Video Frame Interpolation via Asymmetric Refinement with the Event-based Reference

Yuhan Liu, *Graduate Student Member, IEEE,* Yongjian Deng†, Linghui Fu, Hao Chen, Zhen Yang, Youfu Li, *Life Fellow, IEEE,* and Boxin Shi, *Senior Member, IEEE*

***Abstract*—This work focuses on leveraging motion cues from event data to address two fundamental challenges inherent in mainstream motion-based video frame interpolation methods (VFI): the occlusion problem and the challenge of estimating non-rigid motion. To this end, we propose a novel framework named DSER++ (Direct Synthesis via Event-based Reference ++) for event-based video frame interpolation (E-VFI). DSER++ avoids estimation errors induced by complex motion patterns through a three-stage purely synthetic process. Particularly, we encode the event data, which contains precise motion trajectories between keyframes, into a sparse reference assisted by the proposed Event-Aware Reconstruction Strategy. Next, we employ the proposed E-PCD module to achieve bidirectional event-guided alignment between keyframes and the reference. Considering the deficiency of events in color and smooth area encoding, the reconstruction for occlusion regions inevitably contains color fluctuations and semantic incoherence. Thus, we introduce a Bidirectional Asymmetric Selective Refinement (BASR) module to targeted optimize such regions of the reconstruction scene by recognizing and providing precise keyframes' supports. Finally, we equip the multi-scale BASR into an off-the-shelf interpolation refinement network for final predictions. Comprehensive experimental evaluations on both synthetic and real-world datasets underscore the superiority of our approach and its potential to execute high-quality E-VFI tasks. Codes are available at https://github.com/yuhan0802/DSER-pp.**

***Index Terms*—Event camera, Event-based Video Frame Interpolation (E-VFI), Selective Feature Refinement, Multi-Modal Fusion**

I. Introduction

Video frame interpolation (VFI) is a promising computer vision task with widespread applications across various downstream domains, including slow-motion generation [2]–[4], video compression [5], medical imaging [6] and biological motion research [7]. The target of VFI is to obtain intermediate frames between two consecutive keyframes (I_0 & I_1) by leveraging motion and contextual estimation from I_0 and I_1, enabling various effects such as video smoothing and high frame rate conversion. However, blind intervals, where explicit pixel-level correspondences between consecutive frames are absent, pose significant challenges to VFI. To address this issue, existing VFI methods [2], [3], [8]–[15] can be broadly divided into synthesis-based and motion-based approaches. Synthesis-based methods [13], [14], which primarily rely on global representations, often struggle to capture critical pixel-level details. Consequently, they face difficulties in maintaining color consistency and reconstructing the fine details of fast-moving objects. In contrast, despite the inherent disadvantages of motion-based methods (Fig. 1(a)) in accurately handling occlusions and non-rigid motion, their ability to preserve fine-grained spatial details and ensure color consistency has made them the more dominant choice in the field of VFI.

In recent years, the emergence of event cameras has provided new potential solutions for VFI tasks. Event cameras, with high dynamic range advantages, low power consumption and extremely high temporal resolution (micro-second level) [16]–[19], can fill the blind interval between keyframes. Numerous approaches have successfully utilized event data to enhance VFI tasks [20]–[24]. By leveraging event data, researchers can obtain more accurate motion estimation, resulting in higher-quality interpolated frames through a motion-based warping process (Fig. 1(b)). However, though significant progress brought by more accurate motion estimation, another tricky issue in VFI tasks, such as severe occlusions (the rotating disk in Fig. 1) or non-rigid motion (the moving hand in Fig. 2(c)), remains unresolved.

The challenges described above motivate us to rethink the role that events should play in VFI tasks. When revisiting mainstream VFI & E-VFI methods, we find that both motion-based methods [8], [9], [22], [25] and synthesis-based methods [13], [14] treat the keyframes I_0 & I_1 as the core and anchor points of the interpolation task, while other sources of information ($e.g.$, event data) are considered auxiliary enhancement information. This approach is necessary and indispensable for traditional VFI tasks that lack inter-frame messages.

Building upon this premise, our preliminary work [1] introduces **DSER** (**D**irect **S**ynthesis via **E**vent-based **R**eference), a method that fundamentally redefines the roles of events and images within this task. Specifically, DSER treats keyframes as supplementary cues, employed to refine and complete the event-based reference $w.r.t$ the interpolated frames, as shown in Fig. 1(c). The preliminary framework consists of three stages: event-aware reconstruction, bidirectional synthesis, and global refinement.

While our reference-centric paradigm successfully reconstructs underlying structures, it exposes a critical limitation

Y. Liu is with the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen 361005, China.

Y. Deng is with the Tianjin Key Laboratory of Intelligent Robotics, the Academy for Advanced Interdisciplinary Studies and the College of Artificial Intelligence, Nankai University.

L. Fu and Z. Yang are with the College of Computer Science, Beijing University of Technology, Beijing 100124, China.

H. Chen is with Key Lab of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing 211189, China.

Y. Li is with Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR.

B. Shi is with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University.

† Corresponding author: yjdeng@nankai.edu.cn

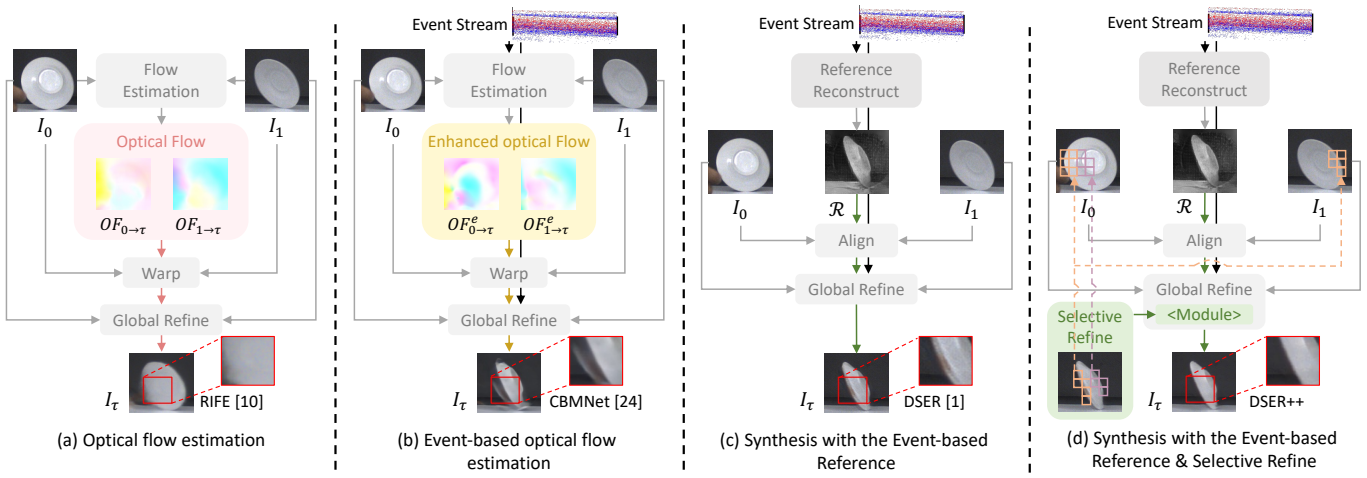

Fig. 1. Visualization comparison of different methods under challenging scenario. (a) and (b) are derived from keyframes: (a) uses frame estimation with entirely incorrect optical flow; (b) corrects the optical flow by combining event and frame-based estimations. (c) utilizes event-based reference for synthesis (DSER [1]). (d) Further selective optimization is applied to under-optimized regions for artifacts elimination (our method), where selected regions from key frames are as reconstruction references to guide interpolation frame optimization.

in the final refinement stage. As illustrated in Fig. 2, standard global decoders struggle to recover accurate textures in complex scenarios, causing texture incompleteness (Fig. 2(a)) and color inconsistency (Fig. 2(b)&(c)). We argue that these phenomena originate from two core theoretical bottlenecks inherent in standard global decoders:

First, *Regression to the Mean.* Optimized with global losses (*e.g.,* $\mathcal{L}_1$, $\mathcal{L}_2$), standard decoders suffer from the regression-to-the-mean problem in highly ill-posed areas [26], [27]. Confronted with extreme displacements in large-scale or non-rigid motions, the network outputs a spatial average rather than retrieving confident color references from keyframes, resulting in pronounced blur and color distortions.

Second, *Information Asymmetry.* Severe occlusions inherently break temporal symmetry, as the valid appearance exists exclusively in a single temporal direction. However, standard decoders employ symmetric receptive fields that blindly fuse bilateral features [28]. Fusing an occluded, invalid patch with a valid one inevitably corrupts clean structural features, causing ghosting artifacts. Consequently, global decoders are intrinsically insufficient for such localized reconstruction failures, necessitating a targeted, asymmetric repair mechanism.

In this paper, to resolve this specific issue, we propose **DSER++**, which extends DSER with a targeted refinement mechanism. Our core innovation is the Bidirectional Asymmetric Selective Refinement (BASR) module, designed as a local "repair" unit rather than a global optimizer. The key insight is to dig and utilize DSER's primary strength step further: BASR uses the already well-reconstructed structural features from the initial synthesis as a high-quality guide to search the keyframes retrospectively and asymmetrically, aiming to recognize and source reliable color and texture support for the problematic patches. Where in such process, performing asymmetric searching is argued as crucial for handling occlusions where visual truth may only exist in one of the two keyframes, *e.g.*, the entire ball and the back of the left hand are visible only in I_1 (Fig. 2). Furthermore, to ensure the seamless integration of these refined patches, we introduce a structural consistency loss function ($\mathcal{L}_{ps}$) to maintain coherence with neighboring regions. Through this targeted repair strategy, DSER++ effectively mitigates event-induced artifacts and further enhances the quality of the interpolated frames in complex motion scenarios (Fig. 2).

A preliminary version of this work was presented in [1]. In [1], we proposed a novel E-VFI framework that directly synthesizes intermediate frames based on an event-based reference without relying on motion estimation. To achieve this, we introduced an event-aware reconstruction strategy to emphasize structural information, alongside a deformable synthesis module (E-PCD) for robust feature alignment. Building upon this foundation, this journal manuscript substantially extends the conference version by addressing the severe event-induced artifacts and structural corruptions in occlusion regions. The novel contributions of this extension (DSER++) are exclusively summarized as follows:

- We propose the Bidirectional Asymmetric Selective Refinement (BASR) module, which introduces a targeted asymmetric refinement mechanism from global averaging to targeted, asymmetric feature retrieval to address event-induced artifacts in occluded regions.
- We introduce a region-based structural consistency loss that enforces semantic coherence between the locally optimized patches and their surrounding global context.
- We provide comprehensive empirical evaluations, including additional evaluations on diverse real-world datasets, region-wise error analyses, and cross-dataset testing under extreme boundary conditions, to rigorously validate the practical effectiveness and architectural superiority of the proposed framework.

II. Related Work

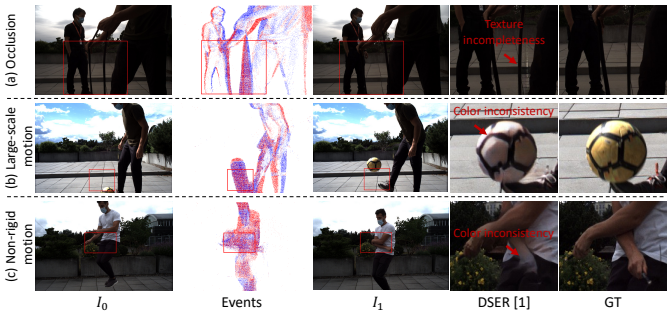

Fig. 2. Visualization of typical failure modes in the original global refinement stage (DSER) compared to the Ground Truth (GT). (a) **Severe Occlusion**: Symmetric fusion with occluded patches corrupts features, causing incomplete textures. (b) **Large-scale Motion**: High uncertainty forces regression to the spatial mean, resulting in severe blur and color inconsistencies on the soccer ball. (c) **Non-rigid Motion**: Complex deformations induce color shifting and ghosting artifacts on the moving arm.

A. Video Frame Interpolation.

Frame-based Video Frame Interpolation (VFI) can be categorized into motion-based [2], [3], [8], [9], [12], [15], [29], [30], synthesis-based [13], [14], kernel-based [2], [31], and phase-based [32] approaches. Among these, motion-based methods dominate, typically warping keyframes via estimated optical flow. To handle complex dynamics and large motions, researchers have developed various strategies, such as coarse-to-fine architectures [30], iterative refinement schemes [25], and global feature interaction [12]. While Hu *et al.* [33] effectively modeled non-linear acceleration, Zhou *et al.* [34] recently achieved state-of-the-art performance by introducing spatiotemporal sparse attention to significantly enlarge the receptive field.

To address these limitations, recent works [13], [14] 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, increasing the burden of training and data acquisition. Wu *et al.* [35] try to generate latent features related to regions that might include artifacts and then directly synthesize intermediate frames using a normalization flow-based generator. While this approach improves perceptual quality, it often produces images that differ from the original in terms of pixel intensity and structural fidelity. In this paper, by leveraging the high temporal resolution of event data, we bridge the knowledge gaps between keyframes, allowing us to synthesize high-quality interpolated frames based on the inter-frame reference provided by event data, without introducing additional sources of knowledge.

B. Event-based Video Frame Interpolation.

Thanks to the microsecond-level temporal resolution, event cameras can fill the inter-frame information blank, giving rise to numerous successful Event-based Video Frame Interpolation (E-VFI) works [20]–[24], [36]–[42]. Mainstream E-VFI models are built on the combination of motion-based warping and synthesis. These methods mainly follow the ideas of previous work, using event data to enhance the estimated optical flow. Initially, Tulyakov *et al.* [20] incorporated event data into their workflow, using a U-Net-like network to generate synthetic images and optical flow, and combining warped and synthesized predictions through image-level attention. Subsequent studies [21], [23], [24], [36]–[39] have all been dedicated to improving the accuracy of motion estimation to achieve better results via the warping processes. Specifically, Tulyakov *et al.* [38] enhanced efficiency and performance by separately calculating motion splines and multi-scale fusion. He *et al.* [21] adopted an unsupervised cyclic consistency strategy, eliminating the need for high-speed training data. Sun *et al.* [41] proposed a bidirectional recurrent network to jointly address interpolation and deblurring problems. Kim *et al.* [23] improved the quality of optical flow estimation by adding multiple cost volumes to enhance the flow estimation module. Ma *et al.* [24] and Liu *et al.* [43] attempted to overcome nonlinear motion by fitting per-pixel multi-segment flows, but this approach inevitably introduced error accumulation and still failed to address the limitations of non-rigid motion. Although performance improvement has been witnessed, these approaches still struggle with the questions that are widespread in real-world scenarios such as how to fit non-rigid motion and alleviate occlusion issues.

Gao *et al.* [40] attempt to handle these issues by providing a slow-fast joint synthesis network, yet the lack of direct reference at the interpolated frame causes their method to easily overfit keyframes' patterns and thereby resulting in unsatisfactory performance. Recently, Liu *et al.* [44] proposed EPA, utilizing feature-level supervision to handle image degradations, but it still struggles with motion-induced occlusions.

In this work, we provide a solution to handle the above issues by proposing a purely synthesis-based E-VFI method with the aid of event-based references containing structural messages of the interpolated frames. The event-based references are reconstructed from the event stream at the time of frame interpolation, which effectively preserves the structural details of the interpolated scenes. Leveraging such references, we no longer rely on pixel-level correspondences to estimate motion. Instead, we directly align the keyframes and references at the feature level, which elegantly avoids the errors introduced by motion estimation.

C. 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 [45]–[48]. With the rise of deep learning, numerous studies have begun to employ neural networks for video reconstruction [49]–[52]. However, these methods are poorly defined because predicting the accurate value of each pixel from the sequence of brightness changes without inputting the initial brightness value is uncertain. Moreover, due to the nature of event cameras that only detect areas with contrast changing, the reconstruction of smooth and static regions poses challenges

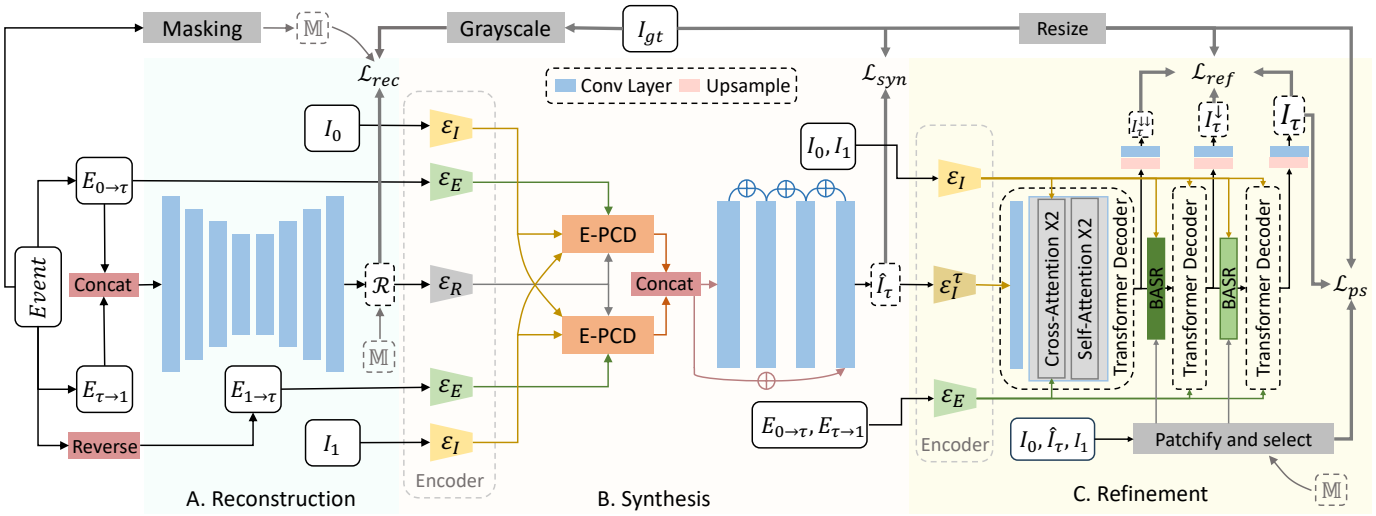

Fig. 3. An overview of the proposed E-VFI framework DSER++. It consists of reconstruction, synthesis and refinement staged. The $\mathcal{L}_{rec}$, $\mathcal{L}_{syn}$ and $\mathcal{L}_{refine}$ are used to supervise the obtained event-based references $\mathcal{R}$, coarse interpolated frames $\hat{I}_\tau$, and final prediction $\hat{I}$, respectively. Modules with the same name share weights.

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.

III. Method

In this section, we describe our event-based synthesis framework DSER++ for E-VFI. Given two input keyframes, I_0 & I_1, and the events ($\mathcal{E}_{0\to1}$) between them, we aim to estimate the intermediate frame I_τ at a specific timestamp between I_0 and I_1, where $\tau \in [0, 1]$. The overall structure of DSER++ is shown in Fig. 3, consisting of the reconstruction, synthesis, and refinement stages.

We denote the event stream as $\mathcal{E} = \{e_i\}_{i=1}^{N}$ where $e_i = (x_i, y_i, p_i, t_i)$. Following [20], [23], we stack events into a spatio-temporal Voxel Grid $E \in \mathbb{R}^{B\times H\times W}$ by linearly interpolating timestamps. Initially, events $\mathcal{E}_{0\to1}$ are divided into two segments corresponding to τ and converted into voxel grids $E_{0\to\tau}$ and $E_{\tau\to1}$ respectively. Then, the reconstruction stage (Fig. 3(A)) utilizes $E_{0\to\tau}$ and $E_{\tau\to1}$ to compute the event-based reference ($\mathcal{R}$), which contains the structural messages of the frame to be interpolated, following the proposed event-aware reconstruction strategy. Next, at the synthesis stage (Fig. 3(B)), we aim to obtain a coarse interpolation by bidirectionally aligning keyframes to $\mathcal{R}$. We perform this process utilizing the proposed E-PCD through explicit guidance from event data.

At the final stage (Fig. 3(C)), we aim to enhance the obtained coarse interpolation to visually pleasant images. In our conference version, we achieve this by simply adopting an off-the-shelf Transformer-based decoder for refinement. Here, observing that the lack of consistency constraints provided by motion-based warping results in color fluctuations and regional blurring, we incorporate a Bidirectional Asymmetric Selective Refinement (BASR) module inside the refinement network. This module can refine the challenging synthetic regions in a targeted manner by detecting motion intensity between interpolated and key frames, ensuring color consistency and detail preservation.

In the following, we first present the Reconstruction phase of Event-Based Reference in Section III-A, the synthesis phase in Section III-B, and finally, we introduce the refinement phase in Section III-C.

A. Reconstruction of Event-Based Reference

Building upon our preliminary work [1], we directly reconstruct an event-based reference ($\mathcal{R}$) to encode the structural features of the interpolated frames. Since event cameras struggle to encode static or low-texture regions and often accumulate a significant amount of noise, standard reconstruction yields severe degradation in these areas. This phenomenon is visually verified in our ablation studies (Fig. 11)

To suppress these low-confidence areas, we compute an event-aware mask ($\mathbb{M}$) utilizing a customized erosion operation with a specific threshold to filter out isolated noise. Next, a standard mathematical morphological dilation operation (using a 5×5 kernel) is applied to connect the remaining high-confidence event points into continuous semantic pathways. The reference masked by $\mathbb{M}$ is then used for subsequent alignment. More detailed information is provided in the *supplementary materials*.

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. 6), 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

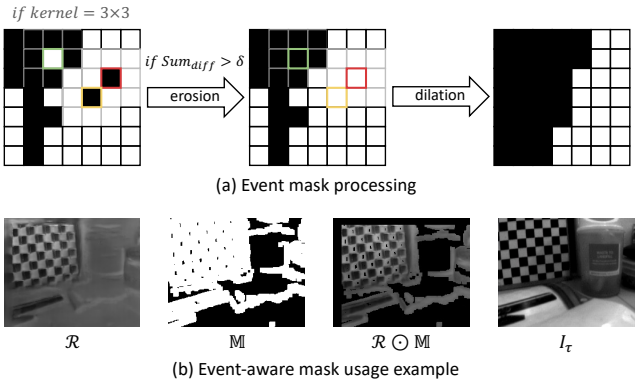

Fig. 4. Illustration of the event-aware masking. (a) shows the generation of the event-aware mask, where yellow, green, and red refer to the center of the kernel. (b) shows a practical example, and the event-aware masking removes the adverse impact from unreliable locations.

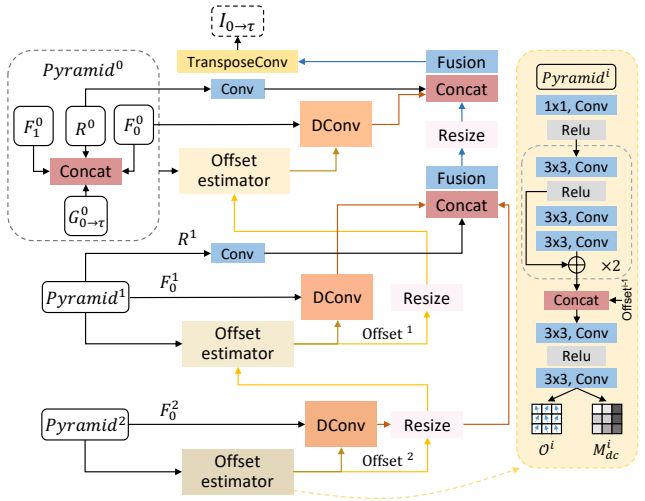

Fig. 5. An overview of the proposed E-PCD module. Resize refers to upsampling using bilinear interpolation.

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. 4(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. Finally, we take the $\mathcal{R}$ masked by the $\mathbb{M}$ as input to the following modules. As shown in Fig. 4.(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.

B. 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. We employ the E-PCD module [1], which utilizes a bidirectional alignment mode guided explicitly by inter-frame events ($E_{0\to\tau}$ and $E_{1\to\tau}$). The E-PCD calculates offsets and masks at multiple pyramid levels to align I_0 and I_1 with $\mathcal{R}$ via deformable convolutions. Finally, a dense residual network fuses the bidirectional features ($\mathcal{I}_{0\to\tau}$ and $\mathcal{I}_{1\to\tau}$) to output the coarse interpolated frame $\hat{I}_\tau$. More detailed information is provided in the *supplementary materials*.

E-PCD. The E-PCD module takes as input features extracted from keyframes, events, and the event-based reference as illustrated in Eq. 1.

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

where $\odot$ denotes the Hadamard Product. Specifically, E-PCD consists of three pyramid feature levels, as shown in Fig. 5. 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. 2.

$$\begin{aligned} Pyramid^i &= \mathcal{C}(F_0^i, F_1^i, V^i, G_{0\to\tau}^i) \\ \mathcal{O}^i, M_{dc}^i &= \mathcal{F}_{OE}(Pyramid^i, \mathcal{O}^{i-1}), \end{aligned} \tag{2}$$

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. 5. Then, we can align I_0 with $\mathcal{R}$ through deformable convolution and fusion operations as formulated in Eq. 3.

$$\mathcal{I}_{0\to\tau}^i = \mathcal{F}_{fu}(\mathcal{C}(\mathcal{F}_{dc}(F_0^i, \mathcal{O}^i, M_{dc}^i), \mathcal{F}_{co}(V^i), \mathcal{I}_{0\to\tau}^{i-1})) \tag{3}$$

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 $\hat{I}_\tau$.

C. Refinement with the BASR Module

In our preliminary work [1], as well as in many synthesis-based interpolation approaches [23], [40], the final stage often employs a global refinement decoder, typically based on Transformer or CNN architectures. While this global approach can mitigate general artifacts like edge blurring and micro-artifacts, we observe that texture incompleteness

and color inconsistency are non-negligible in regions with complex motions or severe occlusions (Fig. 2). We argue that the global optimization paradigms in common decoding architectures are insufficient to resolve these localized, fine-grained reconstruction failures, as their global nature tends to average out details rather than performing targeted repair [53], [54].

To this end, this paper introduces the BASR module that works in conjunction with the originally refinement stage, where selective and asymmetric refinement pattern adopted by the BASR substantially enhancing the model's robustness in handling above challenging regions. This design follows three guiding principles tailored to the event-based synthesis paradigm, detailed in Sec. III-C1.

1) Design Philosophies: First, the refinement needs to be ***Selective***.

To avoid the defects induced by average nature of global refinement, the key is to select regions that require more attention for targeted optimization. In our case, we suggest that regions with large-scale occlusions and non-rigid motions are considered Highly Challenging for Synthesis (HCS). Within the above regions, unpredictable occlusions caused by complex motion result in the synthesis relying primarily on sparse event data, which lacks color and dense semantic supports. This makes it difficult to achieve high-quality reconstruction. As a result, performing optimization to these selected regions with reliable visual supports would further improve the visual quality of our prediction (detailed in Sec. III-C2).

Second, the support from keyframes needs to be sourced ***Asymmetrically***. This principle is a direct response to how occlusion events fundamentally break the symmetry of information between keyframes. For a region undergoing an occlusion, its reliable ground-truth appearance is often exclusively present in one temporal direction (either in I_0 before being occluded or in I_1 after being revealed), rendering the other an invalid source, As shown in the "Visual supports Search" section in Fig. 6. A symmetric mechanism would inevitably corrupt the valid signal by fusing it with features from the occluded source. Our asymmetric design, in contrast, respects this broken symmetry, enabling it to isolate and leverage the single most reliable source of information (detailed in Sec. III-C3).

Third, the local refinement needs to be ***Coherent*** *with its global context.* The process of targeted refinement (BASR) introduces the risk of creating discontinuities at the boundaries between refined patches and their surroundings. Therefore, the structural consistency loss is proposed as a critical constraint that enforces semantic and color coherence, ensuring that our local corrections seamlessly integrate into the global scene's integrity (detailed in Sec. III-C4).

We instantiate these principles in our Bidirectional Asymmetric Selective Refinement (BASR) module. This module is comprised of three key components that work in concert: an Asymmetric Region Selection Strategy (ARS^2) to identify HCS regions, a customized cross attention mechanism for targeted refinement, and a novel loss to ensure the refined patches are locally coherent in structure and color.

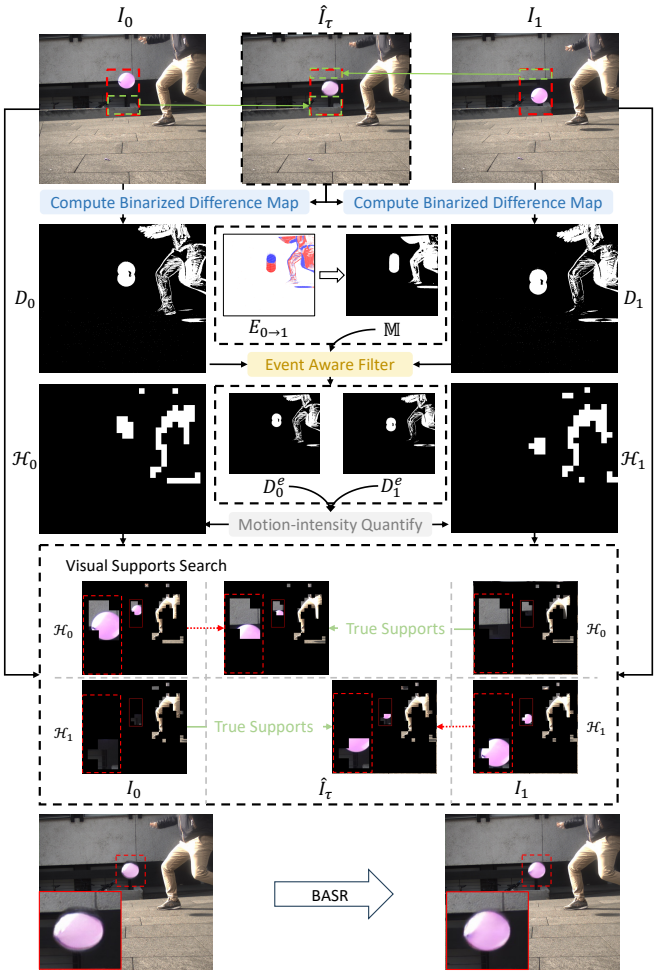

Fig. 6. Overview of the asymmetric area selection mask generation process involved in the BASR module. First, compute the binarized difference map. Then, use the event-aware mask to further filter HCS regions and patch them. Finally, a mask reversal operation is performed to obtain the bidirectional asymmetric region selection maps.

2) Asymmetric Region Selection Strategy: Occlusion regions commonly contain flow discontinuities and event bursts; locations with large-scale motions and strong event activities therefore mark HCS regions with high possibility. Inspired by above observation, we employ a multi-level filtering strategy driven by motion analysis, namely ARS^2, for keyframe regions: a difference map to detect asymmetric differences, an event-aware mask to identify regions potentially affected by events, and motion intensity to assess regional reliability.

We first compute the binarized difference map ($D_{i,i\in0,1}$) $w.r.t$ difference regions between keyframes and $\hat{I}_\tau$ for searching all changed regions, as shown in Eq. 4. We here take the interaction between I_0 and $\hat{I}_\tau$ as an example.

$$D_0 = \mathbb{B}(\sum_{RGB} |I_0 - \hat{I}_\tau|), \tag{4}$$

where $\mathbb{B}$ denotes the binarizing operation, which any value larger than zero is set as one. Then, we multiply the obtained D_0 with the event-aware mask ($\mathbb{M}$), aiming to selectively retain only those regions where imperfect reconstructions may arise due to inherent limitations of event signals for targeted

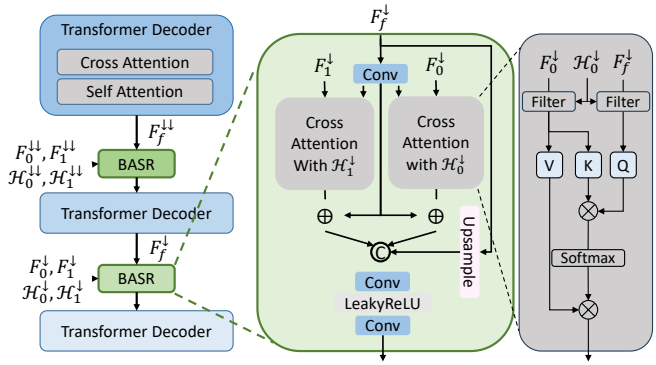

Fig. 7. Overview of the proposed framework with the BASR module.

optimization, while leaving other regions to be processed by the vanilla refinement module.

$$D_0^e = D_0 \odot \mathbb{M}, \tag{5}$$

where $\odot$ denotes the Hadamard product. Next, we aim to further identify synthetically more challenging regions (denote as $\mathcal{H}$) by quantifying local motion intensity. Specifically, a sliding window of size $p \times p$ with stride p is applied to process D_0^e and count the number of active motion pixels within each window. If the cumulative intensity of windows (patches) exceeds our predefined threshold ($\beta \cdot p$), corresponding locations ($\mathcal{H}_0$) will be recorded as HCS, as formulated in Eq. 6.

$$\mathcal{H}_0 = \left\{ (ip, jp) \,\middle|\, \sum_{x=0}^{p-1} \sum_{y=0}^{p-1} D_0^e(ip + x, jp + y) \geq \beta p \right\} \tag{6}$$

where (ip, jp) denotes the top-left coordinates of candidate patches, and $\mathcal{H}_0$ contains potential locations of HCS patches that meet the motion intensity threshold through the corresponding measurement between $\hat{I}_\tau$ and I_0. Similarly, we can achieve the $\mathcal{H}_1$ via computations using $\hat{I}_\tau$ and I_1.

A core problem that we need to address next is how to find appropriate optimization supports for these HCS regions. A straightforward approach might involve using the identified HCS regions to perform refinement based on keyframes used to calculate themselves such as $(I_0, \mathcal{H}_0)$. However, through modeling common extreme scenarios such as occlusion and non-rigid motion, we found that the HCS regions ($\mathcal{H}_0$) located by the interpolation frames and keyframes (I_0) in a particular direction can often be supported by optimization references from keyframes (I_1) in the opposite direction (as shown in Fig. 6). Therefore, we employ a reverse operation. For instance, when optimizing the region corresponding to $\mathcal{H}_0$, we seek visual optimization support from I_1 rather than using the keyframe I_0, and we validated the effectiveness of our design through quantitative analysis.

3) Network Structure of BASR: After obtaining a bidirectional asymmetric selection mask, we can use it to optimize our interpolation. We achieve this by using a simple attention structure, which can maintain a smaller parameter size while effectively supplementing details, as shown in Fig. 7. In specific, we perform bidirectional weight-shared cross-attention as formulated in Eq. 7 (taking the interaction between F_f and F_0 as an example).

$$F_f^{res} = softmax(\frac{Q \cdot K^T}{\sqrt{d}}) \cdot V, \tag{7}$$

Here K, V denotes key and value that are linear projected from $\mathcal{F}_{index}(F_0, \mathcal{H}_1) \in \mathbb{R}^{N_1 \times p^2}$, and Q is obtained from $\mathcal{F}_{index}(F_f, \mathcal{H}_1) \in \mathbb{R}^{N_1 \times p^2}$. The $\mathcal{F}_{index}$ is used for indexing patches $w.r.t$ the location set of HCS regions $\mathcal{H}_1$. Finally, we perform scaling and fusion at the end of the module to adapt to the next layer of the network backbone.

4) Structural Consistency Loss: Leveraging BASR module, the visual details of the targeted regions could be improved. However, such regional operation inevitably leads structural and color inconsistencies between the optimized patches and their neighborhood. To this end, we adopt a structural consistency loss $\mathcal{L}_{ps}$, which uses a region-based SSIM constraint to enforce regional structural consistency. First, we determine all locations in HCS regions ($\mathbb{S}$) that are processed by BASR through Eq. 8.

$$\mathbb{S} = \bigcup_{(ip,jp) \in \mathcal{H}_0 \cup \mathcal{H}_1} \{(x, y) \,|x \in [ip, (i+1)p - 1], \; y \in [jp, (j+1)p - 1]\} \tag{8}$$

Next, we process these regions using the sliding window of SSIM [55], with a commonly adopted window size of 11×11. During the loss calculation, only the center of the sliding window belongs to $\mathbb{S}$ is taken into account, aiming to enhance the color and structural consistency between optimized and non-optimized regions, as described in Eq. 9.

$$\mathcal{L}_{ps} = \frac{\sum_{(x,y) \in \mathbb{S}} \left(1 - \text{SSIM}_{(x,y)} \left(I_{gt}, I_\tau\right)\right)}{|\mathbb{S}|}, \tag{9}$$

where $\text{SSIM}^{(x,y)}$ denotes the SSIM value computed over an 11×11 window centered at (x, y).

IV. Experiment

This section compares our approach with representative VFI methods in both RGB-only and RGB-Event tracks and conducts analysis quantitatively and qualitatively.

A. Datasets

Following the optimization strategy in [20], we train our method on the training set of Vimeo90k-Septuplet [56], where synthetic event data is simulated using the ESIM [57]. For evaluation, we follow the evaluation methods used in [20], [23] and test our method on both synthetic and real-world datasets.

1) Synthetic Event Datasets:

- Vimeo90k-Triplet [56], the test split containing 3,782 triplets at a resolution of 448×256 pixels.
- GoPro [58], the test split containing 11 scenes at a resolution of 1280×720 pixels, with an average of 1,100 images per scene.
- SNU-FILM [59], containing 1,240 triplets with a resolution of approximately 1280×720 pixels. The dataset is

categorized into four subsets based on motion complexity: easy, medium, hard, and extreme. This paper selects two most challenging subsets such as hard and extreme for model evaluation.

2) Real Event Datasets:

- High Quality Frames (HQF) DAVIS240 [51], contains 14 rich textured scenes with a resolution of 240×180 pixels.
- High Speed Event and RGB camera (HS-ERGB) [20]. Fifteen test scenes is included with color images at resolutions greater than 900×800 pixels. The dataset is categorized into two types based on the shooting distance: close and far. It includes challenges such as occlusions and non-rigid motion conditions.
- Beam Splitter Events & RGB (BS-ERGB) dataset [38], contains 45 scenes in the training set, 17 scenes in the validation set, and 26 scenes in the testing set. This dataset includes RGB images with a resolution of 970×625 pixels, characterized by significant noise and non-rigid motion.
- EventAid-F [60], contains 10 test scenes with RGB images at resolutions greater than 954×636 pixels. The dataset includes a wide variety of motion patterns, making it suitable for evaluating methods under diverse motion conditions.

B. Training

1) Settings: We train our proposed pipeline in an end-to-end manner. Our method is optimized by AdamW [61] with weight decay 10^{-4} for 40 epochs using PyTorch [62]. The initial learning rate was set to 10^{-4} and decreased 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.

2) 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. In the BASR module, we set the patch size to 32 and set β to 0.1. The evaluation metrics utilized in our experimental section are SSIM and PSNR [13], [20], [23], which are commonly employed for the VFI task.

3) Loss: Our adopted losses can be divided into three parts corresponding to the proposed framework. As shown in Fig. 3, our approach is trained end-to-end by minimizing as $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{syn} + \mathcal{L}_{ref} + \mathcal{L}_{ps} \tag{10}$$

Reconstruction stage ($\mathcal{L}_{rec}$). We utilize perceptual loss ($\mathcal{L}_{lpips}$) [63] and L1 loss ($\mathcal{L}_1$) for jointly optimizing the structural accuracy of the event-based reference $\mathcal{R}$ in high confidence areas with pseudo ground truth $I_{gt}^{\mathcal{M}\_gray}$, as formulated in Eq.11.

$$\begin{aligned}\mathcal{L}_{rec} =& \mathcal{L}_{lpips}(\mathcal{R} \odot \mathbb{M}, I_{gt}^{\mathbb{M}\_gray}) + \\ & \mathcal{L}_1(\mathcal{R} \odot \mathbb{M}, I_{gt}^{\mathbb{M}\_gray}),\end{aligned} \tag{11}$$

Synthesis stage ($\mathcal{L}_{syn}$). Similarly, we adopt the combination of $\mathcal{L}_{lpips}$ and $\mathcal{L}_1$ for supervising the synthesis stage, as formulated in Eq.12.

$$\mathcal{L}_{syn} = \mathcal{L}_{lpips}(\hat{I}_\tau, I_{gt}) + \mathcal{L}_1(\hat{I}_\tau, I_{gt}) \tag{12}$$

Refinement stage ($\mathcal{L}_{ref}$). We employ multiple losses for jointly optimizing the refinement network:

$$\mathcal{L}_{ref} = \mathcal{L}_{lpips}(I_\tau, I_{gt}) + \mathcal{L}_1^{0,1,2}(I_\tau, I_{gt}) \tag{13}$$

where the $\mathcal{L}_1^{0,1,2}$ denotes the L1 losses applied to three different scales, with weights assigned as 1, 0.5, and 0.1 from the largest to the smallest scale.

C. 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: BMBC [8], RIFE [9] and UPR-Net [25]. (2) Synthesis-based VFI techniques such as VFIT [14]. (3) E-VFI methods with leveraging motion estimation (A^2OF [22], TimeReplayer [21] and TLXNet+ [24]) and synthesis operation (SuperFast [40]) solely, or with using motion and synthesis processes jointly (TimeLens [20], TimeLens++ [38] and CBMNet-L [23]).

Considering various training strategies is adopted by existing approaches, for fairness, we retrain or fine-tune models whose training settings are inconsistent with ours as shown in Table I. Specifically, (1) when testing on HQF and HS-ERGB datasets, we fine-tune our model following the method proposed in [20], [22]; (2) When testing on BS-ERGB and EventAid datasets, which feature more complex motion and noise distributions, we fine-tune our model following the method proposed in [23], [24];

1) Evaluations on Synthetic Datasets:

Quantitative Evaluations. As presented in Table I, our method demonstrates exceptional VFI performance across various synthetic datasets. **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.

We also conduct comparisons on the hard and extreme subsets of the **SNU-FILM dataset**. These two subsets, characterized by large-scale and irregular motion, provide a better evaluation of our method's ability to handle large motion occlusions. Our method holds a substantial performance advantage over other approaches, demonstrating its excellent generalization ability in complex scenes. With our proposed BASR, the fully powered framework can further improve performance across all compared datasets, showing superior robustness and adaptability in handling complex motion and

TABLE I
PSNR(DB)/SSIM RESULTS ON SYNTHETIC DATASETS(VIMEO90K-TRIPLET, GOPRO AND SNU-FILM). THE BEST RESULTS ARE MARKED IN **BOLD** WHILE THE SECOND ONES ARE MARKED WITH UNDERLINES. WE EVALUATE THE PERFORMANCE BASED ON WHOLE FRAMES BETWEEN THE INPUT FRAMES. † DENOTES TRAINED WITH FULL GOPRO TRAINING SET. ++ DENOTES THE DSER MODEL WITH BASR BOOSTED.

Method	Motion	Synthesis	Modal	Vimeo90k-Triplet	GoPro		SNU-FILM	
				1 frame	7 frame	15 frame	Hard	Extreme
BMBC [8]	✓	✗	F	35.06/0.944	25.45/0.755	24.29/0.752	29.23/0.921	23.60/0.833
RIFE [9]	✓	✗	F	34.74/0.957	29.66/0.889	25.14/0.772	30.36/0.920	25.54/0.853
UPR-Net-L [25]	✓	✗	F	36.24/0.966	27.91/0.855	24.61/0.758	30.82/0.928	25.61/0.862
VFIT-B [14]	✗	✓	F	31.94/0.926	-	-	30.95/0.932	27.82/0.880
IQ-VFI [33]	✓	✗	F	36.60/0.982	-	-	30.83/0.938	25.45/0.863
A^2OF [22]†	✓	✗	$F\&E$	-	36.61/0.971	-	-	-
CBMNet-L [23]†	✓	✓	$F\&E$	-	38.15/0.975	37.05/0.969	-	-
TLXNet+ [24]†	✓	✗	$F\&E$	-	-	36.52/0.967	-	-
TimeReplayer [21]	✓	✗	$F\&E$	35.12/0.963	-	-	-	-
A^2OF [22]	✓	✗	$F\&E$	36.54/0.967	34.08/0.954	32.65/0.937	31.72/0.924	28.21/0.890
TimeLens [20]	✓	✓	$F\&E$	36.31/0.962	34.81/0.959	33.21/0.942	37.67/0.969	35.36/0.951
CBMNet-L [23]	✓	✓	$F\&E$	37.69/0.970	36.07/0.972	35.46/0.966	36.52/0.971	35.07/0.958
EPA [44]	✗	✓	$F\&E$	36.95/0.964	35.75/0.968	34.26/0.957	34.44/0.949	30.96/0.899
DSER	✗	✓	$F\&E$	39.17/0.977	36.95/0.975	35.77/0.968	37.39/0.971	35.60/0.958
DSER++	✗	✓	$F\&E$	**39.45/0.983**	**37.26/0.976**	**36.16/0.970**	**37.97/0.974**	**36.07/0.962**

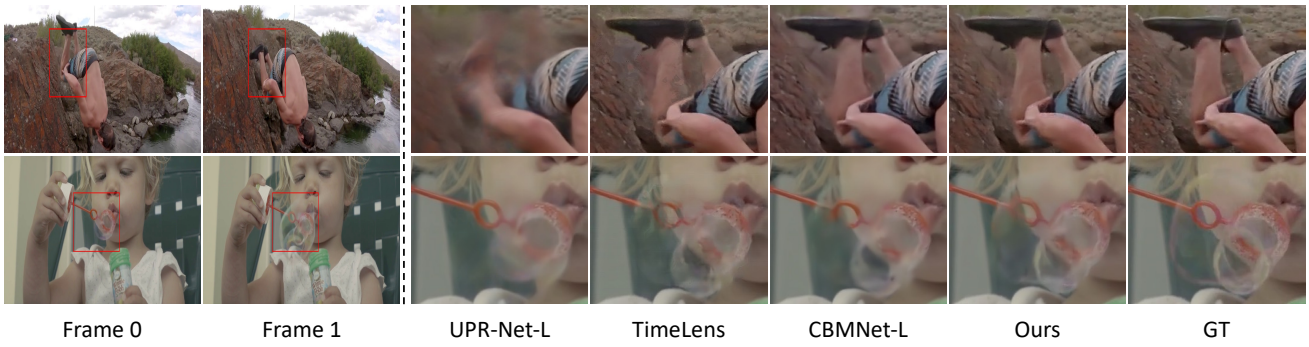

Fig. 8. Visual comparison among different methods on synthetic datasets.

occlusion scenarios. Unlike the conference version [1], we found that the SNU-FILM dataset generated using the original v2e [64] settings suffered from a temporal misalignment between frames and events. We addressed this issue by shifting the event timestamps forward to realign them with the corresponding frames and re-evaluated all methods.

Qualitative Evaluations. As illustrated in Fig. 8, our method adeptly handles challenging issues in the VFI task, *e.g.,* occlusions and non-rigid motion issues introduced by moving legs and bubbles. We attribute this enhancement to the effectiveness of our event-reference-based frame interpolation method, suggesting the advantage of our approach in handling large motion occlusions and non-rigid motion.

2) Evaluations on Real Datasets:

Quantitative Evaluations. In this section, we assess the effectiveness of our method on real-world HQF, HS-ERGB, BS-ERGB and EventAid-F datasets. The event signals in these datasets are closer to real-life scenarios, better reflecting the practical value of the model in real-world applications compared to synthetic datasets.

The comparison results, depicted in Table II and Table III, highlight the leading performance of our model across all conditions and datasets. Notably, even in challenging scenarios like HQF and EventAid-F, 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. Furthermore, the superiority of our method is consistently demonstrated in BS-ERGB and EventAid-F, which feature more diverse motion patterns and scenarios. These results, emphasizing the ability of our model to deliver superior image quality preservation, enhanced structural information reconstruction, and improved robustness to complex motion in real-world scenarios. With our proposed BASR, the fully powered framework can further improve performance across all compared datasets, showing superior robustness and

TABLE II
PSNR(DB)/SSIM RESULTS ON REAL DATASETS(HQF AND HS-ERGB). THE BEST RESULTS ARE MARKED IN **BOLD** WHILE THE SECOND ONES ARE MARKED WITH UNDERLINES. WE EVALUATE THE PERFORMANCE BASED ON WHOLE FRAMES BETWEEN THE INPUT FRAMES. ++ DENOTES THE DSER MODEL WITH BASR BOOSTED.

Method	Motion	Synthesis	Modal	HQF		HS-ERGB			
				1 frame	3 frame	Close		Far	
						5 frame	7 frame	5 frame	7 frame
BMBC [8]	✓	✗	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 [9]	✓	✗	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 [25]	✓	✗	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 [14]	✗	✓	F	31.50/0.882	-	-	-	-	-
TimeReplayer [21]	✓	✗	$F\&E$	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 [22]	✓	✗	$F\&E$	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 [40]	✗	✓	$F\&E$	-	-	-	32.50/0.869	-	27.87/0.845
TimeLens [20]	✓	✓	$F\&E$	33.42/0.934	32.27/0.917	32.19/0.839	31.68/0.835	33.13/0.877	32.31/0.869
TLXNet+ [24]	✓	✗	$F\&E$	33.55/0.916	32.10/0.861	-	33.50/0.847	-	27.34/0.780
CBMNet-L [23]	✓	✓	$F\&E$	34.77/0.953	33.08/0.940	34.17/0.862	33.96/0.857	31.43/0.888	30.47/0.870
EPA [44]	✗	✓	$F\&E$	-	-	34.01/0.854	33.68/0.851	33.41/0.916	32.71/0.906
TimeTracker [43]	✗	✓	$F\&E$	-	-	34.64/0.863	34.21/0.859	33.96/0.920	33.13/0.912
DSER	✗	✓	$F\&E$	35.89/**0.959**	34.27/0.941	34.60/0.865	34.33/0.862	34.19/0.923	33.56/0.921
DSER++	✗	✓	$F\&E$	**35.97/0.959**	**34.32/0.942**	**35.23/0.880**	**34.96/0.878**	**34.51/0.937**	**33.89/0.930**

TABLE III
PSNR(DB)/SSIM RESULTS ON REAL DATASETS(BS-ERGB [38] AND EVENTAID-F [60]). WE EVALUATE THE PERFORMANCE BASED ON WHOLE FRAMES BETWEEN THE INPUT FRAMES. ++ DENOTES THE DSER MODEL WITH BASR BOOSTED.

Methods	BS-ERGB		EventAid-F	
	1 skip	3 skip	7 skip	15 skip
Timelens [20]	28.36/-	27.58/-	34.45/0.942	31.90/0.914
TimeLens++ [38]	28.56/-	27.63/-	-	-
TLXNet+ [24]	29.46/0.815	28.71/0.806	33.24/0.925	30.86/0.891
CBMNet-L [23]	29.43/0.816	28.59/0.808	34.24/0.945	32.92/0.935
EPA [44]	27.94/0.791	27.22/0.782	33.91/0.932	31.99/0.910
TimeTracker [43]	29.41/0.808	28.79/0.803	34.19/0.938	32.96/0.933
DSER	29.09/0.810	28.38/0.802	35.15/0.948	33.76/0.937
DSER++	**29.62/0.820**	**28.88/0.813**	**35.57/0.952**	**34.24/0.943**

TABLE IV
COMPARISON RESULTS OF MODEL COMPLEXITY WITH THE E-VFI METHOD. ALL RUNTIME TESTS WERE CONDUCTED ON A SINGLE NVIDIA RTX 3090 GPU, WITH A RESOLUTION OF 448X256.

Method	Parameters	FLOPs	Modelsize	Runtime	PSNR
TimeLens [20]	79.2M	**18.7B**	302.15MB	**0.033s**	36.31
CBMNet-L [23]	**22.2M**	46.7B	**84.82MB**	0.231s	37.69
DSER	44.66M	40.9B	170.81MB	0.119s	39.17
DSER++	+0.88M	+0.67B	+5.44MB	+0.026s	**39.45**

adaptability in handling real-world application scenarios.

Qualitative Evaluations. Fig. 9 provides a visual comparison between SOTA E-VFI methods [20], [23] and ours on real event datasets. A direct comparison with the outputs of TimeLens and CBMNet-L reveals the superior capability of our approach. We observe that by leveraging the event-based reference, our method robustly reconstructs details in both close-up and distant scenes. It clearly restores challenging elements such as clouds in the distant sky, a rotating disk occluded by a hand, the surface of a rapidly spinning umbrella, and even a fully non-rigid object falling into water. This demonstrates the effectiveness of our approach in extreme occlusion and non-rigid motion scenarios. Moreover, owing to the precise motion tracking provided by events, our method restores certain details, such as the word "*xrite*" with even greater clarity than the ground truth itself. This further substantiates the critical role of events as anchor points for E-VFI tasks.

These observations suggest that while motion-based E-VFI methods perform well with accurate motion estimation, issues like severe occlusion, non-rigid motion, and missing objects remain challenging. In contrast, our pure synthesis-based E-VFI method, aided by event-based reference, effectively addresses these problems.

3) Model complexity analysis: This section aims to evaluate the model complexity of our proposed method. As shown in Table IV, our method with powerful interpolation capability contains only about half the parameters and model size of TimeLens. Moreover, while our method has more parameters and a larger model size compared to CBMNet-L, its inference time is nearly half as long.

To provide deeper insights into the computational distribution, we detail the parameter breakdown across the three

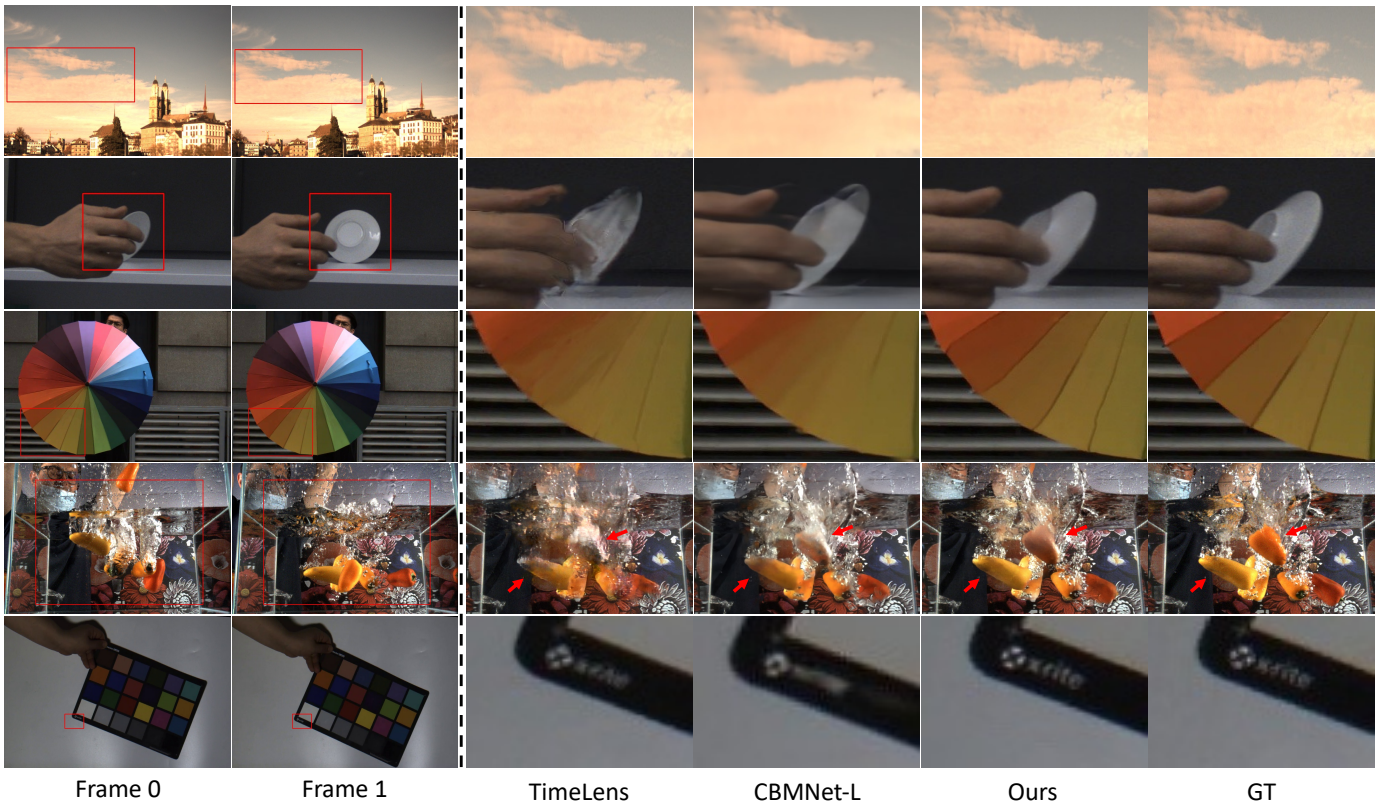

Fig. 9. Visual comparison among different methods on real datasets. For convenience, we have enlarged the details in the red box.

individual stages of our pipeline, as shown in Table V. The reconstruction stage accounts for the largest share at 20.81M parameters, primarily originating from the standard U-Net [65] architecture. The synthesis stage requires 16.22M parameters to perform bidirectional alignment. Finally, the refinement stage utilizes a total of 8.61M parameters. Notably, within this final stage, our newly proposed BASR introduces only 0.88M additional parameters and 0.67B additional FLOPs over the base refinement stage, whose original cost is 7.63M parameters and 16.60B FLOPs. This breakdown explicitly confirms the high efficiency of the BASR module, as it achieves consistent performance gains across all datasets with minimal computational cost. Additionally, since nearly half of the total parameters (20.81M) are concentrated in the initial reconstruction stage, substituting the current network with a more lightweight U-Net alternative [66] could significantly reduce the overall model complexity in future work.

Overall, our method strikes a balance between parameters, model size, and inference speed, making it more suitable for practical applications.

TABLE V
COMPUTATIONAL COMPLEXITY BREAKDOWN BY STAGE.

Stage	Module	Params	FLOPs
Stage A	Reconstruction (U-Net)	20.81M	3.84B
Stage B	Synthesis	16.22M	20.46B
Stage C	Refinement (w/o BASR)	7.63M	16.60B
	BASR module	0.88M	0.67B
Total (DSER++)		**45.54M**	**41.57B**

D. Condition-wise Analysis under Challenging Scenarios

To explicitly characterize the specific data conditions under which the BASR module provides the most substantial gains, we conduct a targeted sequence-level analysis using the BS-ERGB dataset. We select highly representative sequences for four extreme data conditions: high-motion, complex textures, sparse event signals, and severe occlusion. The visual characteristics of these conditions are illustrated in Fig. 10, and the quantitative evaluations are presented in Table VI.

As the results indicate, while the BASR module provides a consistent global average improvement (+0.50 dB) across all scenes, its relative performance gains are highly condition-dependent. The module yields steady, above-average improvements in high-motion (+0.58 dB) and complex texture (+0.70 dB) scenarios by preserving local high-frequency details.

However, the most substantial gains are observed under **Severe occlusion (+1.55 dB)** and **Sparse event signals (+1.33 dB)**—nearly triple the global average. For instance, the pouring water scene (*may29_water_tank_pouring_02*) represents a typical sparse signal condition. In such scenarios, the smooth liquid body fails to provide dense structural events (sparsity), while its dynamic specular reflections and complex refractions induce chaotic, uncorrelated event triggers (noise).

TABLE VI
QUANTITATIVE EVALUATION OF THE BASR MODULE UNDER SPECIFIC CHALLENGING DATA CONDITIONS FROM THE BS-ERGB DATASET (3 SKIP). Δ DENOTES THE RELATIVE PERFORMANCE GAIN.

Data Condition	w/o BASR	w/ BASR	Δ PSNR	Δ SSIM
Global Average	28.38/0.802	28.88/0.813	**+ 0.50**	**+ 0.011**
High-motion	30.00/0.802	30.58/0.807	**+ 0.58**	**+ 0.005**
Complex textures	29.74/0.783	30.44/0.790	**+ 0.70**	**+ 0.007**
Sparse event signals	25.69/0.783	27.02/0.814	**+ 1.33**	**+ 0.031**
Severe occlusion	32.04/0.945	33.59/0.955	**+ 1.55**	**+ 0.010**

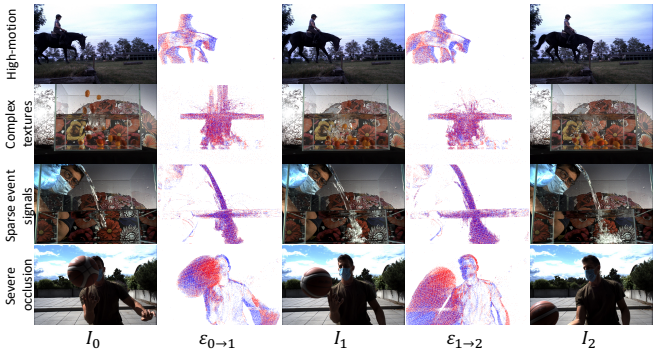

Fig. 10. Visualization of four representative data conditions from the BS-ERGB dataset. From top to bottom: (1) **High-motion** (*horse_13*): Rapid, large-scale displacements challenging temporal alignment. (2) **Complex textures** (*aquarium_08*): Overlapping high-frequency details (*e.g.,* vegetables, bubbles). (3) **Sparse event signals** (*may29_water_tank_pouring_02*): Sparse events corrupted by severe refraction noise. (4) **Severe occlusion** (*basketball_08*): Unpredictable blind spots breaking temporal symmetry.

This combination of structural sparsity and high-frequency noise severely misguides standard temporal fusion.

This structured analysis rigorously validates our architectural motivation: when event signals are highly unreliable or missing (due to occlusion or lack of contrast), standard decoders suffer from severe uncertainty. The proposed BASR module excels in these exact bottlenecks by dynamically bypassing the corrupted regions and asymmetrically retrieving deterministic structural priors from the unoccluded keyframes.

E. DSER Analysis

In this section, we perform experiments on the HS-ERGB dataset to analyze the effectiveness of two core designs in our proposed method DSER (without BASR) such as the Event-Aware Reconstruction Strategy (EARS) and the E-PCD module.

1) 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 Table VII 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.

TABLE VII
EVENT MASK SETTINGS WERE EVALUATED IN THE FAR SCENARIOS OF THE HS-ERGB DATASET.

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

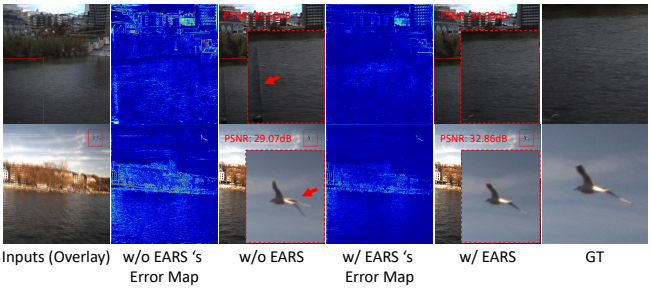

Fig. 11. Visual and quantitative ablation of the Event-Aware Reconstruction Strategy (EARS). The Error Maps highlight how EARS filters noise in static and low-texture regions, resulting in cleaner reconstructions and improved PSNR metrics.

2) The Efficacy of the EARS: We quantitatively verify the effectiveness of the Event-Aware Reconstruction Strategy (EARS) through settings B and C in Table VII. The comparative results indicate that EARS provides a substantial performance gain by mitigating the intrinsic sparsity and noise of event data.

To qualitatively demonstrate this efficacy and explicitly reveal the internal degradation mechanism, we visualize the reconstruction results alongside their corresponding Error Maps in Fig. 11. As shown, the degradation without EARS is particularly severe in static or low-texture regions, such as the water surface (top row) and the sky (bottom row). Because these regions lack dense structural events, the event sensor predominantly generates chaotic, unreliable noise.

As explicitly revealed by the *w/o EARS Error Map*, omitting the event-aware mask forces the network to integrate these chaotic signals. This inevitably leads to severe noise accumulation and structural degradation, manifesting as dense, high-magnitude error spikes in the visualization. Conversely, integrating EARS effectively shields the network from this background noise. As demonstrated by the *w/ EARS Error Map*, the error spikes are significantly suppressed, yielding clean background reconstructions and ensuring the overall structural fidelity of the interpolated frames.

F. BASR Analysis

In this section, we conduct experiments on far scenarios in the HS-ERGB dataset to analyze the effectiveness of the proposed BASR module and its related structural consistency loss.

1) Region-wise Error Analysis: To explicitly investigate whether the BASR module fundamentally alters the local error distribution rather than merely inflating global average metrics through uniform smoothing, we conduct a region-wise error

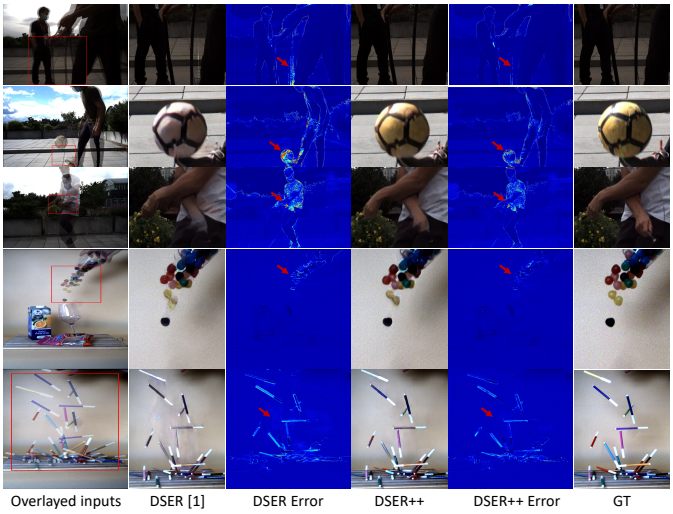

Fig. 12. Visualization of error image comparison. DSER++ has been optimized with a focus on synthetic difficulty areas.

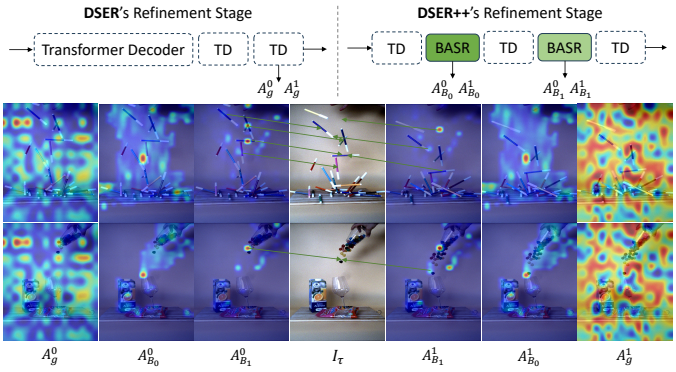

Fig. 13. Activation maps in the refinement stage. **Top:** Architecture comparison between the global refinement (DSER) and cascaded BASR modules (DSER++). **Bottom:** Progressive attention maps for I_0 and I_1, showing the shift from diffuse global attention (A_g) to highly concentrated, asymmetric retrieval (A_{B_0}, A_{B_1}). Maps are averaged across heads and channels.

analysis. We visualize the Error Maps (absolute difference against the Ground Truth) for highly challenging scenarios involving severe occlusion and large-scale motion in Fig. 12.

The visual comparisons clearly demonstrate that the baseline method leaves dense, high-magnitude error clusters concentrated at complex motion boundaries. In contrast, the BASR module accurately targets these challenging areas and significantly flattens the local error spikes. This targeted reduction confirms that the BASR module alters the error distribution by performing precise structural repairs in Highly Challenging for Synthesis (HCS) regions, rather than applying a general global smoothing effect.

2) Internal Attention Mechanism Analysis: To further explore the internal mechanism enabling the targeted repairs demonstrated in Sec. IV-F1, we provide an in-depth architectural analysis before and after integrating the BASR module. As shown in the top panel of Fig. 13, we contrast the standard global refinement stage of the conference version (DSER) with the cascaded BASR architecture of our extension (DSER++). The bottom panel tracks the progressive evolution of attention maps across representative challenging cases, revealing two critical architectural behaviors:

- **Progressive Focus on Intense Motion:** As observed in the outermost columns (A_g^0, A_g^1), the standard global decoder inherently lacks targeted spatial guidance, resulting in diffuse and uniformly scattered activations. In contrast, as features pass through the cascaded BASR modules (moving inwards from A_{B_0} to A_{B_1}), the attention hotspots become increasingly concentrated. The network dynamically locks onto regions undergoing intense, complex motion, allocating maximum representational capability to these HCS zones.
- **Unilateral Attention for Occlusion Regions:** The innermost attention maps ($A_{B_1}^0, A_{B_1}^1$) explicitly demonstrate how the BASR module resolves the occlusion bottleneck. When an object is occluded in one temporal direction but visible in the other, standard decoders blindly fuse features from both sides, inevitably causing ghosting artifacts. The BASR maps vividly demonstrate a unilateral attention behavior, where the module actively suppresses the invalid, occluded direction and asymmetrically retrieves clean structural priors entirely from the unoccluded keyframe.

This deep, multi-stage feature-level comparison rigorously confirms that the BASR module functions as a fundamental architectural enhancement that structurally breaks the global symmetric fusion bottleneck of the preliminary framework.

3) The Impact of Patch Size p and Boundary Analysis: In the BASR module, the patch size (p) fundamentally defines the local receptive field for the cross-attention mechanism during occlusion refinement. To successfully reconstruct an occlusion, the patch must be large enough to encapsulate distinct semantic priors (*e.g.,* structural edges) but small enough to exclude irrelevant, chaotic background signals.

For standard video resolutions ranging from 448×256 (*e.g.,* Vimeo90k) to 1280×720 (*e.g.,* SNU-FILM), typical motion displacements and occlusion regions share a relatively consistent scale in terms of pixel coverage. Within this specific resolution range, setting $p = 32$ consistently strikes the optimal balance between semantic encapsulation and noise exclusion. However, to investigate the failure boundaries and the robustness of this hyperparameter, we conduct additional evaluations on extremely low and ultra-high resolutions, as reported in Table VIII. The results explicitly demonstrate that the optimal p is not a universal constant but must be dynamically adjusted when crossing resolution thresholds:

- **Lower Boundary Test:** At highly reduced resolutions, a 32×32 patch occupies an excessively large proportion of the global image. It over-incorporates irrelevant background context, which dilutes the precise local structural priors. Under this condition, shrinking the receptive field to $p = 16$ yields superior performance.
- **Upper Boundary Test:** Conversely, at ultra-high resolutions, a 32×32 patch becomes overly localized. It often captures only flat textures or fragmented edges without sufficient macro-semantic context for accurate cross-attention matching. In such extreme cases, expanding the

TABLE VIII
BOUNDARY TEST OF DIFFERENT PATCH SIZES ACROSS EXTREME AND STANDARD RESOLUTIONS (PSNR/SSIM). NOTE THAT "EXTREME LOW" AND "ULTRA-HIGH" RESOLUTIONS ARE DOWNSAMPLED AND UPSCALED FROM THE SNU-FILM DATASET, RESPECTIVELY. BOLD VALUES INDICATE THE OPTIMAL PATCH SIZE FOR THE GIVEN RESOLUTION.

Scale	**Resolution**	$p = 16$	$p = 32$	$p = 64$
Extreme Low	224×128	**36.28/0.971**	36.19/0.969	36.16/0.969
Vimeo90k	448×256	39.37/0.983	**39.45/0.983**	39.32/0.982
HS-ERGB(far)	660×732	33.84/0.926	**33.89/0.930**	33.81/0.924
SNU-FILM(Hard)	1280×720	37.92/0.973	**37.97/0.974**	37.93/0.973
Ultra-High	3840×2160	36.13/0.958	36.17/0.958	**36.29/0.959**

TABLE IX
BASR MODULE SETTINGS WERE EVALUATED IN THE FAR SCENARIOS OF THE HS-ERGB DATASET. † REPRESENTS WITHOUT THE REVERSE OPERATION.

Variants	BASR	$\dagger ARS^2$	ARS^2	$\mathcal{L}_{ps}$	HS-ERGB(far)
A					33.56/0.921
B	✓	✓			33.62/0.921
C	✓		✓		33.67/0.921
D	✓	✓		✓	33.82/0.924
E	✓		✓	✓	**33.89/0.930**

receptive field to $p = 64$ becomes highly beneficial to maintain structural coherence.

This boundary analysis confirms that $p = 32$ is the empirically safest and most robust configuration tailored specifically for the resolution range of current mainstream event-based video datasets.

4) The Impact of cumulative intensity β: In our Asymmetric Region Selection Strategy (ARS^2), we quantify local motion intensity by counting active pixels in the masked difference map D^e within a $p \times p$ patch. Since event cameras predominantly trigger along object edges, the expected active pixels scale linearly with the patch dimension p, rather than quadratically (p^2). Thus, the activation threshold for Highly Challenging for Synthesis (HCS) regions is formulated as βp.

To provide a quantitative justification for the optimal intensity coefficient $\beta = 0.1$, we conduct an ablation study across $\beta \in \{0, 0.05, 0.1, 0.2, 0.3\}$, as presented in Table X. The results reveal a clear trade-off governed by the inherent sparsity of event signals. When the threshold is set too high ($\beta \geq 0.2$), the selection criterion becomes excessively conservative, failing to identify genuinely challenging boundaries with sparse event clusters, which degrades the reconstruction quality. Conversely, lowering the threshold ($\beta \leq 0.05$) also leads to a performance drop. Since the Event-Aware Reconstruction Strategy (EARS) has already filtered out inherent noise in preceding stages, an overly sensitive threshold primarily captures trivial pixel-level variations or initial synthesis micro-artifacts. This triggers redundant asymmetric refinements in essentially stable regions, disturbing the structural coherence. Therefore, $\beta = 0.1$ is established as the optimal configuration to accurately locate true structural occlusions while strictly avoiding redundant optimization.

TABLE X
ABLATION STUDY ON THE MOTION INTENSITY THRESHOLD β USED IN THE ASYMMETRIC REGION SELECTION STRATEGY. THE RESULTS DEMONSTRATE THE IMPACT OF DIFFERENT THRESHOLD SETTINGS ON THE FINAL INTERPOLATION PERFORMANCE.

Threshold (β)	**PSNR**	**SSIM**
0	33.78	0.928
0.05	33.86	**0.930**
0.1 (Ours)	**33.89**	**0.930**
0.2	33.85	**0.930**
0.3	33.79	0.929

5) The effectiveness of the mask reverse operation in ARS^2: Here, we verified the effectiveness of the mask reverse operation in the ARS^2 through the settings in Table IX. The comparison results show that the added mask reverse operation effectively improves the quality of the interpolated frame. Thanks to this operation, synthetic frames can find the correct reference patches on key frames to perform detail supplementation. As shown in Fig. 6, the mask generated with reverse operation can better locate the area occluded by the moving balloon.

6) The effectiveness of $\mathcal{L}_{ps}$: Here, we verify the effectiveness of the structural consistency loss $\mathcal{L}_{ps}$ through the settings C and E in Table IX. An obvious gain in model's performance is observed when $\mathcal{L}ps$ is employed to the BASR module, validating the efficacy of our proposed constraints in retaining the coherent semantic and color consistency to a certain extent.

V. CONCLUSION

This study introduces a novel purely synthesis-based framework for the E-VFI task, facilitating the generation of interpolated frames through the alignment of keyframes to the event-based reference. Specifically, the Event-Aware Reconstruction Strategy, tailored to properties of events, is proposed to ensure 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. Then, we propose the E-PCD, a specialized bidirectional alignment module for synthesizing interpolated frames while preserving the structural information in the reference. In addition, considering the limitations of events in encoding color and smooth regions, we integrate the newly introduced BASR module into the refinement stage to address potential color fluctuations and semantic inconsistencies in the reconstructed occluded regions. Our method proffers solutions to two core challenges in VFI tasks such as dealing with non-rigid motion and large-scale occlusion conditions. Extensive experiments on synthetic and real datasets substantiate the superiority of our model and its leading performance.

VI. ACKNOWLEDGMENTS

This work is jointly supported by National Natural Science Foundation of China (62573369), Beijing Natural Science Foundation (4252026).

REFERENCES

[1] Y. Liu, Y. Deng, H. Chen, and Z. Yang, "Video frame interpolation via direct synthesis with the event-based reference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8477–8487.

[2] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, "Depth-aware video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3703–3712.

[3] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super slomo: High quality estimation of multiple intermediate frames for video interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9000–9008.

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

[5] C.-Y. Wu, N. Singhal, and P. Krahenbuhl, "Video compression through image interpolation," in *Proceedings of the European conference on computer vision*, 2018, pp. 416–431.

[6] Z. Tang, Q. Xiong, X. Wu, T. Xu, Y. Shi, X. Xu, J. Xu, and R. Wang, "Radiation reduction for interventional radiology imaging: a video frame interpolation solution," *Insights into Imaging*, vol. 15, p. 42, 2024.

[7] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee, "Learning to generate long-term future via hierarchical prediction," in *International Conference on Machine Learning*, 2017, pp. 3560–3569.

[8] J. Park, K. Ko, C. Lee, and C.-S. Kim, "Bmbc: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Proceedings of the European conference on computer vision*, 2020, pp. 109–125.

[9] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, "Real-time intermediate flow estimation for video frame interpolation," in *Proceedings of the European conference on computer vision*, 2022, pp. 624–642.

[10] Y. Chen, T. Fu, H. Song, D. Xiao, J. Fan, Y. Yu, and J. Yang, "Non-linear motion estimation network for frame interpolation in coronary angiographic sequences," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[11] J. Han, X. Min, J. Jia, Y. Gao, X. Liu, and G. Zhai, "Full-reference and no-reference quality assessment for video frame interpolation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[12] J. Park, J. Kim, and C.-S. Kim, "Biformer: Learning bilateral motion estimation via bilateral transformer for 4k video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1568–1577.

[13] T. Kalluri, D. Pathak, M. Chandraker, and D. Tran, "Flavr: Flow-agnostic video representations for fast frame interpolation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2071–2082.

[14] Z. Shi, X. Xu, X. Liu, J. Chen, and M.-H. Yang, "Video frame interpolation transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 482–17 491.

[15] X. Jin, L. Wu, J. Chen, Y. Chen, J. Koo, C.-H. Hahm, and Z.-M. Chen, "Upr-net: A unified pyramid recurrent network for video frame interpolation," *International Journal of Computer Vision*, vol. 133, no. 1, pp. 16–30, 2025.

[16] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[17] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1964–1980, 2019.

[18] Y. Deng, H. Chen, H. Liu, and Y. Li, "A voxel graph cnn for object classification with event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1172–1181.

[19] B. Yao, Y. Deng, Y. Liu, H. Chen, Y. Li, and Z. Yang, "Sam-event-adapter: Adapting segment anything model for event-rgb semantic segmentation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9093–9100.

[20] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, "Time lens: Event-based video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 155–16 164.

[21] W. He, K. You, Z. Qiao, X. Jia, Z. Zhang, W. Wang, H. Lu, Y. Wang, and J. Liao, "Timereplayer: Unlocking the potential of event cameras for video interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 804–17 813.

[22] S. Wu, K. You, W. He, C. Yang, Y. Tian, Y. Wang, Z. Zhang, and J. Liao, "Video interpolation by event-driven anisotropic adjustment of optical flow," in *Proceedings of the European conference on computer vision*, 2022, pp. 267–283.

[23] T. Kim, Y. Chae, H.-K. Jang, and K.-J. 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*, June 2023, pp. 18 032–18 042.

[24] Y. Ma, S. Guo, Y. Chen, T. Xue, and J. Gu, "Timelens-xl: Real-time event-based video frame interpolation with large motion," in *Proceedings of the European conference on computer vision*, 2025, pp. 178–194.

[25] X. Jin, L. Wu, J. Chen, Y. Chen, J. Koo, and C.-h. Hahm, "A unified pyramid recurrent network for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1578–1587.

[26] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[27] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[28] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 539–14 548.

[29] M. Hu, J. Xiao, L. Liao, Z. Wang, C.-W. Lin, M. Wang, and S. Satoh, "Capturing small, fast-moving objects: Frame interpolation via recurrent motion enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 3390–3406, 2021.

[30] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, "Film: Frame interpolation for large motion," in *Proceedings of the European conference on computer vision*, 2022, pp. 250–266.

[31] X. Cheng and Z. Chen, "Video frame interpolation via deformable separable convolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 607–10 614.

[32] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, and C. Schroers, "Phasenet for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 498–507.

[33] M. Hu, K. Jiang, Z. Zhong, Z. Wang, and Y. Zheng, "Iq-vfi: Implicit quadratic motion estimation for video frame interpolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 6410–6419.

[34] B. Zhou, X. Huang, J. Chen, M. Kaaniche, and P. An, "Capture more, synthesize better: Video frame interpolation with larger receptive field and structural priors," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026.

[35] G. Wu, X. Tao, C. Li, W. Wang, X. Liu, and Q. Zheng, "Perception-oriented video frame interpolation via asymmetric blending," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2024, pp. 2753–2762.

[36] S. Lin, J. Zhang, J. Pan, Z. Jiang, D. Zou, Y. Wang, J. Chen, and J. Ren, "Learning event-driven video deblurring and interpolation," in *Proceedings of the European conference on computer vision*, 2020, pp. 695–710.

[37] Z. Yu, Y. Zhang, D. Liu, D. Zou, X. Chen, Y. Liu, and J. S. Ren, "Training weakly supervised video frame interpolation with events," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 589–14 598.

[38] S. Tulyakov, A. Bochicchio, D. Gehrig, S. Georgoulis, Y. Li, and D. 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*, 2022, pp. 17 755–17 764.

[39] X. Zhang and L. Yu, "Unifying motion deblurring and frame interpolation with events," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 765–17 774.

[40] Y. Gao, S. Li, Y. Li, Y. Guo, and Q. Dai, "Superfast: 200× video frame interpolation via event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7764–7780, 2023.

[41] L. Sun, C. Sakaridis, J. Liang, P. Sun, K. Zhang, J. Cao, Q. Jiang, K. Wang, and L. Van Gool, "Event-based frame interpolation with ad-

hoc deblurring," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 043–18 052.
[42] Y. Liu, L. Fu, H. Chen, Z. Yang, Y. Li, and Y. Deng, "Event-based video interpolation via complementary motion information," *Engineering Applications of Artificial Intelligence*, vol. 162, p. 112606, 2025.
[43] H. Liu, J. Xu, Y. Chang, H. Zhou, H. Zhao, L. Wang, and L. Yan, "Timetracker: Event-based continuous point tracking for video frame interpolation with non-linear motion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 649–17 659.
[44] Y. Liu, L. Fu, Z. Yang, H. Chen, Y. Li, and Y. Deng, "Epa: Boosting event-based video frame interpolation with perceptually aligned learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
[45] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 884–892.
[46] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3d reconstruction and 6-dof tracking with an event camera," in *Proceedings of the European conference on computer vision*, 2016, pp. 349–364.
[47] G. Munda, C. Reinbacher, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *International Journal of Computer Vision*, vol. 126, pp. 1381–1393, 2018.
[48] C. Scheerlinck, N. Barnes, and R. Mahony, "Asynchronous spatial image convolutions for event cameras," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 816–822, 2019.
[49] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3857–3866.
[50] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 156–163.
[51] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, L. K. N. Barnes, and R. Mahoney, "Reducing the sim-to-real gap for event cameras," in *Proceedings of the European conference on computer vision*, 2020, pp. 534–549.
[52] W. Weng, Y. Zhang, and Z. Xiong, "Event-based video reconstruction using transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2563–2572.
[53] A. Li, L. Zhang, Y. Liu, and C. Zhu, "Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 12 514–12 524.
[54] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.
[55] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[56] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, no. 8, pp. 1106–1125, 2019.
[57] D. Gehrig, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Video to events: Recycling video datasets for event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3586–3595.
[58] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.
[59] M. Choi, H. Kim, B. Han, N. Xu, and K. M. Lee, "Channel attention is all you need for video frame interpolation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 663–10 671.
[60] P. Duan, B. Li, Y. Yang, H. Lou, M. Teng, Y. Ma, and B. Shi, "Eventaid: Benchmarking event-aided image/video enhancement algorithms with real-captured hybrid dataset," *arXiv preprint arXiv:2312.08220*, 2023.
[61] I. Loshchilov, F. Hutter *et al.*, "Fixing weight decay regularization in adam," *arXiv preprint arXiv:1711.05101*, vol. 5, p. 5, 2017.
[62] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
[63] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[64] Y. Hu, S.-C. Liu, and T. Delbruck, "v2e: From video frames to realistic dvs events," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1312–1321.
[65] O. Ronneberger, P. Fischer, and T. 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*, 2015, pp. 234–241.
[66] W. Liao, Y. Zhu, X. Wang, C. Pan, Y. Wang, and L. Ma, "Lightm-unet: Mamba assists in lightweight unet for medical image segmentation," *arXiv preprint arXiv:2403.05246*, 2024.

Yuhan Liu (Graduate Student Member, IEEE) received the M.S. degree from Beijing University of Technology, China. He is currently pursuing the Ph.D. degree at Xiamen University, China. His research interests include event-based vision and video processing.

Yongjian Deng received the Ph.D. degree from the City University of Hong Kong in 2021. He is currently with the Tianjin Key Laboratory of Intelligent Robotics, the Academy for Advanced Interdisciplinary Studies, and the College of Artificial Intelligence, Nankai University, Tianjin, China. His research interests include computer vision, deep learning, and event-based vision.

Linghui Fu is currently pursuing the M.S. degree at Beijing University of Technology, China. His research interests include event-based vision and video processing.

Hao Chen received the Ph.D. degree from the City University of Hong Kong in 2019. From 2019 to 2020, he worked as a research fellow at Nanyang Technological University, Singapore. Now, he is an associate professor in the School of Computer Science and Engineering, Southeast University. His research interests include human-inspired computer vision models and machine learning algorithms.

Zhen Yang (Member, IEEE) received the Ph.D. degree in signal processing from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently the Full Professor of computer science and engineering with the Beijing University of Technology, Beijing, China. He has authored or coauthored more than 30 papers in highly ranked journals and top conference proceedings. His research interests include data mining, machine learning, trusted computing, and content security. He is the Senior Member of the Chinese Institute of Electronics.

Youfu Li (Life Fellow, IEEE) received the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, U.K., in 1993. From 1993 to 1995, he was a Research Staff with the Department of Computer Science, University of Wales, Aberystwyth, U.K. He joined the City University of Hong Kong in 1995, where he is currently a Professor with the Department of Mechanical Engineering. He has served as an Associate Editor for the IEEE Transactions on Automation Science and Engineering (T-ASE) and IEEE Robotics and Automation Magazine (RAM), an Editor for Conference Editorial Board (CEB) and the IEEE International Conference on Robotics and Automation (ICRA), and a Guest Editor for RAM. He is a life fellow of IEEE.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Associate Professor (with tenure) and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV/ECCV. He is a senior member of IEEE.

Video Frame Interpolation via Asymmetric Refinement with the Event-based Reference
- Supplementary Material -

Due to the limited space in the main text, we provide more details in the Supplementary Material. This supplementary material is comprised of the following sections.

Sec. 1 presents the detailed network structure of our proposed approach.

Sec. 2 conducts supplementary experiments both in qualitative and quantitative views.

1. Method Details

1.1. Reconstruction stage

We provide a more detailed network structure diagram of the reconstruction module described in the main text, as illustrated in Fig. 1.

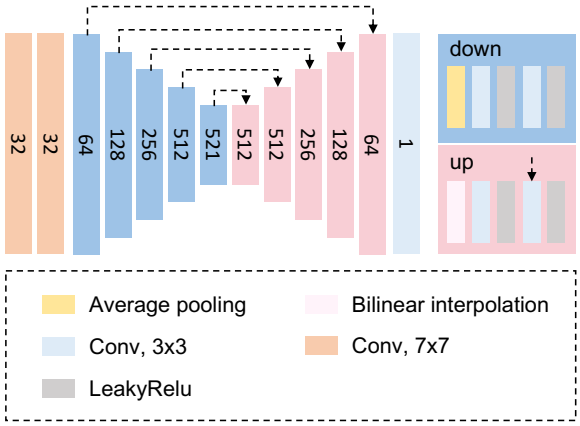

Figure 1. Detailed presentation of the reconstruction network for obtaining the event-based reference.

We adopt an U-shaped network [3] that utilizes pure events for the synthesis of the event-based reference at the interpolated frame position, circumventing the effect of poor keyframe quality on the interpolation results.

1.2. Synthesis stage

The structure of feature encoders ($\mathcal{E}_I$, $\mathcal{E}_E$, $\mathcal{E}_R$) in the synthesis stage is shown in Fig. 2, where we take $\mathcal{E}_I$ as an example.

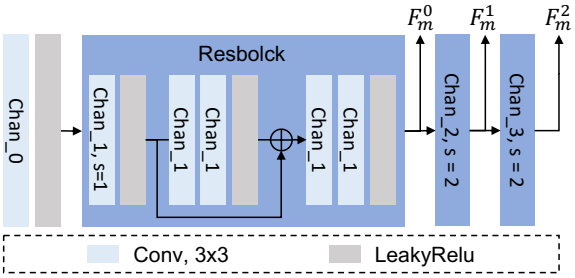

Figure 2. Detailed presentation of the feature *Encoders* adopted in the synthesis and refinement stages.

The symbol $F_m^{0,1,2}$ denotes features obtained from the network at different scales that are supposed to be fed into the E-PCD module. The $\{Chan\_i|i = 0, 1, 2, 3\}$ in the figure represents output channels of the convolutional layer, and s denotes the stride. In specific, output channels are [16, 32, 64, 128], [16, 32, 64, 96] and [8, 16, 32, 64] for $\mathcal{E}_I$, $\mathcal{E}_E$ and $\mathcal{E}_R$, respectively.

Besides, we include four convolutional blocks after E-PCD modules to fuse the bidirectional interpolating features, where the first three consists of a normal 3×3 $Conv$ followed by a $LeakyRelu$ activation function with output channel as 64. The last one holds the similar architecture as the first three blocks but with output channel of 3.

1.3. Refinement stage

We describe detailed network structure of the refinement stage in this section. As illustrated in the main manuscript, the refinement stage comprises of two parts denoted as "*Three Different Encoders*" and "*Transformer Decoder*". Specifically, "*Three Different Encoders*" represent feature encoders that take $\{I_0, I_1, E_{0\to1}, \hat{I}_\tau\}$ as input. We formulate this process in Eq. (1).

$$\begin{aligned} F_{m,*}^{0,1,2} &= \mathcal{E}_I^*(I_m), m \in \{0, 1\}, \\ G_*^{0,1,2} &= \mathcal{E}_E^*(E_{0\to1}), \\ V_*^{0,1,2} &= \mathcal{E}_{pred}^*(\hat{I}_\tau)), \end{aligned} \tag{1}$$

where the structure of these encoders ($\{\mathcal{E}_I^*, \mathcal{E}_E^*, \mathcal{E}_{pred}^*\}$) are similar to feature encoders utilized in the synthesis stage and output channels of their inside layers are set as [8, 16, 32, 64].

After obtaining encoded features from various views ($F_{m,*}^{0,1,2}, G_*^{0,1,2}, V_*^{0,1,2}$), we input them into an interactive Transformer-based decoder (Inspired by [2]). To gain feature refinement at multi-scale, we convey a multi-level Transformer-based decoding to achieve the final interpolation. We illustrate the specific structure of the decoding block at i-th level in Fig. 4 and formulate the input construction of this block in Eq. (2).

$$\begin{aligned} \mathbb{F}_F^i &= \mathcal{C}(F_{m,*}^i, G_*^i), \\ \mathbb{F}_R^i &= \mathcal{C}(V_*^i, F_{m,*}^i, G_*^i, \chi^{i-1}), \\ \mathbb{F}_E^i &= \mathcal{C}(V_*^i, G_*^i), \end{aligned} \tag{2}$$

where the input feature $\mathbb{F}_R$ at i-th level is obtained by fusing the output of previous layer (χ^{i-1}). A total of three different scales of I_τ can be obtained for training, and the generation and computation of Q, K, V follows the standard attentive learning method proposed in [1].

2. Additional Experiments

2.1. Threshold settings for erosion operation

We visualize the erosion operation for different threshold settings, as shown in Fig. 3. We can see that the customized erosion operation is able to get rid of the effect of event noise as much as possible and preventing the significant loss of event data compared to the original erosion operation ($\delta = 1$). In this work, we select δ as 6 for all evaluations.

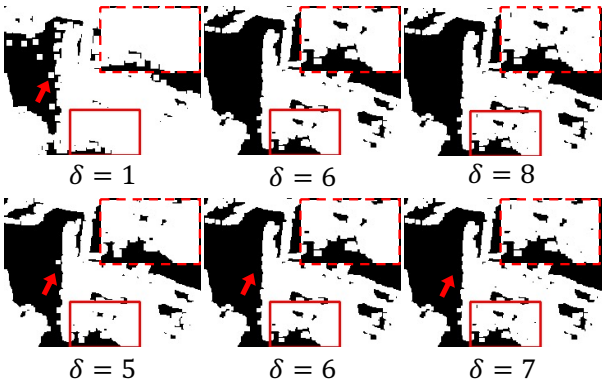

Figure 3. Visualization of event masks generated with different erosion thresholds.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[2] 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. 2

[3] 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. 1

[4] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3

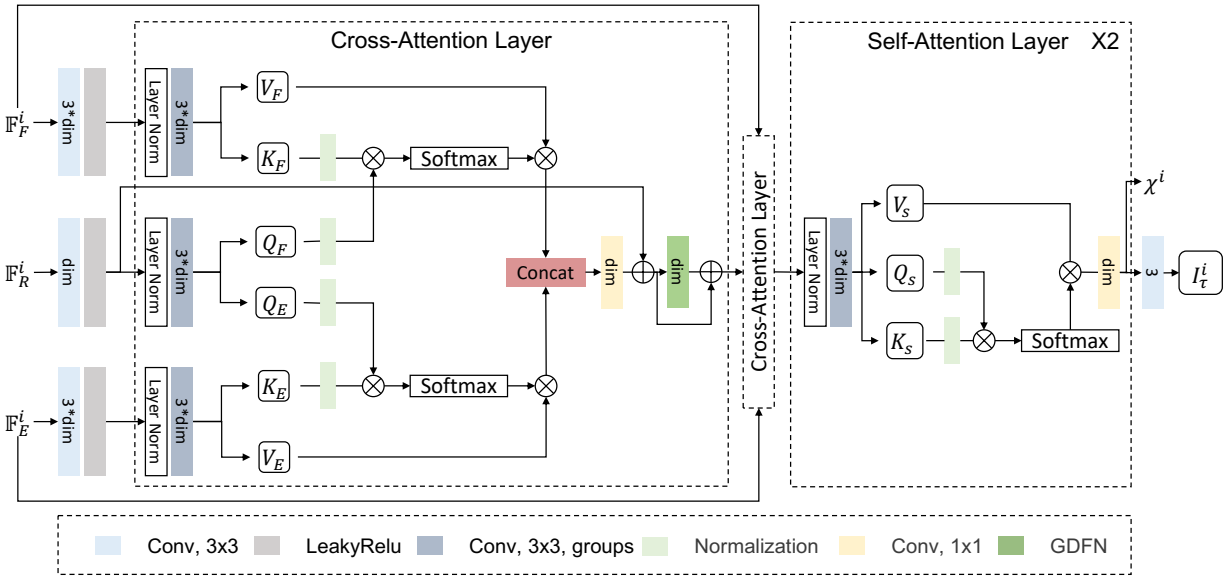

Figure 4. Detailed presentation of the $Transformer\ Decoder$ in the synthesis stage. GDFN represents Gated Deconvolutional Feed-Forward Network designed by [4]. In our work, "dim" is set to 64 for all decoder scales (*e.g.*, $\{i = 0, 1, 2\}$)