Unifying Low-Resolution and High-Resolution Alignment by Event Cameras for Space-Time Video Super-Resolution

Hoonhee Cho, Jae-Young Kang, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon
KAIST
{gnsgnsgml,mickeykang,intelpro,jeongyh98,kjyoon}@kaist.ac.kr

Abstract

Event cameras deliver asynchronous pixel intensity changes, which result in sparse event data that offers the advantages of high temporal resolution. These high temporal characteristics make researchers naturally incorporate event cameras into video frame interpolation (VFI) and video super-resolution (VSR). In this paper, we make the first attempt to solve the space-time video super-resolution (STVSR) task effectively, addressing both VFI and VSR simultaneously, by leveraging temporally dense events. STVSR aims to generate intermediate high-resolution (HR) videos between consecutive low-resolution (LR) frames. To fully exploit the high temporal frequency of events for STVSR, we focus on temporal alignment in two stages, at low-resolution and after up-sampling in high-resolution. In temporal alignment at low-resolution, to upsample spatial dimensions effectively, we leverage high temporal features to preserve spatial context. On the other hand, for temporal alignment at the high-resolution stage, we employ a deformable sampling process from events to achieve accurate alignment with forward and backward directions. In addition, we provide the SuperREST dataset, which features high-frequency details and complex motion in an RGB-Event setup. Experimental results on several datasets demonstrate that our method achieves a significant performance gain on STVSR tasks with low computational cost. Our codes and datasets are available at https://github.com/Chohoonhee/ESTNet.

1. Introduction

Space-time video super-resolution (STVSR) involves enhancing both the spatial and temporal resolutions of videos that have a low frame rate (LFR) and low resolution (LR) to generate high frame rate (HFR) and high-resolution (HR) videos. STVSR [5, 23, 24, 42, 48, 71, 75, 78] has been introduced with the goal of addressing the objectives of both VSR [2,3,25,64,65,68,81] and VFI [1,27,28,33,49,76,82]. In both VSR and VFI, temporal alignment is crucial for information sharing between adjacent frames. However, to achieve high-quality STVSR results, several challenging issues exist. Compared to VSR, the input is LFR, which results in significant pixel displacement between frames. Therefore, retrieving texture information from adjacent frames to the target frame is not straightforward. Compared to VFI, errors from temporal mismatch in the LR accumulate significantly during the super-resolution process. Therefore, accurate temporal alignment is even more important during the STVSR process.

Event cameras [17, 77] are sensors inspired by biology, capable of asynchronously detecting per-pixel intensity changes with low latency. Event cameras are increasingly used alongside frame cameras for vision tasks. In particular, video super-resolution (VSR) [2, 3, 59, 64] and video frame interpolation (VFI) [1, 18, 27, 76] are representative applications that utilize the high temporal frequency characteristics of event cameras to fill in the missing information in video continuity. VSR aims to super-resolve spatial dimensions to enhance the quality of images by preserving detailed structures. As demonstrated in previous works [15, 29, 47, 79], the high temporal information of events strengthens temporal interactions between consecutive frames. On the other hand, VFI [11, 34, 45, 57] aims to increase temporal resolution for the purpose of generating natural-looking videos. When using event cameras, it becomes possible to describe complex and non-linear motion. While there have been studies demonstrating the effectiveness of event cameras in enhancing spatial and temporal dimensions individually, there have been no attempts to elevate both dimensions simultaneously, which could intuitively be more effective.

We propose a method that effectively utilizes the temporal resolution of event cameras for both spatial and temporal upsampling. As shown in Fig. 1, the proposed event-based spatio-temporal interplay network (ESTNet) is composed of two major stages: temporal alignment in low-resolution (TAL) and temporal alignment in high-resolution (TAH). Typical VSR methods [2, 3] effectively enhance the spatial resolution from HFR videos by utilizing information from

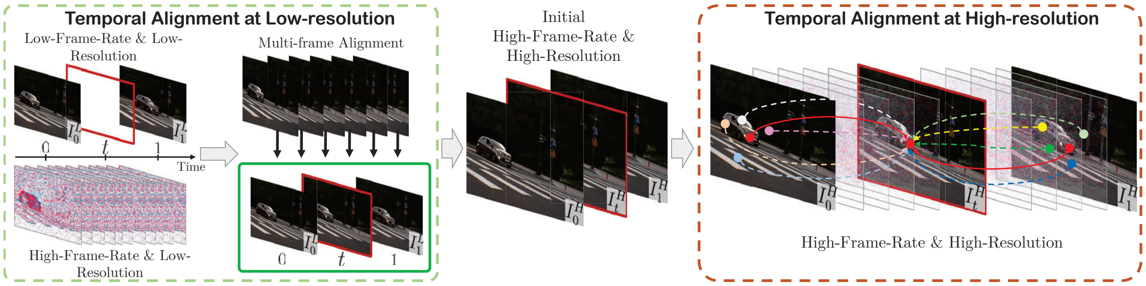

Figure 1. The ESTNet consists of two-stage temporal alignment. In the LR phase, we enhance spatial frequency by leveraging the high temporal frequency of events. Then, in the HR, we achieve more accurate temporal alignment based on the spatially dense information.

adjacent frames. However, frame information aggregation from adjacent frames is less efficient in LFR video due to the temporal inconsistency from large motion between frames. Inspired by the event-based VSR method [29] that super-resolves LR video using temporally dense events, we propose a novel approach to generate HFR virtual inter-frame features. The virtual inter-frame features mimic the adjacent frames of HFR video, allowing effective information aggregation to the target timestamp in LFR settings.

While LFR and LR input is super-resolved using temporally dense events in TAL stage, the limited spatial resolution of the key frames can introduce structural inconsistencies. Therefore, fine-grained alignment is required in the HR domain. However, this HR temporal alignment process is more challenging due to the increased number of pixels. Many pixels lead to a larger number of candidate pixels for temporal alignment, making the process more computationally demanding and prone to errors. To overcome the challenges of HR temporal alignment, we propose a novel bi-directional score-based deformable alignment module.

One of the critical challenges in the STVSR task using event cameras is the lack of high-quality datasets. For STVSR with event cameras, it is necessary to have LR events, LR and LFR videos, and their corresponding HR and HFR videos. However, datasets containing real LR events are missing. Previous works on VSR with event data have used bilinear downsampling of event representation, which is unlikely to replicate the same distribution as real LR events. We acquired the first real-world event STVSR dataset, the SuperREST dataset for effective evaluation of the proposed model. A novel RGB-Event Hybrid system using a beam splitter is designed to collect high-quality HR and HFR frames along with the LR events.

2. Related Works

Space-Time Video Super-Resolution (STVSR). STVSR aims to enhance both the spatial and temporal resolution of low-resolution (LR) and low-frame-rate (LFR) videos simultaneously, which has been widely studied in [6, 19, 21, 35, 42, 48, 54, 71, 75]. There are two approaches to STVSR: two- and one-stage methods. Two-stage networks apply VFI [27, 61] and VSR [3, 64] sequentially to increase frame rate and spatial resolution. However, this strategy overlooks the relationship between temporal interpolation and spatial super-resolution, and often has a slow inference speed. Recent studies [6, 19, 21, 71, 75] have utilized one-stage approaches to effectively make use of spatio-temporal correlation by performing temporal interpolation and super-resolution simultaneously in a unified framework.

Event-based Spatio-Temporal Interpolation. Event cameras [17] asynchronously record per-pixel intensity changes, robust to motion blur [8, 12, 37, 56, 73, 83]. Their high dynamic range enables capturing scenes, such as night environments [13, 26, 39, 70], making them ideal for driving scenarios [14, 58]. As another mainstream line of research, the dense temporal information provided by the event camera has proven beneficial for VFI even in the presence of complex motion [4, 22, 44, 60, 67, 69, 80, 86]. Recent works [15, 20, 29, 31, 32, 47, 72] have demonstrated that incorporating an event camera for VSR can lead to improved performance by amplifying the effect of temporal alignment. While there are approaches to interpolate spatial and temporal dimensions individually, there is no approach that simultaneously performs both in a unified framework. Therefore, to adapt existing event-based methods for STVSR, one must apply VSR and VFI sequentially. However, as shown in previous works [21, 71, 75], sequential tasks use intra-priors but ignore inter-task correlations.

Therefore, to effectively address the STVSR, we design our ESTNet to be learned in an end-to-end manner within a unified framework. As demonstrated in our experiments, simply introducing event cameras into the network can improve performance but may not reach optimal performance. In contrast, ESTNet leverages the temporal high-frequency of event cameras for spatial upsampling and effectively utilizes them during the temporal alignment process.

3. Proposed Method

3.1. Preliminary

The proposed model generates an HR frame, I_t^H, at any arbitrary time t between two consecutive LR and LFR key

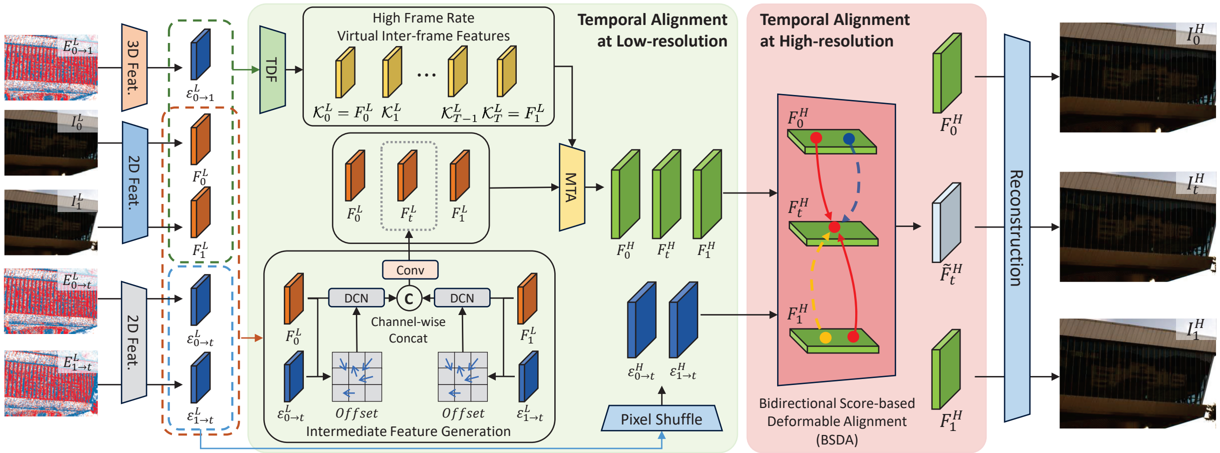

Figure 2. The overall pipeline for event-based spatio-temporal interplay network (ESTNet). ESTNet consists of two main phases: *temporal alignment at low-resolution* phase, which efficiently performs spatial interpolation using high temporal-frequency events, and *temporal alignment at high-resolution* phase, which aims to achieve accurate alignment using improved spatial dimension features.

frames, I_0^L and I_1^L, that were acquired at normalized fractional times of 0 and 1, respectively. We additionally use a LR event sequence, $E_{0\to1}^L$, containing events that occurred during the time interval $[0, 1]$. Similar to prior researches [9, 60, 61], the event stream is represented as a voxel grid of [88]. An event voxel grid between time instances t_a and t_b is denoted as $E_{t_a\to t_b} \in \mathbb{R}^{B\times H\times W}$, where B is bin size. The reversal technique [61] is adapted to reverse event sequence, *e.g.*, $E_{1\to t}^L$.

Figure 2 shows the overall architecture of ESTNet. ESTNet consists of two stages: Temporal Alignment at Low-resolution (TAL) and Temporal Alignment at High-resolution (TAH). TAL is motivated by the observation in existing event-based VSR [29] that high temporal resolution aids spatial super-resolution. Moreover, TAH generates higher-quality intermediate HR results by fine-grained temporal alignment through upsampled features.

The feature extractor makes features $\mathcal{E}_{0\to1}^L$, F_0^L, F_1^L, $\mathcal{E}_{0\to t}^L$ and $\mathcal{E}_{1\to t}^L$ from event and image inputs: the reshaped LR events $E_{0\to1}^L \in \mathbb{R}^{T\times(B/T)\times H\times W}$; the key frames $I_0^L, I_1^L \in \mathbb{R}^{3\times H\times W}$; and events $E_{0\to t}^L, E_{1\to t}^L \in \mathbb{R}^{B\times H\times W}$ up to an arbitrary time point t. $E_{0\to1}^L$ are processed by equally slicing an event voxel $E_{0\to1} \in \mathbb{R}^{B\times H\times W}$ to T voxels along bin axis. Each voxel slices are used to generate inter-frame features. The feature extractor of $E_{0\to1}^L$ only employs 3D residual blocks to incorporate an extra dimension, while other feature extractors utilize 2D blocks.

3.2. Temporal Alignment at Low-resolution (TAL)

Intermediate Feature Generation. The TAL generates high-resolution key frame features: F_0^H; F_1^H; and F_t^H. To generate F_t^H, we first make an intermediate LR feature F_t^L. As shown in the bottom part of the TAL phase in Fig. 2, we estimate the offsets using each key frame feature and event features up to an arbitrary time t. Then, applying DCN [16, 87] with offsets to each key frame, we generate the temporally aligned features. Subsequently, we concatenate the two aligned features from DCN and generate the LR intermediate feature, F_t^L, through a convolution.

Leveraging High-temporal Frequency. In the TAL stage, high temporal frequency in events is transformed into the spatial dimension. Previous event-based VSR methods [20, 29] have directly used event features for alignment to enhance the LR feature. However, as mentioned in [7, 10, 52, 63], the images reconstructed from events are more semantically meaningful and contain finer structural details compared to raw events, leading to better performance in downstream tasks. Our approach aims to utilize events not directly but by generating virtual HFR inter-frame features from events, enabling dense utilization of both temporal and spatial information. Naive approaches such as iterative generation of inter-frame features are possible, but inefficient and time-consuming. We propose a cost-effective design, which makes all inter-frame features at once, based on event-based dynamic filters grounded in synthesis-based interpolation [36, 61]. The temporal dynamic filter (TDF) module simultaneously generates inter-frame features, and the multi-temporal alignment (MTA) module conveys additional structural information from the HFR inter-frame features to the target frame.

Temporal Dynamic Filter (TDF). Figure 3 illustrates the details of Temporal Dynamic Filter (TDF) module, which generates the HFR inter-frame video features. $\mathcal{E}_{0\to1}^L$ is reshaped along the temporal axis, T, with each slice containing information of intermediate frame. Typically, DCN [87] or optical flow methods [51, 67] are commonly used. However, the module employs dynamic filters [38, 50, 84] for simultaneous virtual inter-frame features generation. Specifically, given LR key frame features $F_0^L, F_1^L \in \mathbb{R}^{C\times H\times W}$, which have the spatial dimension of $H \times W$ and chan-

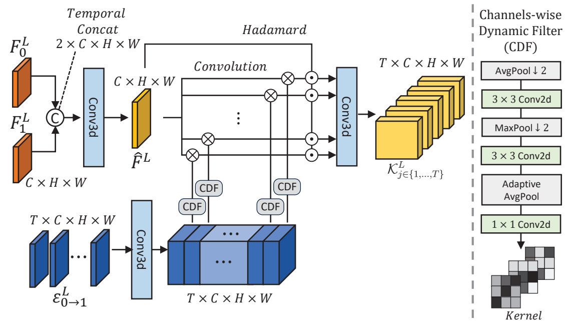

Figure 3. The proposed temporal dynamic filter (TDF) module.

nel dimension of C, we first concatenate with temporal dimension, $\mathbb{R}^{2\times C\times H\times W}$. Then, we apply a 3D convolution with filter size of $3 \times 1 \times 1$ pixels to obtain the features $\hat{F}^L \in \mathbb{R}^{1\times C\times H\times W}$. Next, we apply 3D convolution to event features, $\mathcal{E}^L_{0\to 1} \in \mathbb{R}^{T\times C\times H\times W}$, to perform temporal aggregation, and we estimate the kernel $\{\mathcal{D}_j\}^T_{j=1}$ through channel-wise dynamic filters (CDF), where $\mathcal{D}_j \in \mathbb{R}^{C\times d\times d}$ (d is dynamic filter size). Using dynamic filters, $\mathcal{D}_j$, we generate gated filters $\hat{\mathcal{K}}^L_j$ by:

$$\hat{\mathcal{K}}^L_j = \mathcal{D}_j \otimes \hat{F}^L, \quad j \in \{1, \cdots, T\}, \tag{1}$$

where $\otimes$ denotes the convolution filter operation. Then, we element-wisely multiply gated filters $\hat{\mathcal{K}}^L_j$ with fused key frame feature $\hat{F}^L$ as $\mathcal{K}^L_j = \hat{\mathcal{K}}^L_j \odot \hat{F}^L$. Finally, we apply 3D convolution block to $\mathcal{K}^L_j$ and stack on the F^L_0 to obtain the high temporal frequency features, $\mathcal{K}^L_{[0,\ldots,T]} \in \mathbb{R}^{T'\times C\times H\times W}$, where $T' = T + 1$. Previous VFI [36, 61] methods generated high temporal resolution features through iterative computations, whereas our TDF module is cost-effective as it calculates filters at once from events. We pass through two convolution layers to train the HFR virtual inter-frame feature via supervision, generating each I^L_j from $\mathcal{K}^L_j$ as in Eq. (7) for optimization accordingly.

Multi-Temporal Alignment (MTA). To leverage the benefits of the generated high temporal frequency features [3, 29], a key point is to align the virtual inter-frame features, $\mathcal{K}^L_{[0,\ldots,T]}$, to each frame, $F^L_{i\in\{0,t,1\}}$. While a recurrent one-to-one matching process from multiple features is possible, it overlooks interactions between multiple frames and significantly increases computational overhead. As an effective alternative, we aim to use a 3D flow [40, 74] that defines the displacement from multiple frames to a single target frame. To effectively utilize 3D flow, we design the multi-temporal alignment (MTA) module to perform alignment from multi-temporal features to a single feature. As shown in Fig. 4, for the target i-frame, we first calculate an initial 3D flow [40], $\mathcal{O}_{3d} \in \mathbb{R}^{3\times H\times W}$, using $\mathcal{K}^L_{[0,\ldots,T]}$ and F^L_i through lightweight FlowNet network, which consists of 2D convolutions. Unlike 2D flow which only contains the space motion, 3D flow incorporates spatial and temporal motion. Then, we obtain the initially aligned feature,

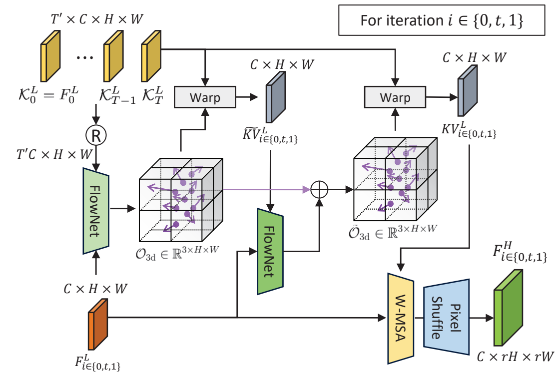

Figure 4. The multi-temporal alignment (MTA) module.

$\widetilde{KV}^L_i \in \mathbb{R}^{C\times H\times W}$, using the space-time sampling operations in [40]. Space-time flow is more challenging to compute than 2D flow. To refine it, we calculate a residual flow using $\widetilde{KV}^L_i$ and then employ space-time sampling to generate the final multi-temporally aligned feature, KV^L_i.

To fuse F^L_i with multi-temporally aligned feature KV^L_i, we adopt the effectively designed transformer structure, W-MSA [46]. We set the target frame F^L_i as query feature and the aligned feature KV^L_i as the key and value features. Then, we generate high spatial features, $F^H_i \in \mathbb{R}^{C\times rH\times rW}$, using pixel shuffle [55] with up-scale factor r. The MTA module utilizes space-time flow to fully exploit global information in high temporal frequency features and employs a transformer structure to perform substantial feature enhancement in high-level representations.

3.3. Temporal Alignment at High-resolution

In the TAL phase, we enhance the spatial dimension using the high temporal resolution of events. However, F^H_t has misalignment caused by low spatial resolution, leading to the accumulation of the artifact. Therefore, additional refinement is required using the improved high-resolution key frame features, F^H_0 and F^H_1. Therefore, we propose the bi-directional score-based deformable alignment (BSDA) module for an alignment in HR features. Before feeding into the BSDA module, we enhance the spatial dimension of $\mathcal{E}^L_{0\to t}$ and $\mathcal{E}^L_{1\to t}$ through pixel shuffle [55].

Bi-directional Score-based Deformable Alignment (BSDA). In contrast to low resolution, at high resolution, where more pixels are present, temporal alignment becomes more challenging due to the increased number of possible candidates. Existing alignment methods [36, 51, 61, 75] all overlook the relationship between forward and backward alignment when aligning frames at arbitrary time t. For example, in previous works, when performing forward alignment from F^H_0 to F^H_t, only F^H_0 and $\mathcal{E}^H_{0\to t}$ (additionally using initially synthesized F^H_t, *e.g.*, [36, 51]) were utilized to define motion vectors, disregarding the correlation with F^H_1. Similarly, backward alignment are defined in the same manner, without considering F^H_0 or

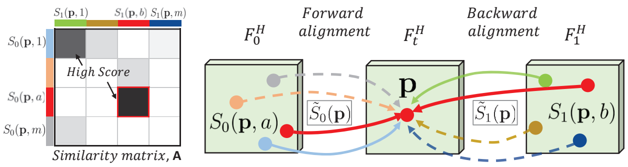

Figure 5. Illustration of bi-directional score-based deformable alignment (BSDA) module.

forward alignment. If the alignment from 0 to t and 1 to t are considered independently, the key frame alignment can become inconsistent. As shown in Fig. 5, the model can confuse multiple alignment candidates that may arise due to repetitive textures or patterns. For instance, at 0, the candidates are red and sky blue, while at 1, they are green and red. This inconsistency can hinder the refinement performance in the HR domain. The proposed method enhances temporal consistency and achieves accurate alignment by calculating the correspondence of multiple candidates between 0 and 1. For example, in Fig. 5, the red point is determined to be the necessary alignment position as it has the highest similarity.

From this motivation, BSDA module simultaneously calculate these interrelated formulations by calculating an affinity similarity matrix. Specifically, we first estimate forward and backward offsets for every pixel in F_t^H using the two key frames and the inter events in between, and deformably sample the pixel candidates, S_0 and S_1 with m sampling locations. The processes of S_0 and S_1 are symmetric, so we only explain with respect to S_0. Let $\mathbf{p} \in \mathbb{R}^2$ denote the spatial locations in features and $S_0(\mathbf{p}) = \{S_0(\mathbf{p}, 1), \cdots, S_0(\mathbf{p}, m)\}$. To effectively aggregate more spatial information, we divide F_0^H into G groups along the channel dimension and calculate G offsets for each sampling location $\mathbf{p}$ as $S_0(\mathbf{p}, a) = \{S_0^g(\mathbf{p}, a)\}_{g=1}^{G}$. The sampled candidates S_0 at each position $\mathbf{p}$, group index g, and sampling index a can be obtained by:

$$S_0^g(\mathbf{p}, a) = F_0^{H,g}(\mathbf{p} + \Delta p_a^g), \quad a \in [1, m], \tag{2}$$

$$\Delta p = f_o(F_0^H, F_t^H, \mathcal{E}_{0\to t}^H), \tag{3}$$

where $\Delta p = \{\Delta p_1, \cdots, \Delta p_m\}$, $\Delta p_a = \{\Delta p_a^1, \cdots, \Delta p_a^g\}$, $F_0^{H,g} \in \mathbb{R}^{C/G \times rH \times rW}$, and f_o consists of few convolution layers to predict the irregular offsets. Next, we compute the similarity matrix, $\mathbf{A} \in \mathbb{R}^{m \times m}$, between the m points sampled from F_0^H and F_1^H, for all cases, as follows:

$$\mathbf{A}(a, b) = \left(\frac{S_0(\mathbf{p}, a)}{\|S_0(\mathbf{p}, a)\|_2^2}\right)^T \cdot \left(\frac{S_1(\mathbf{p}, b)}{\|S_1(\mathbf{p}, b)\|_2^2}\right), \tag{4}$$

where $a, b \in [1, m]$. We compute the attended features:

$$\tilde{S}_0(\mathbf{p}) = \sum_{a=1}^{m}\sum_{b=1}^{m} S_0(\mathbf{p}, a) \cdot \mathbf{A}(a, b) \tag{5}$$

$$\tilde{S}_1(\mathbf{p}) = \sum_{b=1}^{m}\sum_{a=1}^{m} \mathbf{A}(a, b) \cdot S_1(\mathbf{p}, b)$$

The intermediate feature F_t^H is temporally refined as:

$$\tilde{F}_t^H = \mathcal{G}(F_t^H, \tilde{S}_0, \tilde{S}_1), \tag{6}$$

where $\mathcal{G}$ consists of 3×3 convolution layer with channel-wise concatenation. Finally, we utilize two convolutions to generate the output HR video $I^H = \{I_0^H, I_t^H, I_1^H\}$.

3.4. Loss Functions

Our framework is trained in an end-to-end manner by total loss ($\mathcal{L}$) with the hyper-parameter, λ_1. We use the charbonnier loss [30], $\mathcal{L}_c$, for optimization as:

$$\mathcal{L} = \sum_{i \in \{0,t,1\}} \mathcal{L}_c(I_i^H, \hat{I}_t^H) + \lambda_1 \cdot \sum_{j \in \{1,\cdots,T\}} \mathcal{L}_c(I_j^L, \hat{I}_j^L). \tag{7}$$

4. Experiments

4.1. Implementation Details.

We crop 64×64 patches with a batch size of 2. The models are trained on the single NVIDIA TITAN RTX using the Adam optimizer [41], with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to $1 \times e^{-4}$, and decreased by a factor of 0.1 every 100 epochs up to 500 epochs. Our evaluation metrics for STVSR are peak-signal-to-noise (PSNR), structural-similarity-image-metric (SSIM). All networks were trained using a voxel grid [85] , with a bin size of 8. We set the hyper-parameter λ_1 as 0.3 in Eq.(7). We design two models, **ESTNet-S** and **ESTNet-L**, varying the number of residual blocks and feature channels, while keeping everything else the same.

4.2. Datasets.

BS-ERGB and CED Datasets. We utilize two existing RGB-event benchmark datasets: one is BS-ERGB [60] dataset, widely used in VFI task, and the other is the CED dataset [53], actively employed in VSR task. For BS-ERGB, we follow the provided train and test sets. In the case of CED dataset, we radnomly split the dataset into 49 train sequences and 23 test sequences. Consistent with previous STVSR works [19,71,75], we perform spatial $\times 4$ upsampling ($r = 4$) on all datasets. For temporal interpolation, we evaluate on settings of 1-skip and 3-skip for BS-ERGB dataset and 3-skip and 7-skip for the CED dataset. We generate LR frames following [75] and LR events following public code of recent event-based VSR [47].

SuperREST Dataset. From extensive research, it has been found that while RGB data can mimic realistic LR RGB through bicubic downsampling [62], this is not the case for events. When the camera's resolution decreases, not only is there a reduction in spatial resolution, but also the number of triggered events decreases temporally, making it difficult to model the downsampling of events. Existing event-based

Table 1. Quantitative results on CED and BS-ERGB dataset. * denotes the modification in the original only event-based model to incorporate RGB along with event data.

Methods (VFI + VSR / VSR + VFI / STVSR)	CED 3 skips PSNR	CED 3 skips SSIM	CED 7 skips PSNR	CED 7 skips SSIM	BS-ERGB 1 skip PSNR	BS-ERGB 1 skip SSIM	BS-ERGB 3 skips PSNR	BS-ERGB 3 skips SSIM	Param.[M]	FPS
Frame-based Methods										
EMA-VFI [82] + EDVR [64]	25.18	0.7444	24.34	0.7200	19.72	0.5899	18.94	0.5808	66.0+20.7	5.6
EMA-VFI [82] + BasicVSR++ [3]	25.23	0.7453	24.71	0.7253	19.95	0.5946	18.87	0.5746	66.0+6.3	8.1
AMT-G [43] + EDVR [64]	25.44	0.7459	24.13	0.7207	19.33	0.5842	18.55	0.5723	30.6+20.7	4.8
AMT-G [43] + BasicVSR++ [3]	25.69	0.7490	24.03	0.7195	19.88	0.5932	18.52	0.5733	30.6+6.3	7.2
EDVR [64] + EMA-VFI [82]	25.14	0.7441	24.52	0.7203	19.76	0.5901	18.92	0.5793	20.7+66.0	3.9
EDVR [64] + AMT-G [43]	25.61	0.7491	24.32	0.7263	19.54	0.5840	18.82	0.5780	20.7+30.6	3.8
BasicVSR++ [3] + EMA-VFI [82]	25.41	0.7452	24.60	0.7271	19.93	0.5964	18.32	0.5704	6.3+66.0	7.7
BasicVSR++ [3] + AMT-G [43]	25.54	0.7483	24.22	0.7192	19.97	0.5955	18.76	0.5735	6.3+30.6	7.3
Zooming SlowMo [71]	25.42	0.7477	24.01	0.7113	20.17	0.5948	19.05	0.5783	11.1	17.2
TMNet [75]	25.76	0.7483	24.31	0.7149	20.54	0.6053	19.24	0.5842	12.3	17.0
RSTT [19]	25.98	0.7506	24.88	0.7284	20.23	0.6014	19.17	0.5809	7.7	16.6
VideoINR [6]	25.59	0.7488	24.75	0.7202	19.97	0.5893	19.01	0.5797	11.3	15.4
MoTIF [5]	25.71	0.7490	24.81	0.7243	20.12	0.5966	19.07	0.5801	12.6	13.1
Event-based Methods										
TimeLens [61] + E2SRI* [15]	27.84	0.7915	26.12	0.7588	21.28	0.6311	21.01	0.6230	72.2+8.1	7.2
TimeLens [61] + EGVSR [47]	27.77	0.7908	26.33	0.7596	21.55	0.6336	21.08	0.6237	72.2+2.5	6.1
CBMNet [36] + E2SRI* [15]	28.12	0.8112	26.75	0.7624	21.97	0.6415	21.24	0.6263	22.2+8.1	4.8
CBMNet [36] + EGVSR [47]	28.04	0.8096	26.87	0.7657	22.10	0.6433	21.45	0.6274	22.2+2.5	4.0
E2SRI* [15] + ESR [66] + TimeLens [61]	27.99	0.7967	26.25	0.7604	21.63	0.6349	21.07	0.6235	8.1+7.0+72.2	2.9
E2SRI* [15] + ESR [66] + CBMNet [36]	28.25	0.8134	26.90	0.7662	22.18	0.6428	21.53	0.6257	8.1+7.0+22.2	1.7
EGVSR [47] + ESR [66] + TimeLens [61]	27.82	0.7930	26.48	0.7613	21.76	0.6352	21.15	0.6243	2.5+7.0+72.2	2.2
EGVSR [47] + ESR [66] + CBMNet [36]	28.18	0.8124	26.72	0.7701	22.34	0.6437	21.62	0.6285	2.5+7.0+22.2	1.1
ESTNet-S (Ours)	28.73	0.8197	28.04	0.8008	23.50	0.6506	23.03	0.6428	**2.8**	**21.3**
ESTNet-L (Ours)	**29.52**	**0.8360**	**28.62**	**0.8137**	**23.72**	**0.6588**	**23.25**	**0.6513**	4.9	17.6

Table 2. Comparison of our SuperREST dataset against other datasets. Other datasets require post-processing such as bilinear downsampling to synthesize LR events, which are not real events.

Datasets	HR RGB (Ground Truth) Resolution	HR RGB (Ground Truth) FPS	LR Event Resolution	Real LR Event
CED [53]	346 × 260	40	86 × 65	
BS-ERGB [60]	970 × 625	28	242 × 156	
SuperREST	1024 × 720	116	256 × 180	✓

VSR applied bilinear sampling to the event representation to generate LR events, which diverges from the real event data. Hence, to conduct STVSR with events for real application, real LR events paired with HR frames are essential. To design a network for practical STVSR, we have acquired a new dataset, called Super-resolving for Spatio-Temporal of RGB data with Event Data (SuperREST) dataset, consisting of a hybrid dataset of RGB and event data. Previous datasets required downsampling to mimic LR events corresponding to HR RGB images. However, in our case, we acquire real LR events corresponding to RGB images, eliminating the need for such a synthesis process. The SuperREST dataset offers a variety of scenarios, including extremely nonlinear motion, making them more appropriate for real-world applications.

Some methods can only perform interpolation for predetermined time steps, so on the CED and BS-ERGB datasets, we evaluate only at $t = 0.5$. On our SuperREST dataset, we compare models capable of arbitrary time inference between frames, allowing evaluation at all time instances in 7-skip settings. For frames, we downsample them following the previous approach [75].

4.3. Comparison with State-of-the-arts

Comparison methods. We compare our ESTNet with one- and two-stage STVSR methods, following the approach in [19, 71, 75]. For one-stage STVSR methods, we choose the recent STVSR works (Zooming SlowMo [71], TMNet [75], RSTT [19], VideoINR [6], MoTIF [5]). For two-stage approaches, we apply combinations of frame-based VFI (EMA-VFI [82], AMT-G [43]) and frame-based VSR (EDVR [64], BasicVSR++ [3]), as well as event-based VFI (Time Lens [61], CBMNet [36]), event-based VSR (E2SRI [15], EGVSR [47]), and event stream super resolution (ESR [66]). E2SRI relies solely on event-based data, so for a fair comparison, we make slight modifications to incorporate frames, denoted as E2SRI*. We train all networks from scratch using public codes provided by authors because using pre-trained networks resulted in inferior performance due to differences in data distribution.

4.3.1 Experiments on CED and BS-ERGB datasets.

We report the quantitative comparison results in Table 1. To evaluate the speed of inference, we adopted the same input dimensions of 112×64 as previous studies [75]. Event-based methods outperform those using only images by leveraging temporal information, but not as much as our approach. Our method significantly outperforms existing

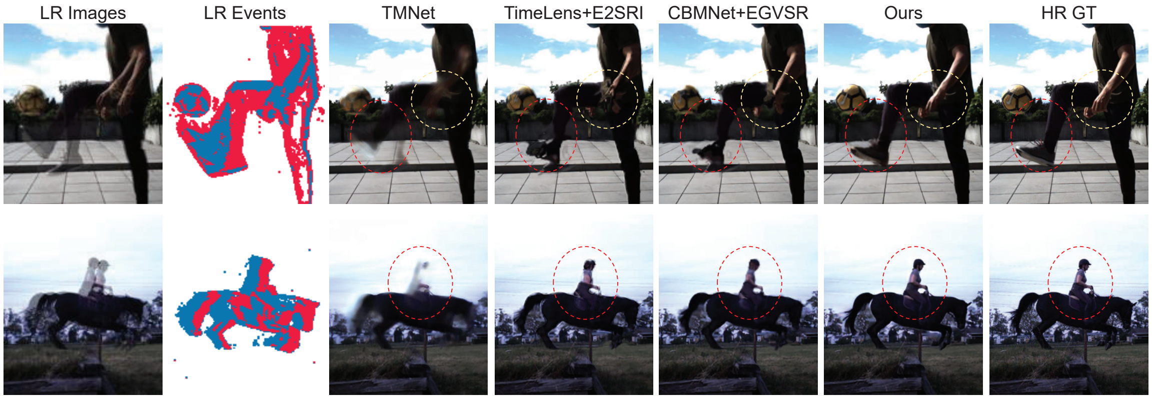

Figure 6. Visual comparisons on the BS-ERGB dataset. Best viewed when zoomed in.

Table 3. Quantitative comparison on SuperREST dataset.

Method	SuperREST			
	7 skip (center)		7 skip (whole)	
VFI+VSR / VSR+VFI / STVSR	PSNR	SSIM	PSNR	SSIM
TMNet [75]	23.96	0.7611	24.93	0.7763
VideoINR [6]	23.84	0.7595	24.23	0.7645
TimeLens [61] + E2SRI* [15]	28.01	0.8206	28.14	0.8315
TimeLens [61] + EGVSR [47]	27.94	0.8184	28.22	0.8275
CBMNet [36] + E2SRI* [15]	29.09	0.8426	29.04	0.8410
CBMNet [36] + EGVSR [47]	29.07	0.8397	28.98	0.8398
E2SRI* [15] + ESR [66] + TimeLens [61]	27.96	0.8216	28.32	0.8342
E2SRI* [15] + ESR [66] + CBMNet [36]	29.15	0.8369	29.27	0.8420
EGVSR [47] + ESR [66] + TimeLens [61]	28.14	0.8265	28.44	0.8382
EGVSR [47] + ESR [66] + CBMNet [36]	29.10	0.8337	29.07	0.8331
ESTNet-S (Ours)	<u>29.94</u>	<u>0.8543</u>	<u>29.91</u>	<u>0.8541</u>
ESTNet-L (Ours)	**30.27**	**0.8600**	**30.23**	**0.8595**

event-based methods. One can see that our ESTNet-L outperforms the third best method (E2SRI* [15] + ESR [66] + CBMNet [36]) by 1.27dB, and 1.72dB on the 3 skip and 7 skip of CED datasets. Similarly, on the challenging BS-ERGB dataset with significant and complex motion, our method demonstrates superior performance. Compared to the best-performing combination (EGVSR [47] + ESR [66] + CBMNet [36]) among the other methods, ESTNet-L achieves a higher PSNR by 1.38 dB and 1.63 dB on 1 skip and 3 skip, respectively. Through this, it is evident that incorporating events into the network alone does not yield excellent performance, and applying it sequentially without considering the correlation between spatial and temporal dimensions does not lead to optimal performance. In contrast, our ESTNet achieves superior performance by using high temporal resolution when upsampling the spatial dimension and by accurately performing temporal alignment using events at high resolution. Moreover, our method exhibits the most favorable values concerning model parameters and run time.

In Fig. 6, we provide a qualitative comparison with other state-of-the-art STVSR methods on BS-ERGB dataset. We select scenes featuring intricate and non-linear motion to showcase the superiority of our approach. Both scenarios involve non-linear motion, making it a highly ill-posed problem where intermediate frames cannot be accurately predicted from frame information alone (*e.g.*, TMNet [75]), leading to artifacts. The event-based method effectively addresses this issue by utilizing event data, but artifacts persist due to error accumulation resulting from the lack of spatial detail preservation and the absence of space-time interaction. On the other hand, our method accomplishes both precise alignment and super-resolution concurrently, even under challenging conditions.

4.3.2 Experiments on SuperREST dataset.

We present the results on SuperREST dataset in Tab. 3. The notable difference in the results compared to the BS-ERGB and CED datasets is the underperformance of frame-based methods, which can be attributed to the combination of both camera ego-motion and object motion, as well as the abundance of high-frequency structures in the SuperREST. When compared to the best-performing existing methods, E2SRI* [15] + ESR [66] + CBMNet [36], and the frame-based method, TMNet [75], there is a notable PSNR difference of 5.19 dB at the center of the 7 skip setting. Our method continues to deliver the best performance and achieves consistently high results across various datasets. The superiority of ours can be better observed in Fig. 7.

5. Ablation Study

Effectiveness of each component. We conduct an ablation study for each module on the CED dataset, as shown in Table 4. To substitute for TDF and MTA, we employ multiple convolution blocks. The results of A, B, and C demonstrate that using TDF and MTA in isolation alone does not yield satisfactory results. In contrast, the result of D shows that these modules interact and provide significant gains when applied together, with a 1.41 dB increase compared to A. Additionally, it's worth noting that while both network A and network E perform temporal alignment, the network E, which additionally performs temporal alignment at high resolution, achieves high performance through precise matching. The best performance is achieved by using all proposed modules.

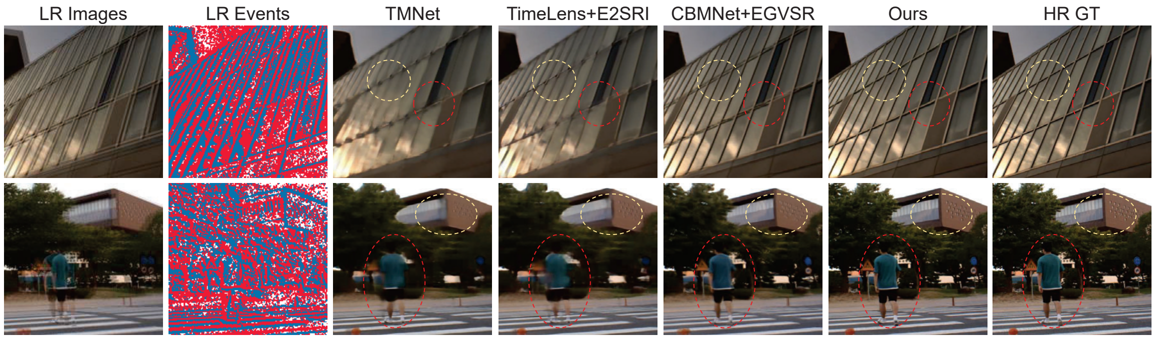

Figure 7. Comparisons on the SuperREST dataset. More results are in the supplementary material.

Table 4. Ablation study of each component.

	A	B	C	D	E	F	G	H
TDF		✓		✓		✓		✓
MTA			✓	✓			✓	✓
BSDA					✓	✓	✓	✓
PSNR	26.35	26.92	27.14	27.76	27.90	28.12	28.34	**28.62**

Table 5. Ablation studies of the number of multi-frames in TDF.

T	1	2	4 (Ours)	8
PSNR	28.01	28.21	**28.62**	28.59

Effectiveness of the number of multi-frames in TDF. In Table 5, we present results for our method with varying numbers of multi-frames, T. We observe that as we increase the number of frames from one to four, the results improve progressively, as it incorporated more temporal information from temporal alignment. However, further increasing the range from four to eight reduce the performance because aligning a greater quantity of frames becomes more challenging for MTA. Consequently, we empirically set the number of multi-frame to four.

Effectiveness of the number of sampling points in BSDA. In Table 6, we provide the performance results based on the number of sampling locations, m, in BSDA. As the number of sampling points increases, the performance improves from 1 to 4 because more points contribute to aggregation in temporal alignment. However, beyond that point, the increasing complexity and the larger number of cases lead to a decrease in performance. We believe the reason for this is that a greater number of matching candidates makes it more difficult for the network to learn. Therefore, we experimentally set the value of m to 4.

Analysis of BSDA. Tab. 7 presents comparisons between our method and other prominent temporal alignments [61, 87] on the SuperREST dataset. DCN achieves better performance than optical flow because it utilizes more sampling points for weight-based aggregation. Nevertheless, it overlooks the potential information provided by backward alignment when dealing with forward alignment. In contrast, our method achieves the best performance by considering the correlation between forward and backward alignments. These advantages can be visually observed in Fig. 8, which includes challenging objects. Both DCN and BSDA perform aggregation with a range of sampling points. However, BSDA emphasizes larger weights at positions corresponding to the reference, while DCN does not. Furthermore, optical flow often exhibits mismatches when dealing with uncertain scenarios involving moving objects, as it relies on a single sampling point.

Table 6. Experiments on the number of sampling points in BSDA.

m	1	2	3	4 (Ours)	5
PSNR	28.33	28.43	28.57	**28.62**	28.60

Table 7. Performance comparison based on various alignments.

Method	Flow [61]	DCN [87]	BSDA (Ours)
PSNR / SSIM	29.86 / 0.8522	30.08 / 0.8576	**30.27 / 0.8600**

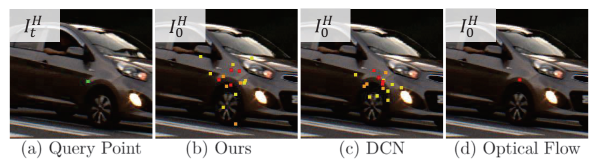

Figure 8. Qualitative comparison of temporal forward alignment. The left green marker represents the reference point in I_t^H, and the three figures to the right depict sampling points in I_0^H. The increasing shades of red indicate higher weights for aggregation.

6. Conclusion

Our work presents a novel method, ESTNet, that fully utilizes a high temporal frequency of events in STVSR. We break new ground by harnessing the high temporal frequency of event cameras to interpolate both spatial and temporal dimensions simultaneously, and extensive experiments demonstrate the effectiveness of ESTNet. The proposed SuperREST dataset is ideal for STVSR tasks, and we expect it to become a valuable resource for future research.

Acknowledgements. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00457882, National AI Research Lab Project) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

References

[1] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 1

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947–4956, 2021. 1

[3] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5972–5981, 2022. 1, 2, 4, 6

[4] Jiaben Chen, Yichen Zhu, Dongze Lian, Jiaqi Yang, Yifu Wang, Renrui Zhang, Xinhang Liu, Shenhan Qian, Laurent Kneip, and Shenghua Gao. Revisiting event-based video frame interpolation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1292–1299. IEEE, 2023. 2

[5] Yi-Hsin Chen, Sihan Chen, Yen-Yu Lin, and Wen-Hsiao Peng. Motif: Learning motion trajectories with local implicit neural functions for continuous space-time video super-resolution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23074–23084, 2023. 1, 6

[6] Zeyuan Chen, Yinbo Chen, Jingwen Liu, Xingqian Xu, Vidit Goel, Zhangyang Wang, Humphrey Shi, and Xiaolong Wang. 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*, pages 2047–2057, 2022. 2, 6, 7

[7] Hoonhee Cho, Jegyeong Cho, and Kuk-Jin Yoon. Learning adaptive dense event stereo from the image domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17797–17807, 2023. 3

[8] Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, and Kuk-Jin Yoon. Non-coaxial event-guided motion deblurring with spatial alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12492–12503, 2023. 2

[9] Hoonhee Cho, Jae-Young Kang, and Kuk-Jin Yoon. Temporal event stereo via joint learning with stereoscopic flow. In *European Conference on Computer Vision*, pages 294–314. Springer, 2025. 3

[10] Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, and Kuk-Jin Yoon. Label-free event-based object recognition via joint learning with image reconstruction from events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19866–19877, 2023. 3

[11] Hoonhee Cho, Taewoo Kim, Yuhwan Jeong, and Kuk-Jin Yoon. Tta-evf: Test-time adaptation for event-based video frame interpolation via reliable pixel and sample estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25701–25711, 2024. 1

[12] Hoonhee Cho and Kuk-Jin Yoon. Event-image fusion stereo using cross-modality feature propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 454–462, 2022. 2

[13] Hoonhee Cho and Kuk-Jin Yoon. Selection and cross similarity for event-image deep stereo. In *European Conference on Computer Vision*, pages 470–486. Springer, 2022. 2

[14] Hoonhee Cho, Sung-Hoon Yoon, Hyeokjun Kweon, and Kuk-Jin Yoon. Finding meaning in points: Weakly supervised semantic segmentation for event cameras. In *European Conference on Computer Vision*, pages 266–286. Springer, 2025. 2

[15] Jonghyun Choi, Kuk-Jin Yoon, et al. Learning to super resolve intensity images from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2768–2776, 2020. 1, 2, 6, 7

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[17] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 1, 2

[18] Yue Gao, Siqi Li, Yipeng Li, Yandong Guo, and Qionghai Dai. Superfast: 200× video frame interpolation via event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:7764–7780, 2022. 1

[19] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17441–17451, 2022. 2, 5, 6

[20] Jin Han, Yixin Yang, Chu Zhou, Chao Xu, and Boxin Shi. Evintsr-net: Event guided multiple latent frames reconstruction and super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4882–4891, 2021. 2, 3

[21] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Space-time-aware multi-resolution video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2020. 2

[22] Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, and Jianxing Liao. Timereplayer: Unlocking the potential of event cameras for video interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17804–17813, 2022.

[23] Mengshun Hu, Kui Jiang, Zhixiang Nie, Jiahuan Zhou, and Zheng Wang. Store and fetch immediately: Everything is all you need for space-time video super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 863–871, 2023. 1

[24] Zhewei Huang, Ailin Huang, Xiaotao Hu, Chen Hu, Jun Xu, and Shuchang Zhou. Scale-adaptive feature aggregation

for efficient space-time video super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4228–4239, 2024. 1

[25] Takashi Isobe, Xu Jia, Xin Tao, Changlin Li, Ruihuang Li, Yongjie Shi, Jing Mu, Huchuan Lu, and Yu-Wing Tai. Look back and forth: Video super-resolution with explicit temporal difference modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17411–17420, 2022. 1

[26] Yuhwan Jeong, Hoonhee Cho, and Kuk-Jin Yoon. Towards robust event-based networks for nighttime via unpaired day-to-night event translation. In *European Conference on Computer Vision*, pages 286–306. Springer, 2025. 2

[27] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9000–9008, 2018. 1, 2

[28] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1578–1587, 2023. 1

[29] Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, and Dacheng Tao. Turning frequency to resolution: Video super-resolution via event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7772–7781, 2021. 1, 2, 3, 4

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[31] Dachun Kai, Jiayao Lu, Yueyi Zhang, and Xiaoyan Sun. Evtexture: Event-driven texture enhancement for video super-resolution. *arXiv preprint arXiv:2406.13457*, 2024. 2

[32] Dachun Kai, Yueyi Zhang, and Xiaoyan Sun. Video super-resolution via event-driven temporal alignment. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2950–2954. IEEE, 2023. 2

[33] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2071–2082, 2023. 1

[34] Onur Selim Kılıç, Ahmet Akman, and A Aydın Alatan. E-vfia: Event-based video frame interpolation with attention. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8284–8290. IEEE, 2023. 1

[35] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Fisr: deep joint frame interpolation and super-resolution with a multi-scale temporal loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11278–11286, 2020. 2

[36] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. 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*, pages 18032–18042, 2023. 3, 4, 6, 7

[37] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Cmta: Cross-modal temporal alignment for event-guided video deblurring. *arXiv preprint arXiv:2408.14930*, 2024. 2

[38] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24966–24976, 2024. 3

[39] Taewoo Kim, Jaeseok Jeong, Hoonhee Cho, Yuhwan Jeong, and Kuk-Jin Yoon. Towards real-world event-guided low-light video enhancement and deblurring. In *European Conference on Computer Vision*, pages 433–451. Springer, 2025. 2

[40] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European conference on computer vision (ECCV)*, pages 106–122, 2018. 4

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[42] Tao Li, Xiaohai He, Qizhi Teng, Zhengyong Wang, and Chao Ren. Space–time super-resolution with patch group cuts prior. *Signal Processing: Image Communication*, 30:147–165, 2015. 1, 2

[43] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 6

[44] Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pages 695–710. Springer, 2020. 2

[45] Yuhan Liu, Yongjian Deng, Hao Chen, Bochen Xie, Youfu Li, and Zhen Yang. Event-based video frame interpolation with edge guided motion refinement. *arXiv preprint arXiv:2404.18156*, 2024. 1

[46] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4

[47] Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, and Lin Wang. 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*, pages 1557–1567, 2023. 1, 2, 5, 6, 7

[48] Uma Mudenagudi, Subhashis Banerjee, and Prem Kumar Kalra. Space-time super-resolution using graph-cut optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):995–1008, 2010. 1, 2

[49] Simon Niklaus, Ping Hu, and Jiawen Chen. Splatting-based synthesis for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 713–723, 2023. 1

[50] Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22191–22200, 2023. 3

[51] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *International Conference on Computer Vision*, 2021. 3, 4

[52] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3857–3866, 2019. 3

[53] Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 5, 6

[54] Eli Shechtman, Yaron Caspi, and Michal Irani. Increasing space-time resolution in video. In *European Conference on Computer Vision*, pages 753–768. Springer, 2002. 2

[55] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[56] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[57] Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, and Luc Van Gool. Event-based frame interpolation with ad-hoc deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18043–18052, 2023. 1

[58] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2

[59] Minggui Teng, Chu Zhou, Hanyue Lou, and Boxin Shi. Nest: Neural event stack for event-based image enhancement. In *European Conference on Computer Vision*, pages 660–676. Springer, 2022. 1

[60] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. *arXiv preprint arXiv:2203.17191*, 2022. 2, 3, 5, 6

[61] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16155–16164, 2021. 2, 3, 4, 6, 7, 8

[62] JD Van Ouwerkerk. Image super-resolution survey. *Image and vision Computing*, 24(10):1039–1052, 2006. 5

[63] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint framework for single image reconstruction and super-resolution with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7657–7673, 2021. 3

[64] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 6

[65] Yingwei Wang, Takashi Isobe, Xu Jia, Xin Tao, Huchuan Lu, and Yu-Wing Tai. Compression-aware video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2012–2021, 2023. 1

[66] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Boosting event stream super-resolution with a recurrent neural network. In *European Conference on Computer Vision (ECCV)*, 2022. 6, 7

[67] Song Wu, Kaichao You, Weihua He, Chen Yang, Yang Tian, Yaoyuan Wang, Ziyang Zhang, and Jianxing Liao. Video interpolation by event-driven anisotropic adjustment of optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3

[68] Bin Xia, Jingwen He, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, and Luc Van Gool. Structured sparsity learning for efficient video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22638–22647, 2023. 1

[69] Lujie Xia, Jing Zhao, Ruiqin Xiong, and Tiejun Huang. Svfi: spiking-based video frame interpolation for high-speed motion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2910–2918, 2023. 2

[70] Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21572–21581, 2023. 2

[71] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. 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*, pages 3370–3379, 2020. 1, 2, 5, 6

[72] Xijie Xiang, Lin Zhu, Jianing Li, Yixuan Wang, Tiejun Huang, and Yonghong Tian. Learning super-resolution reconstruction for high temporal resolution spike stream. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(1):16–29, 2021. 2

[73] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), pages 2583–2592, October 2021. 2

[74] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 4

[75] Gang Xu, Jun Xu, Zhen Li, Liang Wang, Xing Sun, and Ming-Ming Cheng. Temporal modulation network for controllable space-time video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6388–6397, 2021. 1, 2, 4, 5, 6, 7

[76] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[77] Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. *IEEE Journal of Solid-State Circuits*, 50(9):2149–2160, 2015. 1

[78] Chenyu You, Lianyi Han, Aosong Feng, Ruihan Zhao, Hui Tang, and Wei Fan. Megan: memory enhanced graph attention network for space-time video super-resolution. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1401–1411, 2022. 1

[79] Lei Yu, Bishan Wang, Xiang Zhang, Haijian Zhang, Wen Yang, Jianzhuang Liu, and Gui-Song Xia. Learning to super-resolve blurry images with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[80] Zhiyang Yu, Yu Zhang, Deyuan Liu, Dongqing Zou, Xijun Chen, Yebin Liu, and Jimmy S Ren. Training weakly supervised video frame interpolation with events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14589–14598, 2021. 2

[81] Huanjing Yue, Zhiming Zhang, and Jingyu Yang. Real-rawvsr: Real-world raw video super-resolution with a benchmark dataset. In *European Conference on Computer Vision*, pages 608–624. Springer, 2022. 1

[82] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023. 1, 6

[83] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17774, 2022. 2

[84] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021. 3

[85] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 5

[86] Juanjuan Zhu, Zhexiong Wan, and Yuchao Dai. Video frame prediction from a single image and events. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7748–7756, 2024. 2

[87] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 3, 8

[88] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based optical flow using motion compensation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3