

Video Frame Interpolation with Stereo Event and Intensity Cameras

Chao Ding, Mingyuan Lin, Haijian Zhang, Jianzhuang Liu, and Lei Yu

Abstract—The stereo event-intensity camera setup is widely applied to leverage the advantages of both event cameras with low latency and intensity cameras that capture accurate brightness and texture information. However, such a setup commonly encounters cross-modality parallax that is difficult to be eliminated solely with stereo rectification especially for real-world scenes with complex motions and varying depths, posing artifacts and distortion for existing Event-based Video Frame Interpolation (E-VFI) approaches. To tackle this problem, we propose a novel Stereo Event-based VFI (SE-VFI) network (SEVFI-Net) to generate high-quality intermediate frames and corresponding disparities from misaligned inputs consisting of two consecutive keyframes and event streams emitted between them. Specifically, we propose a Feature Aggregation Module (FAM) to alleviate the parallax and achieve spatial alignment in the feature domain. We then exploit the fused features accomplishing accurate optical flow and disparity estimation, and achieving better interpolated results through flow-based and synthesis-based ways. We also build a stereo visual acquisition system composed of an event camera and an RGB-D camera to collect a new Stereo Event-Intensity Dataset (SEID) containing diverse scenes with complex motions and varying depths. Experiments on public real-world stereo datasets, *i.e.*, DSEC and MVSEC, and our SEID dataset demonstrate that our proposed SEVFI-Net outperforms state-of-the-art methods by a large margin. The code and dataset are available at https://dingchao1214.github.io/web_sevfi/.

Index Terms—Stereo event-intensity camera, video frame interpolation, stereo matching, stereo event-intensity dataset.

I. INTRODUCTION

S TEREO event-intensity camera setup allows us to fully perceive the dynamic contents in the scene and has been widely applied in existing depth estimation and stereo matching algorithms [1]–[4]. In this setup, the event camera records continuous motion information with extremely high temporal resolution and low power consumption, while the intensity camera captures precise scene brightness and texture information. However, this setup suffers from the cross-modality parallax issue, especially in real-world scenes with complex non-linear motions and varying depths [5], [6]. As a result, it can significantly degenerate the performance of existing Event-based Video Frame Interpolation (E-VFI) approaches, most of which rely on simulation datasets and require per-pixel spatial alignment between events and frames [7]–[12], leading to

Chao Ding, Mingyuan Lin, Haijian Zhang, and Lei Yu are with the School of Electronic Information, Wuhan University, Wuhan 430072, China. Email:{dingchao, limingyuan, haijian.zhang, ly.wd}@whu.edu.cn.

Jianzhuang Liu is with the Huawei Noah's Ark Lab, Shenzhen 518000, China. Email: liu.jianzhuang@huawei.com.

The research was partially supported by the National Natural Science Foundation of China under Grants 62271354 and 61871297.

Corresponding authors: Lei Yu and Haijian Zhang.

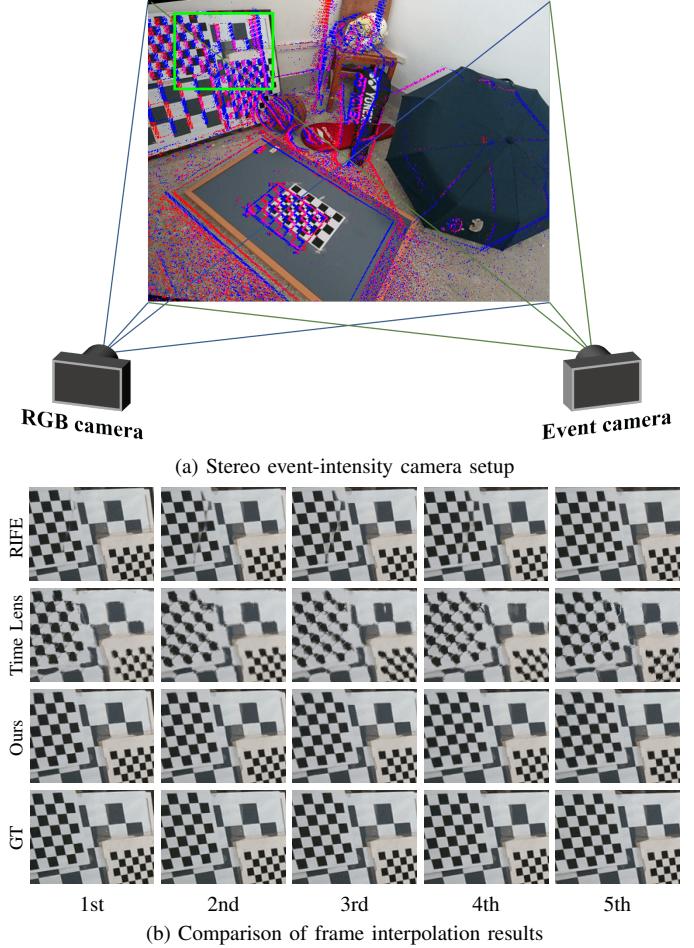


Fig. 1. Illustrative examples of the impact of the misaligned event and frame data. (a) The events and frames captured by a stereo camera setup share the same field of view but are still misaligned at the pixel level. (b) Comparison of video frame interpolation results of RIFE [13], Time Lens [7], and our proposed SEVFI-Net. Our method generates the fewest artifacts and achieves the best visualization performance.

artifacts and distortions with the stereo event-intensity camera setup as shown in Fig. 1.

Achieving the per-pixel cross-modal alignments is extremely challenging for the Stereo E-VFI (SE-VFI) task due to the missing inter-frame intensities and modality differences. Even though existing works [5], [7] directly apply the global homography and stereo rectification, it is only valid for scenes either with a large depth [5] or within a plane [7]. Complex motions in real-world dynamic scenes often lead to temporally varying depths, violating the planar homography and thus bringing cross-modal misalignments. Thus the correspondence

between events and frames is necessary to fulfill the SE-VFI, which is however an ill-posed problem since the missing inter-frame intensities. On the other hand, event cameras work in a completely different mechanism from intensity cameras, only responding to brightness changes and asynchronously emitting binary events. Despite the implicit inclusion of scene structure and texture information in events, they are often triggered at the edges of objects with high-intensity contrast. These modality differences bring another challenge to establishing spatial data correspondence between events and frames.

To handle these problems, we propose a network called SEVFI-Net for the Stereo Event-based Video Frame Interpolation (SE-VFI) task to fully exploit the potential of E-VFI in real-world scenarios. SE-VFI aims to generate high-quality intermediate frames and cross-modal disparities using spatially misaligned events and frames. Our SEVFI-Net consists of four main modules, *i.e.*, *AlignNet*, *SynNet*, *FusionNet*, and *RefineNet*. Specifically, the *AlignNet* is built by composing a weighted dual encoder and a Feature Aggregation Module (FAM) to mitigate the modality gap and spatial misalignment between events and frames. The dual encoder enables us to extract multi-scale features from different modalities separately. To mitigate the spatial misalignment, we employ deformable convolution networks [14] in FAM to estimate the spatial correspondence of different data and obtain aligned features. These features and correspondence enable us to achieve a coarse estimation of motion flows and disparities. Then, we employ motion flows to warp the boundary frames, resulting in flow-based results. Similarly, we utilize disparities to warp the original events, providing contour constraints for synthesis-based results. However, it is important to note that these results are only approximations due to the lack of precise inter-frame intensities and the differences in modalities. As a result, the fused result of them may contain overlaps and distortions. To address these issues, we incorporate a *RefineNet* to eliminate the defects in the fused results and further optimize the estimations from motion flows and disparities. By doing so, we can effectively handle the challenges posed by modality differences and spatial misalignments.

Additionally, due to the limited diversity of scenes captured by current stereo datasets such as the Stereo Event Camera Dataset for Driving Scenarios (DSEC) [5] and the Multi-Vehicle Stereo Event Camera Dataset (MVSEC) [6], they are unable to effectively evaluate the performance of algorithms in various scenes and across a wide range of depth variations. Thus we build a stereo visual acquisition system containing a SilkyEvCam event camera together with an Intel Realsense D455 RGB-D camera and collect a new Stereo Event-Intensity Dataset (SEID). Our SEID captures dynamic scenes with complex non-linear motions and depths variation that pose a challenge for SE-VFI. It provides events with a higher resolution of 640×480 and high-quality RGB frames at the same resolution. Besides, it also provides depth maps synchronized with RGB frames, which can be used for the stereo-matching task.

The main contributions of this paper are three-fold:

- We propose a novel SEVFI-Net framework for the task of E-VFI when frames and events are captured with stereo

camera settings in spatially misaligned scenes.

- We collect a new stereo event-intensity dataset, containing high-resolution events, high-quality RGB frames, and synchronous depth maps captured in various scenes with complex motions and varying depths.
- Extensive experiments show that our SEVFI-Net yields high-quality frames and accurate disparities, and achieves state-of-the-art results on real-world stereo event-intensity datasets.

II. RELATED WORK

A. Frame-based VFI (F-VFI)

Video Frame Interpolation (VFI) is a widely investigated low-level task in computer vision, aiming to restore intermediate frames from the neighboring images in the video sequence [15]. However, it is severely ill-posed and the primary challenge is caused by the missing inter-frame information in terms of motions and textures. To relieve the burden, existing VFI approaches commonly rely on inter-frame motion prediction from neighboring frames [16], and thus can be roughly categorized into *warping-based* and *kernel-based*.

Warping-based methods [17]–[21] utilize optical flow that perceives motion information between consecutive frames and captures dense correspondences. High-quality VFI results often depend on precise motion estimation. Due to the lack of inter-frame motion information, these methods are typically under the assumption of linear motion and brightness constancy between keyframes. Some techniques and information have been utilized to enhance the interpolation performance, *e.g.*, forward warping [17], transformer [18], context [19], depth [20], patch-based [21], [22] and deformable convolution [23], [24]. But due to the assumption of linear motion, these methods encounter challenges when dealing with complex non-linear motions in real-world scenarios. Some approaches design complex high-order motion models, such as cubic [25], [26] and quadratic [27], to address non-linear motions. However, they need more neighboring frames as inputs to estimate motion models and still struggle in tackling real-world complex motions.

Kernel-based methods [15], [28], [29] incorporate both motion estimation and frame reconstruction. They utilize the input frames to estimate convolution kernels, which encode local motions across the input keyframes and generate intermediate frames based on these kernels. Compared to warping-based methods, kernel-based ones are more effective in dealing with occlusion, sudden brightness change, and blurry inputs. But they need expensive costs for computing per-pixel convolution kernels and are hard to capture large motions due to the fixed and small kernel sizes.

The common problem of Frame-based VFI (F-VFI) methods is the missing information between input frames. Thus motion models or convolution kernels can only be estimated by consecutive neighboring frames, which limits the types of scenes that can be coped with by them in the real world.

B. Event-based VFI (E-VFI)

Benefiting from extremely low latency, the event cameras can supplement the inter-frame information by emitting events

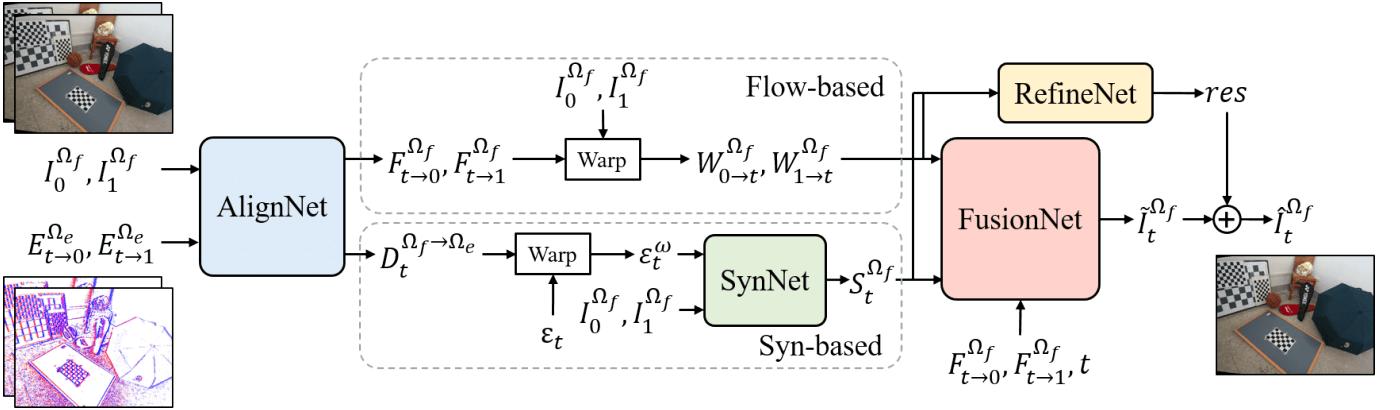


Fig. 2. Overview of the proposed SEVFI-Net, which can be divided into two pathways, i.e., the *Flow-based* and the *Syn-based*, and has four subnets, *AlignNet*, *SynNet*, *FusionNet*, and *RefineNet*.

asynchronously in response to the brightness change [30], [31]. Many works focus on video reconstruction solely with events, which can be seen as a type of VFI. These methods are mainly based on RNN [32]–[34] or GAN [35], [36]. However, it is still ill-posed since events only record the brightness change in the scenes, making it difficult to predict accurate brightness values from them.

Recently, event cameras have been adopted for high-quality VFI [7]–[12]. Thanks to the extremely high temporal resolution, events can provide reliable motion prediction for dynamic scenes even with non-linear motions, resulting in better results than conventional F-VFI approaches. Tulyakov *et al.* [7] propose a network Time Lens that leverages events to compensate for inter-frame motions and textures in a hybrid manner, *i.e.*, warping plus synthesis. Further, Time Lens is extended to Time Lens++ where a motion spline estimator is introduced to predict reliable non-linear continuous flow from sparse events [8]. Instead of supervised learning, He *et al.* [9] design an unsupervised learning framework by applying cycle consistency to bridge the gap between synthesized and real-world events. In addition, Gao *et al.* [11] propose an SNN-based fast-slow joint synthesis framework, *i.e.*, SuperFast, for the high-speed E-VFI task. Xiao *et al.* [12] analyze the drawbacks of existing methods and introduce a novel method named EVA² for E-VFI via cross-modal alignment and aggregation.

However, existing E-VFI approaches largely rely on synthetic data, restricted by per-pixel spatial alignment and ideal imaging without intense motion and sudden brightness change, which is difficult to fulfill in real-world applications since the events and frames are usually captured separately by an event camera and an intensity camera [5], [7]. The misalignment between frames and events cannot be eliminated by global homography or stereo rectification, and inevitably produces cross-modal parallax and poses additional challenges to the VFI task, especially when dynamic scenes with complex motions and varying depths are encountered.

Therefore, our work focuses on handling the parallax between events and frames in SE-VFI effectively utilizing events to compensate for inter-frame information and achieve high-quality video frame interpolation.

C. Stereo Matching

Stereo matching is the process of establishing pixel-to-pixel correspondences between two different views on the epipolar line. Many existing stereo matching methods are based on neural networks and leverage several advanced techniques to improve their performance, *e.g.*, encoder-decoder architecture [37], 3D convolution cost-volume module [38]–[41], adaptive sample aggregation module [42], attention [43], transformer [44], and unsupervised learning [45]. However, frame-to-frame stereo matching methods are based on the assumption of high-quality imaging under ideal conditions, where there are no fast motion and high-dynamic-range scenes.

In recent works, event cameras with high-speed and high-dynamic-range imaging capabilities have been utilized to improve the performance of stereo-matching algorithms in various complex scenes. Due to the fact that event and intensity cameras perceive the same light field, the edge information extracted from events and intensity images can be correlated to calculate the sparse disparity map [3], [4]. Additionally, some learning-based methods are proposed to achieve dense disparity estimation. Specifically, Zou *et al.* [2] introduce an hourglass architecture with a pyramid attention module, extracting multi-scale features and performing fusion by using convolutional kernels of different sizes. Gu *et al.* [1] propose a self-supervised learning framework, achieving data matching and disparity estimation by establishing gradient structure consistency between frames and events.

However, existing stereo-matching methods are limited by the frame rate of the input frame sequence. In contrast, our method aims to estimate the corresponding disparity map during video frame interpolation, allowing the resulting disparity sequence to be decoupled from the frame rate of the input sequence. This enables a more comprehensive and effective representation of depth variations in the scene.

III. PROBLEM FORMUALTION

Conventional Frame-based VFI (F-VFI) methods aim at reconstructing intermediate frames between two keyframes:

$$I_t = \text{F-VFI}(t; I_0, I_1), \quad t \in [0, 1], \quad (1)$$

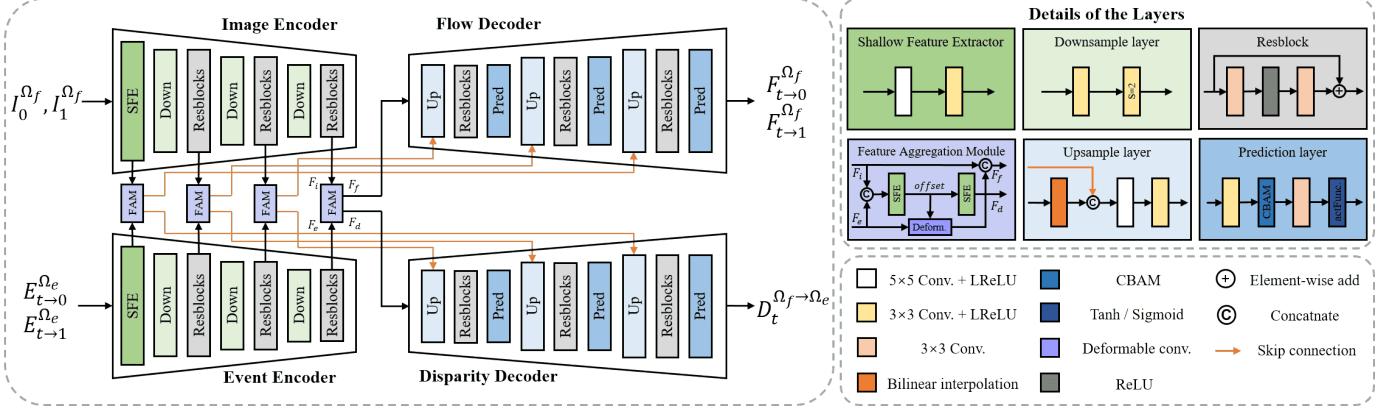


Fig. 3. Structure of *AlignNet*. We use different colors to represent different network layers, and the details of the layers are shown on the right side.

where I_0, I_1 are input keyframes and I_t is the target intermediate frame at a normalized time $t \in [0, 1]$. However, due to the lack of motion information between consecutive input keyframes, most Frame-based VFI (F-VFI) methods are built on simplified assumptions, *e.g.*, linear motions [17]–[20] or local movements [15], [28], leading to performance degradation in real-world scenarios.

To tackle this problem, E-VFI methods are proposed by predicting inter-frame motion information with events [7]–[9], [11], [12], achieving better interpolation performance than F-VFI methods. They generate intermediate frames with the help of events triggered between input keyframes, thus can model motions close to the real trajectory:

$$I_t^{\Omega_f} = \text{E-VFI}(t; I_0^{\Omega_f}, I_1^{\Omega_f}; E_{0 \rightarrow 1}^{\Omega_e}), \quad t \in [0, 1], \quad (2)$$

where Ω_f represents the intensity camera's image plane, implying an assumption that the frames and events are aligned at the pixel level. However, this assumption is not always valid in existing stereo frame-event camera setups. Parallax tends to exist between the two data, especially in foreground regions with small depths. If we mix the unaligned events and frames directly for E-VFI, the parallax of the two data will bring burdens in motion estimation, leading to severe artifacts and distortion in the reconstruction result.

To tackle this problem, we introduce a novel framework for SE-VFI with the misaligned event and frame data:

$$I_t^{\Omega_f}, \mathcal{D}_t^{\Omega_f \rightarrow \Omega_e} = \text{SE-VFI}(t; I_0^{\Omega_f}, I_1^{\Omega_f}; E_{0 \rightarrow 1}^{\Omega_e}), \quad t \in [0, 1] \quad (3)$$

where Ω_f and Ω_e represent different image planes of intensity and event cameras respectively, and $\mathcal{D}_t^{\Omega_f \rightarrow \Omega_e}$ denotes the disparity from Ω_f to Ω_e . The key to SE-VFI is to deal with the parallax between events and frames by establishing a spatial data correlation, complementing the missing motion information between frames using events to realize the reconstruction of intermediate frames, and further estimating the disparity maps between events and frames from the spatial data correlation.

IV. METHOD

In this section, we describe the proposed SEVFI-Net that can generate intermediate frames and disparities from mis-

aligned events and frames.

A. Pipeline Overview

Our goal is to generate one or more intermediate frames $I_t^{\Omega_f}$ from consecutive keyframes $I_0^{\Omega_f}, I_1^{\Omega_f}$ in one view along with concurrent events $E_{0 \rightarrow 1}^{\Omega_e}$ in the other. As shown in Fig. 2, SEVFI-Net contains four subnets, *AlignNet*, *SynNet*, *FusionNet*, and *RefineNet*. The main issue of SE-VFI is to handle the parallax between the input events and frames. So we set *AlignNet* at first, which receives the origin data and mitigates the modality gap and spatial misalignment with the help of the key FAM. Specifically, we use FAM to establish the spatial correlation that is represented as the offset between inputs and obtain the aligned feature. Then the aligned feature and the offset are fed into two decoders to estimate bi-directional optical flow and disparity separately. Afterward, we split the VFI process into two pathways: the *Flow-based* and the *Syn-based*. The flow-based results demonstrate excellent performance in regions with small motions, while the syn-based results excel in regions with sufficient events. Then these results are input to *FusionNet* to produce a high-quality output, which is further refined by *RefineNet* to improve the reconstruction quality.

B. AlignNet

Unlike conventional E-VFI methods, which usually concatenate frames and events as input, SE-VFI focuses more on reducing the parallax effect in the data. Considering the modality discrepancy and the misalignment problem, *AlignNet* is based on an hourglass architecture [46] using two encoders and two decoders with similar structures but unshared weights as illustrated in Fig. 3.

Assuming that we need to reconstruct the image at time t , we first divide the event stream into two parts and operate them by time shift and polarity reversal [7], [47] to get $E_{t \rightarrow 0}^{\Omega_e}, E_{t \rightarrow 1}^{\Omega_e}$, and represent them by the voxel grid [48]. The frames and events are initially fed into a Shallow Feature Extractor (SFE) to obtain full-scale features. Subsequently, they are passed through Downsample layers and Resblocks to obtain features at different scales.

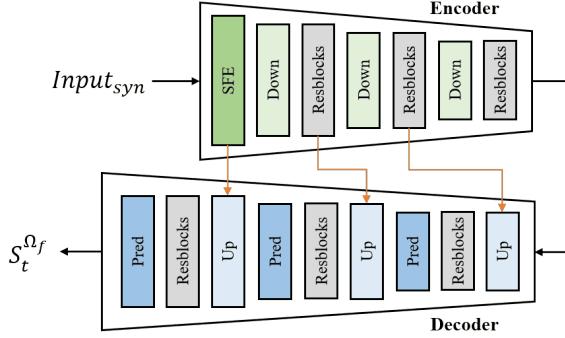


Fig. 4. Architecture of *SynNet*, which is a simple U-Net-like structure. The details of the network layers are described in Fig. 3.

However, the pivotal issue is the parallax between the inputs, so we design the FAM to tackle it. Inspired by [14], [49], in FAM (see Fig. 3), we first concatenate the features F_i and F_e , and then input them to an SFE to compute their spatial data correlation, which is represented as the offset. Since deformable convolution can sample flexible locations, we utilize the offset to guide the deformable convolution layer for feature alignment. We set the feature from the image encoder F_i as the reference. Then the offset provides the data association between F_i and F_e , guiding F_e to align with F_i . Afterward, we concatenate the deformed feature with the input F_i to obtain the aligned feature F_f and pass the offset to another SFE to get the new feature F_d that is used to estimate disparity. The flow decoder and the disparity decoder differ only slightly in the prediction layers. At first, the CBAM [50] block is employed to enhance the features from both channel and spatial perspectives. Then the flow decoder uses Tanh as the activation function to be compatible with negative flow values, while the disparity decoder uses sigmoid since disparities are all positive. The features are fed into the two decoders that output bi-directional optical flows $F_{t \rightarrow 0}^{\Omega_f}, F_{t \rightarrow 1}^{\Omega_f}$ and the disparity $D_t^{\Omega_f \rightarrow \Omega_e}$. In summary, the input and output of AlignNet are represented as:

$$F_{t \rightarrow 0}^{\Omega_f}, F_{t \rightarrow 1}^{\Omega_f}, D_t^{\Omega_f \rightarrow \Omega_e} = \text{AlignNet}(I_0^{\Omega_f}, I_1^{\Omega_f}; E_{t \rightarrow 0}^{\Omega_e}, E_{t \rightarrow 1}^{\Omega_e}). \quad (4)$$

C. SynNet

We divide our SEVFI-Net into two pathways, *i.e.*, the flow-based and the syn-based. In the flow-based way, we utilize bi-directional flows to warp the boundary keyframes $I_0^{\Omega_f}, I_1^{\Omega_f}$ separately to the target time t and get the warped images $W_{0 \rightarrow t}^{\Omega_f}$ and $W_{1 \rightarrow t}^{\Omega_f}$ as:

$$\begin{aligned} W_{0 \rightarrow t}^{\Omega_f} &= \text{Flow-Warp}(I_0^{\Omega_f}; F_{t \rightarrow 0}^{\Omega_f}), \\ W_{1 \rightarrow t}^{\Omega_f} &= \text{Flow-Warp}(I_1^{\Omega_f}; F_{t \rightarrow 1}^{\Omega_f}). \end{aligned} \quad (5)$$

However, the optical flow estimation relies on the assumption of brightness invariance, which can result in inaccuracies in regions where objects move rapidly. Thus we use the syn-based way to fix it.

Since events are emitted with the brightness change at the high temporal resolution, they can record complex motions and

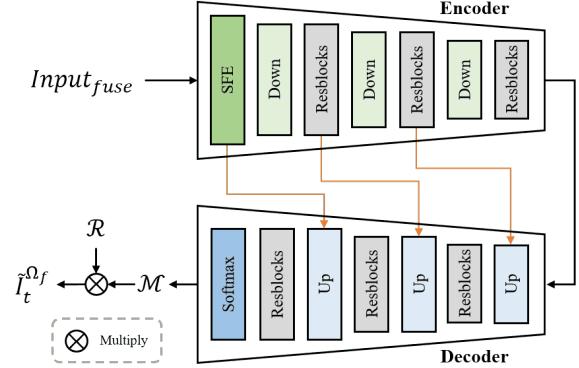


Fig. 5. Architecture of *FusionNet*, which outputs different attention maps to unify the warped and synthesized results together.

help reconstruct intermediate images. But the original events in the stereo camera setup are not aligned with the frames. So we first gather the events $e = \{x_i, y_i, t_i, p_i\}_{i=1}^N$ within a time window Δt around the target time t and then compress them into a tensor, which can be formulated as:

$$\varepsilon_t = \sum_{i \in \{i | t - \frac{\Delta t}{2} \leq t_i \leq t + \frac{\Delta t}{2}\}} p_i \delta(x - x_i, y - y_i). \quad (6)$$

Subsequently, we use the previously learned disparity to warp the event tensors to get new tensors that match the spatial positions with the frames:

$$\varepsilon_t^w = \text{Disp-Warp}(\varepsilon_t; \mathcal{D}_t^{\Omega_f \rightarrow \Omega_e}). \quad (7)$$

Then we concatenate the keyframes with the warped event tensors as the input $Input_{syn} = [I_0^{\Omega_f}, I_1^{\Omega_f}; \varepsilon_t^w]$ of the U-Net-like *SynNet* shown in Fig. 4, and finally we obtain a synthesized result:

$$S_t^{\Omega_f} = \text{SynNet}(Input_{syn}). \quad (8)$$

D. FusionNet and RefineNet

The flow-based method performs well in regions with small motions and stable brightness, while the syn-based method relies more on regions with sufficient events and can handle complex scenes. Therefore, we use FusionNet to combine the strengths of both methods and achieve better results. We use the warped and synthesized results together with the optical flow and target time t as input, *i.e.*, $Input_{fuse} = [W_{0 \rightarrow t}^{\Omega_f}, W_{1 \rightarrow t}^{\Omega_f}, S_t^{\Omega_f}, F_{t \rightarrow 0}^{\Omega_f}, F_{t \rightarrow 1}^{\Omega_f}, t]$, adjusting the proportion of each result by using the optical flow and target time as weights. FusionNet is shown in Fig. 5 and we use the Softmax function at the end to generate three attention maps $\mathcal{M} = \{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2\}$ and then multiply \mathcal{M} by the previous results $\mathcal{R} = \{W_{0 \rightarrow t}^{\Omega_f}, W_{1 \rightarrow t}^{\Omega_f}, S_t^{\Omega_f}\}$ to obtain the fused result $\tilde{I}_t^{\Omega_f}$:

$$\mathcal{M} = \text{FusionNet}(Input_{fuse}), \quad (9)$$

$$\tilde{I}_t^{\Omega_f} = \mathcal{M} \otimes \mathcal{R} = \mathcal{M}_0 \cdot W_{0 \rightarrow t}^{\Omega_f} + \mathcal{M}_1 \cdot W_{1 \rightarrow t}^{\Omega_f} + \mathcal{M}_2 \cdot S_t^{\Omega_f}. \quad (10)$$

Finally, we use a Residual Dense Network (RDN) [51] as RefineNet, as shown in Fig. 2 where the different results are

concatenated and passed to it to learn a residual res , which is added to the fused result $\tilde{I}_t^{\Omega_f}$ to obtain the final result $\hat{I}_t^{\Omega_f}$, which can be formulated as:

$$res = \text{RefineNet}(W_{0 \rightarrow t}^{\Omega_f}, W_{1 \rightarrow t}^{\Omega_f}, S_t^{\Omega_f}), \quad (11)$$

$$\hat{I}_t^{\Omega_f} = \tilde{I}_t^{\Omega_f} + res. \quad (12)$$

E. Loss Functions

Our training loss comprises three parts: *reconstruction loss* \mathcal{L}_{rec} , *flow loss* \mathcal{L}_{flow} , and *disparity loss* \mathcal{L}_{disp} .

The *reconstruction loss* \mathcal{L}_{rec} models the reconstruction quality of the intermediate frames. We denote the ground truth frames by I_t^{gt} and utilize the \mathcal{L}_1 loss to evaluate the similarity between the reconstruction and the ground truth. To achieve better visual quality, we add the perceptual loss [52] to \mathcal{L}_{rec} that is formulated as:

$$\begin{aligned} \mathcal{L}_{rec} = & \mathcal{L}_1(S_t^{\Omega_f}, I_t^{gt}) + \mathcal{L}_1(\tilde{I}_t^{\Omega_f}, I_t^{gt}) + \mathcal{L}_1(\hat{I}_t^{\Omega_f}, I_t^{gt}) \\ & + 0.1 \mathcal{L}_{\text{perceptual}}(\hat{I}_t^{\Omega_f}, I_t^{gt}). \end{aligned} \quad (13)$$

The *flow loss* \mathcal{L}_{flow} consists of the *photometric loss* and the *smoothness loss* used in [53]. The former aims to minimize the difference in intensity between the warped image and the ground truth, while the latter regularizes the output flow by minimizing the flow difference between adjacent pixels in the horizontal, vertical, and diagonal directions. \mathcal{L}_{flow} is formulated as:

$$\begin{aligned} \mathcal{L}_{flow} = & \mathcal{L}_{\text{photometric}}(W_{0 \rightarrow t}^{\Omega_f}, I_t^{gt}) + 0.1 \mathcal{L}_{\text{smoothness}}(F_{t \rightarrow 0}^{\Omega_f}) \\ & + \mathcal{L}_{\text{photometric}}(W_{1 \rightarrow t}^{\Omega_f}, I_t^{gt}) + 0.1 \mathcal{L}_{\text{smoothness}}(F_{t \rightarrow 1}^{\Omega_f}). \end{aligned} \quad (14)$$

The *disparity loss* \mathcal{L}_{disp} models the prediction quality of disparity. We denote the ground truth disparity by D_t^{gt} , and first optimize the prediction by the smooth \mathcal{L}_1 loss since it is more robust, which is defined as:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1, \\ |x| - 0.5 & \text{otherwise.} \end{cases} \quad (15)$$

We also incorporate the edge-aware disparity smoothness loss \mathcal{L}_{ds} used in [54] to promote the local smoothness of disparities by computing the cost using the gradients of both disparities and frames, which is represented as:

$$\mathcal{L}_{ds} = \frac{1}{N} \sum_{i,j} |\partial_x D_{t,i,j}^{\Omega_f \rightarrow \Omega_e}| e^{-||\partial_x I_{t,j}^{gt}||} + |\partial_y D_{t,i,j}^{\Omega_f \rightarrow \Omega_e}| e^{-||\partial_y I_{t,j}^{gt}||}, \quad (16)$$

where i, j are pixel coordinates, x, y represent two directions, and $\partial D, \partial I$ are disparity gradients and image gradients. The disparity loss can be expressed as:

$$\mathcal{L}_{disp} = \text{smooth}_{L_1}(D_t^{\Omega_f \rightarrow \Omega_e}, D_t^{gt}) + 0.1 \mathcal{L}_{ds}. \quad (17)$$

The total training loss combines all of the above terms:

$$\mathcal{L}_{total} = \lambda_r \mathcal{L}_{rec} + \lambda_f \mathcal{L}_{flow} + \lambda_d \mathcal{L}_{disp}. \quad (18)$$

In practice, we adjust the balancing factors for training on different datasets.



Fig. 6. Illustration of the stereo event-intensity camera setup. It features a SilkyEvCam event camera (left) and an Intel Realsense D455 RGB-D camera (right). Both cameras are synchronized with a baseline of 8cm.

TABLE I
DETAILS OF OUR SEID DATASET. IN TOTAL, THE DATASET CONSISTS OF 34 SEQUENCES FOR BOTH INDOOR AND OUTDOOR SCENES.

Split	Subset	#Sequences	#Frames	#Events (M)
Train	Basketball	6	2189	100.6
	Cars	4	1516	152.6
	Checkerboard	4	1189	119.1
	Indoor	5	1736	132.2
	Pedestrians	3	1310	141.8
	Square	6	2433	277.3
		28	10373	923.6
Test	Basketball	1	332	9.7
	Cars	1	338	38.9
	Checkerboard	1	298	20.6
	Indoor	1	512	42.3
	Pedestrians	1	241	34.7
	Square	1	330	28.1
		6	2051	174.3
Total		34	12424	1097.8

V. STEREO EVENT-INTENSITY DATASET

There are two public stereo event datasets, DSEC [5] and MVSEC [6]. DSEC is a large-scale outdoor stereo event dataset especially for driving scenarios, while MVSEC captures a single indoor scene and multi-vehicle outdoor driving scenes. But for the VFI task, these two datasets have limited scene diversity and cannot effectively evaluate the algorithm's generalization ability across different scenes. Additionally, these datasets record depths using LiDAR, which can only provide sparse depth values. Typically, only around 10% of the depth values are valid, while the remaining 90% are empty [55]. Therefore, the LiDAR data as the reference for VFI evaluation would have limited capability.

To address this problem, we build a stereo visual acquisition system containing a SilkyEvCam event camera and an Intel Realsense D455 RGB-D camera with a baseline of 8cm as shown in Fig. 6. We collect a new SEID dataset that captures various indoor and outdoor scenes with varying depths and complex motions. The RGB-D camera provides frames and depth maps that are synchronized in both temporal and spatial domains. We collect the data from different sensors through the ROS system and use the UTC clock of the control computer to timestamp the data for soft synchronization. In order to eliminate possible time delays, we manually calibrate the timestamps of events to achieve precise time synchronization.

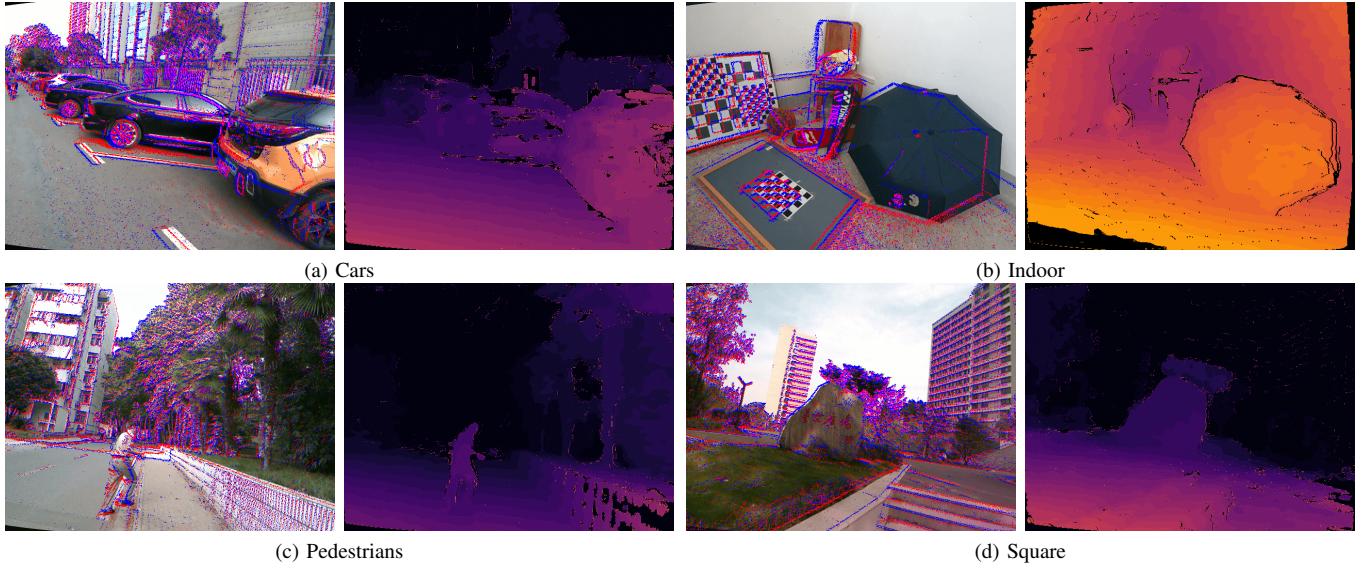


Fig. 7. Illustration of example images rendered with events and their corresponding disparity images. Since we remove the depth values that are smaller than 1 meter to avoid errors during recording, there are holes in the disparity images.

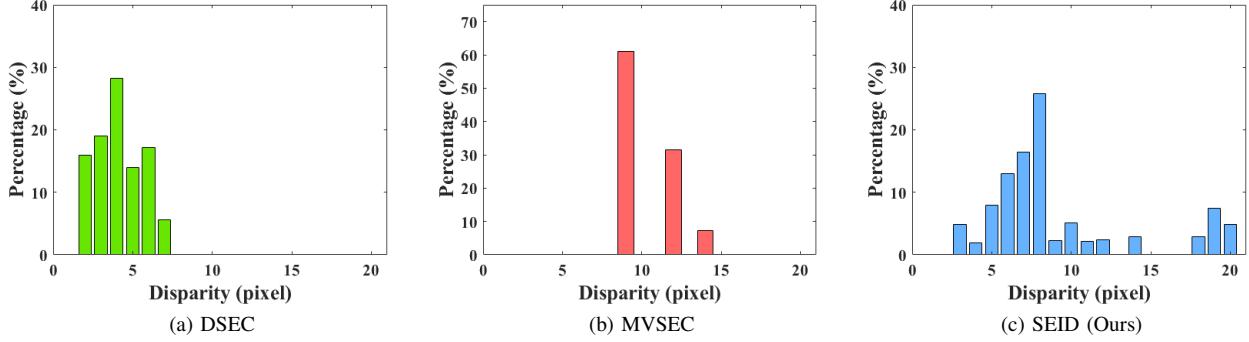


Fig. 8. Illustration of the disparity distribution of each dataset. The disparity values of DSEC [5] and MVSEC [6] are concentrated within a small range, indicating that the scenes they capture are similar and homogeneous. In contrast, our SEID dataset captures a wider range of scenes with richer variations in depth values, resulting in a broader disparity distribution.

TABLE II
COMPARISON OF OUR SEID WITH THE PUBLICLY AVAILABLE DATASETS
MVSEC [6] AND DSEC [5].

	FPS	Resolution	Color	Depth
MVSEC [6]	32	346 × 260	✗	Sparse
DSEC [5]	20	640 × 480	✓	Sparse
SEID (Ours)	60	640 × 480	✓	Dense

To estimate the intrinsic and extrinsic parameters, we utilize the event-based image reconstruction algorithm E2VID [32] to generate high-quality images for calibration. In order to achieve good calibration results, we use the MATLAB calibration toolbox to calibrate the intrinsic parameters of the two cameras first, and then we use the obtained intrinsic parameters to calibrate the extrinsic parameters between the two cameras. Afterward, we employ the calibrated parameters to perform stereo rectification on the data, eliminating the deviations in the y and z directions. Furthermore, we remove the values of less than one meter to avoid invalid depth values captured by the RGB-D camera and then convert the depths to disparities using the calibrated parameters.

Our SEID dataset contains 34 indoor and outdoor sequences that are summarized in Tab. I. The two cameras share a similar field of view (FoV), and both events and frames have a resolution of 640×480 . The dataset features frames at 60 FPS, which is higher than the previous stereo datasets as shown in Tab. II. High spatial resolution helps capture richer texture details, while a high frame rate preserves more temporal and motion information, providing more valid references at various target timestamps for the VFI task. In addition, we use an RGB-D camera to capture the depths of the scenes, which provides dense disparity maps as shown in Fig. 7 different from LiDAR data. This offers more reliable reference values for cross-modal stereo-matching tasks. Our SEID dataset captures richer motions and varying depths, providing a more diverse range of depth variations and scene dynamics. We calculate the average values for each disparity map and further compute the average disparity for each sequence. Then, we conduct a statistical analysis of the disparity distributions in the three datasets, as shown in Fig. 8. The DSEC and MVSEC datasets exhibit a concentration of disparity values within a narrow range, indicating their limited scene diversity. In contrast, our SEID dataset captures a wider range of scenes with richer variations

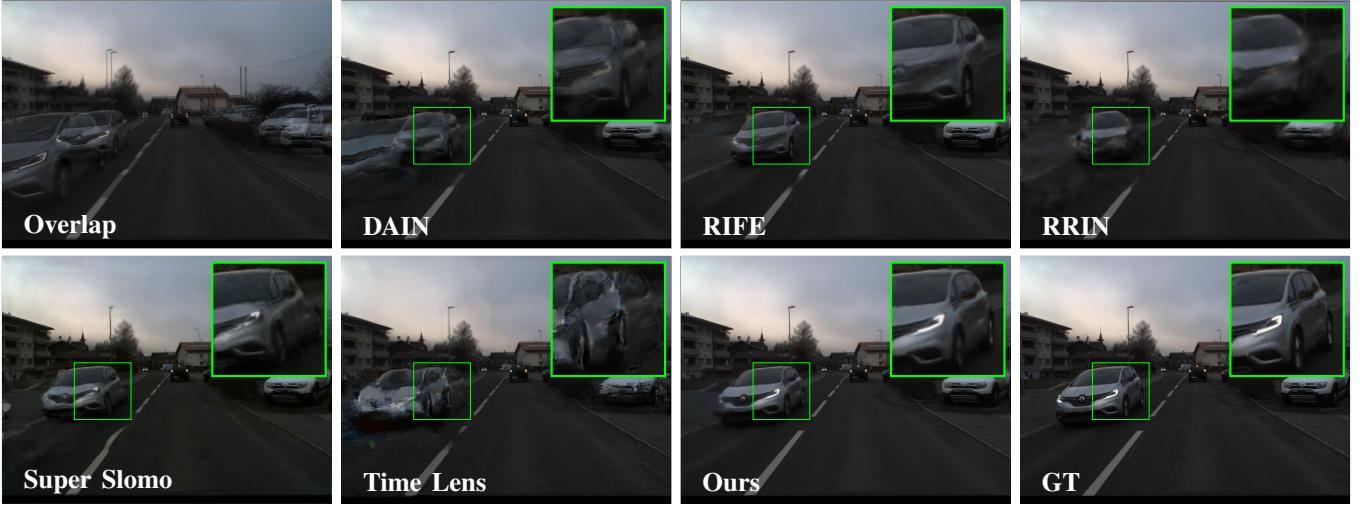


Fig. 9. Qualitative comparisons of SEVFI-Net to state-of-the-art methods on the DSEC [5] dataset. Zoom in on the details for a better view.

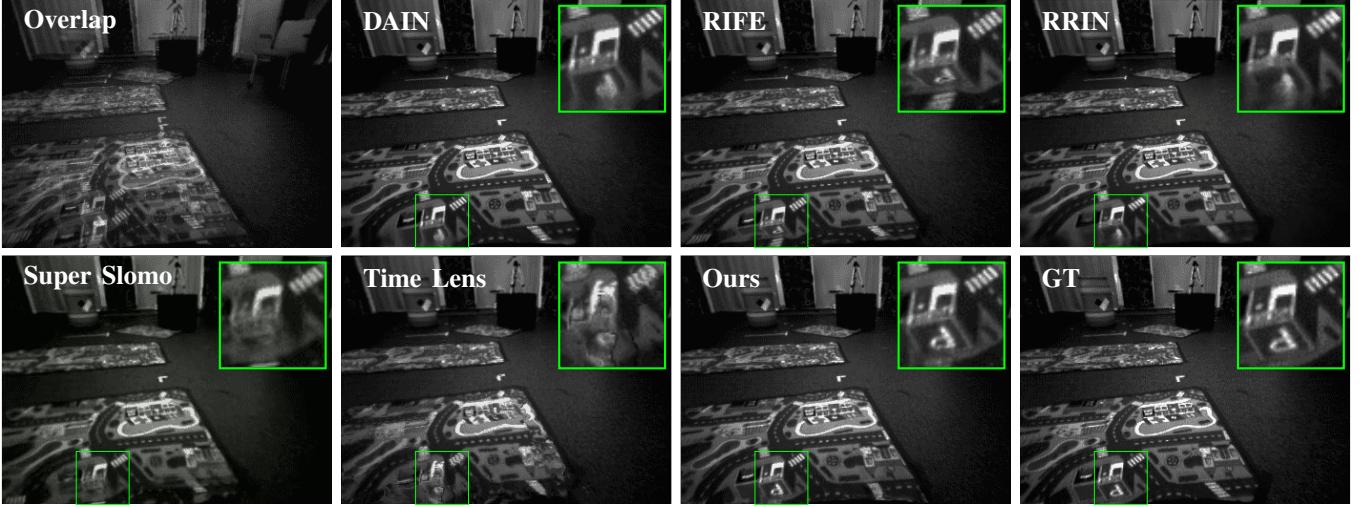


Fig. 10. Qualitative comparisons of SEVFI-Net to state-of-the-art methods on the MVSEC [6] dataset. Zoom in on the details for a better view.

in depths, resulting in a broader distribution of disparities. Therefore, SEID can be used to validate the impact of different disparity conditions on the performance of E-VFI. This enables us to fully exploit the potential of event cameras in stereo tasks and harness their capabilities.

VI. EXPERIMENTS AND ANALYSIS

In this section, we evaluate and analyze the proposed SEVFI-Net. In Sec. VI-A, we first describe the experimental settings, including datasets, training details, and compared methods. Then we compare the VFI performance with the state-of-the-art in Sec. VI-B and analyze the impact of disparity on the E-VFI task. After that, we evaluate the performance of stereo matching with existing cross-modal stereo matching methods in Sec. VI-C. Finally, the importance of the subnets in our SEVFI-Net is discussed in Sec. VI-D.

A. Experimental Settings

1) *Datasets*: We use DSEC [5], MVSEC [6], and our SEID datasets for training and evaluation.

DSEC. We select 23 daylight event-frame sequences for training and 10 for testing, using the data from the left intensity camera and the right event camera. Both events and frames are rectified to eliminate distortions and deviations, and the frames are resized to 640×480 pixels to keep the same resolution as the events.

MVSEC. We use the “*indoor_flying*” scene data to verify the performance of our model. We choose 3 sequences as the training set and 1 sequence as the testing set, also with the left frames and right events. The extrinsic parameters are used to convert the provided depths to disparities, both frames and events are undistorted and rectified with the calibration parameters.

SEID (Ours). We divide SEID into two parts, *i.e.*, the training set containing 28 sequences and the testing set containing 6 sequences. Each part consists of six subsets as illustrated in Tab. I. The events, frames, and disparities are undistorted and rectified using the calibration parameters.

2) *Training Details*: We implement the proposed SEVFI-Net in Pytorch [56] and train three models separately on

TABLE III

QUANTITATIVE COMPARISONS OF THE PROPOSED SEVFI-NET TO STATE-OF-THE-ART METHODS ON THE DSEC [5], MVSEC [6], AND OUR SEID DATASETS. WE COMPUTE PSNR AND SSIM ON THE RECONSTRUCTED FRAMES.

Method	Event	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DSEC							
		1-frame skip		3-frame skip		5-frame skip	
DAIN [20]	✗	30.36	0.8977	27.72	0.8535	25.96	0.8147
RIFE [13]	✗	<u>32.24</u>	0.9008	<u>28.58</u>	<u>0.8541</u>	25.75	0.7969
RRIN [16]	✗	32.15	0.9172	27.25	0.8451	24.99	0.7950
Super Slomo [58]	✗	28.45	0.8507	26.31	0.8101	24.76	0.7726
Time Lens [7]	✓	29.36	0.8661	26.47	0.8144	24.77	0.7735
SEVFI-Net (Ours)	✓	32.97	<u>0.9141</u>	30.29	0.8834	26.69	0.8186
MVSEC							
		3-frame skip		5-frame skip		7-frame skip	
DAIN [20]	✗	32.40	0.9124	30.87	<u>0.8814</u>	29.62	<u>0.8512</u>
RIFE [13]	✗	<u>35.09</u>	0.9159	<u>32.08</u>	0.8648	<u>31.45</u>	0.8440
RRIN [16]	✗	33.57	0.9168	31.40	0.8725	29.83	0.8336
Super Slomo [58]	✗	28.61	0.7843	27.82	0.7611	27.05	0.7383
Time Lens [7]	✓	31.28	0.8692	29.81	0.8341	28.70	0.8049
SEVFI-Net (Ours)	✓	35.23	0.9305	33.79	0.9053	32.04	0.8741
SEID							
		5-frame skip		7-frame skip		9-frame skip	
DAIN [20]	✗	26.38	0.8418	24.84	0.7959	23.50	0.7538
RIFE [13]	✗	<u>27.53</u>	<u>0.8550</u>	<u>26.29</u>	<u>0.8187</u>	<u>24.00</u>	0.7569
RRIN [16]	✗	26.32	0.8470	24.00	0.7831	22.34	0.7307
Super Slomo [58]	✗	26.09	0.8353	24.22	0.7862	22.67	0.7408
Time Lens [7]	✓	25.91	0.8252	24.46	0.7895	23.32	0.7602
SEVFI-Net (Ours)	✓	29.27	0.8881	27.51	0.8533	25.83	0.8138

DSEC, MVSEC, and SEID datasets. To enhance the robustness of SEVFI-Net, we augment the data by randomly cropping the samples into 224×224 patches. It is trained using the Adam optimizer [57] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 100 epochs. The learning rate is initialized at 3×10^{-4} and decayed by 0.8 every 10 epochs. As different datasets record different scenes with varying disparities, we adjust the balancing parameters in Eq. (18) for training. We set $[\lambda_r, \lambda_f, \lambda_d] = [2, 0.005, 0.008]$ for training on DSEC, $[\lambda_r, \lambda_f, \lambda_d] = [1, 0.01, 0.001]$ for MVSEC, and $[\lambda_r, \lambda_f, \lambda_d] = [2, 0.002, 0.005]$ for SEID. The model is trained for 100 epochs on each dataset, using 2 NVIDIA TITAN RTX3090 GPUs.

3) *Benchmark*: We compare SEVFI-Net with one open-sourced E-VFI approach, *i.e.*, Time Lens [7] and four state-of-the-art F-VFI methods, DAIN [20], RIFE [13], RRIN [16], and Super Slomo [58]. The metrics of Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [59] are used for quantitative evaluation. The higher the values of PSNR and SSIM, the better the performance.

Since the datasets have different frame rates, we adaptively evaluate each with different skip values to ensure similar inter-frame motion durations. Specifically, we implement the experiments under the setting of 1-, 3-, and 5-frame skips for sequences in DSEC, 3-, 5-, and 7-frame skips in MVSEC, and 5-, 7-, and 9-frame skips in SEID.

B. Comparisons of Video Frame Interpolation

In this subsection, comparisons of our SEVFI-Net to the state-of-the-art methods are made qualitatively and quantitatively in multi-frame reconstruction. After that, the superiority and robustness of SEVFI-Net compared to other E-VFI methods in dealing with disparity changes are also verified.

1) *Results of F-VFI and E-VFI*: In this part, we analyze the limitations of the F-VFI methods and identify the challenges

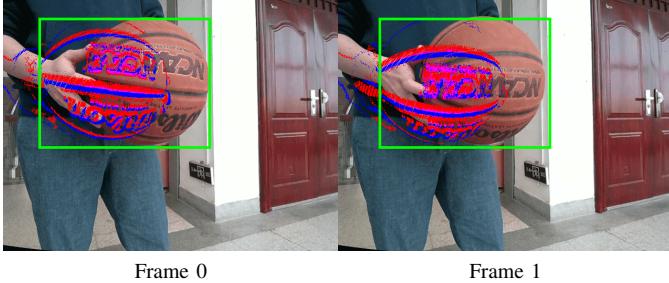
encountered by the E-VFI methods when handling stereo event-frame data. Finally, we highlight the effectiveness of our method. Qualitative and quantitative results are provided to show the superiority of our SEVFI-Net.

Qualitative comparisons. We select two samples from the DSEC and MVSEC datasets to compare the visual performance of each VFI algorithm as illustrated in Figs. 9 and 10.

As displayed in Figs. 9 and 10, the interpolated frames of the F-VFI algorithms fail in dealing with fast and complex motions, leading to blurred or distorted edges in the results. Although they can roughly reconstruct the scene, they estimate inter-frame motion using the input keyframes and are only good at handling simple linear motion. When the scene becomes complex or the motion is intense, the F-VFI algorithms often fail and produce unrealistic structures and textures. This can result in significant degradation of the interpolated frames.

Compared to F-VFI methods, existing E-VFI methods can estimate a more precise inter-frame motion model with the help of events. However, the cross-modal misalignment in stereo camera setups severely disturbs their performance and causes significant distortions and artifacts during the frame interpolation process as illustrated in Figs. 9 and 10. This is because existing E-VFI algorithms often rely on the assumption of pixel-level alignment between events and frames, which poses challenges in the context of stereo camera setups where capturing data from different modalities and aligning them effectively can be difficult.

In contrast, our proposed SEVFI-Net achieves superior visual performance. Compared to existing methods, our method specifically addresses the challenges posed by the stereo camera setups and leverages the misalignment between different data. Benefiting from the multiple encoders and feature aggregation modules, our SEVFI-Net can process events and frames separately, enabling us to effectively handle biases and



Frame 0

Frame 1

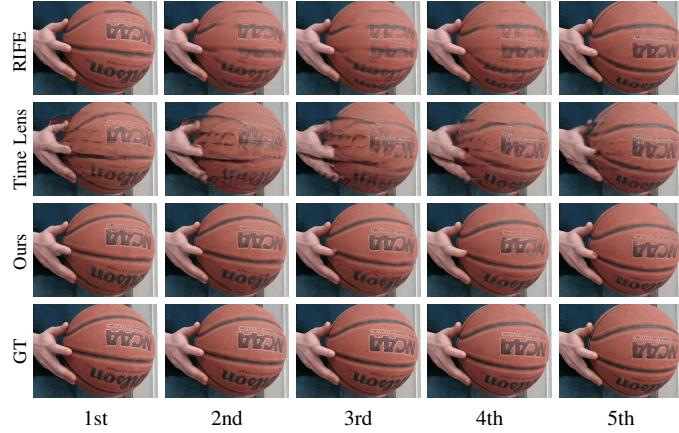


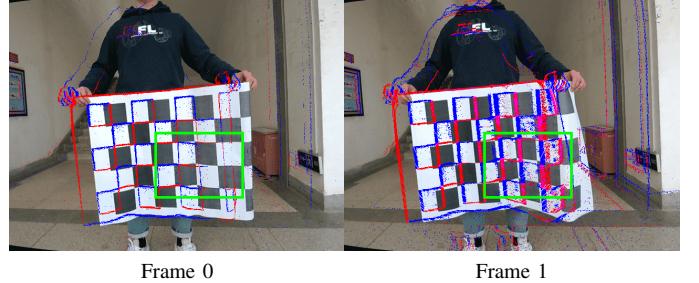
Fig. 11. Comparison of continuous frame reconstruction results by RIFE [13], Time Lens [7] and the proposed SEVFI-Net on SEID. Our SEVFI-Net achieves the best visual quality.

discrepancies between the modalities and achieve data fusion, leading to the best visual performance.

Quantitative comparisons. Quantitative comparisons are given in Tab. III. We first notice that Time Lens, despite its utilization of events to estimate motion information and the integration of synthesis-based and flow-based approaches, exhibits significant performance degradation compared to F-VFI methods when dealing with real-world stereo datasets. This degradation can be attributed to the absence of effective data alignment in Time Lens. We then compare our SEVFI-Net with Time Lens. SEVFI-Net achieves significant improvements over Time Lens, with up to 3.98 dB and 0.0712 enhancements in terms of PSNR and SSIM, respectively. As shown in Figs. 9 and 10, SEVFI-Net showcases superior reconstruction results with fewer artifacts and distortions compared to Time Lens. This demonstrates the effectiveness of our method in mitigating distractions caused by the cross-modal disparity. Moreover, we compare the proposed SEVFI-Net with other F-VFI methods not affected by the parallax. The metrics illustrate that the SEVFI-Net outperforms these state-of-the-art F-VFI methods, which means SEVFI-Net removes the artifacts caused by the cross-modal parallax, achieving precise non-linear motion estimation from misaligned events and frames.

2) Results of Multi-frame Reconstruction: In this part, we compare the performance of our SEVFI-Net with the existing methods in multi-frame reconstruction. We provide qualitative and quantitative results to show the advantages of our method.

SEVFI-Net exhibits unique advantages when it comes to



Frame 0

Frame 1

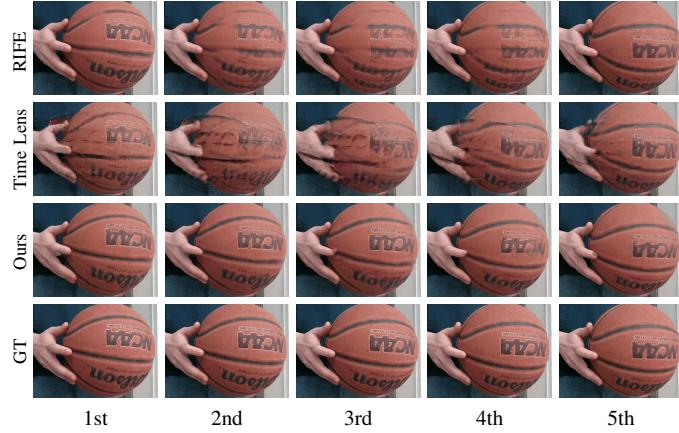


Fig. 12. Comparison of continuous frame reconstruction results by RIFE [13], Time Lens [7] and the proposed SEVFI-Net on SEID. Our SEVFI-Net achieves the best visual quality.

handling non-linear motion, as shown in Figs. 11 and 12, giving the best continuous multi-frame reconstruction and exhibiting excellent ability in reconstructing dynamic scenes, with the motion patterns closely resembling the ground truth. We also conduct experiments of sequence interpolation on MVSEC and our SEID and calculate the average PSNR and SSIM at each frame index separately as shown in Fig. 13, the proposed SEVFI-Net achieves the highest metrics at each frame index. This is due to the high temporal resolution of events enables us to effectively model non-linear and fast-moving motions and generate interpolated frames that are visually appealing while preserving the natural motion characteristics of the scene.

3) Results of Changing Disparity: In addition, we also evaluate the impact of disparity on E-VFI results as illustrated in Fig. 14. We perform experiments on our SEID dataset with the 5-frame skip for each subset, seq1-seq6 in the subfigures represent [“Pedestrians”, “Basketball”, “Cars”, “Square”, “Checkerboard”, “Indoor”] in our SEID respectively. And we calculate the PSNR improvement of our method compared to Time Lens under different disparity conditions, to demonstrate its stability and robustness when dealing with disparity changes. The data misalignment leads to performance degradation for Time Lens, especially when the average disparity value increases. In contrast, our algorithm maintains a relatively stable performance across different disparity values. Therefore, as the average disparity increases, the improvement in PSNR also shows an overall increasing trend, indicating that it is more stable and robust in scenes with disparity variations.

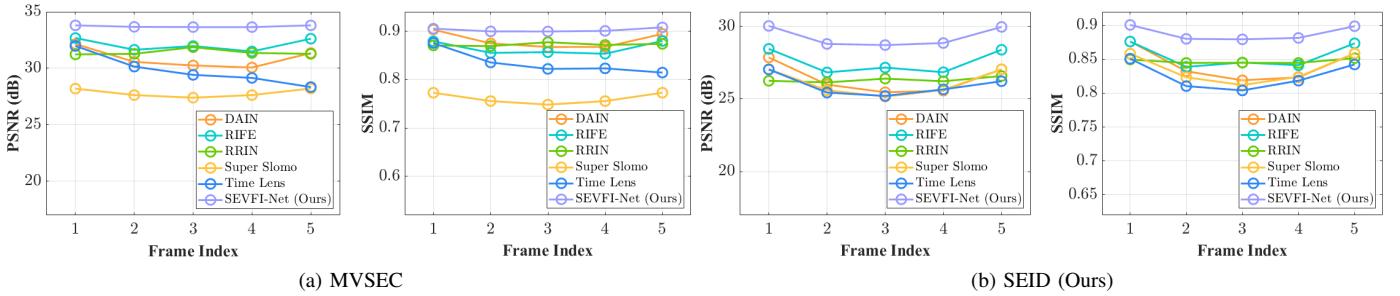


Fig. 13. The ‘‘Rope plot’’ shows the capability of continuous frame reconstruction. We perform frame prediction on the MVSEC [6] and our SEID datasets by generating 5 intermediate frames. We then calculate the average PSNR/SSIM at each frame index separately. The proposed SEVFI-Net achieves the highest metrics at each frame index.

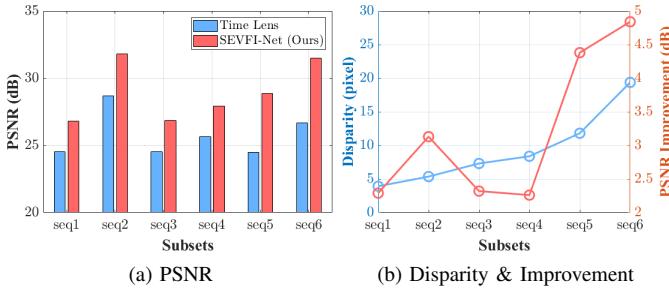


Fig. 14. Illustration of the impact of disparity changes on the E-VFI algorithms. (a) It shows the PSNR of SEVFI-Net and Time Lens [7]. SEVFI-Net outperforms Time Lens over each subset; (b) It shows the average disparity of each subset along with the corresponding PSNR improvement between our SEVFI-Net and Time Lens. As the average disparity increases, the improvement in PSNR also shows an overall increasing trend. This indicates that our method is more robust in scenes with disparity variations.

Overall, our method not only addresses the challenges of the stereo camera setup and misaligned modalities but also achieves the best results in terms of frame interpolation and maintains robustness even in the presence of large depth variations.

C. Comparisons of Stereo Matching

In this subsection, we evaluate the stereo-matching performance of our method with existing cross-modal stereo-matching methods. The qualitative and quantitative experiments demonstrate the superiority of our method.

As formulated in Eq. (4), our SEVFI-Net is capable of not only producing interpolated frames but also estimating the corresponding disparity results. To validate its effectiveness in disparity estimation, we conduct experiments on the DSEC [5] with a 3-frame skip setting and on the MVSEC [6] and SEID with a 5-frame skip setting.

To the best of our knowledge, our SEVFI-Net is the first work that combines stereo matching with video frame interpolation. In our experimental setup, we only have two keyframes and the event stream between them, which cannot be directly applied to existing event-frame stereo-matching algorithms. Therefore, we first choose two representative VFI algorithms, *i.e.*, RIFE [13] and Time Lens [7] to generate intermediate frames and then input them along with the event stream into two existing stereo matching algorithms, *i.e.*, HSM [3] and SSIE [1], to compare their performance with our method.

We utilize the widely-used metrics end-point error (EPE) and 1-pixel, 2-pixel, 3-pixel error, where EPE is the mean disparity error in pixels and the c -pixel error is the average percentage of the pixel whose EPE is larger than c pixels. Compared to the other methods, ours achieves the best visualization performance and gains an 86.6% decrease in EPE and 71.3% decrease in the 3-pixel error as shown in Tab. IV and Fig. 15. This is because the combination of the VFI and stereo-matching methods can be influenced by the quality of frame interpolation. The accuracy and smoothness of the interpolated frames play a crucial role in achieving high-quality results. Additionally, existing cross-modal stereo-matching algorithms are typically designed based on the assumption of static scenes and moving cameras, and may struggle with complex scenes, leading to significant performance degradation.

In contrast, our SEVFI-Net excels at matching events and images effectively, allowing for the recovery of accurate depth information in complex scenes. Importantly, our method overcomes the limitations of traditional methods constrained by frame rates by outputting disparities through interpolation. This approach enables a more flexible and continuous representation of depth information, which greatly enhances the overall performance of the system. The metrics shown in Tab. IV demonstrate that our SEVFI-Net outperforms existing methods by a large margin.

D. Ablation Study

In this section, we investigate the contribution of each subnetwork in our SEVFI-Net to the interpolation results. As illustrated in Tab. V and Fig. 16, we compare the flow-based results, *i.e.*, $W_{0 \rightarrow t}^{\Omega_f}$, $W_{1 \rightarrow t}^{\Omega_f}$, the syn-based result $S_t^{\Omega_f}$, the fused result $\tilde{I}_t^{\Omega_f}$, and the final refined result $\hat{I}_t^{\Omega_f}$ on the DSEC, MVSEC, and SEID datasets.

As displayed in Figs. 16b and 16c, the flow-based results bring artifacts and distortions when the target time is far from input keyframes, while the syn-based results can roughly reconstruct inter-frame motion as shown in Fig. 16d. However, due to the limited information in the warped event data, which only includes the surrounding information of the target frame, it is challenging to effectively incorporate the underlying temporal information and achieve a high level of detail in the reconstruction.

The results in Tab. V demonstrate that the utilization of attention-based fusion maps allows us to harness the ad-

TABLE IV

QUANTITATIVE COMPARISONS OF THE QUALITY OF DISPARITIES ON DSEC [5], MVSEC [6] AND SEID DATASETS. WE COMPUTE AVERAGED END-POINT-ERROR (EPE) AND 1-PIXEL, 2-PIXEL, 3-PIXEL ERROR.

Methods	EPE (px) \downarrow	$> 1\text{px} (\%) \downarrow$	$> 2\text{px} (\%) \downarrow$	$> 3\text{px} (\%) \downarrow$
DSEC				
HSM [3] + RIFE [13]	12.83	77.39	59.80	49.49
HSM [3] + Time Lens [7]	13.94	80.40	63.96	53.77
SSIE [1] + RIFE [13]	16.42	96.79	93.59	90.40
SSIE [1] + Time Lens [7]	16.44	96.79	93.59	90.40
SEVFI-Net (Ours)	4.10	65.27	45.98	35.69
MVSEC				
HSM [3] + RIFE [13]	18.45	74.05	61.35	56.09
HSM [3] + Time Lens [7]	19.54	80.96	68.25	61.63
SSIE [1] + RIFE [13]	13.05	94.59	89.35	84.32
SSIE [1] + Time Lens [7]	13.12	94.63	89.42	84.42
SEVFI-Net (Ours)	2.62	55.38	36.09	24.22
SEID				
HSM [3] + RIFE [13]	21.96	91.80	85.64	79.60
HSM [3] + Time Lens [7]	21.97	91.67	85.52	79.49
SSIE [1] + RIFE [13]	14.00	93.42	87.35	81.62
SSIE [1] + Time Lens [7]	13.94	93.27	87.11	81.37
SEVFI-Net (Ours)	4.32	65.76	52.34	42.47

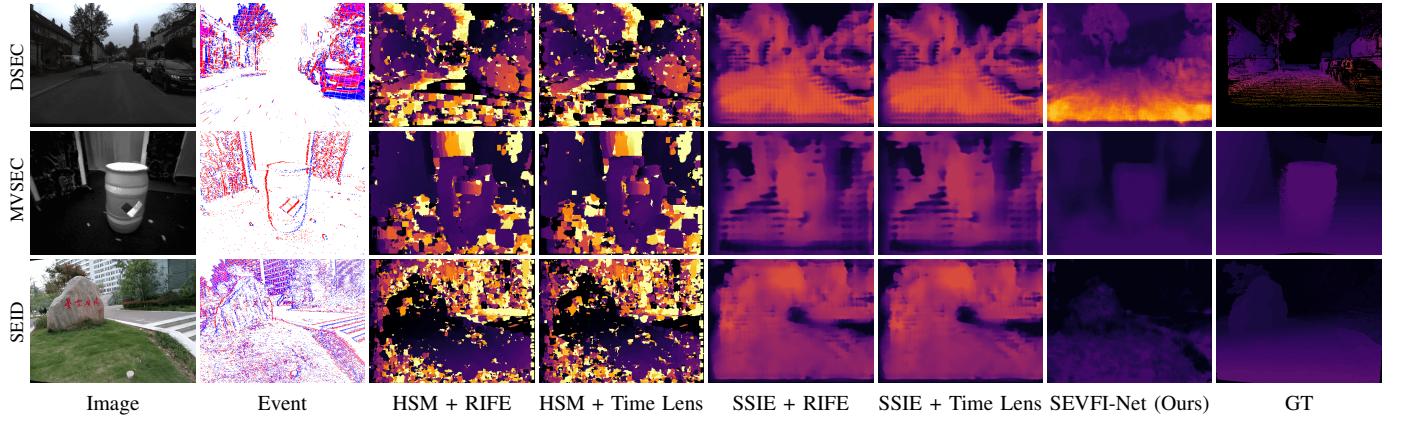


Fig. 15. Illustration of the qualitative results of disparities on DSEC [5] (1st row) MVSEC [6] (2nd row) and SEID (3rd row). The first column is the interpolated frame by our SEVFI-Net, the middle five columns are the disparity results predicted by different methods, and the last column is the ground truth.

vantages of both pathways and enhance the quality of the reconstruction. We observe an average improvement of 2.41 dB in terms of PSNR, and the visual quality is also enhanced as illustrated in Fig. 16e. Furthermore, to improve the fused results by addressing artifacts and inconsistencies, we design RefineNet at the end, which computes the residual maps to capture the differences between the fused result and the ground truth. Tab. V shows an improvement of 0.69 dB in terms of PSNR and Fig. 16f achieves the best visualization quality.

VII. CONCLUSION

In this paper, we tackle the stereo event-based video frame interpolation task with a novel network named SEVFI-Net. Considering the problems of spatial misalignment and modality differences present in stereo camera setups, we design a core FAM, establishing spatial correspondence and achieving data fusion using features extracted from events and frames separately. Additionally, we utilize attention maps to combine

flow-based and syn-based results for better fusion. Our SEVFI-Net can generate high-quality intermediate frames and cross-modal disparities with spatially misaligned events and frames. We also build a stereo visual acquisition system and collect a new SEID containing diverse complex scenes. Extensive experiments demonstrate that the proposed SEVFI-Net establishes state-of-the-art performance on real-world stereo event-intensity datasets.

REFERENCES

- J. Gu, J. Zhou, R. S. W. Chu, Y. Chen, J. Zhang, X. Cheng, S. Zhang, and J. S. Ren, “Self-supervised intensity-event stereo matching,” *arXiv preprint arXiv:2211.00509*, 2022. [1](#), [3](#), [11](#), [12](#)
- Y.-F. Zuo, L. Cui, X. Peng, Y. Xu, S. Gao, X. Wang, and L. Kneip, “Accurate depth estimation from a hybrid event-RGB stereo setup,” in *IEEE Int. Conf. Intell. Robots Syst.*, 2021, pp. 6833–6840. [1](#), [3](#)
- H. Kim, S. Lee, J. Kim, and H. J. Kim, “Real-time hetero-stereo matching for event and frame camera with aligned events using maximum shift distance,” *RAL*, vol. 8, no. 1, pp. 416–423, 2022. [1](#), [3](#), [11](#), [12](#)
- Z. Wang, L. Pan, Y. Ng, Z. Zhuang, and R. Mahony, “Stereo hybrid event-frame (shef) cameras for 3d perception,” in *IROS*. IEEE, 2021, pp. 9758–9764. [1](#), [3](#)

TABLE V

QUANTITATIVE COMPARISONS OF THE OUTPUTS OF EACH SUBNETWORK OF THE PROPOSED SEVFI-NET ON THE DSEC [5], MVSEC [6] AND OUR SEID DATASETS.

Results	DSEC		MVSEC		SEID (Ours)	
	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
Flow-based $W_{0 \rightarrow t}^{\Omega_f}$	28.24	0.8464	32.21	0.8691	27.16	0.8426
Flow-based $W_{1 \rightarrow t}^{\Omega_f}$	27.77	0.8428	31.86	0.8587	26.87	0.8333
Syn-based $S_t^{\Omega_f}$	26.04	0.7988	28.51	0.8008	23.54	0.7345
Fused result $\hat{I}_t^{\Omega_f}$	<u>29.75</u>	<u>0.8722</u>	<u>32.81</u>	<u>0.8834</u>	<u>28.73</u>	<u>0.8711</u>
Refined result $\hat{I}_t^{\Omega_f}$	30.29	0.8834	33.79	0.9053	29.27	0.8881

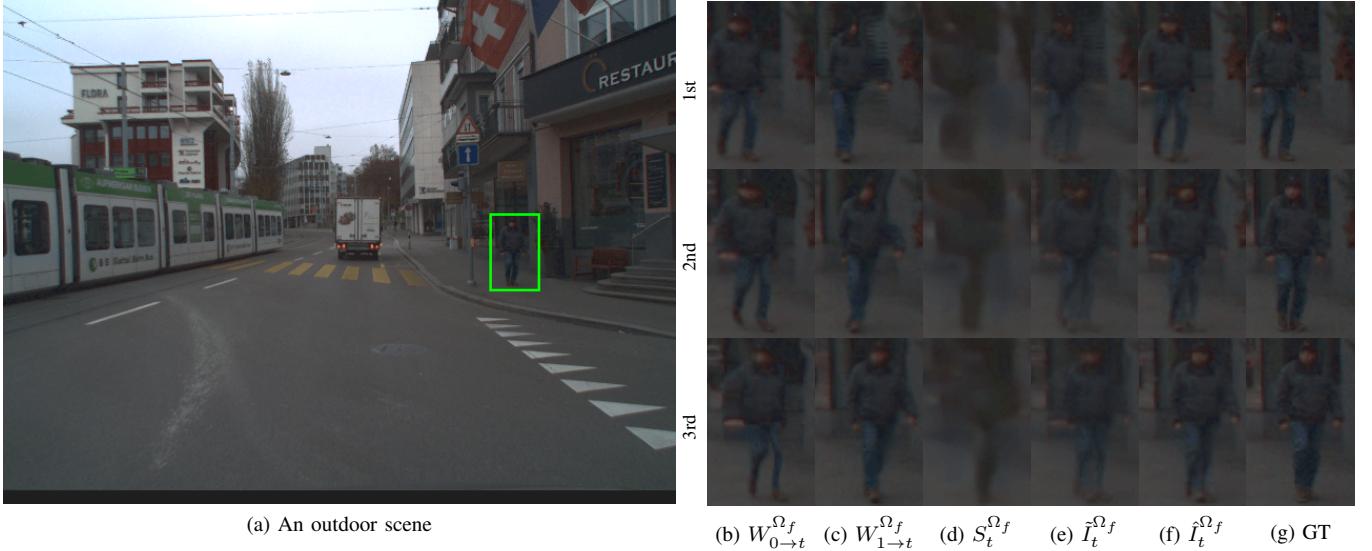


Fig. 16. Qualitative comparisons of the outputs of each subnetwork of the proposed SEVFI-Net on the DSEC dataset.

- [5] M. Gehrig, W. Aarents, D. Gehrig, and D. Scaramuzza, “DSEC: A stereo event camera dataset for driving scenarios,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4947–4954, 2021. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#), [9](#), [11](#), [12](#), [13](#)
- [6] A. Z. Zhu, D. Thakur, T. Özaslan, B. Pfrommer, V. Kumar, and K. Daniilidis, “The multivehicle stereo event camera dataset: An event camera dataset for 3D perception,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2032–2039, 2018. [1](#), [2](#), [6](#), [7](#), [8](#), [9](#), [11](#), [12](#), [13](#)
- [7] S. Tulyakov, D. Gehrig, S. Georgoulis, J. Erbach, M. Gehrig, Y. Li, and D. Scaramuzza, “Time lens: Event-based video frame interpolation,” in *CVPR*, 2021, pp. 16 155–16 164. [1](#), [3](#), [4](#), [9](#), [10](#), [11](#), [12](#)
- [8] 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 *CVPR*, 2022, pp. 17 755–17 764. [1](#), [3](#), [4](#)
- [9] 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 *CVPR*, 2022, pp. 17 804–17 813. [1](#), [3](#), [4](#)
- [10] G. Paikin, Y. Ater, R. Shaul, and E. Soloveichik, “EFI-Net: Video frame interpolation from fusion of events and frames,” in *CVPRW*, 2021, pp. 1291–1301. [1](#), [3](#)
- [11] Y. Gao, S. Li, Y. Li, Y. Guo, and Q. Dai, “SuperFast: 200× video frame interpolation via event camera,” *IEEE TPAMI*, 2022. [1](#), [3](#), [4](#)
- [12] Z. Xiao, W. Weng, Y. Zhang, and Z. Xiong, “EVA²: Event-assisted video frame interpolation via cross-modal alignment and aggregation,” *IEEE Transactions on Computational Imaging*, vol. 8, pp. 1145–1158, 2022. [1](#), [3](#), [4](#)
- [13] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou, “Real-time intermediate flow estimation for video frame interpolation,” in *ECCV*, 2022, pp. 624–642. [1](#), [9](#), [10](#), [11](#), [12](#)
- [14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *ICCV*, 2017, pp. 764–773. [2](#), [5](#)
- [15] S. Niklaus, L. Mai, and F. Liu, “Video frame interpolation via adaptive convolution,” in *CVPR*, 2017, pp. 670–679. [2](#), [4](#)
- [16] H. Li, Y. Yuan, and Q. Wang, “Video frame interpolation via residue refinement,” in *ICASSP*, 2020, pp. 2613–2617. [2](#), [9](#)
- [17] S. Niklaus and F. Liu, “Softmax splatting for video frame interpolation,” in *CVPR*, 2020, pp. 5437–5446. [2](#), [4](#)
- [18] L. Lu, R. Wu, H. Lin, J. Lu, and J. Jia, “Video frame interpolation with transformer,” in *CVPR*, 2022, pp. 3532–3542. [2](#), [4](#)
- [19] S. Niklaus and F. Liu, “Context-aware synthesis for video frame interpolation,” in *CVPR*, 2018, pp. 1701–1710. [2](#), [4](#)
- [20] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang, “Depth-aware video frame interpolation,” in *CVPR*, 2019, pp. 3703–3712. [2](#), [4](#), [9](#)
- [21] Y. Zhang, H. Wang, H. Zhu, and Z. Chen, “Optical flow reusing for high-efficiency space-time video super resolution,” *IEEE TCSVT*, vol. 33, no. 5, pp. 2116–2128, 2022. [2](#)
- [22] H. R. Kaviani and S. Shirani, “Frame rate upconversion using optical flow and patch-based reconstruction,” *IEEE TCSVT*, vol. 26, no. 9, pp. 1581–1594, 2015. [2](#)
- [23] Z. Shi, X. Liu, K. Shi, L. Dai, and J. Chen, “Video frame interpolation via generalized deformable convolution,” *IEEE TMM*, vol. 24, pp. 426–439, 2021. [2](#)
- [24] P. Lei, F. Fang, T. Zeng, and G. Zhang, “Flow guidance deformable compensation network for video frame interpolation,” *IEEE TMM*, 2023. [2](#)
- [25] Z. Chi, R. Mohammadi Nasiri, Z. Liu, J. Lu, J. Tang, and K. N. Plataniotis, “All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling,” in *ECCV*, 2020, pp. 107–123. [2](#)
- [26] S. Y. Chun, T. G. Reese, J. Ouyang, B. Guerin, C. Catana, X. Zhu, N. M. Alpert, and G. El Fakhri, “MRI-based nonrigid motion correction in simultaneous PET/MRI,” *Journal of Nuclear Medicine*, vol. 53, no. 8, pp. 1284–1291, 2012. [2](#)

- [27] X. Xu, L. Siyao, W. Sun, Q. Yin, and M.-H. Yang, "Quadratic video interpolation," *NeurIPS*, vol. 32, 2019. [2](#)
- [28] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *ICCV*, 2017, pp. 261–270. [2](#) [4](#)
- [29] Y. Zhang, D. Zhao, X. Ji, R. Wang, and W. Gao, "A spatio-temporal auto regressive model for frame rate upconversion," *IEEE TCSV*, vol. 19, no. 9, pp. 1289–1301, 2009. [2](#)
- [30] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbrück, "A 240×180 130 db 3 μ s latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014. [3](#)
- [31] 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 TPAMI*, vol. 44, no. 1, pp. 154–180, 2020. [3](#)
- [32] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE TPAMI*, vol. 43, no. 6, pp. 1964–1980, 2019. [3](#) [7](#)
- [33] Y. Zou, Y. Zheng, T. Takatani, and Y. Fu, "Learning to reconstruct high speed and high dynamic range videos from events," in *CVPR*, 2021, pp. 2024–2033. [3](#)
- [34] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *CVPR*, 2019, pp. 3857–3866. [3](#)
- [35] L. Wang, Y.-S. Ho, K.-J. Yoon *et al.*, "Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks," in *CVPR*, 2019, pp. 10081–10090. [3](#)
- [36] M. Mostafavi, L. Wang, and K.-J. Yoon, "Learning to reconstruct hdr images from events, with applications to depth and flow prediction," *IJCV*, vol. 129, pp. 900–920, 2021. [3](#)
- [37] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *CVPR*, 2019, pp. 185–194. [3](#)
- [38] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *ICCV*, 2017, pp. 66–75. [3](#)
- [39] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (pds): Toward applications-friendly deep stereo matching," *NIPS*, vol. 31, 2018. [3](#)
- [40] Y. Guo, Y. Wang, L. Wang, Z. Wang, and C. Cheng, "Cvcnet: Learning cost volume compression for efficient stereo matching," *IEEE TMM*, 2022. [3](#)
- [41] Q. Chen, B. Ge, and J. Quan, "Unambiguous pyramid cost volumes fusion for stereo matching," *IEEE TCSV*, 2023. [3](#)
- [42] H. Xu and J. Zhang, "Aanet: Adaptive aggregation network for efficient stereo matching," in *CVPR*, 2020, pp. 1959–1968. [3](#)
- [43] G. Xu, J. Cheng, P. Guo, and X. Yang, "Attention concatenation volume for accurate and efficient stereo matching," in *CVPR*, 2022, pp. 12981–12990. [3](#)
- [44] Z. Li, X. Liu, N. Drenkow, A. Ding, F. X. Creighton, R. H. Taylor, and M. Unberath, "Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers," in *ICCV*, 2021, pp. 6197–6206. [3](#)
- [45] S. N. Uddin, S. H. Ahmed, and Y. J. Jung, "Unsupervised deep event stereo for depth estimation," *IEEE TCSV*, vol. 32, no. 11, pp. 7489–7504, 2022. [3](#)
- [46] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499. [4](#)
- [47] X. Zhang and L. Yu, "Unifying motion deblurring and frame interpolation with events," in *CVPR*, 2022, pp. 17765–17774. [4](#)
- [48] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *CVPR*, 2019, pp. 989–997. [4](#)
- [49] K. C. Chan, X. Wang, K. Yu, C. Dong, and C. C. Loy, "Understanding deformable alignment in video super-resolution," in *AAAI*, vol. 35, no. 2, 2021, pp. 973–981. [5](#)
- [50] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19. [5](#)
- [51] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *CVPR*, 2018, pp. 2472–2481. [5](#)
- [52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *CVPR*, 2018, pp. 586–595. [6](#)
- [53] A. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in *Proceedings of Robotics: Science and Systems*, 2018. [6](#)
- [54] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017, pp. 270–279. [6](#)
- [55] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion from extremely sparse data: A survey," *IEEE TPAMI*, 2022. [6](#)
- [56] 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," in *NeurIPS*, 2019, pp. 8026–8037. [8](#)
- [57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. [9](#)
- [58] 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 *CVPR*, 2018, pp. 9000–9008. [9](#)
- [59] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004. [9](#)