow + \mathcal{L}_P\right) + \lambda_I \left(\mathcal{L}_I \downarrow + \mathcal{L}_I\right) + \lambda_G \left(\mathcal{L}_G \downarrow + \mathcal{L}_G\right) + \lambda_S \left(\mathcal{L}_S \downarrow + \mathcal{L}_S\right) + \lambda_C \cdot \mathcal{L}_C$$

where $\mathcal{L}_{\downarrow}$ refers to losses of Phase I

- Phase I and II:
  - Perceptual loss: $\mathcal{L}_P = \left\| \phi\left(\hat{L}_j\right) - \phi\left(L_j\right) \right\|_2^2$
  - $\mathcal{L}_1$ loss: $\mathcal{L}_I = \left\| \hat{L}_j - \left(L_j\right) \right\|_1$
  - Gradient loss: $\mathcal{L}_G = -\left(\left\| \nabla_x \left(\hat{L}_j\right) | + | \nabla_y \left(\hat{L}_j\right) \right\|_1 \right)$
  - Temporal consistency loss: $\mathcal{L}_S = \left\| \left(L_{j+1} - L_j\right) - \left(\hat{L}_{j+1} - \hat{L}_j\right) \right\|_1$
- Phase III:
  - Smooth $\mathcal{L}_1$ loss: $\mathcal{L}_C = \mathcal{L}_{1smooth}\left(\hat{ab}_j - ab_j\right)$

Experiments

1. Training Dataset: simulated, generated from transformation:

$$F_0 \xrightarrow{transform} F_1, F_2, \ldots, F_N \xrightarrow{seperate,\ downscale} L_{1\downarrow}, L_{2\downarrow}, \ldots, L_{N\downarrow}$$

2. Event generation:

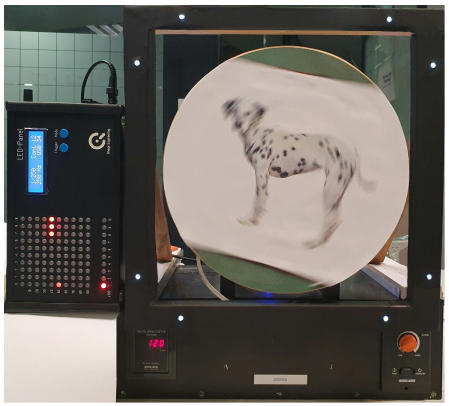
$$E = \begin{cases} 1 & \frac{L_{n+1\downarrow}+c}{L_{n\downarrow}+c} > \tau \\ -1 & \frac{L_{n+1\downarrow}+c}{L_{n\downarrow}+c} < 1/\tau \\ 0 & else \end{cases}$$

- c: positive constant, to avoid noise in dark regions
- τ: threshold, simulating the sensitivity of the event camera

Experiments

Testing Dataset:

- Self-collected Dataset, includes 640 $\times$ 480 DVS event camera and a Samsung Galaxy S10+, 1280 $\times$ 960 at 240FPS.



- Dataset introduced in[1], by a DAVIS240 camera.

(a) Shapes

(b) Wall Poster

[1] Elias Mueggler et al. "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM". In: *The International Journal of Robotics Research* 36.2 (2017), pp. 142–149.

Experiments

Quatitative results

- Comparison on public dataset[1]

Sequence	SSIM							MSE						
	EV [54]	EG [64]	SR [7]	SSM [22]	DAIN [2]	EDI [46]	EFI-Net	EV [54]	EG [64]	SR [7]	SSM [22]	DAIN [2]	[46]	EFI-Net
dynamic 6dof	0.46	0.44	0.48	0.79	0.8	0.64	0.78	0.14	0.05	0.03	0.006	0.006	0.01	0.002
boxes 6dof	0.62	0.61	0.45	0.59	0.61	0.39	0.71	0.04	0.02	0.03	0.009	0.009	0.017	0.003
poster 6dof	0.62	0.63	0.61	0.58	0.6	0.42	0.74	0.06	0.02	0.01	0.009	0.009	0.014	0.002
shapes 6dof	0.8	0.79	0.56	0.84	0.84	0.75	0.86	0.04	0.01	0.03	0.006	0.006	0.011	0.001
office zigzag	0.54	0.68	0.67	0.83	0.82	0.64	0.79	0.03	0.01	0.01	0.006	0.007	0.011	0.002
slider depth	0.58	0.59	0.54	0.91	0.91	0.61	0.8	0.05	0.02	0.02	0.005	0.005	0.012	0.002
calibration	0.7	0.71	0.67	0.91	0.9	0.73	0.85	0.02	0.01	0.01	0.005	0.006	0.011	0.002
Mean	0.617	0.636	0.569	0.778	0.783	0.597	0.79	0.054	0.02	0.02	0.007	0.008	0.013	0.002

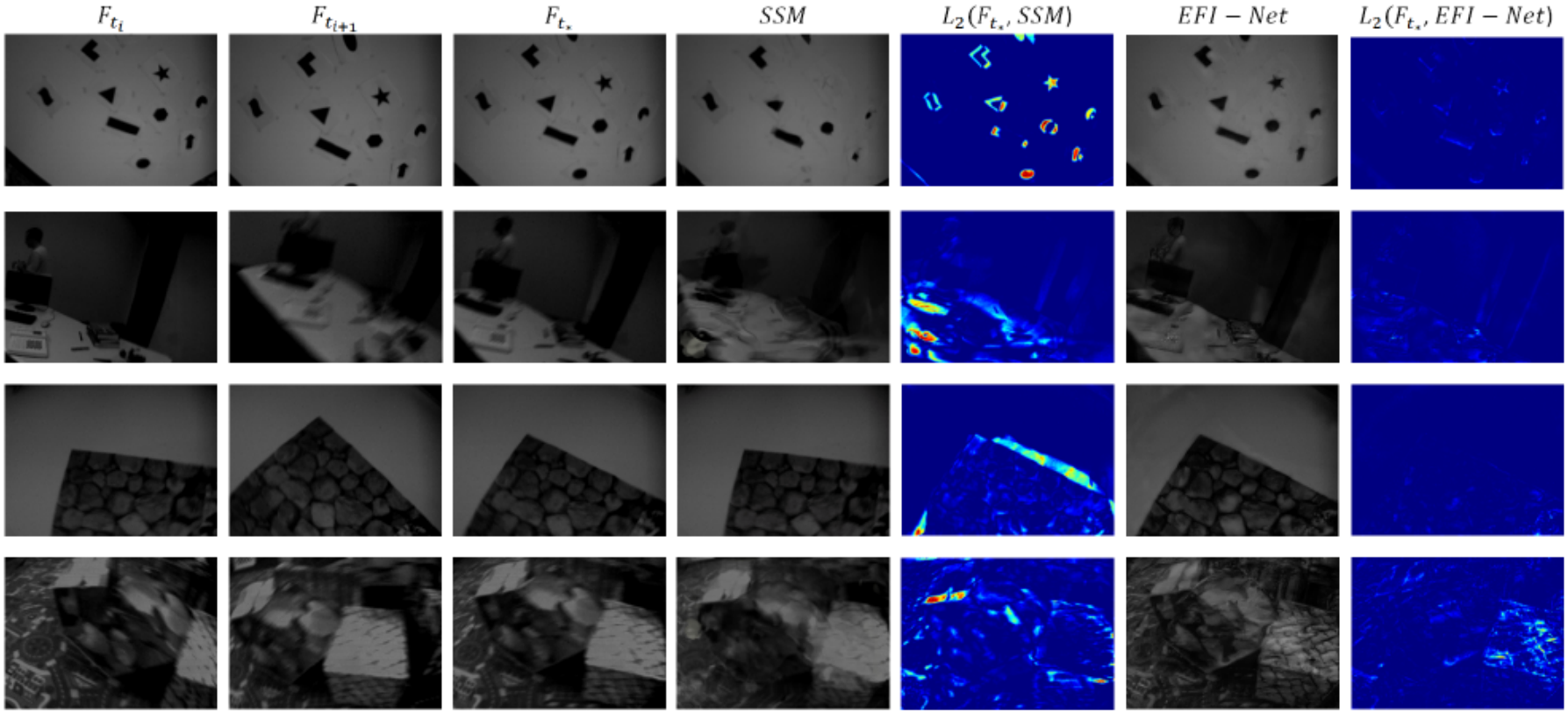
- Comparison on self-collected dataset

Sequence	SSIM								PSNR							
	SSM [22]		DSM8 [45]		DAIN [2]		EFI-Net		SSM [22]		DSM8 [45]		DAIN [2]		EFI-Net	
Input\Output FPS	60\240	30\120	60\240	30\120	60\240	30\120	60\240	30\120	60\240	30\120	60\240	30\120	60\240	30\120	60\240	30\120
star 100 RPM	0.93	0.86	0.94	0.92	0.88	0.85	0.92	0.91	29.19	20.49	30.89	29.2	25.48	20.09	29.03	26.62
star 150 RPM	0.91	0.86	0.94	0.91	0.91	0.79	0.92	0.91	23.9	20.61	31	28.85	26.05	19.31	28.18	25.83
star 200 RPM	0.89	0.85	0.94	0.91	0.88	0.8	0.91	0.90	22.3	19.86	30.66	28.29	21.97	19.4	26.57	23.99
dog 100 RPM	0.94	0.89	0.94	0.92	0.94	0.9	0.92	0.92	30.91	25.06	32.09	30	31.31	27.26	28.45	27.95
car 100 RPM	0.93	0.87	0.94	0.92	0.93	0.83	0.91	0.91	29.86	23.61	31.29	29.47	30.42	23.14	28.95	27.94
Mean	0.92	0.866	0.94	0.916	0.91	0.83	0.916	0.91	27.232	21.926	31.186	29.162	27.05	21.84	28.236	26.466

[1] Elias Mueggler et al. "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM". In: *The International Journal of Robotics Research* 36.2 (2017), pp. 142–149.

Experiments

Qualitative results



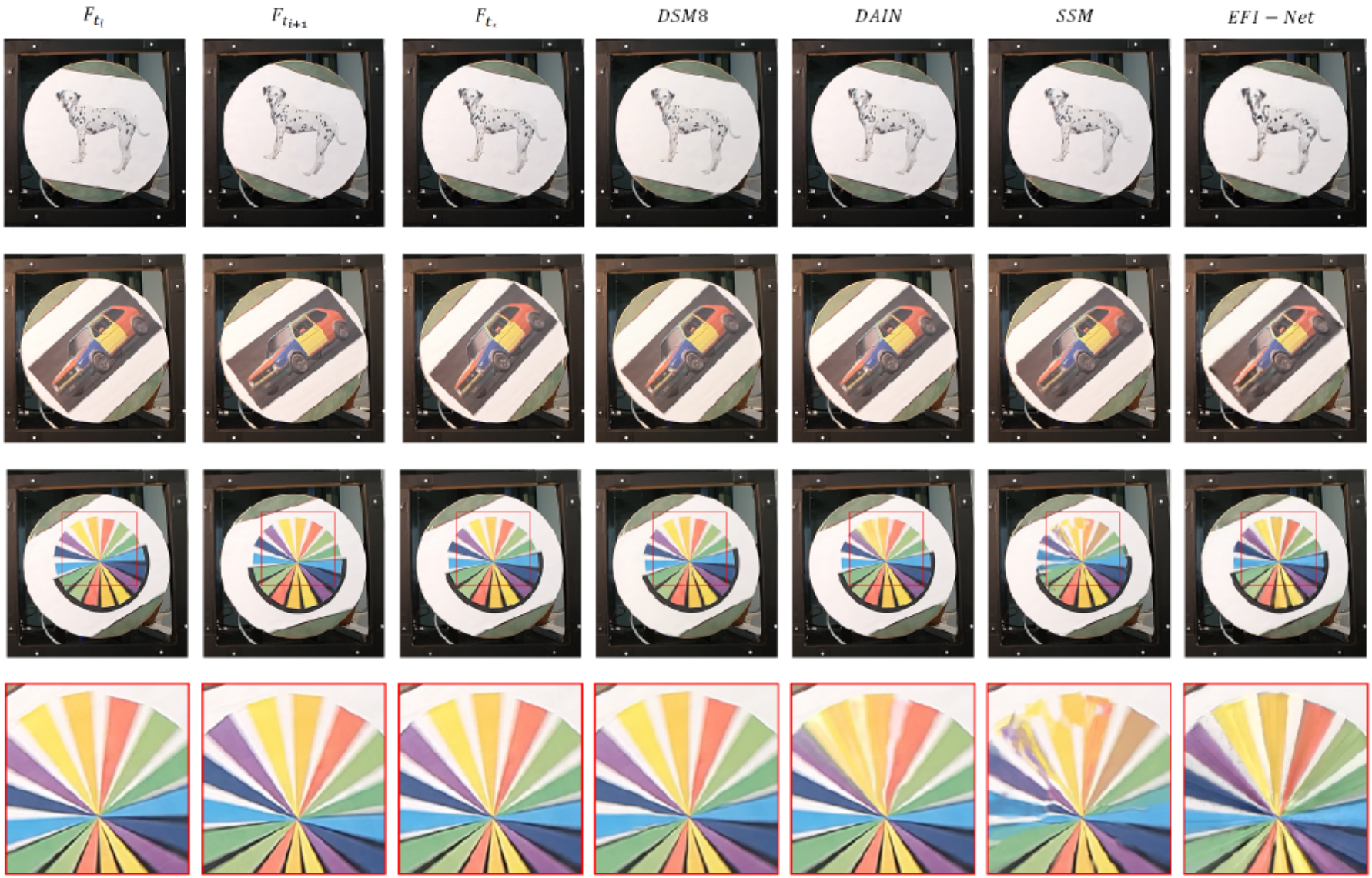
Comparison with SSM[1] on the dataset[2]

[1] Huaizu Jiang et al. "Super slomo: High quality estimation of multiple intermediate frames for video interpolation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9000–9008.

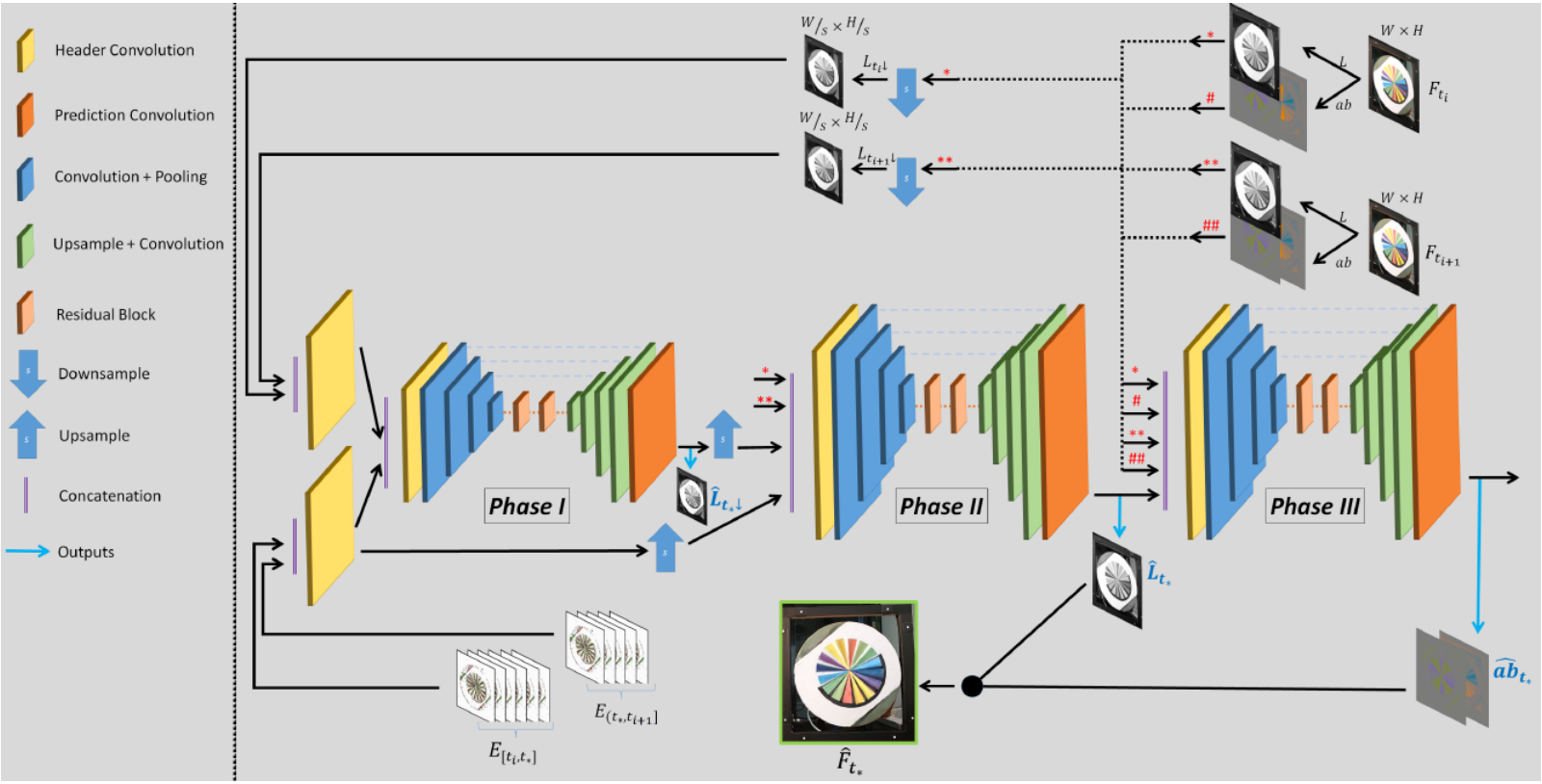
[2] Elias Mueggler et al. "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM". In: *The International Journal of Robotics Research* 36.2 (2017), pp. 142–149.

Experiments

Qualitative results



Conclusions



- Propose a fusion pipeline, combine conventional frames with events for *VFI*.
- Simulate events in a simple way, and verify the generalization to multiple dataset.
- Contribute a novel dataset with spatio-temporal alignment of frames and events.

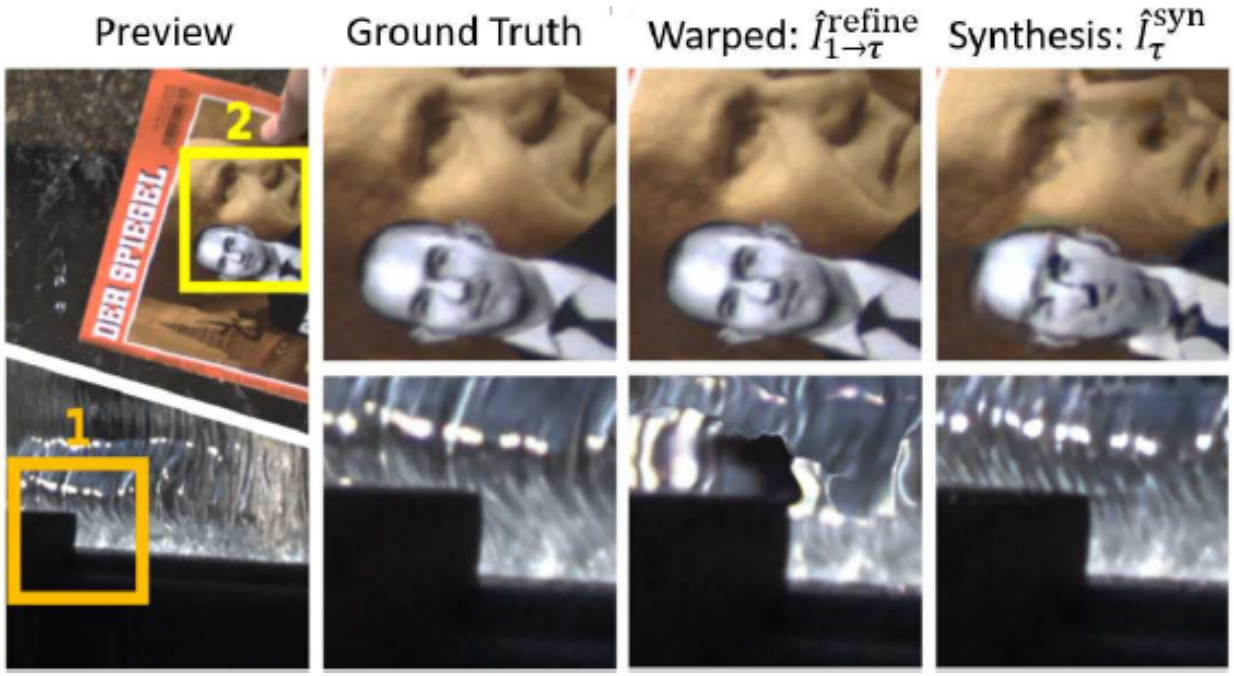
Paper 2: Time Lens*(CVPR2021)*

Time Lens: Event-based Video Frame Interpolation

Stepan Tulyakov*,[1] Daniel Gehrig*,[2] Stamatios Georgoulis[1] Julius Erbach[1]
Mathias Gehrig[2] Yuanyou Li[1] Davide Scaramuzza[2]
[1]Huawei Technologies, Zurich Research Center
[2]Dept. of Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich

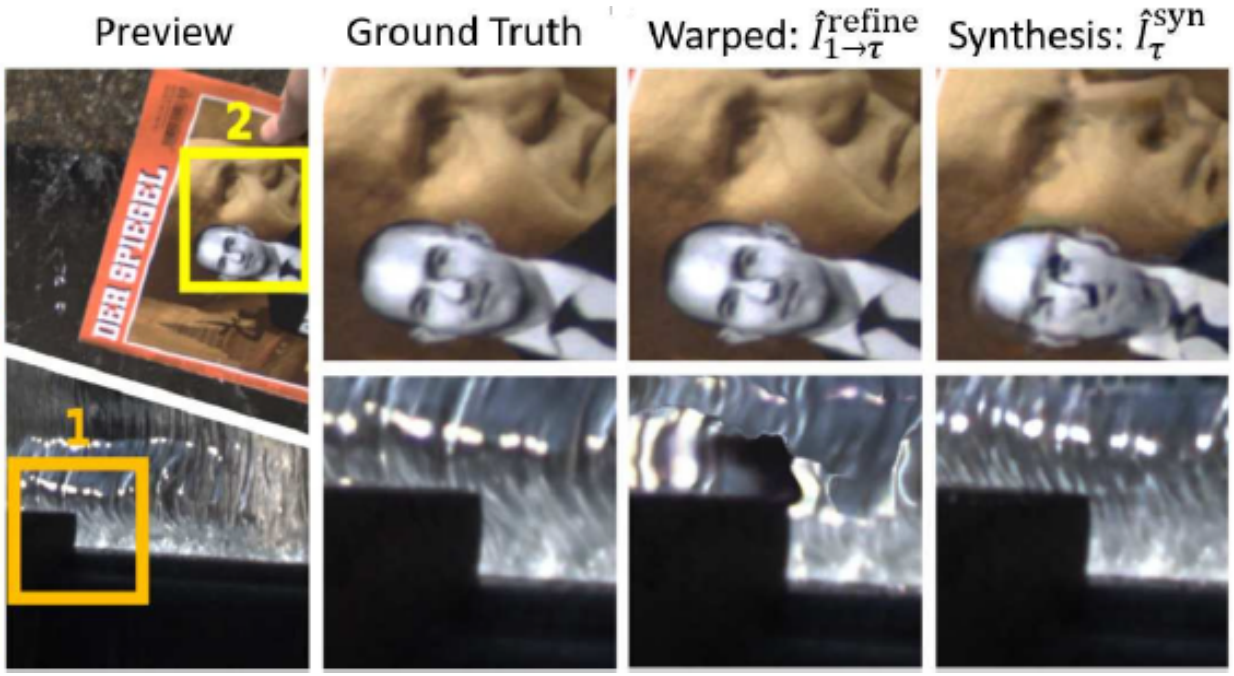
Motivation

- Warping-based methods: compute optical flow to warp input frames, can't solve non-linear motion.
- Event-based methods: learn frame residual, can't solve low-texture surface.



Motivation

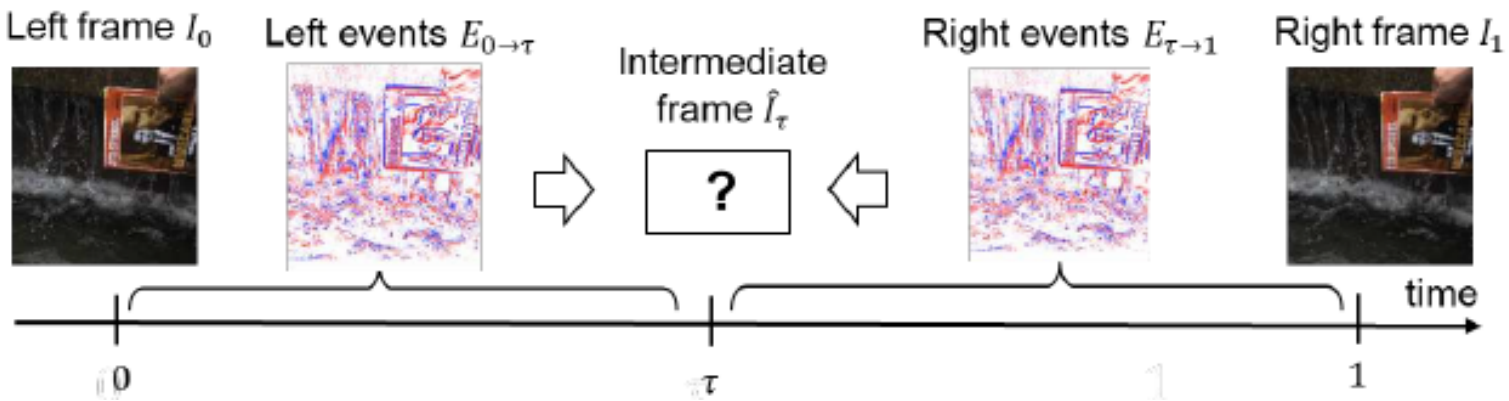
- Warping-based methods: compute optical flow to warp input frames, can't solve non-linear motion.
- Event-based methods: learn frame residual, can't solve low-texture surface.



- Warping-based and Event-based methods each has its advantages and disadvantages, so we can complement both methods.

Problem Formulation

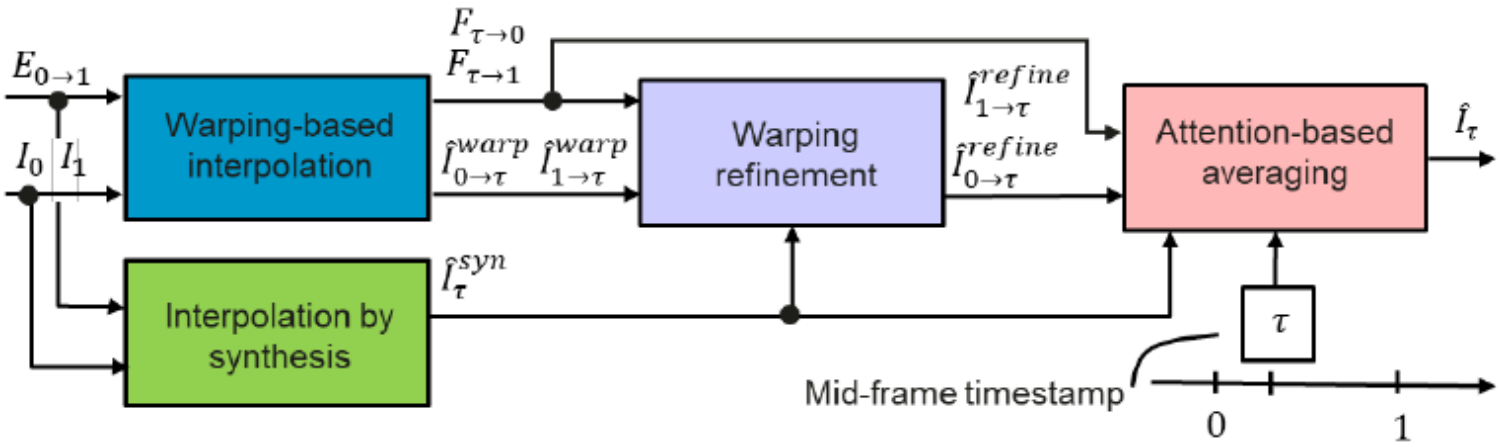
Input: left I_0 and right I_1 RGB frames, and the left $E_{0\to\tau}$ and right $E_{1\to\tau}$ event sequences.



Output: interpolate $\hat{I}_\tau$ at random timestamp

Proposed Method

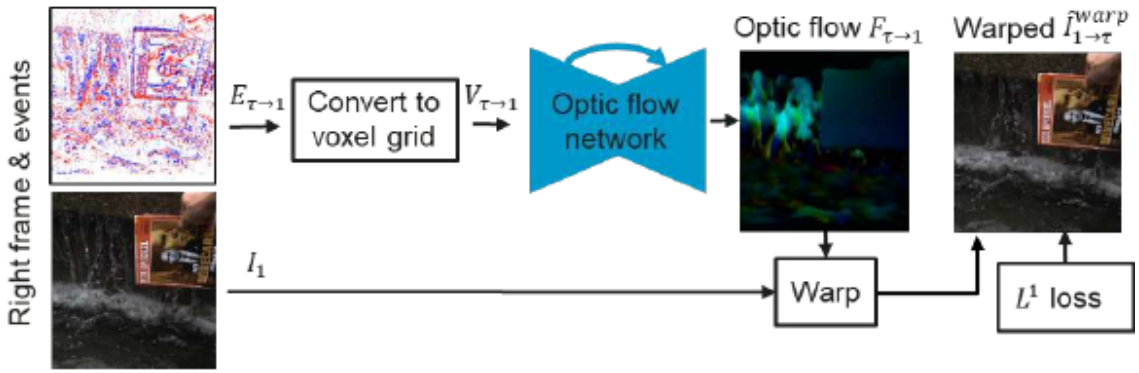
WorkFlow



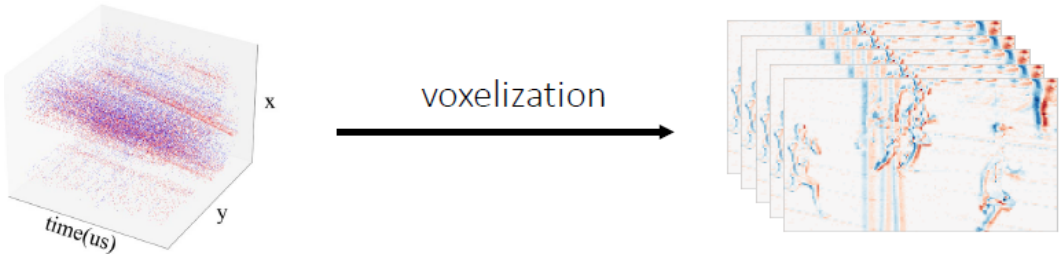
1. Warping-based interpolation: use events to compute flow, and warp input frames
2. Interpolation by synthesis: fuse frames and events
3. Warping refinement module: compute residual flow to refine
4. Attention-based averaging: combine warping-based and synthesis-based results

Proposed Method

Warping-based interpolation

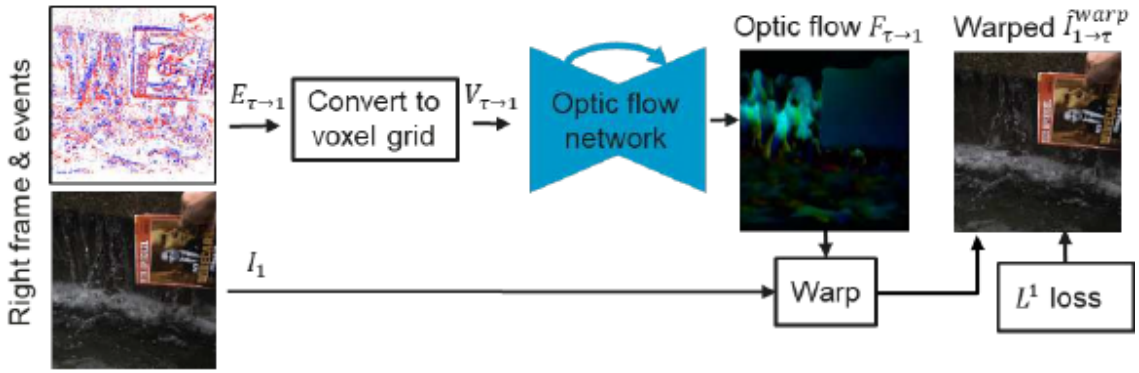


- Event representation: convert to voxel grid



Proposed Method

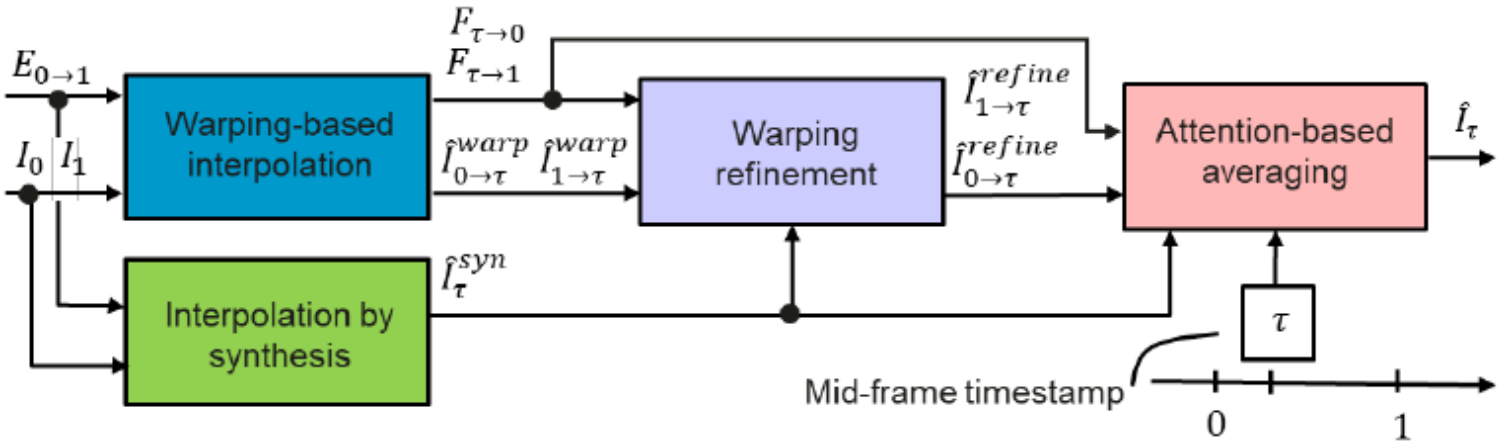
Warping-based interpolation



- Event representation: convert to voxel grid
- Estimate optical flow by events $E_{\tau\to0}$ and $E_{\tau\to1}$
- Warp the input frames
- Loss Function: $\mathcal{L}_1$ loss

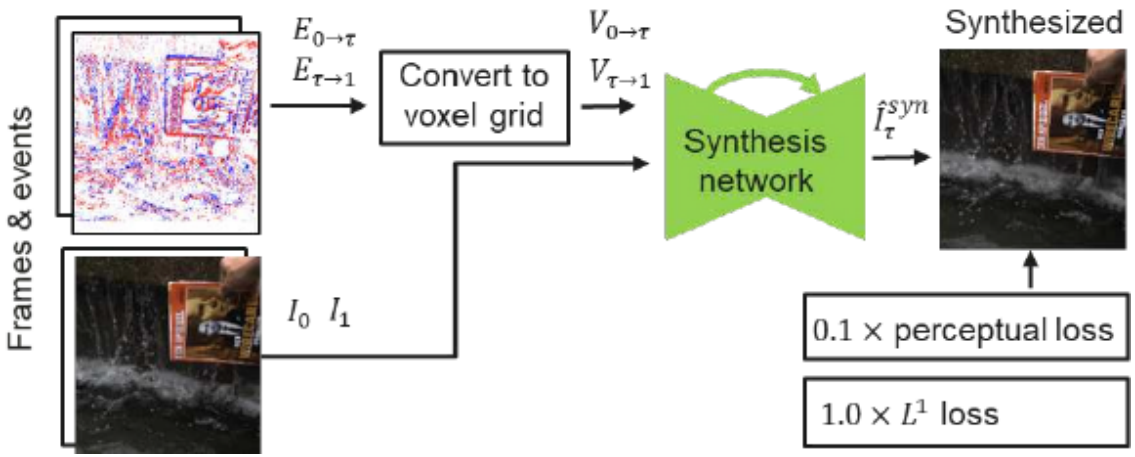
Proposed Method

WorkFlow



Proposed Method

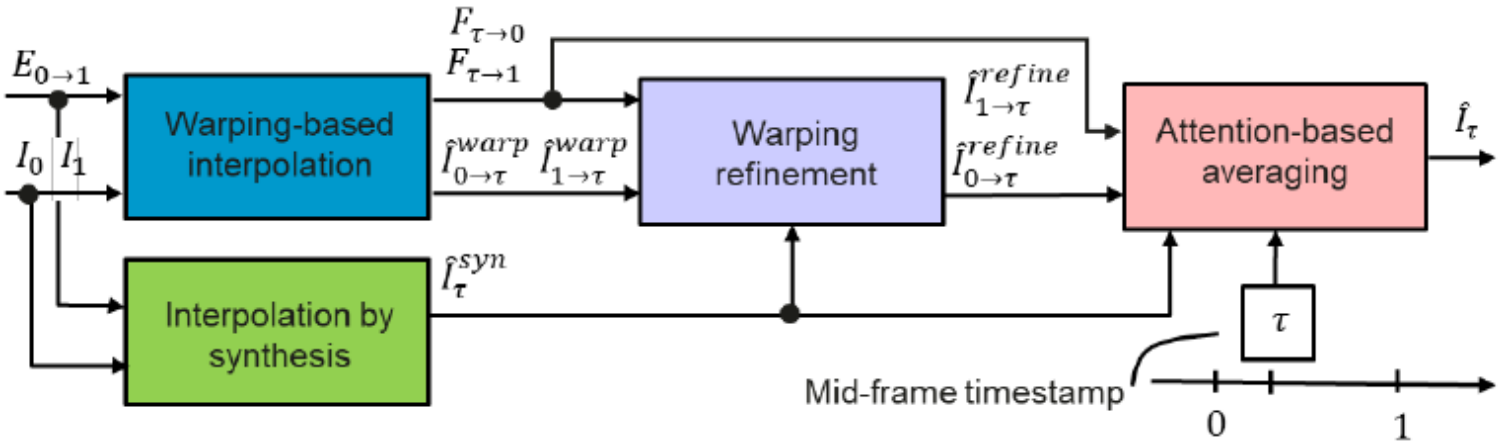
Interpolation by synthesis



- Event representation: convert to voxel grid
- Synthesis I_0, I_1, $V_{0\to\tau}$, $V_{\tau\to 1}$ directly
- Loss Functions: $0.1\times$perceptual loss $+$ $1.0\times\mathcal{L}_1$ loss

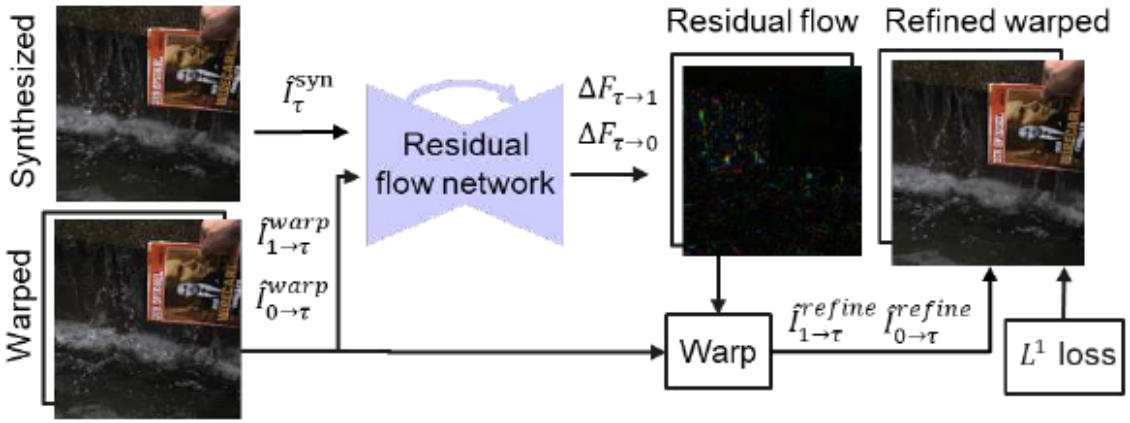
Proposed Method

WorkFlow



Proposed Method

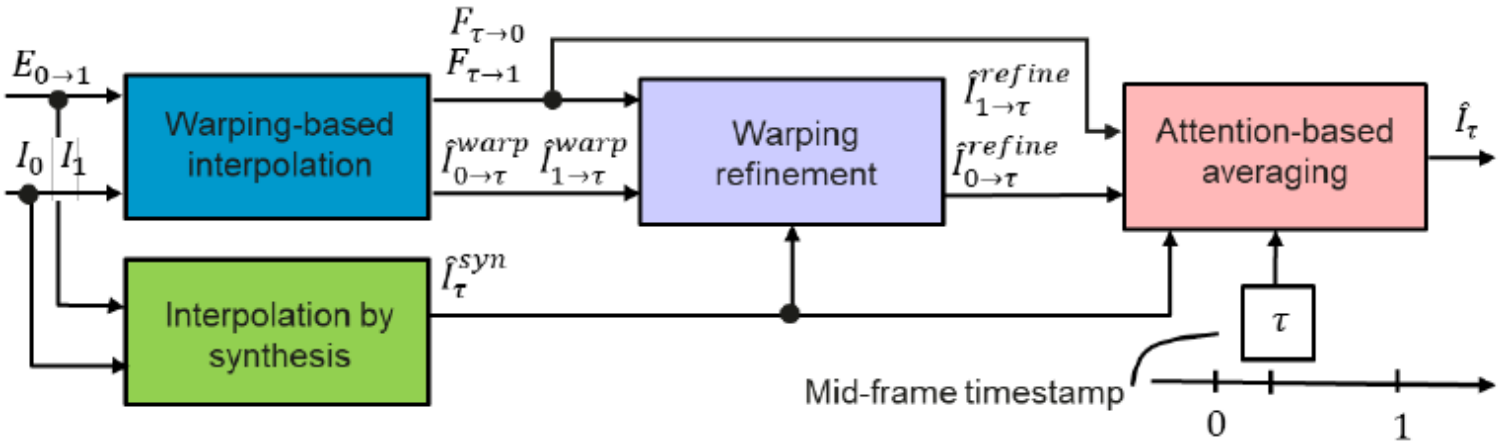
Warping refinement



- Compute residual flow $\Delta F_{\tau\to0}$ and $\Delta F_{\tau\to1}$
- Warp $\hat{I}_{0\to\tau}^{warp}$ and $\hat{I}_{1\to\tau}^{warp}$ by residual flow
- Loss Function: $\mathcal{L}_1$ loss

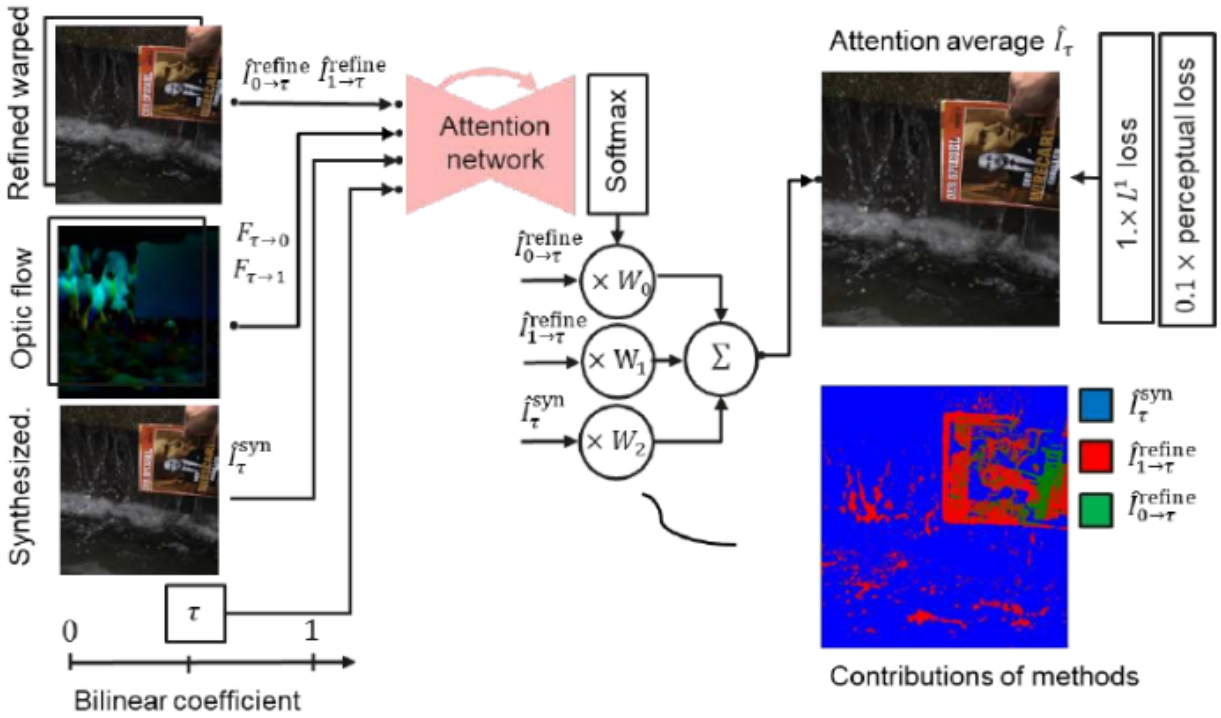
Proposed Method

WorkFlow



Proposed Method

Attention Averaging



- Compute blending coefficients, via $\hat{I}^{\text{refine}}_{0\to\tau}$, $\hat{I}^{\text{refine}}_{1\to\tau}$ and $\hat{I}^{\text{syn}}$, $F_{\tau\to 0}$, $F_{\tau\to 1}$ and bi-linear coefficient τ, which depends on the new frame position.
- Blend $\hat{I}^{\text{refine}}_{0\to\tau}$, $\hat{I}^{\text{refine}}_{1\to\tau}$ and $\hat{I}^{\text{syn}}$

Experiments

Training Details:

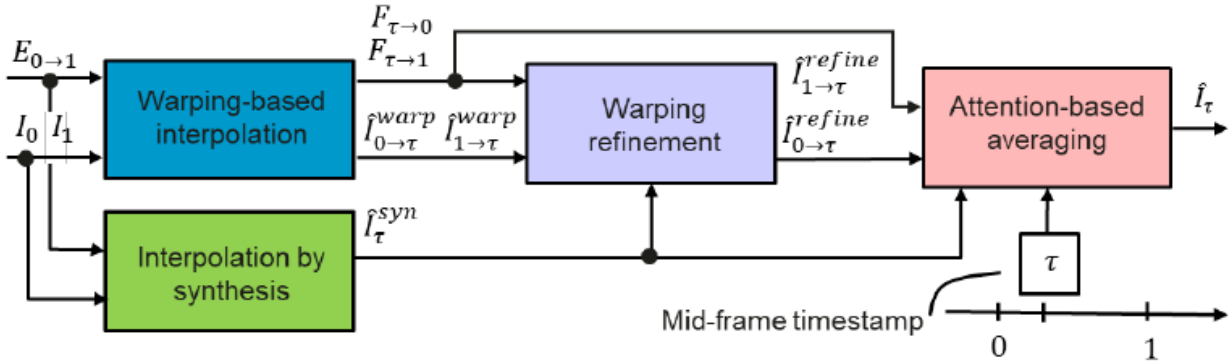
- Training Dataset: *Vimeo90k* septuplet dataset[1] frames, and its systhetic events.

[1] Tianfan Xue et al. "Video enhancement with task-oriented flow". In: *International Journal of Computer Vision* 127.8 (2019), pp. 1106–1125.

Experiments

Training Details:

- Training Dataset: *Vimeo90k* septuplet dataset[1] frames, and its systhetic events.



- Training Process: synthesis-based interpolation $\rightarrow$ warping-based interpolation $\rightarrow$ warping refinement $\rightarrow$ attention averaging module.

[1] Tianfan Xue et al. "Video enhancement with task-oriented flow". In: *International Journal of Computer Vision* 127.8 (2019), pp. 1106–1125.

Experiments

Testing Benchmarking:

- Synthetic datasets: *Vimeo90k*[2] and *Middlebury*[3], skip 1 or 3 frames and reconstruct them.
- High Quality Frames(HQF)[4] dataset: video&events dataset, without blur and saturation, collected by DAVIS240.

[2] Tianfan Xue et al. "Video enhancement with task-oriented flow". In: *International Journal of Computer Vision* 127.8 (2019), pp. 1106–1125.

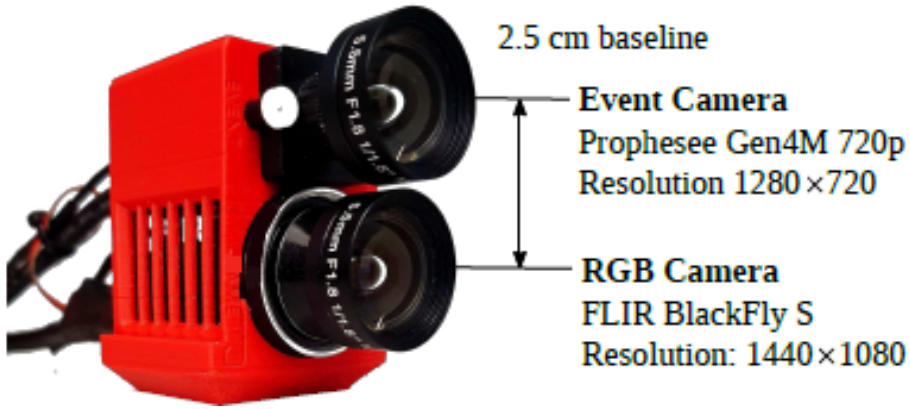
[3] Simon Baker et al. "A database and evaluation methodology for optical flow". In: *International journal of computer vision* 92.1 (2011), pp. 1–31.

[4] Timo Stoffregen et al. "Reducing the sim-to-real gap for event cameras". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer. 2020, pp. 534–549.

Experiments

Testing Benchmarking:

- Synthetic datasets: *Vimeo90k*[2] and *Middlebury*[3], skip 1 or 3 frames and reconstruct them.
- High Quality Frames(HQF)[4] dataset: video&events dataset, without blur and saturation, collected by DAVIS240.
- High Speed Event-RGB dataset: Self-collected, Higher frame rate 225FPS, Higher event resolution 1280×720.



[2] Tianfan Xue et al. "Video enhancement with task-oriented flow". In: *International Journal of Computer Vision* 127.8 (2019), pp. 1106–1125.

[3] Simon Baker et al. "A database and evaluation methodology for optical flow". In: *International journal of computer vision* 92.1 (2011), pp. 1–31.

[4] Timo Stoffregen et al. "Reducing the sim-to-real gap for event cameras". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer. 2020, pp. 534–549.

Experiments

Quatitative results

- Comparison on *Vimeo90k*[5] and *Middlebury*[6] dataset

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**

[5] Tianfan Xue et al. "Video enhancement with task-oriented flow". In: *International Journal of Computer Vision* 127.8 (2019), pp. 1106–1125.

[6] Simon Baker et al. "A database and evaluation methodology for optical flow". In: *International journal of computer vision* 92.1 (2011), pp. 1–31.

Experiments

Quatitative results

- Comparison on HQF[5] dataset

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**

- *Time Lens-syn*: train only on synthetic data
- *Time Lens-real*: train on synthetic data and fine-tuned on real event data

[5] Timo Stoffregen et al. "Reducing the sim-to-real gap for event cameras". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer. 2020, pp. 534–549.

Experiments

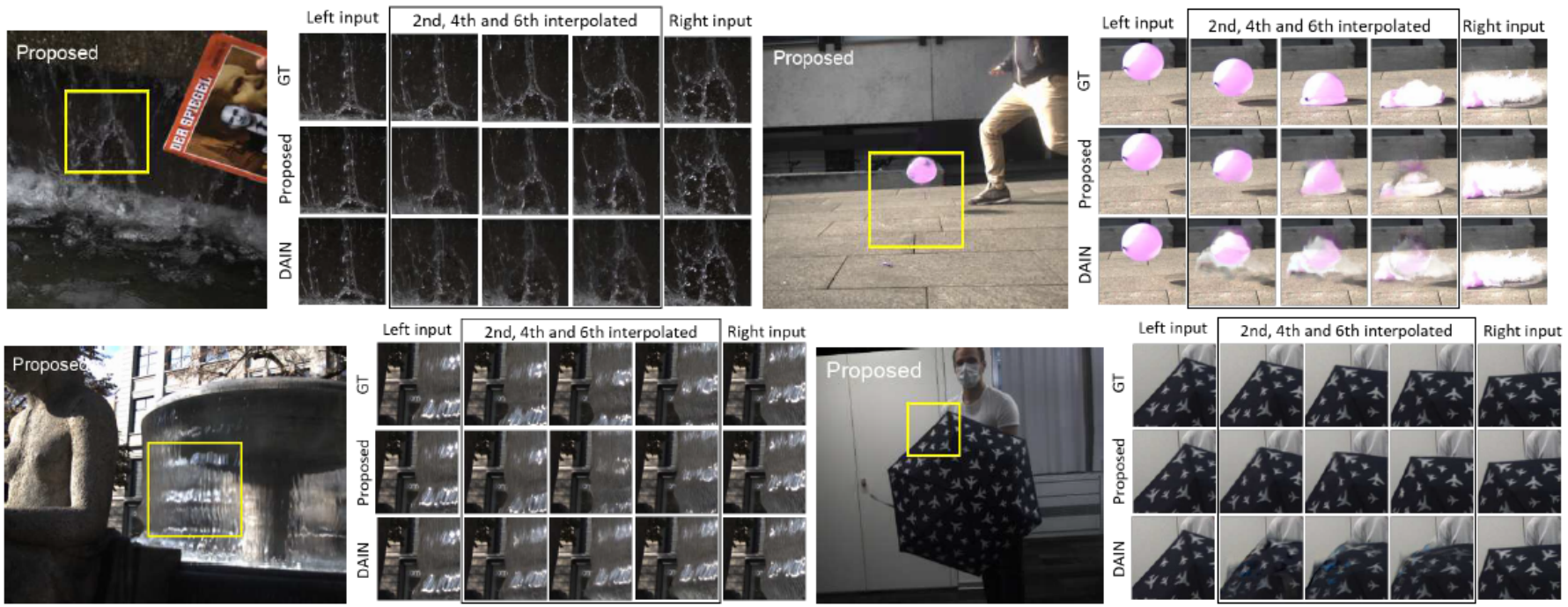
Quatitative results

- Comparison on HS-ERGB(self-collected) dataset

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**

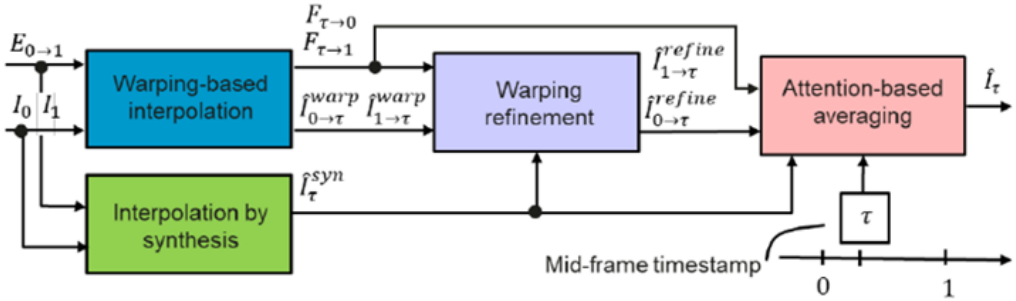
Experiments

Qualitative results



Comparison on non-linear and complex scenarios

Conclusions



- Leverage the advantages of event-based and flow-based respectively.
- Release a large-scale HS-ERGB dataset, pushing the limits of *VFI* in complex scenarios.

Thanks, Q & A