E2HQV: High-Quality Video Generation from Event Camera via Theory-Inspired Model-Aided Deep Learning

Qiang Qu[1], Yiran Shen[2*], Xiaoming Chen[3*], Yuk Ying Chung[1], Tongliang Liu[1]

[1]School of Computer Science, The University of Sydney.
[2]School of Software, Shandong University.
[3]School of Computer Science and Engineering, Beijing Technology and Business University.
{vincent.qu, vera.chung, tongliang.liu}@sydney.edu.au; yiran.shen@sdu.edu; xiaoming.chen@btbu.edu.cn

Abstract

The bio-inspired event cameras or dynamic vision sensors are capable of asynchronously capturing per-pixel brightness changes (called event-streams) in high temporal resolution and high dynamic range. However, the non-structural spatial-temporal event-streams make it challenging for providing intuitive visualization with rich semantic information for human vision. It calls for events-to-video (E2V) solutions which take event-streams as input and generate high quality video frames for intuitive visualization. However, current solutions are predominantly data-driven without considering the prior knowledge of the underlying statistics relating event-streams and video frames. It highly relies on the non-linearity and generalization capability of the deep neural networks, thus, is struggling on reconstructing detailed textures when the scenes are complex. In this work, we propose **E2HQV**, a novel E2V paradigm designed to produce high-quality video frames from events. This approach leverages a model-aided deep learning framework, underpinned by a theory-inspired E2V model, which is meticulously derived from the fundamental imaging principles of event cameras. To deal with the issue of state-reset in the recurrent components of E2HQV, we also design a temporal shift embedding module to further improve the quality of the video frames. Comprehensive evaluations on the real world event camera datasets validate our approach, with E2HQV, notably outperforming state-of-the-art approaches, e.g., surpassing the second best by over 40% for some evaluation metrics.

Introduction

Inspired by the human visual system, Silicon Retina (Mahowald 1991) has pioneered an approach to perceptual sensing with event cameras or Dynamic Vision Sensors (DVS) (Lichtsteiner, Posch, and Delbruck 2008; Posch, Matolin, and Wohlgenannt 2010; Berner et al. 2013) and gained significant interests from both academia and industry. Unlike traditional cameras, event cameras detect microsecond-level intensity changes, generating an asynchronous stream of 'events', termed as event-stream. Event cameras offer several advantages over conventional CCD/CMOS cameras, including high temporal resolution, high dynamic range of up to 140dB (Lichtsteiner, Posch, and Delbruck 2008), and low resource consumption due to the sparse nature of event-streams. For example, the DVS128 sensor platform consumes 150 times less energy than a conventional CMOS camera (Lichtsteiner, Posch, and Delbruck 2008).

Despite the appealing advantages of event cameras, the non-structural event-streams are not inherently compatible with traditional computer vision methodologies (Scheerlinck et al. 2020) and the visualization is not intuitive for human users to understand. To address the above issue, the research on events-to-video (E2V), which aims to generate video frames from pure event-streams, has been raised to provide convenient and intuitive access to the rich information encapsulated in the sparse and non-structure event-streams. There have been a number of successful approaches for E2V task, such as E2VID (Rebecq et al. 2019), FireNet (Scheerlinck et al. 2020), SPADE-E2VID (Cadena et al. 2021), and ET-Net (Weng, Zhang, and Xiong 2021).

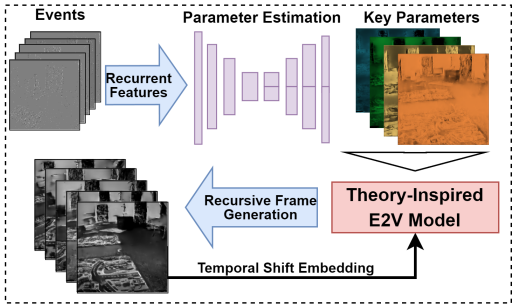

Figure 1: Conceptual Overview of the Proposed Model-Aided Learning Framework.

However, the quality of the video frames generated by the existing E2V approaches is still not satisfactory and fail to recover detailed texture for the complex scenes (Ercan et al. 2023). This issue is predominantly attributed to the fact that many of these approaches, such as E2VID and ET-Net, primarily adopt a purely data-driven approach to learn the mapping from event-streams to video frames directly. However, the purely data-driven approaches are lack of in-

*Corresponding author. The implementation of our work is publicly available at https://github.com/VincentQQu/E2HQV.

terpretability and flexibility (Shlezinger et al. 2023), and they do not take into account the prior knowledge of the underlying statistics relating event-streams and video frames. Therefore, their performance is largely dependent on the non-linearity and generalization capability of the neural networks, which poses significant challenges when the scenes to be reconstructed are complex (Jarrett et al. 2009).

To address the aforementioned challenges, we introduce **E2HQV**, a novel E2V paradigm designed to produce high-quality video frames from events. This is achieved through a model-aided deep learning framework that integrates a theory-inspired E2V model. Rooted in the fundamental imaging principles of event cameras, this theory-inspired E2V model elucidates the relationship between consecutive frames and their associated inter-frame event-streams, offering valuable prior knowledge that enhances the learning efficacy of our deep learning framework. As shown in Figure 1, instead of generating video frames in a pure data-driven approach, E2HQV estimates a number of intermediate key parameters defined by the theory-inspired E2V model then reconstructs the video frames accordingly. The contributions of this work can be summarized as:

- We propose E2HQV, a novel high-quality video frames generation approach from event-streams by facilitating a model-aided learning framework which learns the key parameters defined by a theory-inspired E2V model and generates high quality video frames accordingly.
- According to the imaging principle of event camera and relation between video frames and event-stream, a theory-inspired E2V model is derived to guide the design of the model-aided learning framework.
- A new temporal shift embedding module is designed to deal with the perturbation introduced by the state-reset mechanism of the recurrent components in the framework and ensuring seamless fusion of events and reconstructed frames.
- Through extensive experiments on mainstream event-based video reconstruction datasets, E2HQV consistently exhibits superior performance over state-of-the-art (SOTA) approaches. Remarkably, for certain evaluation metrics, E2HQV surpasses the next best approach by a substantial margin of over 40%.

Related Works

With the advent of deep learning, several methods have been proposed that use neural networks for event-based video reconstruction. Rebecq et al. (Rebecq et al. 2019) proposed a ConvLSTM-based model that leverages the spatiotemporal representation of events for video reconstruction, providing high speed and dynamic range video with an event camera. FireNet (Scheerlinck et al. 2020) offers fast image reconstruction with a lightweight network. Stoffregen et al. (Stoffregen et al. 2020) proposed an augmentation method on simulated training data that improves the performance of E2VID (E2VID+) and FireNet (FireNet+). SPADE-E2VID (Cadena et al. 2021) introduces spatially-adaptive denormalization for event-based video reconstruction. SSL-E2VID (Paredes-Vallés and de Croon 2021) focuses on self-supervised learning of image reconstruction via photometric constancy. Lastly, ET-Net (Weng, Zhang, and Xiong 2021) utilizes a vision transformer for event-based video reconstruction, suffering from computational burden.

In our pursuit of a suitable deep learning framework to serve as the backbone for the proposed parametric frame generator, we conducted a review of several extant neural network-based architectures. Over the past decade, myriad architectures have been developed, including but not limited to ResNet (Szegedy et al. 2017), MobileNet (Howard et al. 2017), SENet (Hu, Shen, and Sun 2018), EfficientNet (Tan and Le 2019), and the more recently introduced EfficientNetV2 (Tan and Le 2021). These models have consistently set performance benchmarks, advancing the field substantially. Among these models, we opted for EfficientNetV2 (Tan and Le 2021), which represents the SOTA in the field, due to its optimal balance between training time and parameter efficiency. Elaborating on our choice, EfficientNetV2 (Tan and Le 2021) has been crafted employing a blend of training-aware neural architecture search and scaling. This results in a model that demonstrates superior training speed and parameter efficiency compared to its predecessors. For the construction of our proposed parametric frame generator's backbone, we incorporated the MBConv layers and Fused MBConv layers from EfficientNetV2.

Methodology

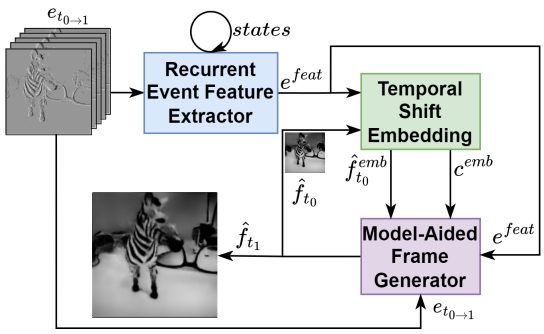

Figure 2: Overview of the Proposed Model-Aided Learning Framework.

Overview of E2HQV

As shown in Figure 2, the pipeline of E2HQV can be vastly categorized into three major components, including a Recurrent Event Feature Extraction (REFE) module, a Temporal Shift Embedding (TSEM) module and a Model-Aided Frame Generation (MAFG) module. The **REFE** module is implemented by leveraging a lightweight ConvLSTM UNet (Shi et al. 2015) and it takes the event-stream $e_{t^i_{0\to1}}$, represented in VoxelGrid (Zhu et al. 2019) following SOTA E2V approaches (Rebecq et al. 2019; Scheerlinck et al. 2020; Weng, Zhang, and Xiong 2021), to extract the features effectively from the spatial-temporal event-stream. The **TSEM** module is proposed to deal with the issue of state-reset in recurrent models. It takes the features extracted

from the REFE module and the last frame generated from the MAFG module as input, and provide a new embedding for the following the MAFG module to deal with imprecise state-reset. The **MAFG** module is the backbone of E2HQV. It exploits the theoretical relation between consecutive frames and event-stream in-between to determine the key parameters needed for E2V generation. Then a model-aid learning framework is proposed to estimate the key parameters and then generate video frames accordingly.

Problem Formulation. Figure 2 also presents the inputs, outputs and intermediate results passed within E2HQV. In the context of E2V generation, a pair of consecutive generated frames can be expressed as $\hat{p}_i = \{\hat{f}_{t_0^i}, \hat{f}_{t_1^i}\}$, corresponding to timestamp t_0^i and t_1^i, respectively. The concurrent event-stream is denoted as $e_{t_{0\to1}^i} = \{e_t | t \in [t_0^i, t_1^i]\}$, are the set of events triggered during the time interval $[t_0^i, t_1^i]$. Provided the last generated frame $\hat{f}_{t_0^i}$ and the following event-stream $e_{t_{0\to1}^i}$, the objective of E2V generation is to learn an optimal model H^* such that,

$$H^* = \arg\min \frac{1}{N} \sum_{i=1}^{N} loss[H\{e_{t_{0\to1}^i}, \hat{f}_{t_0^i}\}, f_{t_1^i}], \tag{1}$$

where N is the total number of frame pairs. Model H takes event-stream $e_{t_{0\to1}^i}$ and the last reconstructed frame $\hat{f_{t_0^i}}$ as input to generate the next frame $\hat{f}_{t_1^i}$. The $loss$ function measures the difference between the reconstructed frame $\hat{f}_{t_1^i}$ and the ground truth $f_{t_1^i}$.

As shown in Figure 2, in each cycle of frame generation, the REFE module takes the event-stream $e_{t_{0\to1}}$ as input to extract recurrent event features, denoted as e^{feat}. Then the event features e^{feat} and the last generated frame $\hat{f}_{t_0}$ are fed to TESM module to produce the embedded frame $\hat{f}_{t_0}^{emb}$ and a new embedding channel c^{emb}. The new channel indicates the relative distance to the last reset frame to aid the following MAFG module for generating high-quality video frames recursively with the integration of different features:

$$\Psi = \{\hat{f}_{t_0}^{emb}, c^{emb}, e_{t_{0\to1}}, e^{feat}\}, \tag{2}$$

Model-Aided Frame Generation

The backbone of our proposed E2HQV is a Model-Aided Frame Generator (MAFG). According to the theory-inspired E2V model derived from the theoretical relation between video frames and event-stream, it is designed as two major components: the key parameters estimation and recursive video generation.

Theory-Inspired E2V Model Intuitively, event-streams signify brightness changes, while video frames directly record brightness levels, suggesting an intrinsic relationship between them. In this section, leveraging the imaging principles of event cameras, we derive a relationship between consecutive frames and their associated event-streams. This prior knowledge can then bolster the effectiveness of our deep learning framework, detailed in the next section.

According to the event generation model(Lichtsteiner, Posch, and Delbruck 2008), an event is triggered when a log-intensity change is detected (over a threshold). The log-intensity $d_i^{x,y}$ at pixel (x,y) can be expressed as,

$$d_i^{x,y} = logI_{i+1}^{x,y} - logI_i^{x,y}. \tag{3}$$

where $I_i^{x,y}$ and $I_{i+1}^{x,y}$ are the intensity values of pixel (x, y) when two consecutive events are generated. The accumulated log-intensity change within $[t_0, t_1]$ (time interval between two consecutive video frames f_{t_0} and f_{t_1}) can be approximated as,

$$\begin{aligned}\sum_{i=0}^{n_{x,y}-1} d_i^{x,y} &= logI_n^{x,y} - logI_{n-1}^{x,y} + logI_{n-1}^{x,y} \\ &\quad ... - logI_1^{x,y} + logI_1^{x,y} - logI_0^{x,y}.\end{aligned} \tag{4}$$

By cancelling out the same items, the equation above can be rewritten as,

$$\sum_{n=0}^{n_{x,y}-1} d_i^{x,y} = logI_n^{x,y} - logI_0^{x,y}, \tag{5}$$

Then $d_i^{x,y}$ can be decomposed as,

$$d_i^{x,y} = \begin{cases} +\theta_{+i}^{x,y} & \text{if } d_i^{x,y} \geq 0 \\ -\theta_{-i}^{x,y} & \text{if } d_i^{x,y} < 0 \end{cases}, \tag{6}$$

where $\theta_{(+/-)i}^{x,y}$ is the polarity-wise absolute value of the threshold to trigger an event at pixel (x, y). As the period of $[t_0, t_1]$ is typically only tens of milliseconds, we assume $\theta_{(+/-)}^{x,y}$ is a constant threshold for pixel $p_{x,y}$ during the specific period. The positive events (denoted as "+") and negative events (denoted as "−") are counted over $[t_0, t_1]$ at pixel (x, y), i.e.,

$$\mathcal{E}_{(+/-)}(x, y) = \sum_{i=0}^{n_{x,y}-1} (+/-)_i^{x,y}. \tag{7}$$

Combing Eq. (6) and Eq. (7), the accumulative log-intensity change within $[t_0, t_1]$ can be expressed as,

$$\begin{aligned}\sum_{i=0}^{n_{x,y}-1} d_i^{x,y} &= \theta_{+i}^{x,y} \sum_{i=0}^{n_{x,y}-1} +_i^{x,y} - \theta_{-i}^{x,y} \sum_{i=0}^{n_{x,y}-1} -_i^{x,y} \\ &= \theta_+^{x,y}\mathcal{E}_+^{x,y} - \theta_-^{x,y}\mathcal{E}_-^{x,y}.\end{aligned} \tag{8}$$

By incorporating Eq. (5) into Eq. (8), we have

$$\theta_+^{x,y}\mathcal{E}_+^{x,y} - \theta_-^{x,y}\mathcal{E}_-^{x,y} = log\frac{I_n^{x,y}}{I_0^{x,y}}. \tag{9}$$

Then we assume a linear relationship between intensity I and a corresponding normalized frame f, i.e.,

$$I = af + b, \tag{10}$$

where $a > 0$ and the pixel values of $f(x, y) \in [0, 1]$. Then,

$$\frac{I_n^{x,y}}{I_0^{x,y}} = \frac{af_{t_1}^{x,y} + b}{af_{t_0}^{x,y} + b} = \frac{f_{t_1}^{x,y} + k}{f_{t_0}^{x,y} + k}, \tag{11}$$

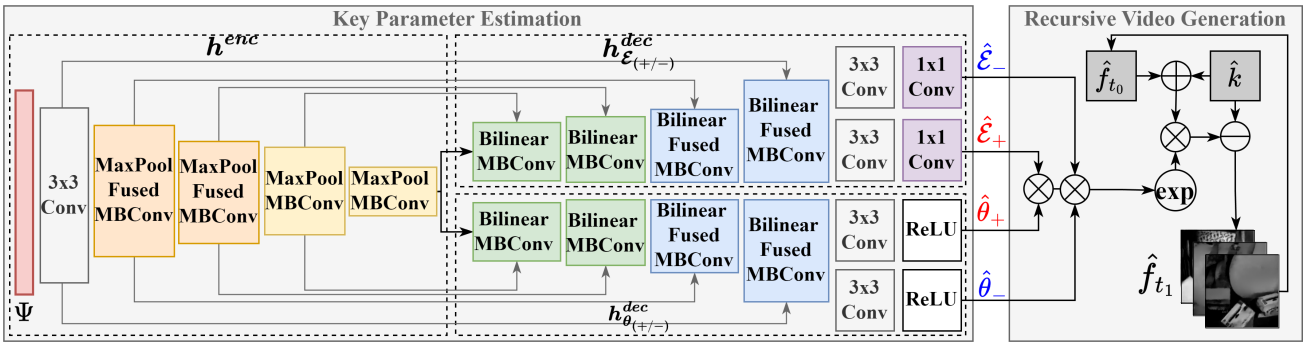

Figure 3: The detailed settings of the model-aided frame generator (MAFG). The generator accepts multimodal features Ψ integrated from the output of the REFE module and the TSEM module, which are then input into a shared downsampling encoder, h^{enc}. This is followed by two upsampling decoders, $h^{dec}_{\mathcal{E}_{(+/-)}}$ and $h^{dec}_{\theta_{(+/-)}}$, and four output branches. These branches are meticulously designed for the estimation of the key parameters, $\mathcal{E}_{(+/-)}$ and $\theta_{(+/-)}$, respectively. Provided the estimated parameters, video frames are recursively generated from events according to Equation (12) derived from the theoretical relation between frame and event-stream.

where $k = b/a$, f_{t_0} and f_{t_1} are normalized frames at timestamps t_0 and t_1. Finally, based on Eq. (9) and Eq. (11), the APS frame f_{t_1} can be represented as,

$$f^{x,y}_{t_1} = exp(\theta^{x,y}_{+}\mathcal{E}^{x,y}_{+} - \theta^{x,y}_{-}\mathcal{E}^{x,y}_{-})(f^{x,y}_{t_0} + k) - k. \quad (12)$$

The equation above is our **Theory-Inspired E2V Model** derived from imaging principal of event camera, and the relationship between frames and events. According to the model, the video frames can be generated recursively provided the key parameters $\{\theta_{(+/-)}, \mathrm{k}\}$ and the counts of positive and negative events $\mathcal{E}_{(+/-)}$.

However, because the thresholds of event camera are unknown and changing overtime, it is challenging to obtain the thresholds accurately. Then due to the noisy nature of event camera (Wu et al. 2020; Baldwin et al. 2020; Guo and Delbruck 2022), $\mathcal{E}_{(+/-)}$ obtained from simply counting the number of events overtime may introduce significant interference to the theory-inspired E2V model.

Detailed Design of the MAFG To address the challenges above, we design a Model-Aided Frame Generator (MAFG) to estimate the key parameters and generate high-quality video frames from events according to the theory-inspired E2V model. As shown in Figure 3, the MAFG facilities a deep neural network for key parameters estimation and then, provided estimated parameters and events, generates video frames recursively through simple addition and multiplication operations.

As shown in Fig. 3, the backbone of the MAFG starts with taking the combination of features from other modules (refer to Figure 2) as input to the encoder h^{enc}. The encoder consists of multiple convolutional and (Fused) MBConv (Sandler et al. 2018; Tan and Le 2021) layers followed by MaxPooling for feature extraction. The MBConv utilizes the inverted bottleneck structure (Sandler et al. 2018) and depthwise convolutional layers (Howard et al. 2017) to improve memory efficiency. In addition, a squeeze-and-excitation unit (Hu, Shen, and Sun 2018) is inserted in the middle of MBConv in order to adaptively recalibrate channel-wise feature responses. Then two non-sharing decoders $h^{dec}_{\mathcal{E}_{(+/-)}}$ and $h^{dec}_{\theta_{(+/-)}}$ with the same network architecture (except for the last two layers as shown in Fig. 3) are adopted to decode the low-dimensional features with upsampling layers to obtain the estimates $\hat{\mathcal{E}}_{(+/-)}$ and $\hat{\theta}_{(+/-)}$. The decoder consists of the blocks of a combination of spatial bilinear interpolation and (Fused) MBConv layers. A 1×1 convolutional layer without activation function is applied to finalize $h^{dec}_{\mathcal{E}_{(+/-)}}$ and ReLU activation function is adopted to ensure the non-negative property of $\theta_{(+/-)}$. At last, according to the theory-inspired E2V model (Eq. (12)), the estimate of the frame, $\hat{f}_{t_1}$, can be generated provided the estimated $\{\hat{\mathcal{E}}_{(+/-)}, \hat{\theta}_{(+/-)}\}$ and last generated frame $\hat{f}_{t_0}$. The training loss is the Mean Absolute Error between the generated frames and the ground truths.

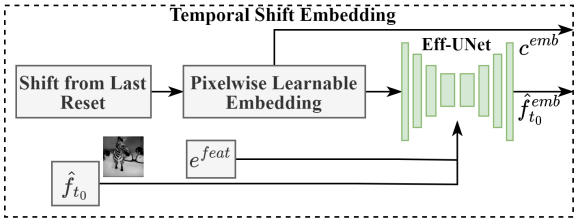

Figure 4: Structure of the Temporal Shift Embedding.

Temporal Shift Embedding Module The theory-inspired E2V model, as depicted in Eq. (12), recursively takes previously synthesized frames and concurrent events to generate succeeding frames. Periodic state-resets, commencing at the outset and recurring after each predefined frame generation interval, are imperative due to issues like interference between unrelated sequences and the inconsistency of the starting state (Chung et al. 2014; Le, Jaitly, and Hinton 2015). This could potentially impact the performance of the

	IJRR			MVSEC			HQF		
	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓
E2VID	0.212	0.424	0.350	0.337	0.206	0.705	0.127	0.540	0.382
FireNet	0.131	0.502	0.320	0.292	0.261	0.700	0.094	0.533	0.441
E2VID+	0.070	0.560	0.236	0.132	0.345	0.514	0.036	0.643	**0.252**
FireNet+	0.063	0.555	0.290	0.218	0.297	0.570	0.040	0.614	0.314
SPADE-E2VID	0.091	0.517	0.337	0.138	0.342	0.589	0.077	0.521	0.502
SSL-E2VID	0.046	0.364	0.425	0.062	0.345	0.593	0.126	0.295	0.498
ET-Net	0.047	0.617	0.224	0.107	0.380	0.489	0.032	0.658	0.260
E2HQV (Ours)	**0.028**	**0.682**	**0.196**	**0.032**	**0.421**	**0.460**	**0.019**	**0.671**	0.261

Table 1: Quantitative comparisons of the evaluated SOTA methods on IJRR, MVSEC, and HQF. The best results are in bold while the second best results are underlined (the same for the rest tables).

proposed method. The MAFG module, devoid of knowledge pertaining to its relative temporal distance from the last reset frame, may lack the capability for adaptive fusion of events with the most recent reconstructed frame.

As shown in overview of the E2HQV framework (Figure 2), to resolve issue of state-reset, we introduce a Temporal Shift Embedding (TSEM) module designed to embed the previous frame and generate a new embedded channel to provide state-reset information for the MFGA module. This allows MFGA module to adjust its reconstruction behavior based on the position of current timestamp within the reset interval. Fig. 4 presents the detailed structure of the TSEM module. This module maintains a sequence of pixelwise learnable embedding tensors, each of which corresponds to a specific shift from the last state-reset within a reset interval. When provided with the shift from the last reset, the TSEM module identifies the relevant pixelwise embedding. This embedding is then integrated with the recurrent event feature e^{feat} and the previously generated frame $\hat{f}_{t_0}$, and fed into a lightweight and efficient UNet (Shi et al. 2015), namely Eff-UNet, to generate the embedded frame $\hat{f}^{emb}_{t_0}$ and produce a new embedding channel c^{emb}. The Eff-UNet features an encoder-decoder structure reminiscent of h^{enc} and $\mathcal{E}_+$ with significantly reduced convolutional filters at each layer, rendering it highly efficient.

Experiments

Training on Simulated Dataset

We utilize the simulated event-based dataset (Rebecq et al. 2019) converted from MS-COCO (Lin et al. 2014) for training which is pervasively adopted in a number of E2V approaches (Rebecq et al. 2019; Scheerlinck et al. 2020; Cadena et al. 2021; Weng, Zhang, and Xiong 2021). The dataset is generated using the ESIM event simulator (Mueggler et al. 2017), which produces synthetic events by rendering images from the MS-COCO dataset along a simulated camera trajectory. The camera sensor size is set to be 240 × 180 pixels, matching the resolution of the DAVIS240C sensor (Berner et al. 2013). Data enrichment is achieved by assigning different positive and negative contrast thresholds to each simulated scene, sampled from a normal distribution. This prevents the network from merely integrating events and ensures better generalization to real event-streams. The dataset comprises 1,000 sequences of 2-seconds event-streams, totaling approximately 35 minutes. For all the experiments, we utilize a desktop computer with an AMD 5950X processor, RTX 3090 GPU, and 32GB RAM, running on Ubuntu 20.04 and implemented using the PyTorch (Paszke et al. 2019).

Evaluation on Real World Datasets

Our approach E2HQV is evaluated on three publicly available datasets consisting of real event-streams collected by event cameras, i.e., IJRR (Rebecq et al. 2019), MVSEC (Zhu et al. 2018), and HQF (Stoffregen et al. 2020) which are commonly used to benchmark the accuracy of E2V approaches (Rebecq et al. 2019; Scheerlinck et al. 2020; Stoffregen et al. 2020; Paredes-Vallés and de Croon 2021; Weng, Zhang, and Xiong 2021). To ensure strict consistency between the timestamps of the reconstruction and ground truth, we utilize the events between two consecutive frames to generate the later frame. Following the most comprehensive work on E2V benchmarking (Ercan et al. 2023), E2HQV is compared with seven SOTA methods: E2VID (Rebecq et al. 2019), FireNet (Scheerlinck et al. 2020), E2VID+ (Stoffregen et al. 2020), FireNet+ (Stoffregen et al. 2020), SPADE-E2VID (Cadena et al. 2021), SSL-E2VID (Paredes-Vallés and de Croon 2021), and ET-Net (Weng, Zhang, and Xiong 2021). All the approaches are solely trained with simulated dataset and tested directly on the three real-world datasets without any further fine-tuning. The accuracy of the generated frames are evaluated by comparing with the ground truths using the following metrics: Mean Squared Error (MSE), Structural Similarity (SSIM) (Wang et al. 2004), and Perceptual Similarity (LPIPS) (Zhang et al. 2018) which are the same as those in the literature.

Quantitative Evaluation The quantitative evaluation results are presented in Table 1. From the results we can find, E2HQV demonstrates superior performance over the seven SOTA methods across all three metrics on the IJRR and MVSEC datasets. It also surpasses all other methods in terms of MSE and SSIM. Specifically, E2HQV reduces the MSE by approximately 40.4%, 70.1%, and 40.6% across all the three datasets, compared with the second best approach, SSL-E2VID. In terms of SSIM, E2HQV significantly outperforms the second best approach, ET-Net, achieving scores of 0.682, 0.421, and 0.671 on the HQF, IJRR, and MVSEC datasets, respectively. For the Learned

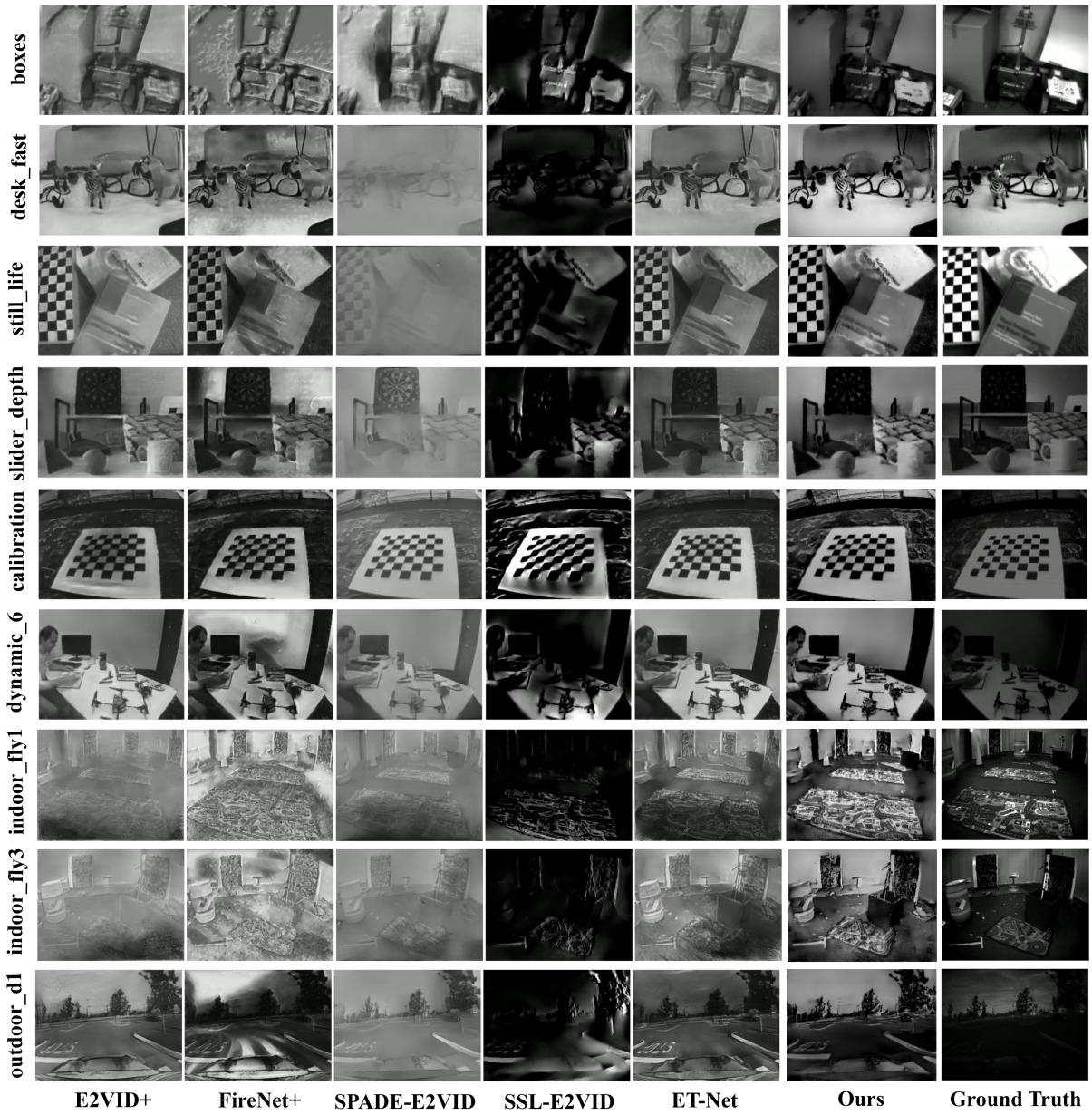

Figure 5: Qualitative Analysis across Datasets. Comparative visualizations of sequence data from HQF (rows 1-3), IJRR (rows 4-6), and MVSEC (rows 7-9). The evaluated baseline methods often exhibit limitations such as diminished contrast, noticeable blur, and prominent artifacts. In contrast, our reconstructions offer high contrast and are adept at maintaining sharp edge details, while manifesting minimal artifacts in regions devoid of texture.

LPIPS metric, E2HQV either exceeds or is close to the best performance of SOTA methods.

Qualitative Evaluation Figure 5 provides qualitative comparison of the generated video frames from E2HQV and other competing methods across the three evaluated datasets. The ground truth is presented in the rightmost column as reference. By comparing the examples from different approaches, we can observe E2HQV shows superior performance on reconstructing complex scenes: it effectively reduces foggy artifacts (as evident in the first, second, third, fourth, and penultimate rows), enhances the details in the context (as seen in the tree details of the last row), and improves contrast (as observed in the third, fifth, penultimate, and last rows).

Ablation Study

Investigating State-Reset Intervals for REFE and TSEM The state-reset interval, as introduced in Section **Temporal Shift Embedding**, serves as a pivotal hyperparameter in

Reset Interval		IJRR			MVSEC			HQF		
REFE	TSEM	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓
20	20	0.032	0.665	0.193	0.040	**0.447**	**0.442**	0.020	0.670	0.273
20	40	0.029	0.656	**0.190**	0.049	0.419	0.463	0.023	0.663	0.277
40	20	**0.028**	**0.682**	0.196	**0.032**	0.421	0.460	**0.019**	**0.671**	**0.261**
40	40	0.041	0.648	0.205	0.043	0.416	0.475	0.021	0.667	0.271

Table 2: Ablation study on different combination of reset intervals.

	IJRR			MVSEC			HQF		
Selected Modules	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓	MSE↓	SSIM↑	LPIPS↓
MAFG	0.080	0.575	0.263	0.062	0.384	0.465	0.033	0.631	0.310
MAFG+REFE	0.051	0.639	0.218	0.064	0.383	0.500	0.027	0.653	0.290
MAFG+TSEM	0.032	0.652	0.202	0.072	0.400	0.478	0.025	0.639	0.309
MAFG+REFE+TSEM	**0.028**	**0.682**	**0.196**	**0.032**	**0.421**	**0.460**	**0.019**	**0.671**	**0.261**

Table 3: Ablation study on different combination of modules.

TSEM. Furthermore, the REFE, which incorporates ConvLSTM units, also relies on the state-reset interval as a hyperparameter. This is due to the fact that these recurrent units are initialized to zero at each state-reset interval. Given the presence of these two hyperparameters, it is of significant interest to conduct an ablation study to explore the impact of different combinations of these parameters to find the optimal configuration. The results, presented in Table 2, show that a combination of 40 and 20 for the REFE and the TSEM modules respectively achieves the best overall accuracy.

Contributions of Different Modules To evaluate the effectiveness of each module of E2HQV, we conduct ablation studies on different combinations of the three modules. The results are reported in Table 3. By comparing different configurations, the highest performance is achieved when all three modules are deployed and each module contributes to the video frames generation with a good margin. Especially, the comparison between the "MAFG" and "MAFG+TSEM" configurations reveals significant performance uplift, underscoring the substantial contributions of the TSEM module to the efficacy of E2HQV.

	Params (M) ↓	GFLOPs ↓
E2VID	10.71	21.2
FireNet	**0.04**	**1.8**
SPADE-E2VID	11.46	71.78
ET-Net	22.18	145.12
E2HQV (Ours)	7.82	18.35

Table 4: Quantitative comparisons of model complexity

Computational Complexity Analysis We also analyse the computational complexity of E2HQV, compared with the competing methods. To quantify the complexity of the approaches, we consider two salient computational metrics delineated in Table 4, which are the number of model parameters and GFLOPs. Despite FireNet exhibiting the lowest model complexity, our proposed model attains a superior trade-off, balancing commendable accuracy against a judicious use of parameters and computational resources (visualized in Fig. 6).

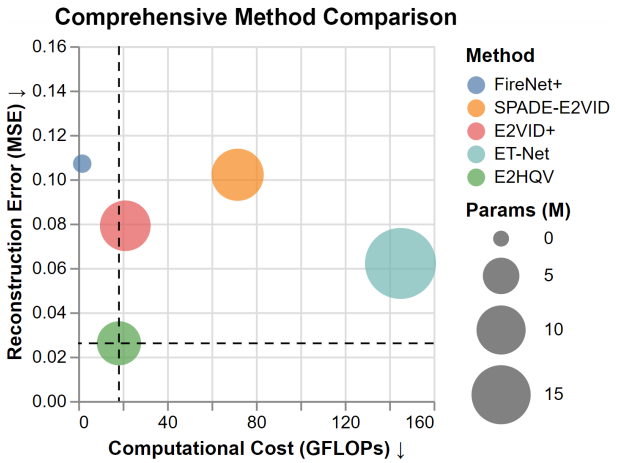

Figure 6: Trade-off Between Accuracy and Complexity. The reconstruction error (MSE) is shown on the y-axis and the computational cost (GFLOPs) is depicted on the x-axis. The number of parameters for each model is visually represented by the size of the corresponding circle.

Conclusion

In this study, we propose E2HQV to improve the quality of generated video frames significantly for E2V task via model-aided learning with a theory-inspired E2V model derived from the imaging principle of event cameras. E2HQV, which integrates theoretical insights with data-driven learning, has shown superior performance over SOTA E2V approaches in extensive evaluations on real world datasets. The introduction of the temporal shift embedding module further enhances the robustness of our approach, ensuring seamless event and frame fusion. The comparison of results of different approaches in evaluation shows E2HQV can generate high quality video frames.

Acknowledgments

This work was supported in part by Beijing Natural Science Foundation (No. 4222003), National Natural Science Foundation of China (No. 62032006 and No. 62177001), and Shandong Provincial Natural Science Foundation (No.2022HWYQ-040).

References

Baldwin, R.; Almatrafi, M.; Asari, V.; and Hirakawa, K. 2020. Event probability mask (epm) and event denoising convolutional neural network (edncnn) for neuromorphic cameras. In *CVPR*, 1701–1710.

Berner, R.; Brandli, C.; Yang, M.; Liu, S.-C.; and Delbruck, T. 2013. A 240× 180 10mw 12us latency sparse-output vision sensor for mobile applications. In *Symposium on VLSI Circuits*, C186–C187. IEEE.

Cadena, P. R. G.; Qian, Y.; Wang, C.; and Yang, M. 2021. Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE TIP*, 30: 2488–2500.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Ercan, B.; Eker, O.; Erdem, A.; and Erdem, E. 2023. EVREAL: Towards a Comprehensive Benchmark and Analysis Suite for Event-based Video Reconstruction. In *CVPR*, 3942–3951.

Guo, S.; and Delbruck, T. 2022. Low Cost and Latency Event Camera Background Activity Denoising. *IEEE TPAMI*.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*, 7132–7141.

Jarrett, K.; Kavukcuoglu, K.; Ranzato, M.; and LeCun, Y. 2009. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, 2146–2153. IEEE.

Le, Q. V.; Jaitly, N.; and Hinton, G. E. 2015. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

Lichtsteiner, P.; Posch, C.; and Delbruck, T. 2008. A 128 ×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. *IEEE JSSC*, 43(2): 566–576.

Lin, T.; Maire, M.; Belongie, S. J.; Bourdev, L. D.; Girshick, R. B.; Hays, J.; Perona, P.; Ramanan, D.; Doll'a r, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312.

Mahowald, M. 1991. MeAd c.(1991). the silicon retina. *Scientific American*, 264(5): 76–83.

Mueggler, E.; Rebecq, H.; Gallego, G.; Delbruck, T.; and Scaramuzza, D. 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. *Int. J. Robot. Res.*, 36(2): 142–149.

Paredes-Vallés, F.; and de Croon, G. C. 2021. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *CVPR*, 3446–3455.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurPS*, 8024–8035. Curran Associates, Inc.

Posch, C.; Matolin, D.; and Wohlgenannt, R. 2010. A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS. *IEEE JSSC*, 46(1): 259–275.

Rebecq, H.; Ranftl, R.; Koltun, V.; and Scaramuzza, D. 2019. High speed and high dynamic range video with an event camera. *IEEE TPAMI*, 43(6): 1964–1980.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 4510–4520.

Scheerlinck, C.; Rebecq, H.; Gehrig, D.; Barnes, N.; Mahony, R.; and Scaramuzza, D. 2020. Fast image reconstruction with an event camera. In *WACV*, 156–163.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28.

Shlezinger, N.; Whang, J.; Eldar, Y. C.; and Dimakis, A. G. 2023. Model-based deep learning. *Proceedings of the IEEE*.

Stoffregen, T.; Scheerlinck, C.; Scaramuzza, D.; Drummond, T.; Barnes, N.; Kleeman, L.; and Mahony, R. 2020. Reducing the sim-to-real gap for event cameras. In *ECCV*, 534–549. Springer.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 6105–6114. PMLR.

Tan, M.; and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. In *ICML*, 10096–10106. PMLR.

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

Weng, W.; Zhang, Y.; and Xiong, Z. 2021. Event-based video reconstruction using transformer. In *ICCV*, 2563–2572.

Wu, J.; Ma, C.; Li, L.; Dong, W.; and Shi, G. 2020. Probabilistic undirected graph based denoising method for dynamic vision sensor. *IEEE TMM*, 23: 1148–1159.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhu, A. Z.; Thakur, D.; Özaslan, T.; Pfrommer, B.; Kumar, V.; and Daniilidis, K. 2018. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robotics and Automation Letters*, 3(3): 2032–2039.

Zhu, A. Z.; Yuan, L.; Chaney, K.; and Daniilidis, K. 2019. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 989–997.