Paper Reading

Dachun Kai

USTC

December 17, 2021

WEVI*(ICCV2021)*

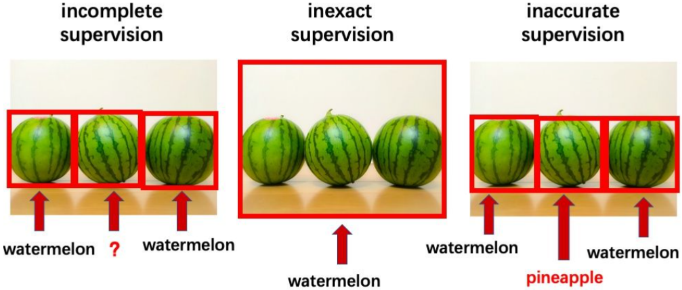
Training Weakly Supervised Video Frame Interpolation with Events

Zhiyang Yu[† 1, 2], Yu Zhang[* 2, 3], Deyuan Liu[† 2, 5], Dongqing Zou[2, 4],
Xijun Chen[* 1], Yebin Liu[3], and Jimmy Ren[2, 4]

[1]Harbin Institute of Technology, [2]SenseTime Research and Tetras.AI, [3]Tsinghua University
[4]Qing Yuan Research Institute, Shanghai Jiao Tong University, [5]Peking University

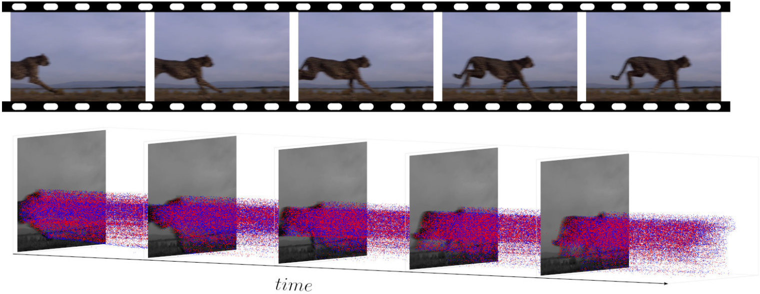
Preliminaries
Weakly Supervised learning

- Incomplete supervision: subset labels → trained on low frame-rate videos
- Inexact supervision: coarse-grained labels
- Inaccurate supervision: wrong labels



Preliminaries

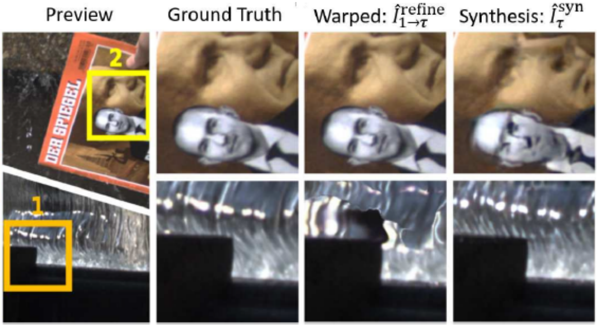
Great performance of Event Camera



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

Motivation

- Parameterized motion model falls into an ill-posed problem when input frames are sparse.
- Event-based methods: learn frame residual, can't solve low-texture surface, always need high frame-rate auxilary videos.

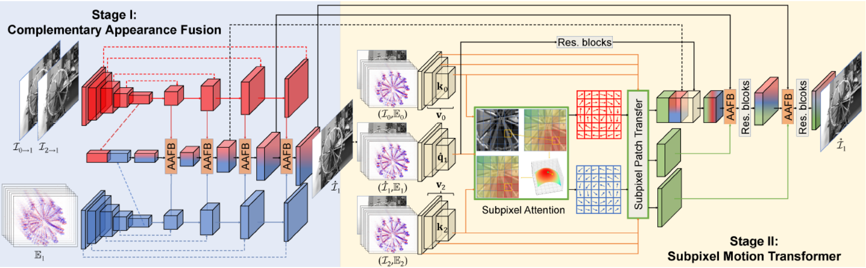


Contributions

- Weakly supervised training, and perform SOTA, generalize well.
- Complementary appearance fusion *(CAF)* for aggregates image and event appearance.
- Subpixel Motion Transformer *(SMT)*, for super resolution *(SR)* reconstruction.
- Contribute an event dataset, but not open source yet.

Proposed Method

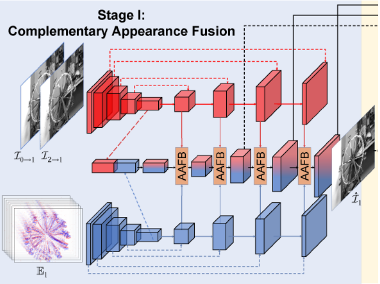
Overview



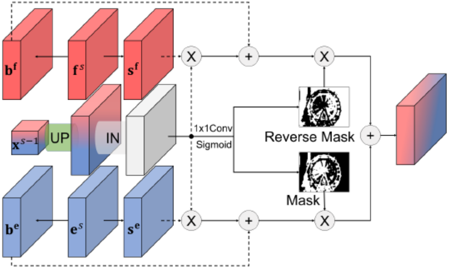
- Two stages: complementary appearance fusion *(CAF)* and subpixel motion transfer *(SMT)*
- CAF stage: fuse images and events, get initial $\hat{I}_1$.
- SMT stage: refine results with subpixel attention mechanism.

Proposed Method

Stage I: Complementary appearance fusion(CAF)



Stage I



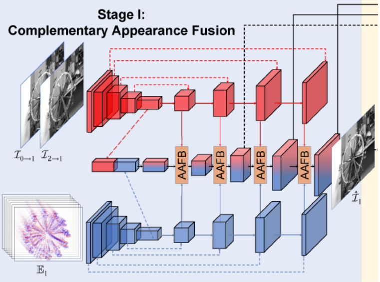
AAFB

- Prepare works:
  - Forward warping $\mathcal{I}_0$ and $\mathcal{I}_2$ with PWC-Net[1] $\rightarrow \mathcal{I}_{0\rightarrow 1}$, $\mathcal{I}_{2\rightarrow 1}$
  - Event representation $\mathbb{E}_1$

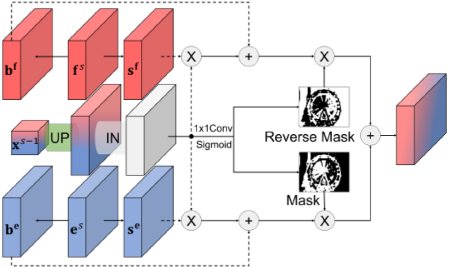
[1] Deqing Sun et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8934–8943.

Proposed Method

Stage I: Complementary appearance fusion(CAF)



Stage I



AAFB

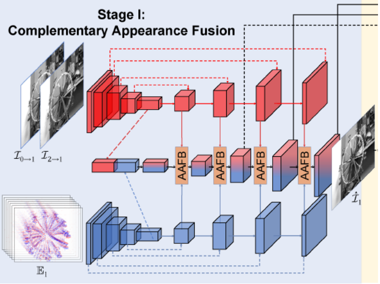
- Adaptive Appearance Fusion Block(AAFB):

$$\mathrm{x}^s = g\left(\mathrm{x}_{\uparrow}^{s-1}; \mathrm{f}^s, \mathrm{e}^s\right), s \in \{1,2,3,4,5\} \tag{1}$$

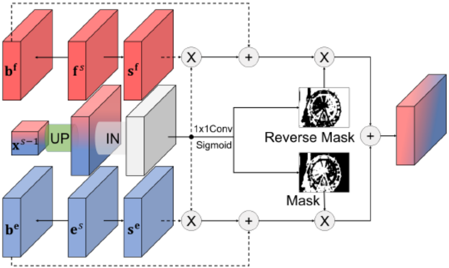
$\mathrm{x}_{\uparrow}^{s-1}$ denotes the $2\times$ upsampled version of x^{s-1}
$\mathrm{f}^s, \mathrm{e}^s$ denotes 2 branches features.

Proposed Method

Stage I: Complementary appearance fusion(CAF)



Stage I



AAFB

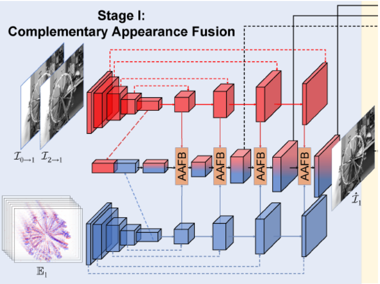
- Adaptive Appearance Fusion Block(AAFB):

$$y^e = \left(\frac{x_{\uparrow}^{s} - \mu\left(x_{\uparrow}^{s}\right)}{\sigma\left(x_{\uparrow}^{s}\right)} \right) \odot s^e + b^e \qquad (1)$$

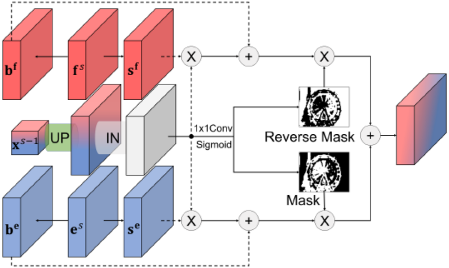
Break up f^s, e^s with scalings and biases s^f, b^f and s^e, b^e

Proposed Method

Stage I: Complementary appearance fusion(CAF)



Stage I



AAFB

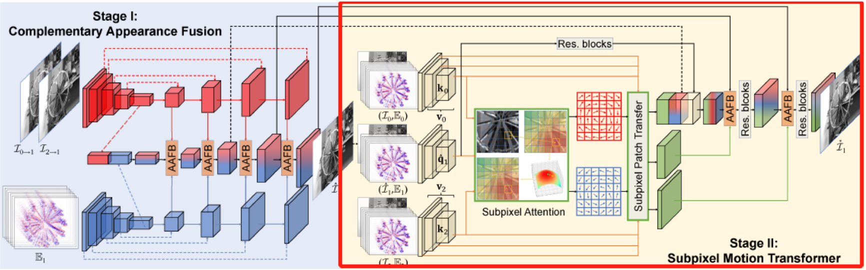
- Adaptive Appearance Fusion Block(AAFB):

$$y^e = \left(\frac{x_{\uparrow}^{s} - \mu\left(x_{\uparrow}^{s}\right)}{\sigma\left(x_{\uparrow}^{s}\right)} \right) \odot s^e + b^e \tag{1}$$

$$y = y^e \odot \mathbf{m} + y^f (1 - \mathbf{m}) \tag{2}$$

Proposed Method

Stage II: Subpixel Motion Transformer(SMT)



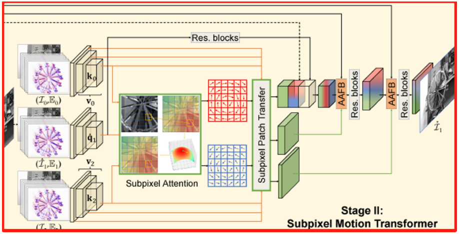
- Prepare works:
  - $(\mathcal{I}, \mathbb{E}) \xrightarrow{conv} \{\mathrm{v}^s \mid s \in \{0,1,2\}\}$ as values
  - $\mathrm{v}^2 \xrightarrow{clone} \mathrm{k}_0,\ \mathrm{k}_2$ as keys, $\hat{\mathrm{k}}_1$ as query
  - relevance measure:

$$D_0(\mathrm{i}, \mathrm{p}) = \left\| \frac{\hat{\mathrm{k}}_1(\mathrm{i})}{\left\|\hat{\mathrm{k}}_1(\mathrm{i})\right\|_2} - \frac{\mathrm{k}_0(\mathrm{i}+\mathrm{p})}{\|\mathrm{k}_0(\mathrm{i}+\mathrm{p})\|_2} \right\|_2^2 \tag{1}$$

$\mathrm{p} \in [-m, m]^2$ represents a spatial offset.

Proposed Method

Stage II: Subpixel Motion Transformer(SMT)



- Subpixel attention learning:

$$\mathrm{j} = \mathrm{i} + \mathrm{p}^* \text{ where } \mathrm{p}^* = \arg\min_{\mathrm{p}} \mathrm{D}_0(\mathrm{i}, \mathrm{p}) \tag{1}$$

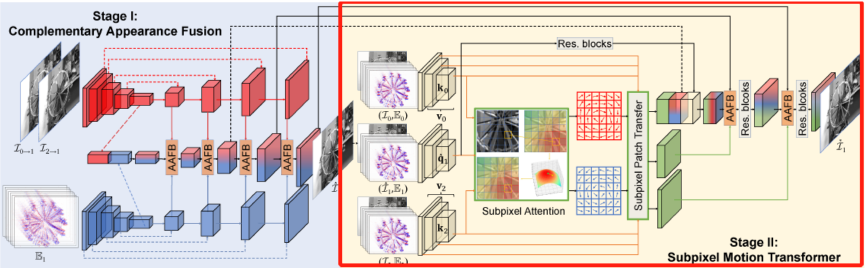
$$\mathrm{d}(\mathrm{u}) = \mathrm{D}_0\left(\mathrm{i}, \mathrm{p}^* + \mathrm{u}\right), \mathrm{u} \in \mathbb{Z}^2 \cap [-n, n]^2 \tag{2}$$

$$\mathrm{d}(\mathrm{u}) \approx \hat{\mathrm{d}}(\mathrm{u}) = \frac{1}{2}\mathrm{u}^{\mathrm{T}}\mathrm{A}\mathrm{u} + \mathrm{b}^{\mathrm{T}}\mathrm{u} + c \tag{3}$$

$$\min_{\mathrm{A},\mathrm{b},c} \sum_{\mathrm{u}} \mathrm{w}(\mathrm{u}) \|\hat{\mathrm{d}}(\mathrm{u}) - \mathrm{d}(\mathrm{u})\|^2 \tag{4}$$

Proposed Method

Stage II: Subpixel Motion Transformer(SMT)

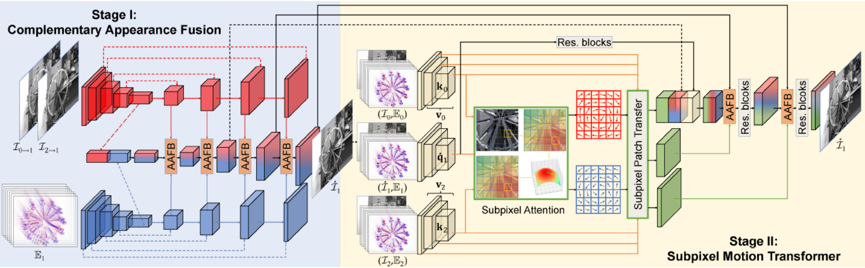


- Subpixel patch transfer:
  - distance means relevalce($K \cdot Q$), then mutiply with V, get transferred values z_0^s and z_2^s
  - select transferred values

$$z_1^s(\mathrm{i}) = \begin{cases} z_0^s(\mathrm{i}), & \text{if } \mathrm{D}_0\left(\mathrm{i}, \mathrm{p}_0^*\right) < \mathrm{D}_2\left(\mathrm{i}, \mathrm{p}_2^*\right) \\ z_2^s(\mathrm{i}), & \text{otherwise.} \end{cases} \tag{1}$$

  - fuse with previous features

Proposed Method

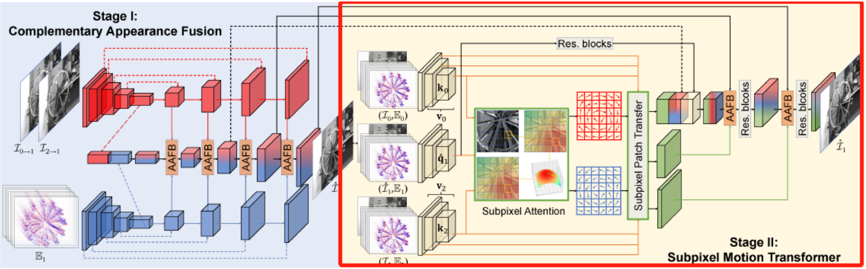


- Settings:
  - Datasets: GroPro(720p, 240FPS), SloMo-DVS dataset(self-collected) Adopt 10×, so sample 1th, 11th, 21th frames to form a sparse training triplet to train.
  - Loss: Charbonnier Loss(superior to ℓ_1, ℓ_2 loss)

$$\text{Charbonnier\_loss} = \sqrt{x^2 + \epsilon^2} \tag{1}$$

Proposed Method

Thoughts



- Frame info still rely on optical flow, not end-to-end.
- Process event in image format, point cloud or 3D conv works?
- 240FPS with sample rate 10 $\xLeftrightarrow{\textit{versus}}$ 25FPS with sample rate 1, which is weakly supervision?

Experiments

Quatitative results

Table 1. Comparing models on GoPro dataset, measured in PSNR and SSIM. Bold indicates the top place while underline the second.

Supervision	High FPS videos					High FPS videos + events					Low FPS videos + events	
Methods	SloMo[14]	QVI[48]	DAIN[3]	TAMI†[7]	FLAVR[16]	ETV[38]	SloMo*[14]	QVI*[48]	EMD[15]	EDVI[21]	BHA[33]	Proposed
PSNR	27.79	29.54	27.30	32.91	31.10	32.25	32.79	<u>33.07</u>	29.67	30.90	28.49	**33.33**
SSIM	0.838	0.872	0.836	**0.943**	0.917	0.925	<u>0.940</u>	**0.943**	0.927	0.905	0.920	<u>0.940</u>

† TAMI also adopts external private datasets for training. ∗ Enhanced variants with events added into the inputs of network.

Table 2. Comparing models on SloMo-DVS dataset, measured in PSNR and SSIM. Bold indicates the top place while underline the second.

Supervision	High FPS videos				High FPS videos + events				Low FPS videos + events	
Methods	SloMo[14]	QVI[48]	DAIN[3]	FLAVR[16]	ETV[38]	SloMo*[14]	QVI*[48]	EDVI[21]	BHA[33]	Proposed
PSNR	30.69	30.93	30.38	30.79	32.06	33.46	<u>33.70</u>	33.60	22.95	**34.17**
SSIM	0.915	0.920	0.914	0.917	0.936	0.950	**0.953**	0.948	0.828	<u>0.952</u>

Experiments

Abalation Study

Table 3. Performance in PSNR with low frame-rate training.

Method	SloMo*[14]		QVI*[48]		Proposed
Frame rate	High	Low	High	Low	Low
GoPro	32.79	31.40	33.07	29.88	33.33
SloMo-DVS	33.46	32.76	33.70	31.80	34.17

Table 4. Analysing the performance of CAF network.

Setting	PSNR	SSIM
Replacing AFFB with cat.+conv.	32.27	0.930
Using image branch only	29.43	0.882
Using event branch only	31.37	0.927
Full model	32.47	0.929

Table 5. Analysing the performance of SMT network.

ID	Key type	Value type	Att. type	Fused stage	PSNR
1	img.+evt.	image	subpix.	both	32.72
2	img.+evt.	event	subpix.	both	32.91
3	image	img.+evt.	subpix.	both	33.01
4	event	img.+evt.	subpix.	both	33.03
5	img.+evt.	img.+evt.	subpix.	first	33.00
6	img.+evt.	img.+evt.	subpix.	second	32.56
7	img.+evt.	img.+evt.	hard	both	33.02
8	img.+evt.	img.+evt.	soft	both	32.50
9	img.+evt.	img.+evt.	subpix.	both	33.33

Experiments

Qualitative results

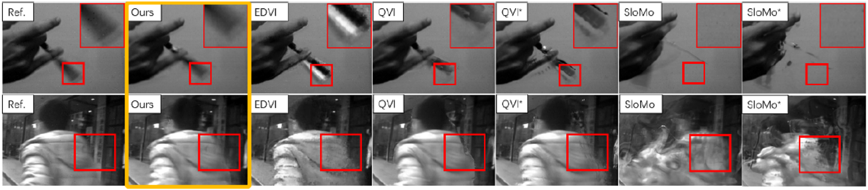


Figure 5. Qualitative comparisons on real data. In the first column (Ref.) we visualize the nearest input frame as reference since there is no groundtruth. We suggest the readers to watch our supplementary video for more qualitative comparisons on real-world video interpolation.

Experiments

Qualitative results

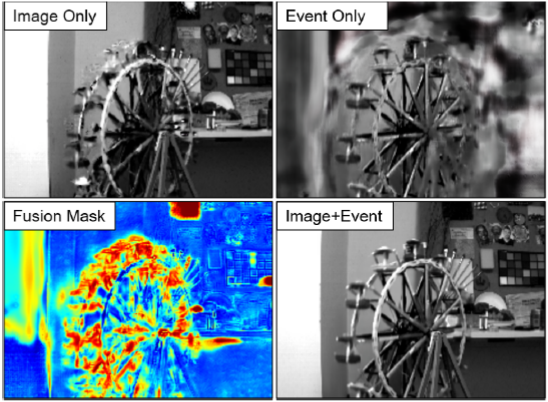


Figure 6. Visualizing the impact of adaptively fusing image and event appearance features in the CAF network.

Experiments

Qualitative results

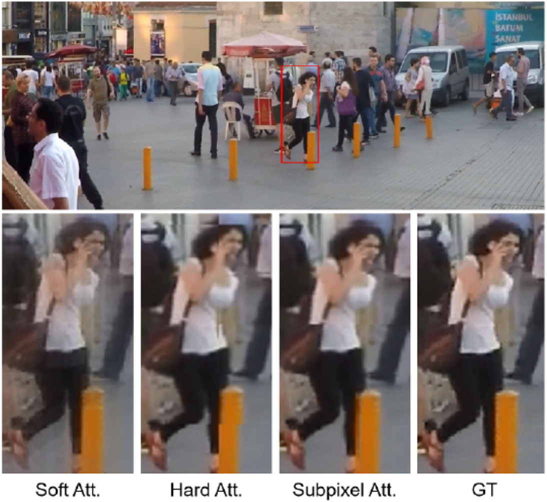
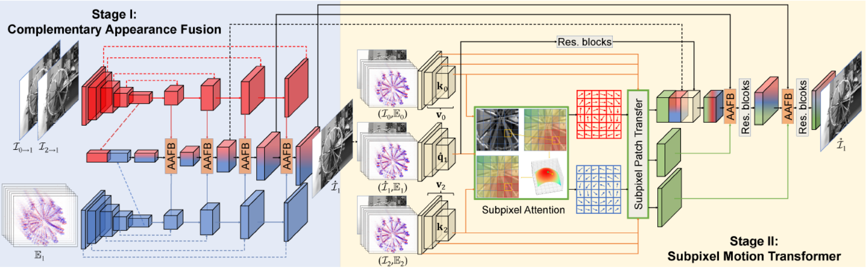


Figure 7. Patch transfer results with different types of attention.

Conclusions



- Train with low frame-rate videos, but outperform others and generalize well.
- Attention mechanism, no direct motion modelling, eg optical flow.

Thanks, Q & A