Video Frame Interpolation via Direct Synthesis with the Event-based Reference

Yuhan Liu[1] Yongjian Deng[1*] Hao Chen[2] Zhen Yang[1]
[1]College of Computer Science, Beijing University of Technology
[2]Key Lab of Computer Network and Information Integration, Southeast University
{liuyuhan@emails., yjdeng@, yangzhen@}bjut.edu.cn, haochen303@seu.edu.cn

Abstract

Video Frame Interpolation (VFI) has witnessed a surge in popularity due to its abundant downstream applications. Event-based VFI (E-VFI) has recently propelled the advancement of VFI. Thanks to the high temporal resolution benefits, event cameras can bridge the informational void present between successive video frames. Most state-of-the-art E-VFI methodologies follow the conventional VFI paradigm, which pivots on motion estimation between consecutive frames to generate intermediate frames through a process of warping and refinement. However, this reliance engenders a heavy dependency on the quality and consistency of keyframes, rendering these methods susceptible to challenges in extreme real-world scenarios, such as missing moving objects and severe occlusion dilemmas.

This study proposes a novel E-VFI framework that directly synthesize intermediate frames leveraging event-based reference, obviating the necessity for explicit motion estimation and substantially enhancing the capacity to handle motion occlusion. Given the sparse and inherently noisy nature of event data, we prioritize the reliability of the event-based reference, leading to the development of an innovative event-aware reconstruction strategy for accurate reference generation. Besides, we implement a bi-directional event-guided alignment from keyframes to the reference using the introduced E-PCD module. Finally, a transformer-based decoder is adopted for prediction refinement. Comprehensive experimental evaluations on both synthetic and real-world datasets underscore the superiority of our approach and its potential to execute high-quality VFI tasks.

1. Introduction

Video frame interpolation (VFI) is an important direction in current computer vision research and finds widespread applications in various domains, including slow-motion generation [1, 16, 46], and video compression [44]. It is primarily used to generate non-existent intermediate frames in video sequences, enabling various effects such as video smoothing and high frame rate conversion. Existing VFI methods mostly employ motion-based strategies, estimating pixel-level motion from keyframes and using warping techniques to obtain interpolated frames. However, these methods have certain limitations when dealing with occlusion and non-linear motion, often struggling to accurately predict motion in complex scenes, thus affecting the quality of interpolated frames.

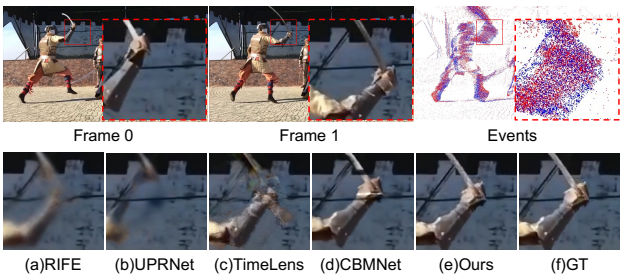

Figure 1. Qualitative comparison of the occlusion handling. (a) and (b) estimate occlusion mappings by frames, (c) use events for occlusion judgments, and (d) make occlusion judgments for optical flow at the feature level. Our method (e) achieves the best results by no longer estimating the occlusion mapping but giving a direct structural reference.

In recent years, the emergence of event cameras has brought new possibilities to address this challenge. Event cameras, detecting brightness changes asynchronously at the micro-second level [8–10, 34], can fill the information blank of the inter-frame gap. By leveraging event data, researchers can obtain more accurate motion estimation, resulting in higher-quality interpolated frames through a motion-based warping process [13, 21, 40, 45]. However, despite significant progress brought by more accurate motion estimation, another tricky issue in VFI tasks, namely occlusion problems such as severe occlusions (Fig. 1) and situations that the interpolation scene is not visible in the keyframes (Fig. 6), still remains unresolved.

We suggest the core challenge in achieving reliable

*: Corresponding author

frame interpolation lies in the unknown nature of inter-frame information. Whether it is traditional synthesis-based approaches [19, 38] or motion-based methods [15, 17, 30, 45], the generation process heavily relies on the information provided by the keyframes, such as scene content and temporal correlations. This raises us a question: If we could directly provide reliable references for intermediate frames, would it be possible to overcome the common limitations in frame interpolation tasks?

In response to this question, this study proposes a novel approach by directly synthesizing interpolated frames based on the inter-frame reference reconstructed from event data. While current E-VFI works leverage inter-frame cues from events as well, they typically treat this information as an augmentation to the keyframes, such as assisting in motion estimation of keyframes. Instead, our work inverts the role of events and keyframes. We regard keyframes as auxiliary information for completing the event-based reference, circumventing the challenges of fitting nonlinear motion and the occlusion issues introduced by warping operations. Meanwhile, by utilizing event data as a reference, we avoid the reliance on additional keyframes and long-term temporal messages commonly adopted in traditional synthesis-based methods [19, 38].

In specific, our E-VFI model consists of three stages: reconstruction, synthesis, and refinement. In the reconstruction stage, we use events solely to reconstruct a structural reference for the moving foreground at the interpolated positions. We introduce an event-aware reconstruction strategy built on a customized erosion and dilation operation to force the model to focus on enhancing the structural information of the regions related to moving objects as well as reducing the impact of noise except for these regions. In the synthesis stage, we extend the PCD module [42] to the E-VFI field, named E-PCD, which is designed to synthesize coarse interpolated frames by aligning keyframes to the computed reference leveraging explicit event-based guidance. Finally, we adopt an off-the-shelf Transformer-based decoder for refinement.

Our contributions can be summarized as follows:

- We propose a novel E-VFI framework that eliminates the need for motion estimation and directly synthesizes intermediate frames based on the event-based reference.
- We introduce a reconstruction strategy for obtaining the event-based reference, namely event-aware reconstruction strategy, aiming to emphasize precise structural information in the interpolated frame while suppressing regions with low confidence.
- A deformable synthesis module E-PCD is proposed to align useful keyframes' features with the event-based reference for information completion , resulting in coarse interpolated frames.
- Extensive experiments on both synthetic and real E-VFI datasets demonstrate the effectiveness and generalization of the proposed framework.

2. Related Work

Video Frame Interpolation (VFI). Four research paths can be distinguished. (1) Motion-based approaches [1, 2, 14–16, 18, 25, 30]; (2) Synthesis-based methods [19, 38]; (3) Kernel-based [1, 2, 4] and (4) phase-based [26] methodologies. Due to the strong constraints imposed by motion estimation, the motion-based method becomes a choice of most works. For example, Jiang *et al*. [16] and Niklaus *et al*. [29] employ an optical flow estimation network and compute occlusion maps to generate multiple intermediate frames. Park *et al*. [30] and Jin *et al*. [18] introduce bi-directional motion estimation strategies for handling more complex motion conditions. Building upon these works, studies in [17, 22] achieve state-of-the-art (SOTA) performance through synchronized updates of motion and synthesized images with the iterative feature learning schedule. However, these methods generally struggle to address extreme occlusion issues, holding defects in performing reliable VFI in the real world. To this end, recent works [19, 38] have revived research on synthesis-based methods aiming to avoid the occlusion issues brought by the warping process. However, without the assistance of motion estimation, these approaches can only achieve comparable performance by leveraging additional keyframes and temporal information, enlarging the training pressure. In this paper, we draw inspiration from synthesis-based approaches and instead of relying on additional temporal information, we employ the inter-frame events as references for the VFI task.

Event-based VFI (E-VFI). Thanks to the microsecond-level temporal resolution, event cameras can fill the inter-frame information blank, giving rise to numerous successful E-VFI works [11, 13, 21, 23, 40, 41, 45, 48, 50]. Mainstream E-VFI models are built on the combination of motion-based warping and synthesis. Initially, Tulyakov *et al*. [40] incorporate events into their workflow, using image-level attention to merge the warped and synthesized predictions. Following studies [13, 23, 41, 48, 50] have all been dedicated to improving the accuracy of motion estimation to achieve better results via the warping process. In this regard, Kim *et al*. [21] achieve SOTA performance with the help of motion predictions on both image- and feature-level. Yet, though achieving more accurate motion estimation, the natural limitations of motion-based approaches in handling large motion trajectories and extreme occlusion problems still remain unsolved. Gao *et al*. [11] attempt to handle these issues by providing a slow-fast joint synthesis network, yet the lack of direct reference at the interpolated frame causing their method easy to overfitting keyframes' patterns and thereby resulting in unsatisfied performance. In this work, we provide a solution to handle the above is-

sues by proposing a purely synthesis-based E-VFI method with the aid of event-based references containing structural messages of the interpolated frames.

Event-based Video Reconstruction. Event cameras can report pixel-level intensity changes between frames, theoretically allowing for the reconstruction of inter-frame details. Early reconstruction efforts depend on handcrafted features to estimate the intensity of events [3, 20, 27, 36]. With the rise of deep learning, numerous studies have begun to employ neural networks for video reconstruction [33, 34, 37, 39, 43]. However, due to the nature of event cameras that only detect areas with contrast changing, the reconstruction of smooth and static regions poses challenges and instability. Therefore, we focus on the reconstruction of regions with clear events and propose an event-aware reconstruction strategy to assist in restoring high-confidence structural information of moving scenes, improving the credibility of the reconstructed event-based reference.

3. Event Representation

The i-th event, denoted as e_i, in an event stream can be represented as (x_i, y_i, p_i, t_i). Here, x_i and y_i represent the spatial coordinates, p_i and t_i denote the polarity and the timestamp of events respectively. To handle the unstructured format of event data, a common approach [6, 7, 21, 40, 51] is to discretize the time dimension into B consecutive temporal bins and then integrate the events into a 3D spatio-temporal Voxel Grid ($E \in \mathbb{R}^{B\times H\times W}$) linearly. The integration of a specific temporal bin can be formulated as Eq. (1).

$$E(k) = \sum_i p_i \max\left(0, 1 - \left|k - \frac{t_i - t_0}{t_{N_e} - t_0}(B-1)\right|\right), \quad (1)$$

where t_0 and t_{N_e} respectively denote the start time and end time of the integrated event stream, and N_e represents the number of event data. The range of k is in $[0, B-1]$.

4. The Proposed Method

Given two input keyframes, I_0&I_1, and the events between them, we aim to estimate the intermediate frame I_τ at a specific timestamp between I_0 and I_1, where $\tau \in [0, 1]$ represents the timestamp of the interpolated frame. The overall structure of our E-VFI framework is shown in Fig. 2. It consists of the reconstruction, synthesis, and refinement stages.

We first divide events at τ into two segments and convert them into voxel grids $E_{0\to\tau}$ and $E_{\tau\to1}$ before feeding them to the network. At the reconstruction stage, we directly utilize $E_{0\to\tau}$ and $E_{\tau\to1}$ to reconstruct an event-based reference ($\mathcal{R}$), which contains the structural messages of the frame to be interpolated, following the proposed event-aware reconstruction strategy. Then, at the synthesis stage, we aim to synthesize a coarse interpolation by bidirectionally aligning keyframes to the event-based reference. We perform this process utilizing the proposed E-PCD through explicit guidance proffered by event data. Finally, we simply adopt an off-the-shelf Transformer-based decoder to refine the obtained coarse interpolation.

The following sections will describe the core components of our method in detail. *Please refer to the supplementary for the exact architecture of our method.*

4.1. Reconstruction of Event-Based Reference

State-of-the-art E-VFI methods [13, 15, 21, 45] primarily obtain the intermediate frame through warping process using motion flow estimated from events and keyframes. To achieve reliable interpolated frames, not only high-quality keyframes but accurate occlusion detection are necessary for these approaches. Specifically, keyframes lacking information about moving objects can adversely affect motion estimation, while poor occlusion detection can result in the presence of artifacts in the interpolated images. Estimating accurate occlusion in real-world scenarios with large motion and complex motion patterns is indeed a challenging task. To address these issues, we leverage the high temporal resolution of events and directly reconstruct an event-based reference, encoding the structural features of the interpolated frames precisely. Based on the event-based reference, we no longer need motion estimation or occlusion detection but can directly perform feature alignment and inter-frame synthesis.

Normally, the event-based reference ($\mathcal{R}$) can be simply reconstructed from event data using a U-Net [35]. However, we notice that event cameras are challenging to encode regions that are static or with low textures and often carry a significant amount of noise, resulting in inaccurate reconstruction of these places. Therefore, to address this issue and prevent error amplification through network flow, we introduce an event-aware reconstruction strategy that aims to emphasize precise structural information in the interpolated frame while suppressing regions with low confidence.

Event-Aware Reconstruction Strategy. In detail, we first compute an event-aware mask ($\mathbb{M}$) where each pixel records whether events occurred at its location, *e.g.* if an event has occurred, the corresponding pixel is set to 1, otherwise it is set to 0. Then, we apply a customized erosion operation to the $\mathbb{M}$ for removing the impact of low event confidence area in $\mathcal{R}$ reconstruction. Considering the extreme sparsity of events (as shown in Fig. 1), direct application of erosion operations, even with the smallest kernel size, could cause a significant loss of events. Hence, we define a kernel of size $n \times n$ and calculate the number of surrounding pixels whose values differ from the center pixel. If this number exceeds the threshold δ, we invert the value of the center pixel, as shown in Fig. 3.(a). Next, we use the dilation operation to connect the sparse event points in $\mathbb{M}$ into pathways for preserving the semantic coherence of the $\mathcal{R}$ to a certain extent.

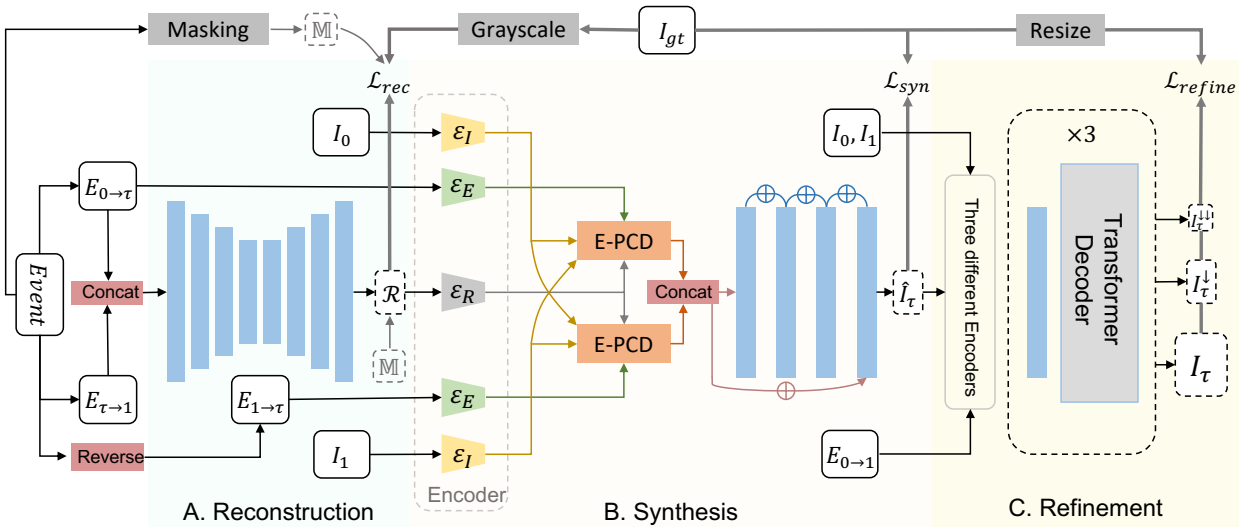

Figure 2. An overview of the proposed E-VFI framework. It consists of reconstruction, synthesis and refinement staged. The $\mathcal{L}_{rec}$, $\mathcal{L}_{syn}$ and $\mathcal{L}_{refine}$ are used to supervise the obtained event-based references $\mathcal{R}$, coarse interpolated frames $\hat{I}_\tau$, and final prediction $\hat{I}$, respectively. Modules with the same name share weights. The blue modules represent convolutional blocks.

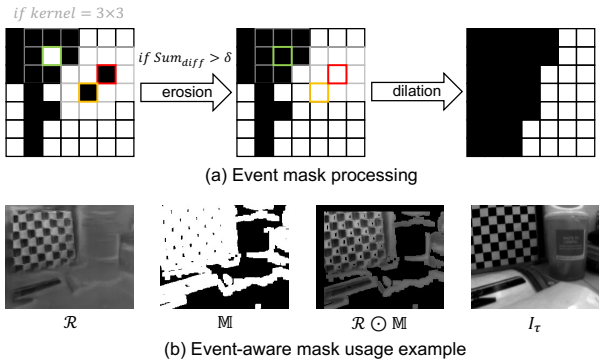

Figure 3. Illustration of the event-aware masking. (a) shows the generation of the event-aware mask, where yellow, green, and red refer to the center of the kernel. (b) shows a practical example, and the event-aware masking removes the adverse impact from unreliable locations.

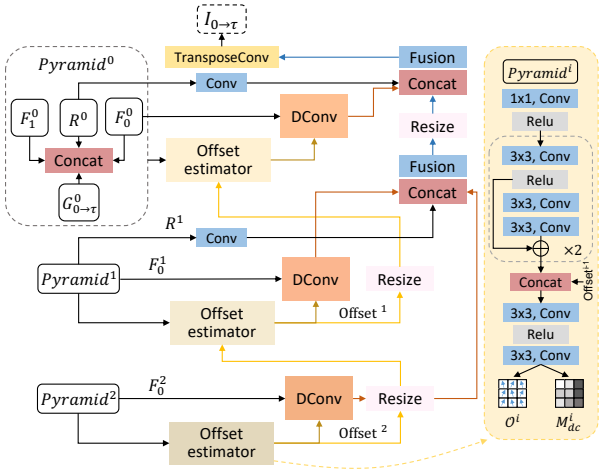

Figure 4. An overview of the proposed E-PCD module. Resize refers to up-sampling using bilinear interpolation.

Finally, we take the $\mathcal{R}$ masked by the $\mathbb{M}$ as input to the following modules. As shown in Fig. 3.(b), the failing reconstruction of the bottle's logo has been masked thereby alleviating wrong messages conveyed in the unmasked event-based reference.

During training, we introduce a pseudo ground truth ($I_{gt}^{\mathbb{M}\_gray}$) for supervising the reconstruction process. In particular, we first apply the same $\mathbb{M}$ to the I_{gt}. Furthermore, due to no color and precise value of pixels can be extracted from event data, we adopt the gray-scale version of the masked ground truth $I_{gt}^{\mathbb{M}\_gray}$ as the final supervision for the reconstruction stage. Leveraging the $I_{gt}^{\mathbb{M}\_gray}$, we aim to prevent the network from performing reference generation based on inductive biases introduced by whole images but focus on the specific task that only reconstructs the structural information from event data.

4.2. Synthesis via Event-Based Reference

The primary objective during the synthesis stage is to complete the event-based reference ($\mathcal{R}$) by utilizing the effective messages of keyframes through an aligning process.

As depicted in Fig. 2.B, the core at this stage is the synthesis module E-PCD, which is proposed by customizing the PCD (Pyramid, Cascading, and Deformable convolutions) module [42] to our task. Specifically, we introduce a bidirectional alignment mode, employing two E-PCD modules with shared weights to align I_0 and I_1 with the reference $\mathcal{R}$, respectively. Concurrently, we utilize the inter-

frame events ($E_{0\to\tau}$ and $E_{1\to\tau}$) to provide temporal and motion cues for this alignment process explicitly, assisting the network in discerning moving and static regions in the keyframes during alignment. Ultimately, we fuse alignment results with $\mathcal{R}$ at each scale, emphasizing the consistency and guidance of the structural information in $\mathcal{R}$. In the following, we will delineate the workflow of the E-PCD module. For clarity, we take the E-PCD module guided by $E_{0\to\tau}$ as an example.

E-PCD. The E-PCD module takes as input features extracted from keyframes, events, and the event-based reference as illustrated in Eq. (2).

$$\begin{aligned} F_m^{0,1,2} &= \mathcal{E}_I(I_m), m \in \{0,1\}, \\ G_o^{0,1,2} &= \mathcal{E}_E(E_o), o \in \{0\to\tau, 1\to\tau\}, \\ V^{0,1,2} &= \mathcal{E}_R(\mathcal{R}\odot\mathbb{M})), \end{aligned} \tag{2}$$

where $\odot$ denotes the Hadamard Product. Specifically, E-PCD consists of three pyramid feature levels, as shown in Fig. 4. At level i, we first calculate an offset and a mask using features of keyframes ($F_0^i \& F_1^i$), the event-based reference (V^i) and corresponding events ($G_{0\to\tau}^i$), as indicated in Eq. (3).

$$\begin{aligned} Pyramid^i &= \mathcal{C}(F_0^i, F_1^i, V^i, G_{0\to\tau}^i) \\ \mathcal{O}^i, M_{dc}^i &= \mathcal{F}_{OE}(Pyramid^i, \mathcal{O}^{i-1}), \end{aligned} \tag{3}$$

where $\mathcal{C}$ denotes the concatenation operation, $\mathcal{O}^i$, and M_{dc}^i denote the learned offset and mask for deformable convolution usage, $\mathcal{F}_{OE}$ represents the offset estimator in Fig. 4. Then, we can align I_0 with $\mathcal{R}$ through deformable convolution and fusion operations as formulated in Eq. (4).

$$\mathcal{I}_{0\to\tau}^i = \mathcal{F}_{fu}(\mathcal{C}(\mathcal{F}_{dc}(F_0^i, \mathcal{O}^i, M_{dc}^i), \mathcal{F}_{co}(V^i), \mathcal{I}_{0\to\tau}^{i-1})) \tag{4}$$

where $\mathcal{F}_{dc}$ represent the deformable convolution operation, $\mathcal{F}_{co}$ and $\mathcal{F}_{fu}$ denote feature projection and fusion modules built on convolution blocks, and $\mathcal{I}_{0\to\tau}^{i-1}$ denotes the output from the previous level. Regarding the alignment from I_1 to the reference $\mathcal{R}$, a symmetrical operation is performed. Finally, we use a dense residual network to fuse the obtained bidirectional features ($\mathcal{I}_{0\to\tau} \& \mathcal{I}_{1\to\tau}$) for achieving coarse interpolated frame $\hat{I}_\tau$.

4.3. Refinement

This stage is employed for solving silhouette issues and slight artifacts caused by the last two stages, and we achieve this target by adopting an off-the-shelf transformer-based decoder commonly used in previous works [21, 25, 32]. In specific, the coarse interpolation ($\hat{I}_\tau$), keyframes (I_0, I_1), and the inter-frame events ($E_{0\to1}$) are utilized for interactive learning within Transformer blocks to obtain final interpolation I_τ. *Please refer to the Supplementary for details.*

4.4. Loss

As shown in Fig. 2, our approach adopts a hybrid loss for each network stage as formulated in Eq. (5).

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{syn}\mathcal{L}_{syn} + \lambda_{refine}\mathcal{L}_{refine} \tag{5}$$

At the reconstruction stage, we utilize $\mathcal{L}_{rec}$ to optimize the structural accuracy of the event-based reference $\mathcal{R}$ in high confidence areas, supervised by the pseudo ground truth $I_{gt}^{\mathcal{M}\_gray}$, as depicted in Eq. (6).

$$\begin{aligned} \mathcal{L}_{rec} =& \mathcal{L}_{lpips}(\mathcal{R}\odot\mathbb{M}, I_{gt}^{\mathbb{M}\_gray}) + \\ & \mathcal{L}_1(\mathcal{R}\odot\mathbb{M}, I_{gt}^{\mathbb{M}\_gray}), \end{aligned} \tag{6}$$

where the $\mathcal{L}_{lpips}$ is the perceptual loss introduced in [49], which excels at measuring perceptual similarity between images, providing a more human-eye aligned evaluation compared to traditional pixel-wise losses. The function $\mathcal{L}_1$ denotes the L1 loss that measures pixel-level distances between two images.

Similarly, we adopt the combination of $\mathcal{L}_{lpips}$ and $\mathcal{L}_1$ for supervising the other two stages, as formulated in Eq. (7)

$$\begin{aligned} \mathcal{L}_{syn} &= \mathcal{L}_{lpips}(\hat{I}_\tau, I_{gt}) + \mathcal{L}_1(\hat{I}_\tau, I_{gt}) \\ \mathcal{L}_{refine} &= \mathcal{L}_{lpips}(I_\tau, I_{gt}) + \mathcal{L}_1^{0,1,2}(I_\tau, I_{gt}), \end{aligned} \tag{7}$$

where the $\mathcal{L}_1^{0,1,2}$ denotes the computation of losses at three different scales. The proposed model is trained end-to-end by minimizing $\mathcal{L}_{total}$, where $\lambda_{rec}, \lambda_{syn}$ and λ_{refine} are set as 1, 1, 1 respectively.

5. Experiment

5.1. Setup

Dataset. Following the optimization strategy in [40], we train our method on the training set of Vimeo90k-Septuplet [47], where synthetic event data is simulated using the ESIM [12]. For evaluation, we follow the evaluation methods used in [21, 40] and test our method on both synthetic and real-world datasets. For instance, we choose Vimeo90k-Triplet [47], GoPro[28], and SNU-FILM (Hard & Extreme) [5] as the evaluated synthetic datasets, where SNU-FILM is deemed to be the most challenging one since it only contains samples with complex motion conditions. Besides, we utilize two commonly adopted real-world datasets for testing, including High Quality Frames (HQF) DAVIS240 [39] and High Speed Event and RGB camera (HS-ERGB) [40]. These real-world datasets can measure the generalization and practical value of the proposed E-VFI approaches more accurately.

Training Settings. Our method is optimized by AdamW [24] with weight decay 10^{-4} for 40 epochs using PyTorch [31]. The initial learning rate was set to 10^{-4} and decreased

Table 1. PSNR(dB)/SSIM results on synthetic datasets. The best results are marked in **Bold** while the second ones are marked with underlines. We reconstructed all skipped frames for GoPro. †: Trained with full GoPro training set. ‡: The training strategy is identical to ours.

Method	Motion	Synthesis	Modal	Vimeo90k-Triplet	GoPro		SNU-FILM	
				1 frame	7 frame	15 frame	Hard	Extreme
BMBC[30]	✓	✗	F	35.06/0.944	25.45/0.755	24.29/0.752	29.23/0.921	23.60/0.833
RIFE[15]	✓	✗	F	34.74/0.957	29.66/0.889	25.14/0.772	30.36/0.920	25.54/0.853
UPR-Net-L[17]	✓	✗	F	36.24/0.966	27.91/0.855	24.61/0.758	30.82/0.928	25.61/0.862
VFIT-B[38]	✗	✓	F	31.94/0.926	-	-	30.95/0.932	27.82/0.880
A^{2}OF[45]†	✓	✗	$F\&E$	-	36.61/0.971	-	-	-
CBMNet-L[21]†	✓	✓	$F\&E$	-	38.15/0.975	37.05/0.969	-	-
TimeReplayer[13]‡	✓	✗	$F\&E$	35.12/0.963	-	-	-	-
A^{2}OF[45]‡	✓	✗	$F\&E$	36.54/0.967	34.08/0.954	32.65/0.937	31.72/0.924	28.21/0.890
TimeLens[40]‡	✓	✓	$F\&E$	36.31/0.962	34.81/0.959	33.21/0.942	31.75/**0.935**	28.64/0.889
CBMNet-L[21]‡	✓	✓	$F\&E$	37.69/0.970	36.07/0.972	35.46/0.966	29.59/0.885	28.99/0.858
Ours	✗	✓	$F\&E$	**39.17/0.977**	**36.95/0.975**	**35.77/0.968**	**33.04**/0.914	**31.46/0.893**

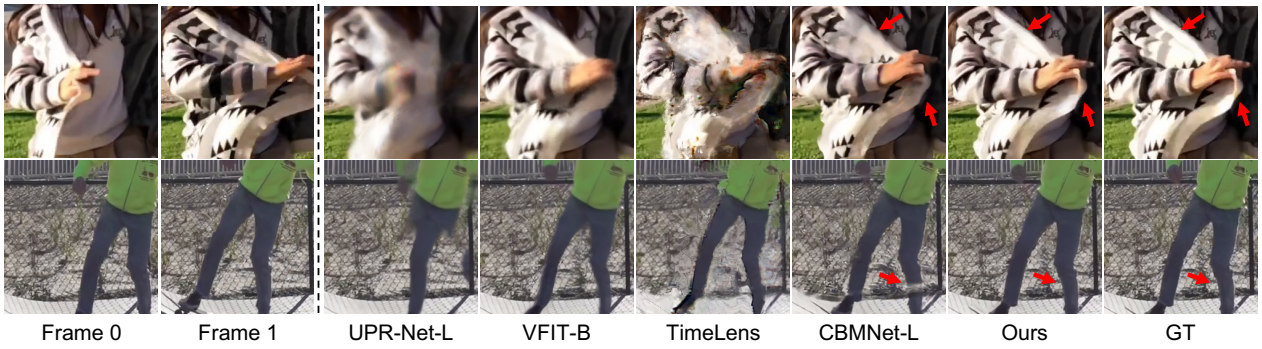

Figure 5. Visual comparison among different methods on synthetic datasets.

gradually to 10^{-6} using cosine annealing. The batch size for each training step was set to 6. We randomly select 3 frames from a set of 7, where the first and the third frames are keyframes (I_0, I_1), and the second frame is chosen as the ground truth frame (I_{gt}) to be interpolated. As for data augmentation, we crop the input frames and their paired event voxel grids to a size of 256×256 and randomly apply rotation and flipping. Also, we include a small set (10%) of GoPro training data for training to enlarge the capability of our method in handling large-resolution inputs.

Particulars. We set B as 8 for all event voxel grids construction. At the reconstruction stage, we set the kernel sizes of erosion and dilation operations as 3 and 5, respectively, and a threshold δ used in the erosion process as 6. The evaluation metrics utilized in our experimental section are SSIM and PSNR [19, 21, 40], which are commonly employed for the VFI task.

5.2. Comparison to State-of-the-Art Methods

In evaluating the effectiveness of our proposed method, we conduct a comprehensive comparison with state-of-the-art techniques in both VFI and E-VFI fields, categorizing them as follows: (1) Motion-based VFI approaches: This category includes BMBC [30], RIFE [15] and UPR-Net [17]. (2) Synthesis-based VFI techniques: Encompassing FLAVR [19]. (3) E-VFI methods containing both motion estimation and synthesis: Including TimeLens [40] and CBMNet-L [21]. (4) Motion-based E-VFI approaches: Comprising A^{2}OF [45] and TimeReplayer [13]. (5) Synthesis-based E-VFI method: SuperFast [11].

These studies employ diverse training strategies $w.r.t$ synthetic datasets, *e.g.* A^{2}OF and CBMNet-L are optimized using full GoPro training set. To ensure a fair comparison, we retrain these models with identical training strategy to ours as shown in Tab. 1. For the evaluation on real datasets, we fine-tune our model following the method proposed in [40, 45]. It is worth noting that CBMNet-L [21] does not test HS-ERGB in their paper. Therefore, we fine-tune their well pretrained model on this specific dataset for an accurate and fair analysis.

5.2.1 Evaluations on Synthetic Datasets

Quantitative Evaluations. In our initial assessment, we focus on the performance of each method on synthetic datasets, as presented in Tab. 1. Our method demonstrates exceptional VFI performance across various synthetic datasets. First, we conduct evaluations on the Vimeo90k-Triplet [47] and GoPro [28] datasets, comparing

Table 2. PSNR(dB)/SSIM results on real datasets. The best results are marked in **Bold** while the second ones are marked with underlines.

Method	Motion	Synthesis	Modal	HQF		HS-ERGB			
				1 frame	3 frame	Close		Far	
						5 frame	7 frame	5 frame	7 frame
BMBC[30]	✓	✗	F	30.72/0.881	27.09/0.741	29.22/0.820	27.98/0.807	25.62/0.741	24.14/0.710
RIFE[15]	✓	✗	F	31.70/0.889	27.93/0.796	33.12/0.857	32.32/0.846	29.47/0.849	27.20/0.801
UPR-Net-L[17]	✓	✗	F	32.15/0.915	27.96/0.863	32.22/0.841	31.01/0.829	28.85/0.841	26.27/0.787
VFIT-B[38]	✗	✓	F	31.50/0.882	-	-	-	-	-
TimeReplayer[13]	✓	✗	$F\&E$	31.07/0.931	28.82/0.866	31.21/0.818	29.83/0.816	31.98/0.861	30.07/0.834
A^{2}OF[45]	✓	✗	$F\&E$	33.94/0.945	31.85/0.932	33.21/**0.865**	32.55/0.852	33.64/0.891	33.15/0.883
SuperFast[11]	✗	✓	$F\&E$	-	-	-	32.50/0.869	-	27.87/0.845
TimeLens[40]	✓	✓	$F\&E$	33.42/0.934	32.27/0.917	32.19/0.839	31.68/0.835	33.13/0.877	32.31/0.869
CBMNet-L[21]	✓	✓	$F\&E$	34.77/0.953	33.08/0.940	34.17/0.862	33.96/0.857	31.43/0.888	30.47/0.870
Ours	✗	✓	$F\&E$	**35.89/0.959**	**34.27/0.941**	**34.60/0.865**	**34.33/0.862**	**34.19/0.923**	**33.56/0.921**

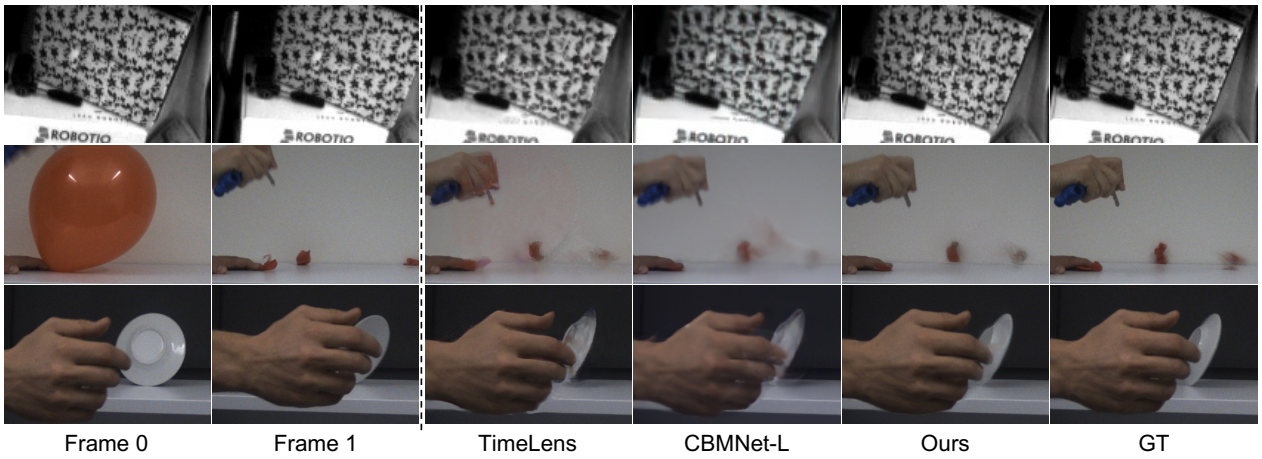

Figure 6. Visual comparison among different methods on real datasets.

our results with those of other methods. **Vimeo90k-Triplet Dataset:** Our method exhibits superior performance, surpassing other methods by a significant margin in various metrics. **GoPro Dataset:** With the same training strategy, our method reaches the leading performance among all compared approaches. However, we notice that our approach holds lower PSNR than CBMNet-L that trained on the full GoPro dataset. Considering the discrepancy between Vimeo-90K and GoPro in resolution (448×256 versus 1280×720) and scene types, We speculate the performance gap is caused by the different training strategies usage. To further emphasize the robustness of our method in handling extreme conditions, we conduct tests on the challenging scenarios presented in the **SNU-FILM dataset**. These scenarios involve highly demanding VFI tasks characterized by large and irregular motions. Our method holds a substantial performance advantage over other approaches, showcasing its resilience and adaptability under both normal and extreme motion conditions across diverse content types.

Qualitative Evaluations. Visual comparison results on synthetic datasets are illustrated in Fig. 5, showcasing the performance of various methods in challenging scenarios. As evident in Fig. 5, our method adeptly handles challenging issues in the VFI task, *e.g.* complex non-linear motion and occlusion issues introduced by moving clothes and legs. Compared to other methods, without explicit motion estimation, our approach accurately synthesizes moving objects with finer details while preserving the still background unchanged based on guidance from the event-based reference. These visual comparisons substantiate the efficacy of our proposed event-based reference for frame interpolation, affirming its ability to accurately structural information reconstruction and occlusion handling.

5.2.2 Evaluations on Real Datasets

Quantitative Evaluations. In this section, we assess the effectiveness of our method on real-world HQF and HS-ERGB datasets. These datasets present different sample distributions than synthetic datasets and include more intricate motion patterns. Also, these datasets, containing event data and their paired videos that collected in the real world, could better value the practical potential of E-VFI methods compared to synthetic datasets.

The test results, depicted in Tab. 2, highlight the leading performance of our model across all conditions and datasets. Notably, even in challenging scenarios like 3 skips in HQF, our model achieves 1dB PSNR higher than the SOTA E-VFI approach CBMNet-L. In addition, remarkable improvements are observed in both close and far scenarios of HS-ERGB, evident in both PSNR and SSIM metrics. These results underscore the superior image quality preservation, enhanced structural information reconstruction capabilities, and heightened robustness to complex motion in real-world scenarios offered by our model.

Qualitative Evaluations. Fig. 6 provides a visual comparison between different E-VFI approaches on real event datasets. A direct comparison with the outputs of TimeLens and CBMNet-L reveals the better capabilities of our approach, *e.g.* our method exhibits more accurate occlusion detection and delivers clearer depiction of plain pattern, exploded balloon fragments and the rotation of the plate. Interestingly, we find that with the aid of the event-based reference, we can even obtain the interpolation that does not occurred in keyframes such as the plate sides with clear edges. From these observations, we suggest that though these E-VFI methods perform well in most cases thanks to the accurate motion estimation, the severe occlusion issues introduced by the warping process or missing objects still cannot be handled well. Instead, with the help of event-based reconstruction, our pure synthesis-based E-VFI method shows capability in addressing this problem.

5.3. Model Analysis

In this section, we perform experiments on the HS-ERGB dataset to analyze the effectiveness of two core designs in our proposed method such as the event-aware reconstruction strategy and the E-PCD module.

The effectiveness of E-PCD. To validate the efficacy of the proposed E-PCD module, we introduce a baseline model that adopts the original PCD architecture, only keep our proposed bidirectional aligning mode but excluding the event-based guidance and multi-level reference fusion designed in E-PCD. The settings A and C in Tab. 3 demonstrate that our customized designs for events and event-based references, significantly enhance the PCD module's aligning capabilities from keyframes to the event-based reference, thereby improving the quality of the synthesized interpolated frames.

The Efficacy of the Reconstruction Strategy. Here, we verify the effectiveness of the Event-Aware Reconstruction Strategy (EARS) through settings B and C in Tab. 3. Comparative results indicate that the proposed EARS enhances the quality of synthesized interpolated frames effectively. This improvement is attributed to our reconstruction strategy's fitness with the intrinsic properties of event data, namely its extreme sparsity and noise. The event-aware mask we computed can mitigate these disadvantages of event data. Furthermore, the efficacy of EARS was also evaluated from a qualitative perspective. As shown in Fig. 7, the results with EARS are markedly superior to those without. This disparity is particularly pronounced in scenes with small motions such as areas of water or sky. These regions do not generate dense events but do produce considerable noise, resulting in inaccurate references in these areas and, consequently, impacting the interpolated frames.

Table 3. Event mask settings were evaluated in the spinning umbrella scenario of the HS-ERGB dataset.

Variants	EACS	PCD	E-PCD	HS-ERGB (far)	
				PSNR	SSIM
A	✓	✓		29.92	0.852
B			✓	32.47	0.906
C	✓		✓	**33.56**	**0.921**

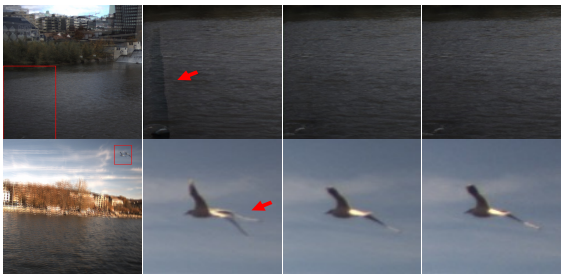

Figure 7. Comparison of the effect of using event-aware reconstruction strategy.

6. Conclusion

This study introduces a novel pure synthesis-based framework for the E-VFI task, facilitating the generation of interpolated frames through the alignment of keyframes to the event-based reference. We have incorporated an Event-Aware Reconstruction Strategy, tailored to properties of events, ensuring that the event-based reference retains the structural cues of high-confidence regions while mitigating the adverse effects of sparsity and noise prevalent in event data on interpolation. Furthermore, we propose the E-PCD, a specialized bidirectional alignment module for synthesizing interpolated frames while preserving the structural information in the reference. Our method addresses the occlusion challenges in VFI tasks by directly generating an event-based reference for the frames to be interpolated. Extensive experiments on synthetic and real datasets substantiate the superiority of our model and its leading performance.

Acknowledgments. This work is jointly supported by National Key R&D Program of China (2022YFF0610000), the National Natural Science Foundation of China (62203024, 92167102, 62102083, 62173286, 61875068, 62177018, 62261160576), the R&D Program of Beijing Municipal Education Commission (KM202310005027), the Natural Science Foundation of Jiangsu Province (BK20210222).

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

[2] Wenbo Bao, Wei-Sheng Lai, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Memc-net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):933–948, 2019. 2

[3] Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892, 2016. 3

[4] Xianhang Cheng and Zhenzhong Chen. Video frame interpolation via deformable separable convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10607–10614, 2020. 2

[5] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10663–10671, 2020. 5

[6] Yongjian Deng, Hao Chen, Huiying Chen, and Youfu Li. Learning from images: A distillation learning framework for event cameras. *IEEE Transactions on Image Processing*, 30: 4919–4931, 2021. 3

[7] Yongjian Deng, Hao Chen, and Youfu Li. Mvf-net: A multi-view fusion network for event-based object classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8275–8284, 2022. 3

[8] Yongjian Deng, Hao Chen, Hai Liu, and Youfu Li. A voxel graph cnn for object classification with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1172–1181, 2022. 1

[9] Yongjian Deng, Hao Chen, Bochen Xie, Hai Liu, and Youfu Li. A dynamic graph cnn with cross-representation distillation for event-based recognition. *arXiv preprint arXiv:2302.04177*, 2023.

[10] 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

[11] 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(6):7764–7780, 2023. 2, 6, 7

[12] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 5

[13] 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. 1, 2, 3, 6, 7

[14] Ping Hu, Simon Niklaus, Stan Sclaroff, and Kate Saenko. Many-to-many splatting for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3553–3562, 2022. 2

[15] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 624–642. Springer, 2022. 2, 3, 6, 7

[16] 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

[17] 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. 2, 6, 7

[18] Xin Jin, Longhai Wu, Guotao Shen, Youxin Chen, Jie Chen, Jayoon Koo, and Cheul-hee Hahm. Enhanced bi-directional motion estimation for video frame interpolation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5049–5057, 2023. 2

[19] 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. 2, 6

[20] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016. 3

[21] 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 (CVPR)*, pages 18032–18042, 2023. 1, 2, 3, 5, 6, 7

[22] 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. 2

[23] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 695–710. Springer, 2020. 2

[24] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2017. 5

[25] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3532–3542, 2022. 2, 5

[26] Simone Meyer, Abdelaziz Djelouah, Brian McWilliams, Alexander Sorkine-Hornung, Markus Gross, and Christopher Schroers. Phasenet for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 498–507, 2018. 2

[27] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126:1381–1393, 2018. 3

[28] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 5, 6

[29] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1710, 2018. 2

[30] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 109–125. Springer, 2020. 2, 6, 7

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[32] Markus Plack, Karlis Martins Briedis, Abdelaziz Djelouah, Matthias B Hullin, Markus Gross, and Christopher Schroers. Frame interpolation transformer and uncertainty guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9811–9821, 2023. 5

[33] 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

[34] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 3

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[36] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. *IEEE Robotics and Automation Letters*, 4(2):816–822, 2019. 3

[37] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 156–163, 2020. 3

[38] Zhihao Shi, Xiangyu Xu, Xiaohong Liu, Jun Chen, and Ming-Hsuan Yang. Video frame interpolation transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17482–17491, 2022. 2, 6, 7

[39] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, L. Kleeman N. Barnes, and R. Mahoney. Reducing the sim-to-real gap for event cameras. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 5

[40] 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. 1, 2, 3, 5, 6, 7

[41] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022. 2

[42] 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. 2, 4

[43] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 3

[44] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 416–431, 2018. 1

[45] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 267–283. Springer, 2022. 1, 2, 3, 6, 7

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

[47] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127 (8):1106–1125, 2019. 5, 6

[48] 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

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[50] 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

[51] 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. 3