EvSTVSR: Event Guided Space-Time Video Super-Resolution

Haojie Yan[1,2], Zhan Lu[3], Zehao Chen[1,2], De Ma[1,2*], Huajin Tang[1,2], Qian Zheng[1,2*], Gang Pan[1,2]

[1]The State Key Lab of Brain-Machine Intelligence, Zhejiang University, China
[2]College of Computer Science and Technology, Zhejiang University, China
[3] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
{hjyan,zehao,made,htang,qianzheng,gpan}@zju.edu.cn, zhan007@e.ntu.edu.sg

Abstract

In the domain of space-time video super-resolution, it is typically challenging to handle complex motions (including large and nonlinear motions) and varying illumination scenes due to the lack of inter-frame information. Leveraging the dense temporal information provided by event signals offers a promising solution. Traditional event-based methods typically rely on multiple images, using motion estimation and compensation, which can introduce errors. Accumulated errors from multiple frames often lead to artifacts and blurriness in the output. To mitigate these issues, we propose EvSTVSR, a method that uses fewer adjacent frames and integrates dense temporal information from events to guide alignment. Additionally, we introduce a coordinate-based feature fusion upsampling module to achieve spatial super-resolution. Experimental results demonstrate that our method not only outperforms existing RGB-based approaches but also excels in handling large motion scenarios.

Code — https://github.com/hjyyyd/EvSTVSR.

Introduction

Visual information in the real world is inherently continuous in both spatial and temporal dimensions. However, image sensor size, cost, transmission bandwidth, and storage capacity limitations often result in recorded visual data with low spatiotemporal resolution. Acquiring high-resolution(HR) visual information is crucial for both practical applications and enhancing downstream vision tasks. But how can we reconstruct high space-time resolution visual data from low-resolution(LR) inputs?

Several RGB-based approaches have recently explored this challenge, highlighting the mutual benefits of jointly addressing temporal and spatial super-resolution tasks. For instance, Zooming SlowMo (Xiang et al. 2020), and TMNet (Xu et al. 2021) use Bi-directional Deformable ConvLSTM for spatiotemporal feature fusion, while VideoINR (Chen et al. 2022) and MoTIF (Chen Y H et al. 2023) achieve alignment through spatial and temporal implicit neural representations(INR). Their critical distinction lies in their alignment strategies: the former employs deformable convolutions for implicit alignment, whereas the latter uses optical flow for explicit alignment. Aligning reference frames to the target frame is essential for video frame interpolation(VFI) (Kim et al. 2023), video super-resolution(VSR) (Liu et al. 2022), and space-time video super-resolution(STVSR) (Chen et al. 2022). However, the lack of inter-frame information in RGB-based methods hinders accurate alignment under complex motions, such as large and nonlinear movements. Inaccurate alignment introduces errors, often leading to artifacts and blurring(as shown in (c) of Figure 1).

To tackle these challenges, recent studies (Tulyakov et al. 2021, 2022; He et al. 2022; Kim et al. 2023) in VFI have introduced the use of event cameras to address the issue of motion loss between frames. However, there is still a lack of research utilizing event alignment in VSR and STVSR to accomplish these tasks. As an emerging type of visual sensor, the event camera possesses microsecond-level temporal resolution (Chen et al. 2021, 2024). It is sensitive to edges, allowing it to accurately record motion trajectories even under large movements (see Figure 1(b)). Consequently, using event signals for alignment is an intuitive approach. Moreover, event cameras capture dense temporal information (see Figure 1(f)). This dense temporal data can be transformed into dense spatial information (Jing et al. 2021), offering finer details to enhance RGB frames for the reconstruction of high-quality video results (see Figure 1(h)).

Therefore, in this paper, we present a method that leverages two RGB frames along with the events between them to accomplish the task. Our contributions are as follows,

- We propose a framework that addresses both super-resolution and frame interpolation in large-motion scenarios using fewer input frames. Our results demonstrate that reducing the number of input frames, even without optical flow, mitigates blurring and ghosting effects common in multi-frame and event-based methods while maintaining competitive performance.
- To handle super-resolution and frame interpolation with limited frame input, we introduce a method that combines optical flow prediction from dense images and sparse events, coupled with an implicit sampling strategy that fuses features based on positional coordinates. This approach effectively solves the space-time super-resolution task.

[*]Corresponding author.

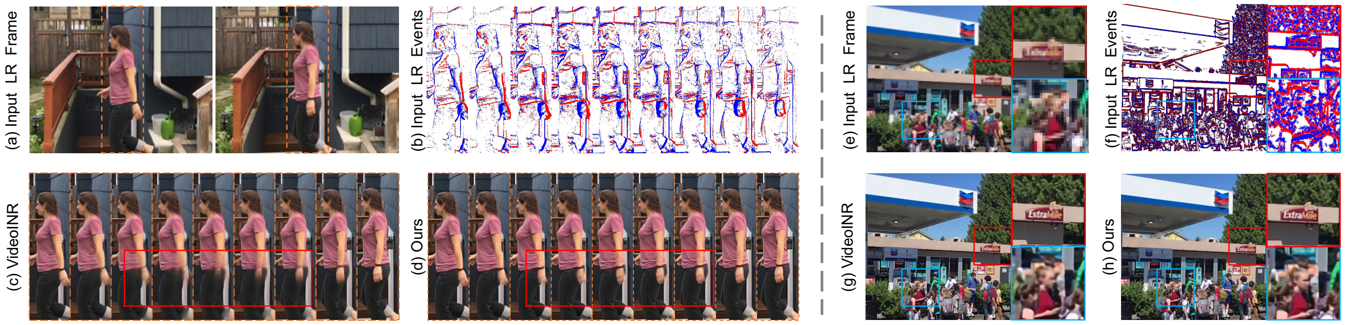

Figure 1: Qualitative comparison between our method and the RGB-based approach, **VideoINR**. The results demonstrate that in scenes with extensive object movement, our method is adept at handling large, nonlinear motions, as illustrated by the movement of the woman's hand in (d). Furthermore, in scenes with slow changes, by incorporating dense temporal information from event signals, our approach efficiently enhances the recovery of spatial details, such as the text on the wall and the girl's facial features in (h).

- Our experiments show that our method achieves state-of-the-art performance in event-guided space-time super-resolution, as well as in both temporal and spatial super-resolution tasks.

Related Work

Video Frame Interpolation

RGB-based Video Frame Interpolation (VFI) has made significant progress. Both supervised methods (Jiang et al. 2018) and unsupervised methods (Zhu et al. 2019) utilize optical flow to learn motion. However, due to the blind time between consecutive frames, they typically artificially add a space-time directional smoothness prior, such as the linear motion assumption, which performs poorly in complex motion scenarios.

Event-based VFI methods (Tulyakov et al. 2021; Kim et al. 2023), achieves more accurate motion estimation in complex motion scenarios. TimeLens (Tulyakov et al. 2021) directly estimates optical flow from events and warps adjacent frames using the estimated optical flow. However, due to the lack of image information, the optical flow estimation is inaccurate in regions with sparse events. TimeLens++ (Tulyakov et al. 2022) uses parameterized equations to estimate motion, and CBMNet (Kim et al. 2023) estimates asymmetric inter-frame motion fields using edge information from events and texture information from images. These methods can achieve excellent results in complex motion and varying illumination scenarios.

Video Super-Resolution

The main difference between video super-resolution (VSR) and image super-resolution (ISR) is that VSR leverages temporal information to enhance spatial details. Traditional VSR methods, which rely on motion estimation and compensation, often use multiple frames, but this complicates optical flow estimation in large motion scenarios, leading to performance degradation as the number of low-resolution frames increases.

Event-based VSR methods (Kai et al. 2024; Xiao et al. 2024b,a; Han et al. 2021) enhance RGB-based approaches by incorporating event signals as auxiliary inputs. EBVSR (Kai 2023) and EvTexture (Kai et al. 2024) uses a pre-trained optical flow model, SPyNet, for spatial alignment and pixel-shuffle for upsampling, while (Jing et al. 2021) integrates high temporal resolution events into the VSR framework through adaptive threshold learning without explicit motion learning. Another approach, (Lu et al. 2023), employs a sliding-window method to learn a space-time implicit representation for VSR, also without explicit motion learning. However, these methods either require more than five input frames or lack explicit motion learning, resulting in poor performance in scenarios involving large motions.

Space-Time Video Super-Resolution.

RGB-based Space-Time Video Super-Resolution(STVSR) methods learn motion by estimating optical flow. This work (Chen et al. 2022) achieves motion compensation through backward warping, while (Chen Y H et al. 2023) uses forward warping. However, due to the lack of inter-frame information, optical flow estimation becomes deficient in large motion scenarios, leading to artifacts in the final results under significant motion.

To the best of our knowledge, the HR-INR (Lu et al. 2024) approach is the only event-based method for STVSR. HR-INR utilizes multiple frames (four) along with event data to extract space-time information, which is then decoded using an INR-based method. However, its failure to explicitly capture motion from events leads to artifacts such as ghosting and color distortion. Moreover, while multiple frames improve detail, they complicate alignment in large-motion scenarios, often resulting in blurring and ghosting.

In contrast, our method reduces the risk of ghosting by using fewer input frames, mitigating the error accumulation seen in multi-frame approaches, especially in cases involving large, nonlinear motion. By integrating event signals, we enhance motion estimation and preserve high-frequency details. Our approach combines optical flow-based warping and synthesis techniques, effectively compensating for the

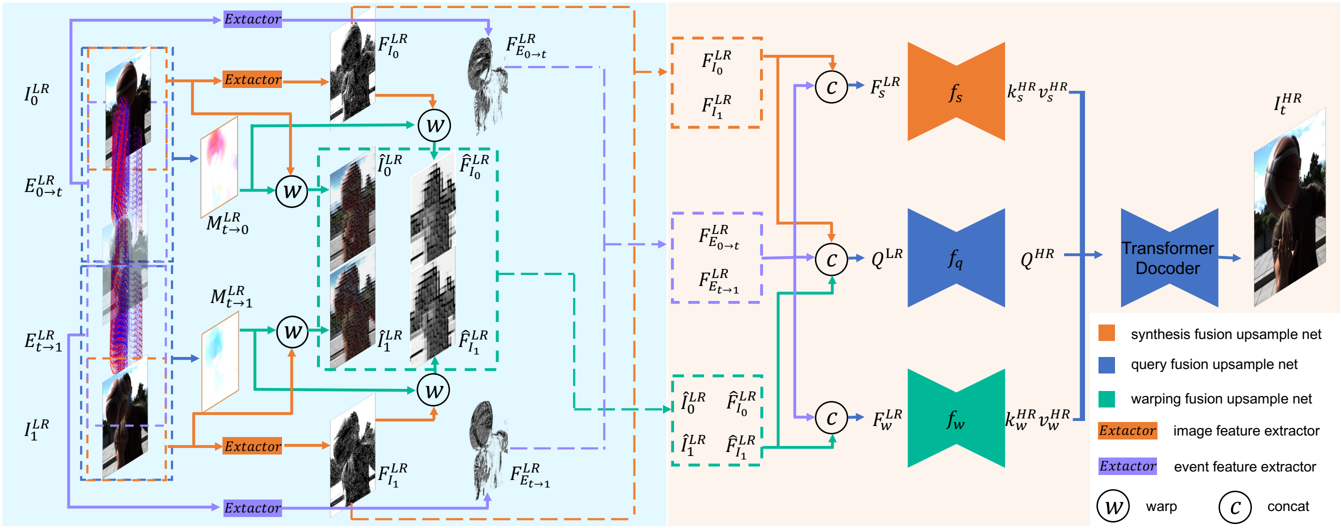

Figure 2: Overall architecture of the Warp and Synthesis Fusion Network with a positional-coordinate-based upsampling network for space-time super-resolution. The left side shows the event-guided alignment of images and image features to the target time. The right side depicts the synthesis of the final high-resolution frame, where the aligned images and event features are processed through a positional-coordinate-based upsampling network and a transformer decoder.

reduced number of frames and avoiding the noise and artifacts caused by multi-frame fusion.

Preliminary

Event Voxel Representation

Event signals are composed of a stream of quadruples $\mathcal{E} = \{e_i\}_{i=1}^{N_e}$, where $e_i = (t_i, x_i, y_i, p_i)$, t_i represent the timestamp, (x_i, y_i) represent coordinate, $p_i \in (-1, 1)$ represent the positive or negative polarity. Follow (Zhu et al. 2019), we convert event streams into a voxel grid representation $E \in \mathbb{R}^{B \times H \times W}$,

$$\mathbf{E}(x_l, y_m, t_n) = \sum_{\substack{x_i = x_l \\ y_i = y_m}} p_i \max(0, 1 - |t_n - t_i^*|), \qquad (1)$$

where $t_i^* \triangleq \frac{B-1}{\Delta T}(t_i - t_0)$, is the normalized event timestamp, B represents the number of bins in the temporal dimension. We select bin size B = 16 to balance the efficiency and temporal precision.

Proposed Methods

Problem Formulations and Overview

Problem Formulations. To address challenges such as handling large motions and complex temporal dynamics, we introduce events into the Space-Time Video Super-Resolution (STVSR) task. Specifically, our method takes two low-resolution video frames, $I_0^{LR}, I_1^{LR} \in \mathbb{R}^{3 \times H \times W}$, and low-resolution events, $\mathcal{E}_{0\to 1} = \{e_i\}_{i=1}^{N_e}$, as input. The task is to obtain a high-resolution frame $I_t^{HR} \in \mathbb{R}^{3 \times H' \times W'}$ at any arbitrary time $t \in [0, 1]$ with predefined super-resolution scale $s = H'/H = W'/W \geq 1$. We divide the events at any arbitrary time t into two parts and represent them as two voxel grids $E_{0\to t}^{LR}$ and $E_{t\to 1}^{LR}$. Thus, our solution for STVSR can be expressed as

$$I_t^{HR} = f(I_0^{LR}, I_1^{LR}, E_{t\to 0}^{LR}, E_{t\to 1}^{LR}, s). \qquad (2)$$

Overview. Our framework consists of two main components, as shown in Figure 2. The first module is for spatial alignment, where we obtain aligned images ($\hat{I}_0^{LR}, \hat{I}_1^{LR}$) and aligned image features ($\hat{F}_{I_0}^{LR}, \hat{F}_{I_1}^{LR}$) from the first and second reference frames. The second module integrates feature fusion using two distinct paths: one based on synthesis and the other on warping. It employs three positional-coordinate-based implicit upsampling networks to generate high-resolution query features and attention features. Finally, this module leverages an interactive attention-based frame synthesis network to produce the target high-resolution frame from the derived features.

Cross-Modal Spatial Alignment

In this module, as shown in Figure 2 (left), we extract the low-resolution frame feature $F_{I_0}^{LR}, F_{I_1}^{LR}$ and event feature $F_{E_{0\to t}}^{LR}, F_{E_{t\to 1}}^{LR}$ using weight-sharing frames and event encoders. The low-resolution motion flow fields $M_{t\to 0}^{LR}$ and $M_{t\to 1}^{LR}$ are obtained(inspired by the approach in (Kim et al. 2023)) from

$$(M_{t\to 0}^{LR}, M_{t\to 1}^{LR}) = \mathcal{M}(I_0^{LR}, I_1^{LR}, E_{t\to 0}^{LR}, E_{t\to 1}^{LR}). \qquad (3)$$

We use the motion information contained in the events and images to align the reference frames I_0^{LR}, I_1^{LR}, along with their features $F_{I_0}^{LR}, F_{I_1}^{LR}$ to the target time t. In formulation, $\hat{I}_0^{LR} = \mathcal{W}(I_0^{LR}, M_{t\to 0}^{LR})$, $\hat{I}_1^{LR} = \mathcal{W}(I_1^{LR}, M_{t\to 1}^{LR})$, $\hat{F}_{I_0}^{LR} = \mathcal{W}(F_{I_0}^{LR}, M_{t\to 0}^{LR})$, $\hat{F}_{I_1}^{LR} = W(F_{I_1}^{LR}, M_{t\to 1}^{LR})$, where $\mathcal{W}$ denotes backward warping operation.

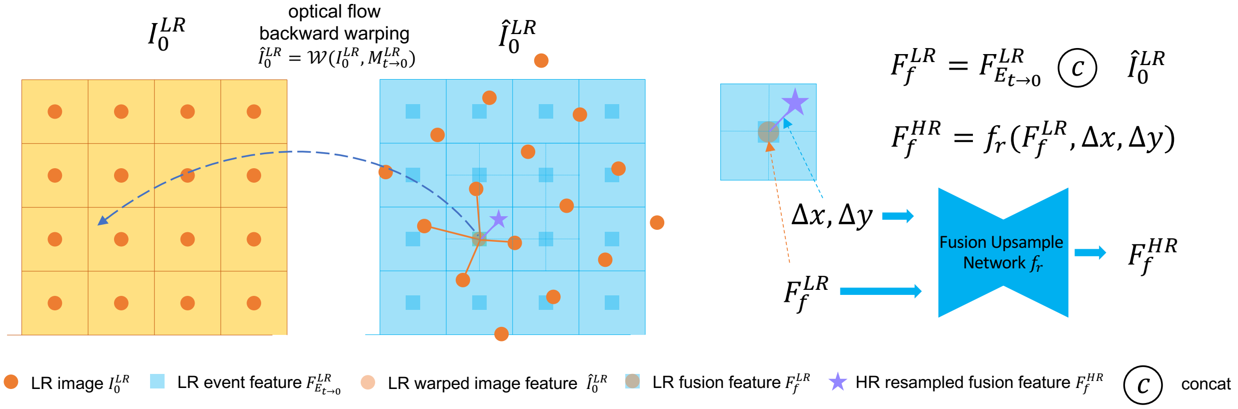
I_0^{LR}
optical flow
backward warping
$\hat{I}_0^{LR} = \mathcal{W}(I_0^{LR}, M_{t\to 0}^{LR})$
$\hat{I}_0^{LR}$
$F_f^{LR} = F_{E_{t\to 0}}^{LR}$ Ⓒ $\hat{I}_0^{LR}$
$F_f^{HR} = f_r(F_f^{LR}, \Delta x, \Delta y)$
$\Delta x, \Delta y$
Fusion Upsample Network f_r
F_f^{HR}
F_f^{LR}
● LR image I_0^{LR}
■ LR event feature $F_{E_{t\to 0}}^{LR}$
● LR warped image feature $\hat{I}_0^{LR}$
■ LR fusion feature F_f^{LR}
★ HR resampled fusion feature F_f^{HR}
Ⓒ concat

Figure 3: Using optical flow estimated from images and event data, low-resolution reference frames are warped to align with key frames. This process is further enhanced by concatenating low-resolution event features and utilizing a fusion upsampling network based on relative positional coordinates (Δx, Δy) with a SIREN architecture. This approach facilitates the effective integration and upsampling of features. Leveraging the superior high-frequency information processing capabilities of the SIREN architecture, the position-based fusion upsampling method preserves more high-frequency details compared to traditional bilinear interpolation. Such an enhancement is particularly advantageous for STVSR tasks.

Position-Based INR Resampling

We upsample the aligned features and decode them to generate high-resolution frames. First, we fuse the feature with two fusion modules, *i.e.*, warping-based fusion and synthesis-based fusion. Warping-based fusion relies on deformation alignment to handle large motions but struggles with occlusions and brightness variations. While synthesis-based fusion integrates events with images to address occlusions and lighting changes but might fail to manage large motions and can introduce noise due to the sparsity of events. We combine these two modules to complement each other.

Inspired by CBMNet (Kim et al. 2023), we adopt its feature-level fusion strategy, leveraging multi-head self-attention and cross-attention mechanisms to enhance feature aggregation across modalities. This enables more effective capture and integration of complex space-time features.

Spatial upsampling is essential in decoding. While traditional methods such as deconvolution and pixel shuffling are effective, INR-based upsampling, inspired by LIIF (Chen, Liu, and Wang 2021), has demonstrated superior performance. Therefore, we propose a fusion upsampling scheme before the transformer decoder. As shown in Figure 2 right, three independent upsampling modules are applied to the query, warp, and synthesis features, respectively, before they are fed into the decoder.

As illustrated in Figure 3, we demonstrate our strategy for obtaining high-resolution (HR) features using positional-based INR resampling with input features. We concatenate the input features and then feed them, along with the pixel's relative coordinates $\Delta x, \Delta y$, into a SIREN-based MLP (Sitzmann et al. 2020). This process enables us to query the fused feature of the pixel in the high-resolution frame. Formally, this can be expressed as,

$$\begin{aligned}
F_s^{LR} &= \text{Concat}(F_{E_{0\to t}}^{LR}, F_{E_{t\to 1}}^{LR}, F_{I_{0,1}}^{LR}),\\
F_w^{LR} &= \text{Concat}(F_{E_{0\to t}}^{LR}, F_{E_{t\to 1}}^{LR}, \hat{F}_{I_{0,1}}^{LR}, \hat{I}_{0,1}^{LR}),\\
Q^{LR} &= \text{Concat}(F_{E_{0\to t}}^{LR}, F_{E_{t\to 1}}^{LR}, \hat{F}_{I_{0,1}}^{LR}, \hat{I}_{0,1}^{LR}, F_{I_{0,1}}^{LR}),\\
k_s^{HR} v_s^{HR} &= f_s(F_s^{LR}, \Delta x, \Delta y),\\
k_w^{HR} v_w^{HR} &= f_w(F_w^{LR}, \Delta x, \Delta y),\\
Q^{HR} &= f_q(Q^{LR}, \Delta x, \Delta y),
\end{aligned} \tag{4}$$

where f_s, f_w, and f_q are synthesis fusion, warping fusion, and query fusion upsample networks, respectively.

The upsampled HR features are then decoded by a transformer decoder to produce the estimated HR image I_t^{HR}.

Loss Functions

We employ a two-stage training approach. In the first stage, we train the optical flow network with charbonnier loss $\rho(\cdot)$ from (Charbonnier et al. 1994) and an edge-aware smoothness loss L_{smooth} from (Wang et al. 2018),

$$\begin{aligned}
\mathcal{L}_{flow} &= \lambda_1(\rho(I_{GT}^{LR} - \hat{I}_0^{LR}) + \rho(I_{GT}^{LR} - \hat{I}_1^{LR}))\\
&\quad + \lambda_2 L_{smooth}(\rho(I_{GT}^{LR}, M_{t\to 0})\\
&\quad + L_{smooth}(\rho(I_{GT}^{LR}, M_{t\to 1})).
\end{aligned} \tag{5}$$

The second stage involves joint training of the position-based INR upsampling network and the frame synthesis network with using charbonnier loss and SSIM loss,

$$\mathcal{L}_{total} = \lambda_1 \rho(I_{GT}^{HR} - I_t^{HR}) + \lambda_2 L_{ssim}(I_{GT}^{HR} - I_t^{HR}). \tag{6}$$

Experiments

Experiments Setup

STVSR Datasets. Similar to previous methods that addressed the STVSR task, we followed the training and testing protocols of VideoINR (Chen et al. 2022) to validate

VFI Method	VSR Method	Input Type	GoPro-Center	GoPro-Average	Adobe-Center	Adobe-Average
SuperSloMo	Bicubic	I	27.04/0.7937	26.06/0.7720	26.09/0.7435	25.29/0.7279
SuperSloMo	EDVR	I	28.24/0.8322	26.30/0.7960	27.25/0.7972	25.95/0.7682
SuperSloMo	BasicVSR	I	28.23/0.8308	26.36/0.7977	27.28/0.7961	25.94/0.7679
QVI	Bicubic	I	26.50/0.7791	25.41/0.7554	25.57/0.7324	24.72/0.7114
QVI	EDVR	I	27.43/0.8081	25.55/0.7739	26.40/0.7692	25.09/0.7406
QVI	BasicVSR	I	27.44/0.8070	26.27/0.7955	26.43/0.7682	25.20/0.7421
DAIN	Bicubic	I	26.92/0.7911	26.11/0.7740	26.01/0.7461	25.40/0.7321
DAIN	EDVR	I	28.01/0.8239	26.37/0.7964	27.06/0.7895	26.01/0.7703
DAIN	BasicVSR	I	28.00/0.8227	26.46/0.7966	27.07/0.7890	26.23/0.7725
Zooming SlowMo		I	30.69/0.8847	−/−	30.26/0.8821	−/−
TMNet		I	30.14/0.8692	28.83/0.8514	29.41/0.8524	28.30/0.8354
VideoINR-fixed		I	30.73/0.8850	−/−	30.21/0.8805	−/−
VideoINR		I	30.26/0.8792	29.41/0.8669	29.92/0.8746	29.27/0.8651
MoTIF		I	31.04/0.8877	30.04/0.8773	30.63/0.8839	29.82/0.8750
HR-INR		I+E	31.97/0.9298	32.13/**0.9371**	31.26/**0.9246**	31.11/**0.9216**
Ours		I+E	**32.50**/**0.9340**	**32.23**/0.9320	**31.79**/0.9200	**31.61**/0.9194

Table 1: Quantitative metrics on ×8 VFI and ×4 VSR in terms of PSNR/SSIM. Center and Average means evaluate the average metrics of the center frames (*i.e.*, the 1^{st}, 4^{th}, 9^{th} frames) and all 9 output frames.

our approach on the **Adobe240** (Su et al. 2017) and **GoPro** (Nah, Hyun Kim, and Mu Lee 2017) datasets.

Both datasets have a resolution of 1280×720 and a frame rate of 240 fps. We generated events between consecutive frames using vid2e (Gehrig et al. 2020) to simulate realistic event noise, showcasing our method's robustness to noise. The Adobe240 dataset includes 100 training, 16 validation, and 17 testing videos, while the GoPro dataset contains 22 training and 11 testing videos. We trained our model on Adobe and tested it on both Adobe and GoPro, following VideoINR's approach. We used a sliding window of 9 frames, with the 1st and 9th frames, along with intermediate events, as inputs, down-sampled by a factor of 4. The high-resolution frames served as the ground truth.

VFI and VSR Datasets. Since our method can independently perform both super-resolution and interpolation tasks, we conducted experiments on two real event datasets to validate their performance thoroughly. Specifically, we performed Video Frame Interpolation (VFI) experiments on the **BS-ERGB** dataset (Tulyakov et al. 2022). BS-ERGB is widely used for event-guided VFI tasks and is characterized by complex motions, including non-linear and large movements. We trained and tested our method on this dataset and compared the results with previous methods. Additionally, we performed video super-resolution(VSR) experiments on the **CED** dataset (Scheerlinck et al. 2019), and compared our results with those of prior approaches.

Low-Resolution Data Generation. For both the synthetic event and the real-world event dataset, to align with the low-resolution (LR) images (which are downsampled from the high-resolution (HR) images via bilinear interpolation), we first convert the HR events into a voxel grid (Zhu et al. 2019) and then downsample it using bilinear interpolation. This approach preserves the key spatial features of the event data while minimizing potential artifacts that might arise from resampling misalignments.

Implementation Details. For all experiments, the Adam optimizer (Kingma 2014) was employed with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate was set at 4×10^{-4} and was systematically reduced to 1×10^{-7} through cosine annealing every 150k iterations. The training was conducted over 600k iterations with a batch size of 8. Data augmentation strategies, including random rotations and random cropping, were applied. The experiments were executed on four NVIDIA RTX 3090 GPUs.

Evaluation. During testing, for ease of description, we denote scaling configurations as xAxB, where A represents the spatial up-sampling scale and B is the temporal up-sampling scale. For all experiments, the performance of our model is assessed across three RGB channels, employing the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) (Wang et al. 2004). This approach is consistent with methodologies outlined in prior works (Chen et al. 2022; Chen Y H et al. 2023).

Comparison to State-of-the-arts

Event-guided STVSR. Following the experimental setups of previous works (Chen et al. 2022), we conducted training on the Adobe240 dataset and performed testing on both Adobe240 and GoPro test sets.

We categorized comparative methods into four distinct groups: **(1)** RGB-Based Two-Stage Method: This category includes combinations of Video Frame Interpolation (VFI) methods such as SuperSloMo (Jiang et al. 2018), QVI (Xu et al. 2019) and DAIN (Bao et al. 2019), and Video Super-Resolution (VSR) methods such as EDVR (Wang et al. 2019) and BasicVSR (Chan et al. 2021). **(2)** RGB-Based One-Stage F-STVSR Methods: This group comprises methods like Zooming SlowMo and TMNet. **(3)** RGB-Based One-Stage C-STVSR Methods: This includes methods such

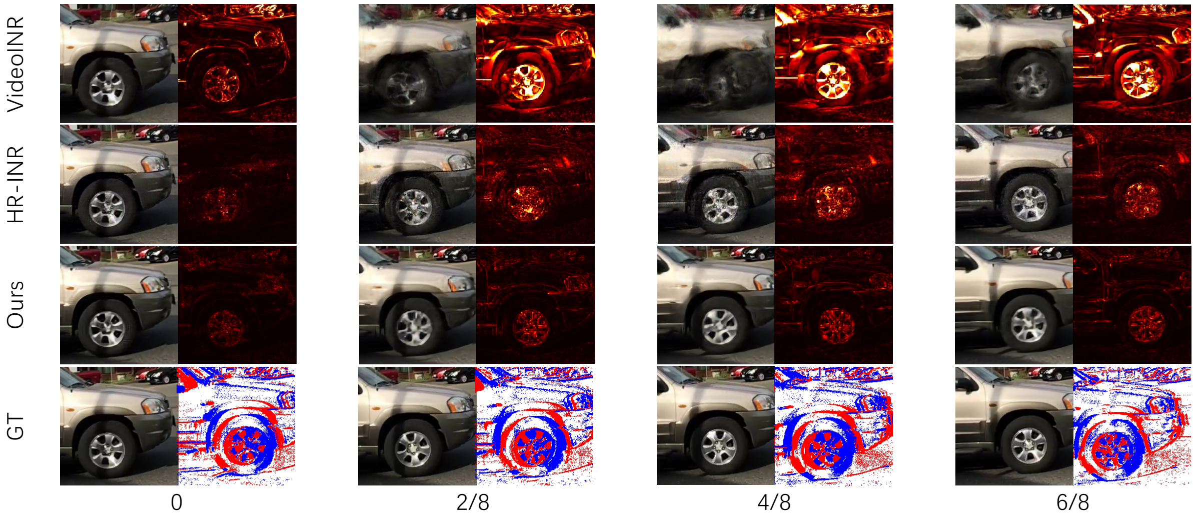

Figure 4: Quality comparison in Time x8 Space x4 super-resolution tasks. Both our approach and HR-INR utilize event data to accurately capture inter-frame dynamics, whereas RGB-based VideoINR exhibits voids due to a lack of motion capture. Notably, HR-INR, which does not employ explicit optical flow for warping, tends to introduce substantial noise from event data, impacting the quality of the reconstructed images.

as VideoINR (Chen et al. 2022) and MoTIF (Chen Y H et al. 2023). **(4)** Event-Based One-Stage C-STVSR Method: HR-INR(Lu et al. 2024).

Table 1 presents a comparison between our method and previous results. Several patterns emerge from the data: (1) One-Stage Methods Outperform Two-Stage Methods: This suggests that integrating frame interpolation and super-resolution tasks leads to superior results. (2) Event-Based Methods Surpass RGB-Based Methods: The additional inter-frame information from events enhances performance. For instance, our method outperforms the best RGB-based method, MoTIF, with significant gains in PSNR and SSIM on both the GoPro and Adobe datasets, especially in average frame performance. This highlights the limitations of RGB-based methods in mid-moment space-time super-resolution due to insufficient inter-frame data. (3) Superior to HR-INR: Our method consistently outperforms the event-based HR-INR in PSNR across various evaluations, with reasons for this advantage discussed in the following sections.

In Figure 4, both our method and HR-INR demonstrate proficient modeling of nonlinear motions, such as the rotation of a wheel. In contrast, the RGB-based VideoINR struggles with this task, resulting in considerable artifacts. Furthermore, our approach shows a superior capability to suppress noise induced by events compared to HR-INR. This advantage arises because HR-INR employs a feature-extractor and space-time decoding technique, which essentially integrates event information by ”adding” it to the sequential frames, naturally introducing event-related noise. In contrast, our method utilizes a synthesis-based and warping-based framework. Since the warping is applied directly to the images of preceding and succeeding frames, it effectively mitigates event noise.

Methods	Input	1skip		3skips	
	Type	PSNR↑	SSIM↑	PSNR↑	SSIM ↑
FLAVR	I	25.95	-	20.90	-
DAIN	I	25.20	-	21.40	-
Super SloMo	I	-	-	22.48	-
QVI	I	-	-	23.20	-
TimeLens	I+E	28.36	-	27.58	-
TimeLens++	I+E	28.56	-	27.63	-
CBMNet	I+E	29.32	0.815	28.46	0.806
CBMNet-Large	I+E	29.43	0.816	28.59	0.808
HR-INR	I+E	29.66	0.828	28.59	0.814
Ours	I+E	**29.73**	**0.83**	**28.64**	**0.817**

Table 2: Quantitative metrics of VFI on the BS-ERGB dataset. 1skip represents Time x2, 3skip represents Time x4.

In Table 1, our method shows slightly lower SSIM scores than HR-INR for certain metrics, primarily due to its focus on large-motion areas, which constitute a small fraction of the total video. Unlike HR-INR, which uses four image inputs, our model relies on just two, limiting its ability to capture detailed structural information in smaller motion regions. Since events provide sparse differential information, they cannot fully replace the structural data from additional images.

Event-guided VFI and Event-guided VSR. We compared our method, which performs interpolation and super-resolution independently, against individual VFI and VSR methods under similar conditions.

Table 2 presents a quantitative comparison between our method and other VFI approaches, demonstrating that our method outperforms existing ones. *Please refer to the Suppl.*

Methods	Input	Space ×4		Space ×2	
	Type	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DUF	I	24.43	0.8177	31.83	0.9183
TDAN	I	27.88	0.8231	33.74	0.9398
SOF	I	27.00	0.8050	31.84	0.9226
RBPN	I	29.80	0.8975	36.66	0.9754
BasicVSR	I	32.93	0.9001	39.57	0.9778
VideoINR	I	25.53	0.7871	26.77	0.7938
E-VSR	I+E	30.15	0.9052	37.32	0.9783
EG-VSR	I+E	31.12	0.9211	38.69	0.9771
HR-INR	I+E	32.15	0.9658	42.01	0.9905
EvTexture	I+E	33.68	0.9112	40.52	0.9813
EvTexture+	I+E	33.71	0.9126	40.57	0.9815
Ours	I+E	**33.83**	**0.9282**	**42.47**	**0.9906**

Table 3: Quantitative metrics of VSR on the CED dataset.

	Adobe-Center		Adobe-Average	
	PSNR↑	SSIM↑	PSNR↑	PSNR↑
Ours	**31.79**	**0.9200**	**31.61**	**0.9194**
Ours $(-f)$	30.92	0.9123	30.88	0.9121

Table 4: Ablation study of optical flow alignment on the Adobe240 dataset.

Mat. for qualitative results.

Table 3 presents a quantitative comparison between our method and other VSR approaches, demonstrating that our results on the CED dataset outperform both HR-INR and EvTexture (Kai et al. 2024). *Please refer to the Suppl. Mat. for qualitative results.*

For the CED dataset, we computed optical flow using RAFT (Teed and Deng 2020) on HR frames from 10 scenes. To quantify motion, we summed pixel displacements, applying a clipping threshold C to exclude background motion from camera movement, as defined by $y = \sum_{i\in H, j\in W}(d_{i,j}$ if $d_{i,j} > C$ $else$ $0)$.

A higher y value indicates scenes with greater motion. On the left side of Figure 5, the x-axis shows 10 scenes sorted by displacement (largest to smallest), while the y-axis shows the corresponding y values. The right side aligns with the left, with the x-axis matching scene indices and the y-axis showing our method's PSNR improvement over the other four methods. Our method achieves higher gains in scenes with larger motions (left side of the x-axis), highlighting its effectiveness with significant motion.

Ablation Study

Motion Flow Estimation. In our method, event-guided optical flow alignment is a crucial component that aligns features from preceding and succeeding frames to the target frame. As shown in Table 4, disabling explicit optical flow alignment (by setting the estimated bidirectional flows to zero) results in performance degradation. Nevertheless, even with only two input frames, our method surpasses HR-INR (Lu et al. 2024), which relies on multiple input frames. Visual results indicate that, under significant motion, our method generates fewer artifacts than HR-INR, likely due to the reduced number of input frames. *Refer to the supplementary materials for qualitative results.*

	Ev	RAFT	Input LR Index	Output HR Index	PSNR↑	SSIM↑
Ours $(-e)$	×	✓	$0, 1, 2$	1	32.57	.9085
Ours (-1)	✓	×	$0, 2$	1	33.79	.9264
Ours $(+2)$	✓	×	$0, 1, 2, 3, 4$	2	33.70	.9254
Ours	✓	×	$0, 1, 2$	1	**33.83**	**.9282**

Table 5: Ablation study of event signal and input frame numbers on the CED dataset.

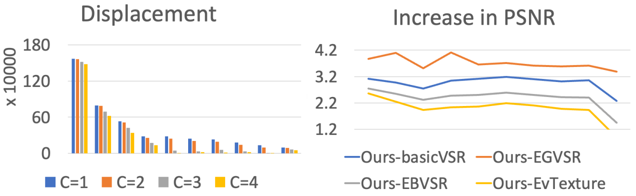

Figure 5: With the increasing scale of motion, our method demonstrates a more pronounced performance improvement relative to other methods.

Other Settings. For the VSR task on the CED dataset, we used three input frames, which slightly improved performance over using two. In Table 5, "Ours" refers to inputting the 1st, 2nd, and 3rd LR images to generate the 2nd HR image. However, increasing the input to five frames did not improve results and even caused a slight decline, likely due to the challenges of aligning multiple frames with complex motion and the increased computational cost. The high temporal resolution of events ensures that a few RGB frames with adjacent event signals are sufficient for super-resolution—more frames are not necessarily better.

In our method, events are a vital complement to RGB images. As shown in Table 5, removing the event signals ("ours$(-e)$") and replacing our flow estimation with pre-trained RAFT (Teed and Deng 2020) models leads to a performance drop. This indicates that events aid in alignment and enhance detailed features post-alignment.

Conclusion

In this paper, we introduce the Event-guided STVSR method, which effectively combines a few images and events. The event-guided optical flow suppresses noise, and using fewer images enhances robustness in complex motions, including nonlinear and large movements. Our approach shows competitive performance compared to both RGB-based and event-based STVSR methods.

Limitations and Future Work. The training process is time-consuming, and improving the generalization to real events remains a challenge.

Acknowledgments

This work was supported in part by the STI 2030 Major Projects under Grant 2021ZD020040, in part by the National Natural Science Foundation of China (62376247, 61925603, U20A20220, and 62334014), in part by the grant from Key R&D Program of Zhejiang (2022C01048), and in part by the Fundamental Research Funds for the Central Universities. We express special thanks to Prof. Xudong Jiang for his useful suggestions and guidance.

References

Bao, W.; Lai, W.-S.; Ma, C.; Zhang, X.; Gao, Z.; and Yang, M.-H. 2019. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3703–3712.

Chan, K. C.; Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2021. BasicVSR: The Search for Essential Components in Video Super-Resolution and Beyond. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; and Barlaud, M. 1994. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of 1st international conference on image processing*, volume 2, 168–172. IEEE.

Chen, Y.; Liu, S.; and Wang, X. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8628–8638.

Chen, Z.; Chen, Y.; Liu, J.; Xu, X.; Goel, V.; Wang, Z.; Shi, H.; and Wang, X. 2022. Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2047–2057.

Chen, Z.; Lu, Z.; Ma, D.; Tang, H.; Jiang, X.; Zheng, Q.; and Pan, G. 2024. Event-ID: Intrinsic Decomposition Using an Event Camera. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10095–10104.

Chen, Z.; Zheng, Q.; Niu, P.; Tang, H.; and Pan, G. 2021. Indoor lighting estimation using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14760–14770.

Chen Y H, Y.-H.; Chen, S.-C.; Lin, Y.-Y.; and Peng, W.-H. 2023. MoTIF: Learning motion trajectories with local implicit neural functions for continuous space-time video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23131–23141.

Gehrig, D.; Gehrig, M.; Hidalgo-Carrió, J.; and Scaramuzza, D. 2020. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3586–3595.

Han, J.; Yang, Y.; Zhou, C.; Xu, C.; and Shi, B. 2021. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4882–4891.

He, W.; You, K.; Qiao, Z.; Jia, X.; Zhang, Z.; Wang, W.; Lu, H.; Wang, Y.; and Liao, J. 2022. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17804–17813.

Jiang, H.; Sun, D.; Jampani, V.; Yang, M.-H.; Learned-Miller, E.; and Kautz, J. 2018. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9000–9008.

Jing, Y.; Yang, Y.; Wang, X.; Song, M.; and Tao, D. 2021. Turning frequency to resolution: Video super-resolution via event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7772–7781.

Kai, D. 2023. Video Super-Resolution Via Event-Driven Temporal Alignment. In *2023 IEEE International Conference on Image Processing (ICIP)*, 2950–2954. IEEE.

Kai, D.; Lu, J.; Zhang, Y.; and Sun, X. 2024. EvTexture: Event-driven Texture Enhancement for Video Super-Resolution. In *Forty-first International Conference on Machine Learning*.

Kim, T.; Chae, Y.; Jang, H.-K.; and Yoon, K.-J. 2023. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18032–18042.

Kingma, D. 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, H.; Ruan, Z.; Zhao, P.; Dong, C.; Shang, F.; Liu, Y.; Yang, L.; and Timofte, R. 2022. Video super-resolution based on deep learning: a comprehensive survey. *Artificial Intelligence Review*, 55(8): 5981–6035.

Lu, Y.; Wang, Z.; Liu, M.; Wang, H.; and Wang, L. 2023. Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1557–1567.

Lu, Y.; Wang, Z.; Wang, Y.; and Xiong, H. 2024. HR-INR: Continuous Space-Time Video Super-Resolution via Event Camera. *arXiv preprint arXiv:2405.13389*.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Scheerlinck, C.; Rebecq, H.; Stoffregen, T.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2019. CED: Color event camera dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33: 7462–7473.

Su, S.; Delbracio, M.; Wang, J.; Sapiro, G.; Heidrich, W.; and Wang, O. 2017. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1279–1288.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Tulyakov, S.; Bochicchio, A.; Gehrig, D.; Georgoulis, S.; Li, Y.; and Scaramuzza, D. 2022. 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*, 17755–17764.

Tulyakov, S.; Gehrig, D.; Georgoulis, S.; Erbach, J.; Gehrig, M.; Li, Y.; and Scaramuzza, D. 2021. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16155–16164.

Wang, X.; Chan, K. C.; Yu, K.; Dong, C.; and Change Loy, C. 2019. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 0–0.

Wang, Y.; Yang, Y.; Yang, Z.; Zhao, L.; Wang, P.; and Xu, W. 2018. Occlusion aware unsupervised learning of optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4884–4893.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xiang, X.; Tian, Y.; Zhang, Y.; Fu, Y.; Allebach, J. P.; and Xu, C. 2020. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3370–3379.

Xiao, Z.; Kai, D.; Zhang, Y.; Sun, X.; and Xiong, Z. 2024a. Asymmetric Event-Guided Video Super-Resolution. In *ACM MM*.

Xiao, Z.; Kai, D.; Zhang, Y.; Zha, Z.-J.; Sun, X.; and Xiong, Z. 2024b. Event-Adapted Video Super-Resolution. In *ECCV*.

Xu, G.; Xu, J.; Li, Z.; Wang, L.; Sun, X.; and Cheng, M.-M. 2021. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6388–6397.

Xu, X.; Siyao, L.; Sun, W.; Yin, Q.; and Yang, M.-H. 2019. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 989–997.