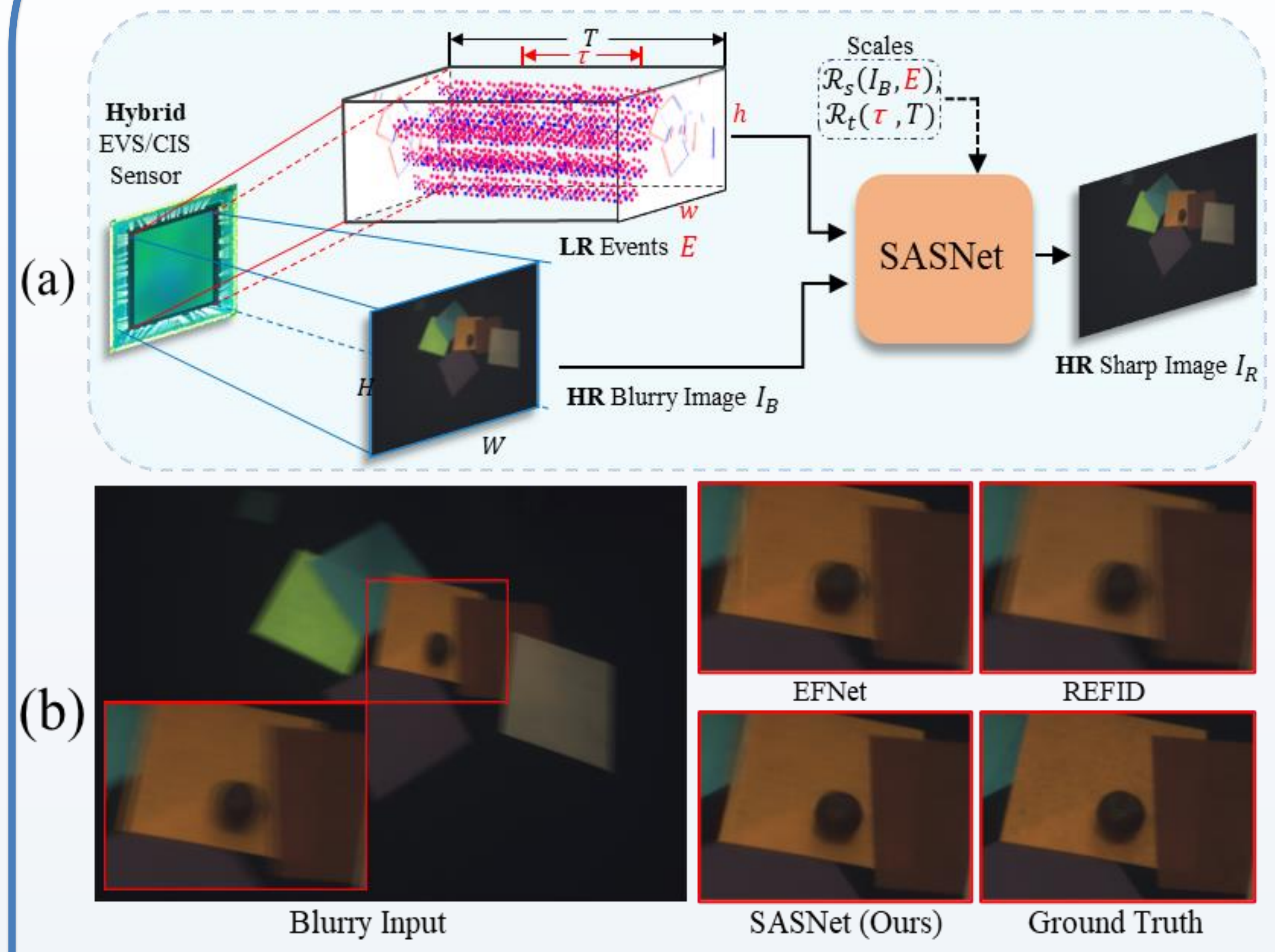
Learning Scale-Aware Spatio-temporal Implicit Representation for Event-based Motion Deblurring

Wei Yu[1], Jianing Li[2], Shengping Zhang[1], Xiangyang Ji[3],
[1]Harbin Institute of Technology, [2]Peking University, [3]Tsinghua University

ICML International Conference On Machine Learning

Introduction



Limitation:

- **Dataset Limitation**: Current event-based deblurring datasets are usually collected from cameras with low spatial resolution (e.g., DAVIS346 with 346×260) or binocular camera systems using beam splitters, which are cumbersome and inaccurate due to the artificial spatial alignment and time synchronization of CIS and EVS.
- **Algorithm Limitation**: Existing algorithms always assume that the inputs of CIS images and EVS events have the same spatial (i.e., resolution) and temporal (i.e., exposure duration) scales, which are confined by the scale differences of different shooting equipment and environments in practice, as shown in Fig (b).

Contributions:

- We build a real event-based motion deblurring dataset, High-resolution Hybrid Deblur (H2D), with naturally spatially aligned and temporally synchronized events at various scales using a novel hybrid EVS/CIS sensor in Fig (a).
- We first investigate the arbitrary-scale event-based motion deburring problem and propose a Scale-Aware Spatio-temporal Deblurring Network (SASNet) to restore unknown highly blurred areas and eliminate global motion blur with varying magnitudes.

Methodology

Arbitrary-scale Event-based Motion Deblurring:

The input HR blurry image within the shutter period T

$$I_B(t) = \frac{1}{T}\int_{t\in(0,T)} I(t)dt$$

The input LR event streams within the exposure duration τ

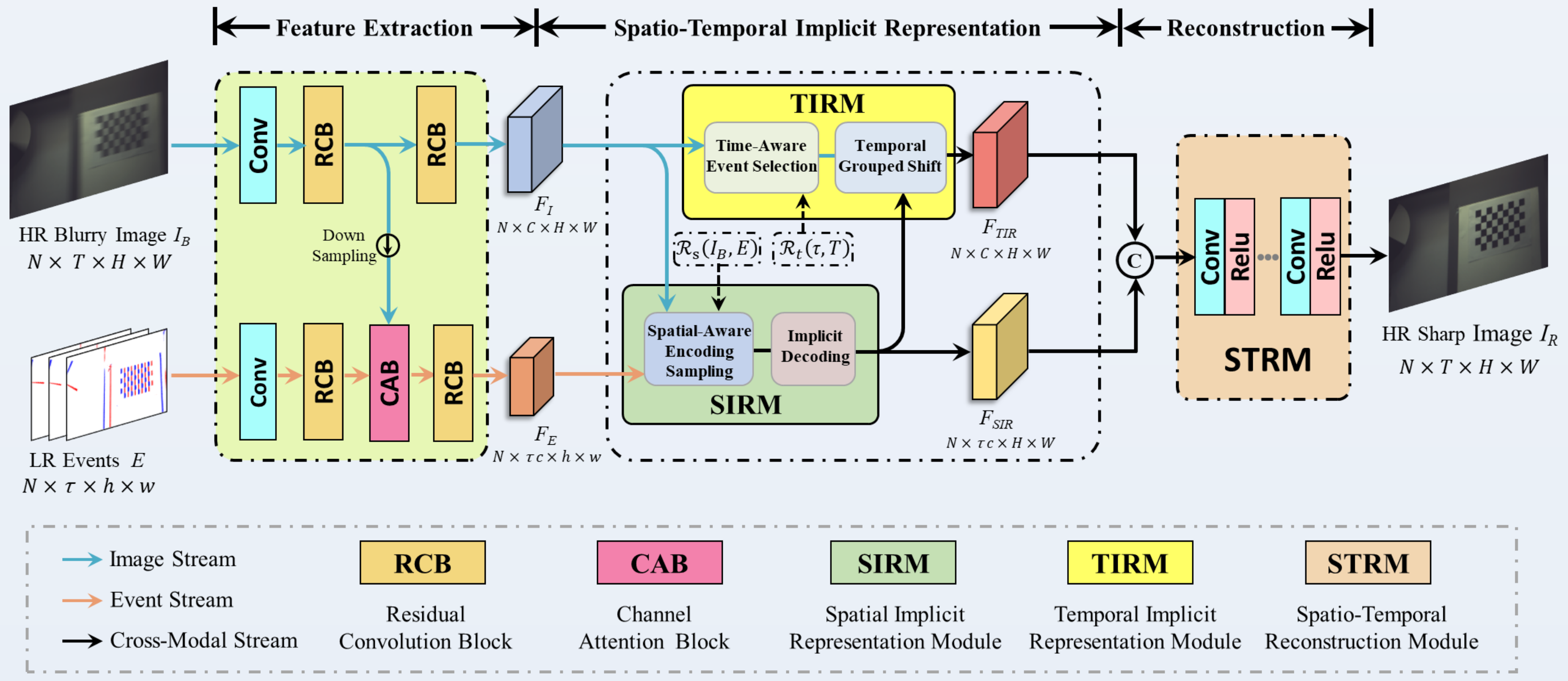
$$E(t,\tau) = \frac{1}{\tau}\int_{t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} \exp(c\int_{t-\Delta t}^{t} p(s)ds)dt \quad \forall t,\tau\in(0,T)$$

The output HR sharp image restored from arbitrary spatial and temporal scales

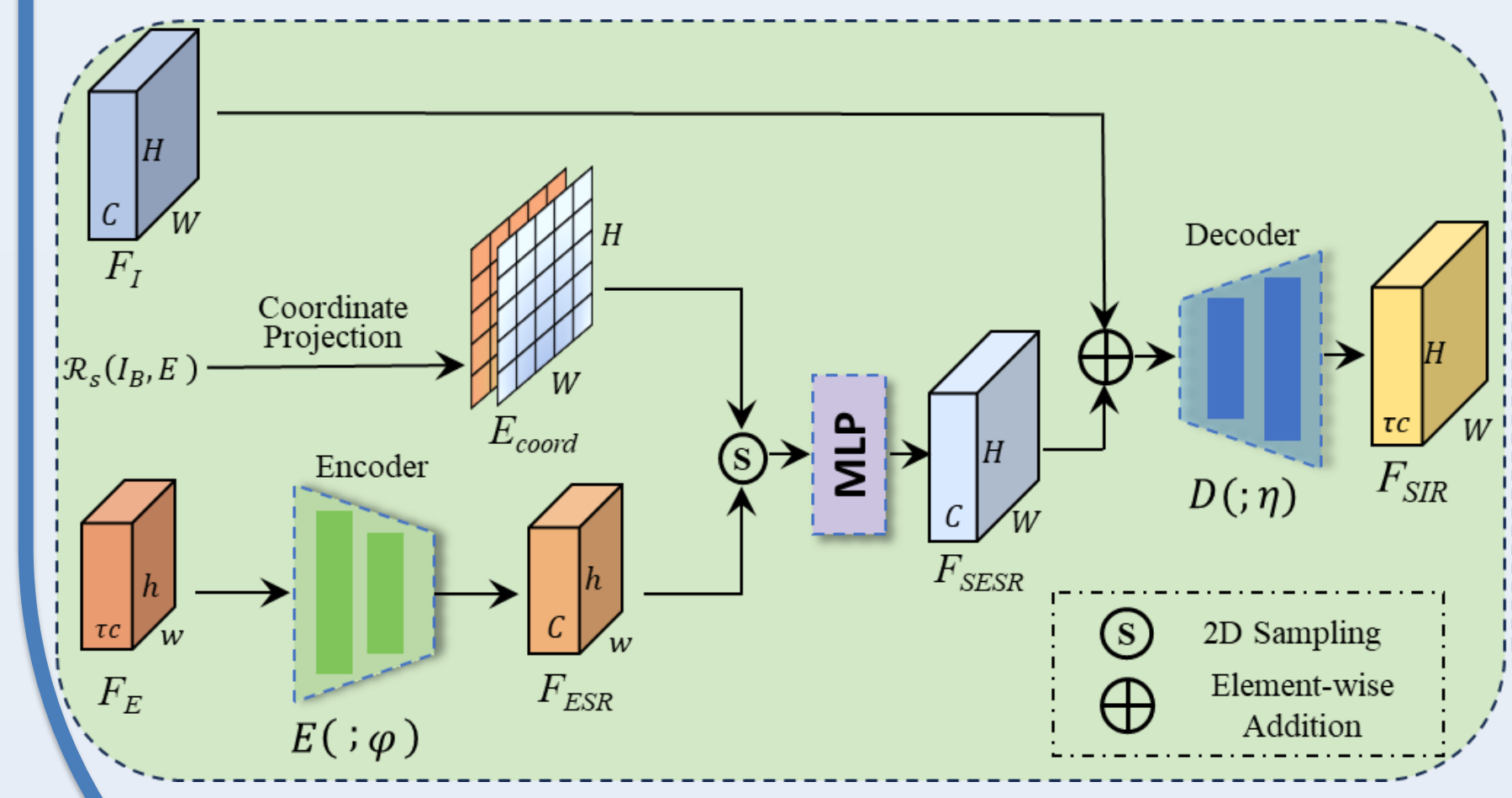
$$I_R(t) = \mathrm{Deblur}(I_B(t), E(t,\tau); \mathcal{R}_s, \mathcal{R}_t)$$
$$\forall t,\tau \in T, \quad \mathcal{R}_s \in [1,4], \quad \mathcal{R}_t \in (0,1]$$

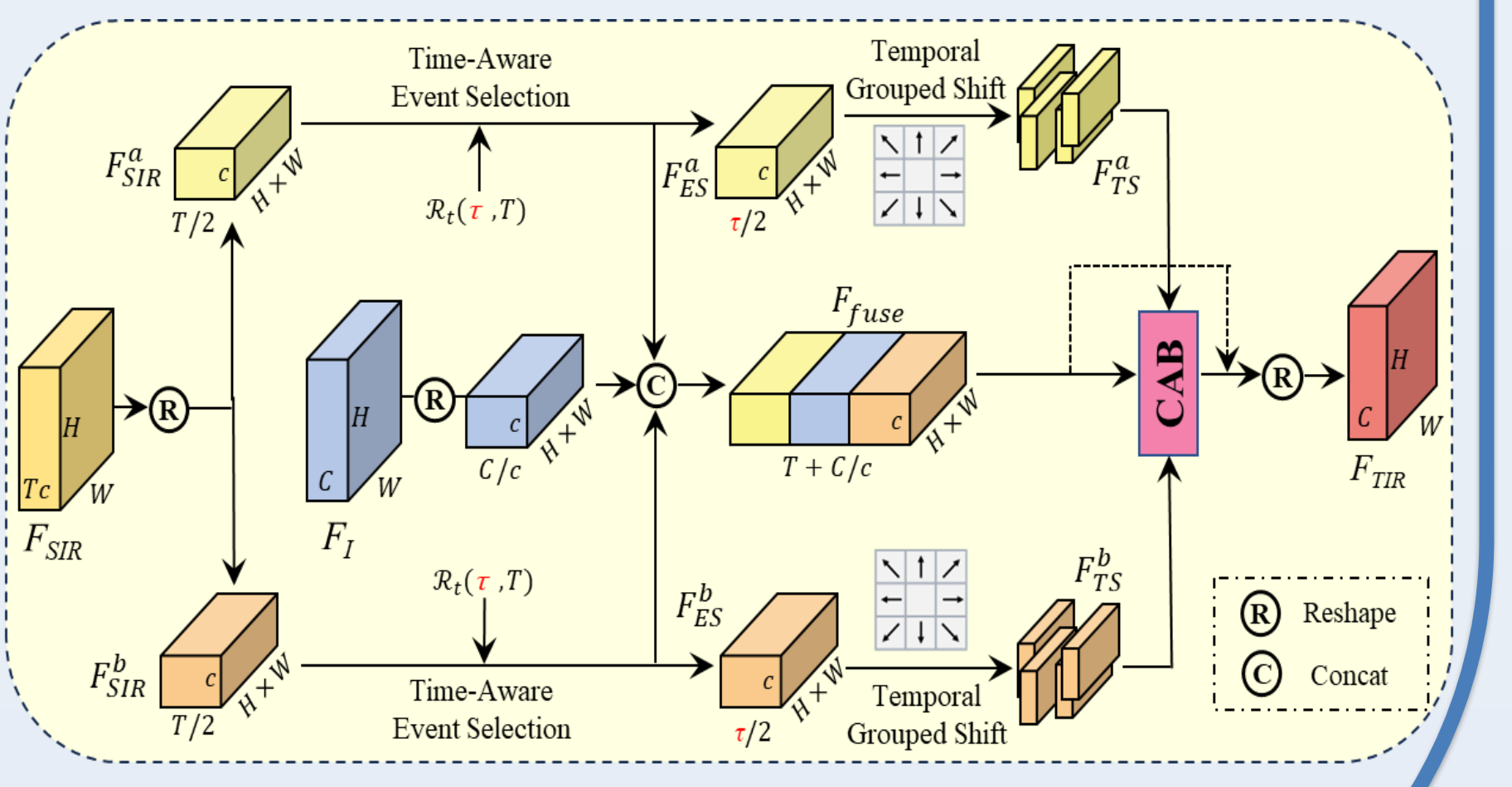
Framework:

- SASNet implicitly aggregates both spatial and temporal correspondence features of images and events to generalize at continuous scales.
- SIRM aggregates spatial correlation at any resolution through event encoding sampling to restore highly blurred local areas.
- TIRM learns temporal correlation via temporal shift operations with long-term aggregation to tackle global motion blur.


Feature Extraction
Spatio-Temporal Implicit Representation
Reconstruction
HR Blurry Image I_B
$N\times T\times H\times W$
Conv
RCB
RCB
Down Sampling
Conv
RCB
CAB
RCB
LR Events E
$N\times\tau\times h\times w$
F_I
$N\times C\times H\times W$
F_E
$N\times\tau c\times h\times w$
TIRM
Time-Aware Event Selection
Temporal Grouped Shift
$\mathcal{R}_s(I_B,E)$
$\mathcal{R}_t(\tau,T)$
Spatial-Aware Encoding Sampling
Implicit Decoding
SIRM
F_{TIR}
$N\times C\times H\times W$
F_{SIR}
$N\times\tau c\times H\times W$
C
Conv
Relu
Conv
Relu
STRM
HR Sharp Image I_R
$N\times T\times H\times W$
Image Stream
Event Stream
Cross-Modal Stream
RCB
Residual Convolution Block
CAB
Channel Attention Block
SIRM
Spatial Implicit Representation Module
TIRM
Temporal Implicit Representation Module
STRM
Spatio-Temporal Reconstruction Module

Overview of our SASNet

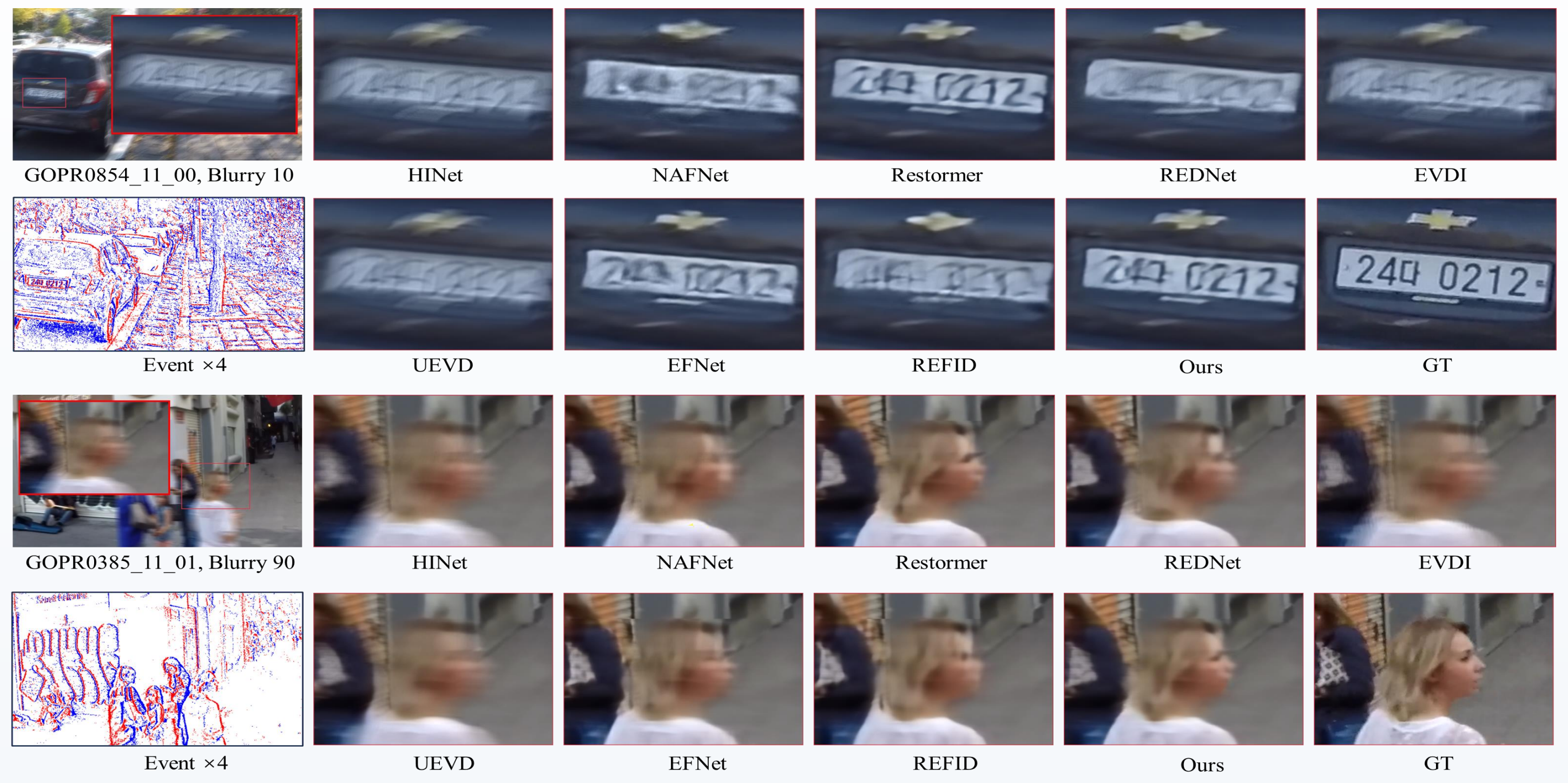




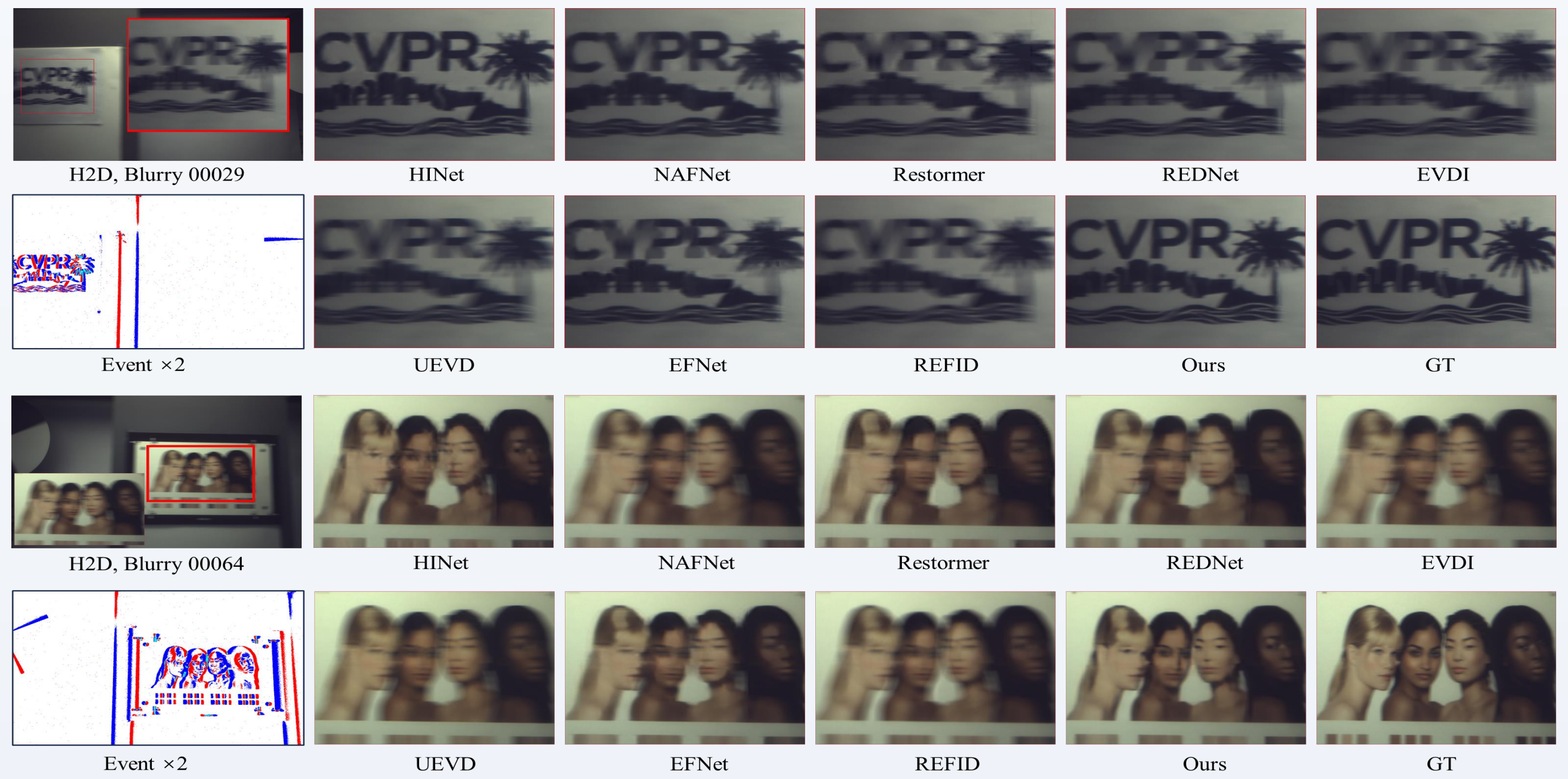
Spatial Implicit Representation Module （SIRM）

Temporal Implicit Representation Module （TIRM）

Qualitative Results



Synthetic GoPro Dataset



Real H2D Dataset

Quantitative Results

Method	Input	GOPRO ($R_s=4, R_t=1$) PSNR↑	GOPRO ($R_s=4, R_t=1$) SSIM↑	H2D ($R_s=2, R_t=1$) PSNR↑	H2D ($R_s=2, R_t=1$) SSIM↑	Complexity #Params	Complexity #FLOPs
HINet (Chen et al., 2021a)	Image	25.41	0.7958	32.57	0.9394	88.67M	170.55G
NAFNet (Chen et al., 2022)	Image	27.31	0.8426	32.81	0.9414	16.01M	16.06G
Restormer (Zamir et al., 2022)	Image	28.37	0.8731	33.39	0.9426	26.13M	140.99G
RED-Net (Xu et al., 2021)	Image + Events	27.19	0.8382	33.98	0.9458	9.70M	159.01G
EVDI (Zhang & Yu, 2022)	Image + Events	25.84	0.8069	32.94	0.9432	0.39M	35.54G
UEVD (Kim et al., 2022)	Image + Events	25.69	0.8231	31.98	0.9377	14.23M	101.60G
EFNet (Sun et al., 2022)	Image + Events	28.08	0.8661	34.59	0.9501	8.47M	111.06G
REFID (Sun et al., 2023)	Image + Events	27.51	0.8473	32.61	0.9347	88.96M	208.98G
Ours	Image + Events	28.82	0.8811	35.72	0.9541	1.46M	43.35G

Quantitative comparison

Method	Type	PSNR↑/SSIM↑	#Params	#FLOPs
Interpolation	Explicit	28.36/0.8699	1.461M	43.26G
Transposed Conv	Explicit	28.12/0.8639	1.470M(+0.009)	43.85G(+0.59)
Pixel Shuffle	Explicit	28.43/0.8701	1.610M((+0.149))	43.87G(+0.61)
Learnable Upsample	Implicit	28.61/0.8741	1.688M(+0.227)	43.86G(+0.60)
SIRM (Ours)	Implicit	28.82/0.8811	1.462M(+0.001)	43.35G(+0.09)

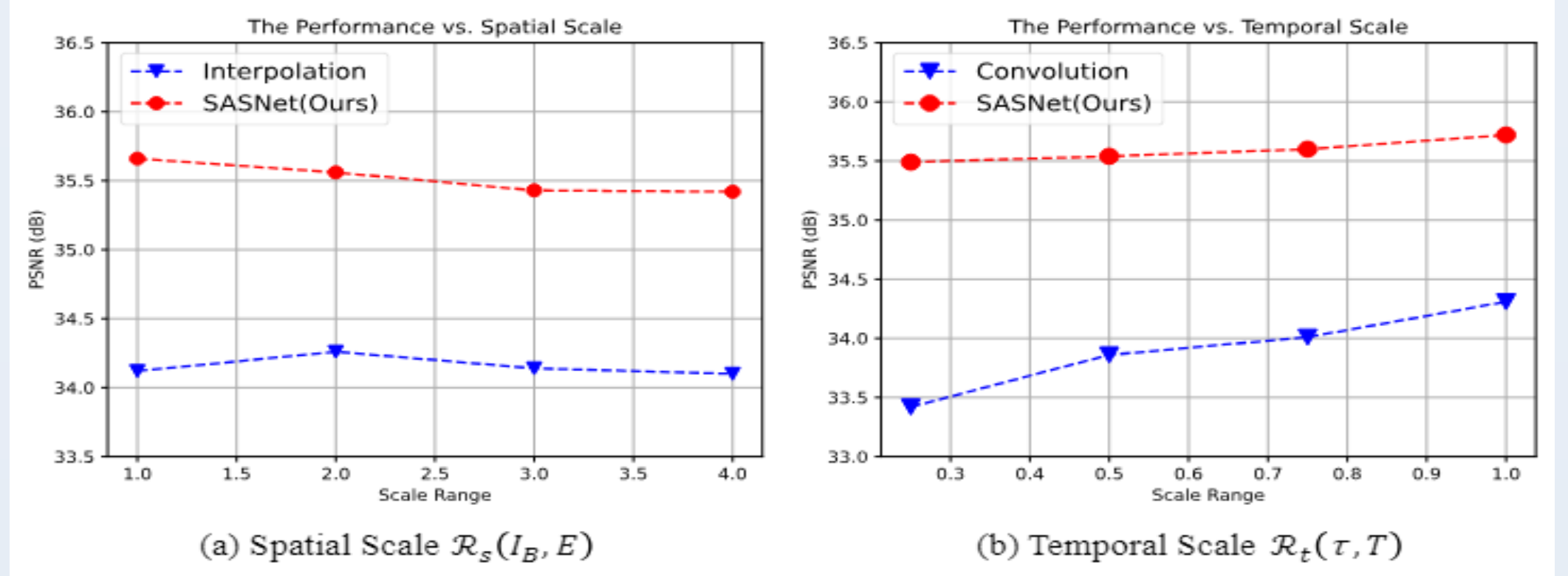
Comparison of different spatial representation

Method	Type	PSNR↑/SSIM↑	#Params	#FLOPs
Convolution	Implicit	27.72/0.8631	1.372M	40.39G
Optical Flow	Explicit	27.99/0.8678	1.5387M(0.1667)	51.30G(+10.91)
Deformable Conv	Explicit	28.26/0.8759	1.423M(+0.051)	40.44G(+0.05)
VIT Attention	Implicit	28.84/0.8803	2.514M(+1.142)	77.83G(+37.44)
TIRM (Ours)	Implicit	28.82/0.8811	1.462M(+0.090)	43.35G(+2.96)

Comparison of different temporal representation

Dataset	Color	Camera	Image Resolution	Event Resolution	Type of Scenes	SA	TS	HR
BS-ERGB	RGB	FLIR + Prophesee Gen4	970 × 625	970× 625	Low Speed	✗	✗	✓
THU-HSEVI	Gray	EoSens + DAVIS346	340× 260	340× 260	High Speed	✗	✗	✗
DAVIS 240C Dataset	Gray	DAVIS246	240× 180	240× 180	Low Speed	✓	✓	✗
HQF	Gray	DAVIS240C	346× 260	346× 260	Low Speed	✓	✓	✗
Ours (H2D)	RGB	OV60B	1920× 1080	960× 540	High Speed	✓	✓	✓

Comparison of H2D with other event-based deblurring datasets



(a) Spatial Scale $\mathcal{R}_s(I_B,E)$

(b) Temporal Scale $\mathcal{R}_t(\tau,T)$

Quantitative results at different spatial and temporal scales