Revisit Event Generation Model: Self-Supervised Learning of Event-to-Video Reconstruction with Implicit Neural Representations

Zipeng Wang[1], Yunfan Lu[1], and Lin Wang*[1,2]

[1] Artificial Intelligence Thrust, HKUST(GZ)
[2] Dept. of Computer Science and Engineering, HKUST
{zwang253,ylu066}@connect.hkust-gz.edu.cn, linwang@ust.hk

Abstract. Reconstructing intensity frames from event data while maintaining high temporal resolution and dynamic range is crucial for bridging the gap between event-based and frame-based computer vision. Previous approaches have depended on supervised learning on synthetic data, which lacks interpretability and risk over-fitting to the setting of the event simulator. Recently, self-supervised learning (SSL) based methods, which primarily utilize per-frame optical flow to estimate intensity via photometric constancy, has been actively investigated. However, they are vulnerable to errors in the case of inaccurate optical flow. This paper proposes a novel SSL event-to-video reconstruction approach, dubbed **EvINR**, which eliminates the need for labeled data or optical flow estimation. Our core idea is to reconstruct intensity frames by directly addressing the event generation model, essentially a partial differential equation (PDE) that describes how events are generated based on the time-varying brightness signals. Specifically, we utilize an implicit neural representation (INR), which takes in spatiotemporal coordinate (x, y, t) and predicts intensity values, to represent the solution of the event generation equation. The INR, parameterized as a fully-connected Multi-layer Perceptron (MLP), can be optimized with its temporal derivatives supervised by events. To make EvINR feasible for online requisites, we propose several acceleration techniques that substantially expedite the training process. Comprehensive experiments demonstrate that our EvINR surpasses previous SSL methods by **38**% *w.r.t.* Mean Squared Error (MSE) and is comparable or superior to SoTA supervised methods. Project page: `https://vlislab22.github.io/EvINR/`.

1 Introduction

Event cameras [13,55] are novel sensors that offer numerous advantages over traditional frame-based cameras, including low power consumption, high dynamic range (HDR), and high temporal resolution [34]. However, their unique imaging paradigm presents a challenge when applying vision algorithms designed for frame-based cameras. To address this challenge and bridge the gap between

* Corresponding author

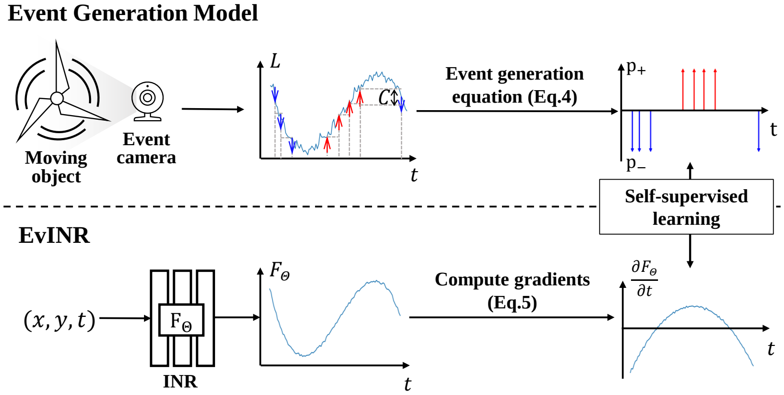


Fig. 1: Connection between event generation model and EvINR: Event generation model reveals the relation between discrete events and continuous temporal intensity changes, described as the event generation equation (Eq. 4). EvINR utilizes an INR to solve Eq. 4 and recovering a continuous function of intensity w.r.t. time, implicitly parameterized with a fully connected MLP.

event-based and standard computer vision [10,15,17,18,36], many methods have been proposed to reconstruct intensity frames from events.

Early methods primarily rely on hand-crafted integrators or filters, resulting in significant loss of detail in reconstructed results [3,22,28,40,59]. More recently, deep learning-based methods [6,35,41,49,53,64] have emerged, typically supervised by large-scale synthetic datasets generated using an event simulator [14,19]. However, the interpretability of these methods is limited due to the 'black-box' nature of the neural networks. Furthermore, the performance of these methods is constrained by the domain gap between synthetic and real-world data [49], leading to suboptimal results if the event simulator's settings do not accurately match those of the inference data. To address these issues, some researchers have explored self-supervised learning (SSL) frameworks [31,60], aiming to eliminate the dependency on labeled synthetic data. Nonetheless, these approaches still depend on event-based optical flow estimation, which is prone to over-fitting, occlusions, and non-convergence issues [44,45]. As a result, the reconstructed frames often suffer from the loss of textural details and various artifacts.

To address these challenges, we revisit the event generation model [13], which forms a fundamental link between events and intensity. The event generation model can be expressed as a partial differential equation (PDE), describing how events are triggered by logarithmic intensity changes that exceed a certain threshold. This PDE, known as the event generation equation, establishes a direct connection between discrete events and the partial derivative of the continuous intensity function w.r.t. time. Our key insight is that, *solving the event generation equation offers an ideal self-supervised solution for event-to-video reconstruction, thereby eliminating the need for synthetic data or optical* *flow estimation*. However, several properties of event data make solving the equation nontrivial: **1**) events can be triggered at extremely high frequencies (up to $10^8/s$), which results in a large volume of data and a heavy computational burden. **2**) events are inherently noisy, particularly in extreme visual conditions [51], which pose challenge to the robustness of solvers. **3**) events do not capture the initial intensity, which makes determining the boundary values challenging.

Recently, implicit neural representations (INRs) have gained popularity for solving the inverse problems in 3D reconstruction or image super-resolution [8, 26, 46] by parameterizing complex signals via deep neural networks. In this work, we find that INRs possess several key advantages that render them particularly suitable for solving the event generation equation: ***they inherently accommodate a large volume of event data, exhibit high noise tolerance, and are flexible to add additional loss terms to regulate initial value***.

In light of this, we propose a novel SSL framework, termed **EvINR**, that employs an INR to represent the solution of the event generation equation. Our EvINR can be directly optimized by minimizing the residual between its temporal derivatives and local intensity changes estimated from event data based on the event generation model (Sec. 3.2). Moreover, we incorporate a spatial regularization term that regularizes the relative values of adjacent pixels by constraining the magnitude of spatial gradients, which effectively reduces noise in the reconstruction process (Sec. 3.2). Although the basic implementation of EvINR yields acceptable results, its convergence on seconds-long event sequences takes minutes, limiting its real-world applicability, especially in online scenarios. To expedite EvINR's training process, we introduce several acceleration techniques, including frame-based optimization, coarse-to-fine training, and model ensembling. These approaches reduce the training time from minutes to seconds while not compromising the reconstruction quality (Sec. 3.4).

Moreover, most approaches are typically evaluated on event datasets captured by DAVIS sensors [27, 42, 49, 61], making it difficult to evaluate their stability and robustness to other types of event camera [1, 2, 39]. For this reason, we collected a new event dataset using an ALPIX-Eiger event camera [1], with well-aligned events and intensity frames.

In summary, our paper makes three key contributions: **(I)** We propose EvINR, a concise SSL framework that solves the event generation equation via implicit neural representations for event-to-video reconstruction. **(II)** Our EvINR substantially outperforms previous SSL methods [31, 40, 59] and attains comparable, or even superior, performance compared to state-of-the-art supervised methods [35, 49, 53]. **(III)** We collect a real-world dataset with an ALPIX-Eiger event camera, complementary to the datasets captured by DAVIS cameras.

2 Related Works

Implicit Neural Representations (INRs) have emerged as a powerful tool for parameterizing signals, such as images, videos, and audio, in a continuous manner using neural networks [46]. Compared to traditional discrete sig-

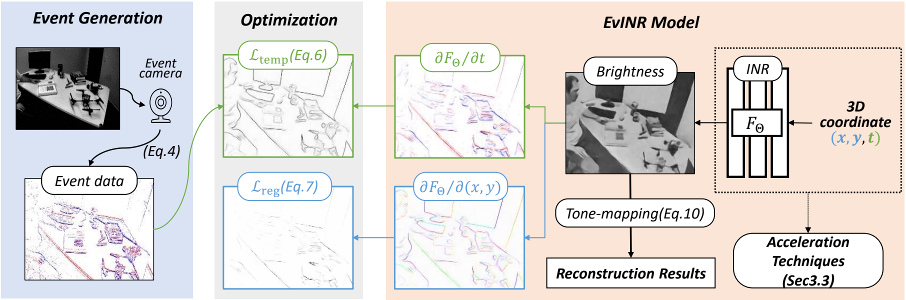

Event Generation
Event camera
(Eq.4)
Event data
Optimization
$\mathcal{L}_{\text{temp}}$(Eq.6)
$\mathcal{L}_{\text{reg}}$(Eq.7)
EvINR Model
$\partial F_\Theta/\partial t$
$\partial F_\Theta/\partial(x,y)$
Brightness
INR
F_Θ
3D coordinate
(x,y,t)
Tone-mapping(Eq.10)
Reconstruction Results
Acceleration Techniques (Sec3.3)

Fig. 2: Overview of EvINR. A fully connected MLP is used to implicitly solve the event generation equation. The temporal gradient of the MLP is supervised by temporal intensity changes of events, and the spatial gradient is penalized to reduce noise.

nal representations, INRs offer the advantage of being able to be sampled at arbitrary resolutions with fixed memory requirements. As a result, INRs have found widespread applications in various fields, including 3D scene representation [25, 26, 32], video representation [7], generative models [29, 43, 47], and model compression [50]. A unique property of INRs is that they can be effectively learned from the derivatives of signals, such as the normals of 3D shapes [30,46,58] and the gradients of images [46]. Such a characteristic has motivated us to approach the task of event-to-video reconstruction by optimizing the INR of video from its temporal derivatives. Previous research has investigated the potential of INR for novel view synthesis using event data [20,24,38]. These approaches reconstruct 3D neural radiance fields using multiple event sequences with known camera poses from a stationary scene. Most similar to our work, E-CIR [48] uses polynomial to regress the intensity function, which may only represent a short time interval (e.g., the exposure time of an image). In contrast, our method employs INR, which can represent longer time of intensity change.

Event-based Video Reconstruction has been a hot topic in the literature. Early attempts [3, 9, 21] address this problem based on the photometric constancy, which describes the relationship between intensity gradients and optical flow. Other approaches [28, 40] are based on direct event integration without estimating optical flow. Rebecq *et al.* [35] developed the first DL-based framework, called E2VID, to reconstruct intensity frames from events in an end-to-end manner, outperforming earlier techniques by a significant margin. E2VID was updated by some following research, trying to tackle problems of inference speed [41], the cold start problem [6], and training strategy [49]. These methods are supervised and learned using the synthetic dataset obtained via the event simulators [14,19] due to the lack of real-world datasets with well-aligned events and intensity frames. Therefore, the generalization capability of these methods is limited by the simulator-to-real gap [49].

To address this problem, Federico *et al.* [31] proposed an SSL framework with a network to estimate optical flow and another network to reconstruct intensity frames based on the photometric constancy. Zhang *et al.* [60] updated this idea into a linear inverse problem that can be solved using modern linear solvers without using deep learning. However, these methods either assume optical flow is known or estimate the optical flow using contrast maximization [62, 63]. Consequently, their performance cannot be guaranteed unless high-quality optical flow can be obtained. *In contrast, our approach is based solely on the physical event generation model, leading to a more straightforward solution that also demonstrates significantly improved performance and greater flexibility.*

3 Method

Overview: Our objective is to reconstruct intensity frames from events in a self-supervised manner, without the need for end-to-end training or optical flow estimation. To accomplish this, we have reformulated event-based video reconstruction as solving the event generation equation. We employ an INR that is supervised solely by event data to represent the intensity function. An overview of our approach can be seen in Fig.2. In Sec. 3.1, we explain the event generation model and its connection to event-based video reconstruction. In Sec. 3.2, we detail how to train an EvINR by supervising its spatial and temporal gradients. Techniques to speed up the training of EvINR for online applications are discussed in Sec. 3.3. Finally, in Sec. 3.4, we describe our collected dataset using the ALPIX event camera [1].

3.1 Preliminary: Event Generation Model

We begin by providing a brief overview of the event generation model [13], which forms the theoretical foundation for our approach. Let $I(x, y, t)$ denote the intensity of the spatial location (x, y) at time t in a video. Since event cameras operate with logarithmic intensity, we denote $L(x, y, t) = \log I(x, y, t)$.

An event camera comprises a frame of independent pixels that respond to changes in the logarithmic intensity signal and produce sequences of sparse and asynchronous events. An event (x_i, y_i, t_i, p_i) is triggered when the logarithmic intensity change surpasses a threshold C since the previous event was triggered at the same pixel. where (x, y) is the spatial location of the pixel, t is the timestamp, and $p \in \{-1, 1\}$ is the polarity of the logarithmic intensity change.

For simplicity, let us consider the temporal changes in logarithmic intensity of a single pixel with fixed spatial coordinates and disregard the spatial terms (x, y). We can describe an event e_i on that pixel using the Dirac delta function as follows:

$$e_i(t) = p_i \cdot C \cdot \delta(t - t_i). \tag{1}$$

Therefore, the logarithmic intensity increment $\mathit{\Delta} L = L(t_2) - L(t_1)$ in a time interval $[t_1, t_2]$ can be represented by the accumulation of events, which can be expressed as:

$$\mathit{\Delta} L = \int_{t_1}^{t_2} \sum_i e_i(t) \mathrm{d}t. \tag{2}$$

Assuming a short time interval and ignoring noise, we can approximate the logarithmic intensity increment by its first-order temporal derivative using Taylor expansion:

$$\frac{\mathit{\Delta} L}{t_2 - t_1} = \frac{\partial L((t_1 + t_2)/2)}{\partial t}. \tag{3}$$

By substituting Eq. 2 into Eq. 3, we derive **the event generation equation** (Eq. 4), which bridges the discrete event data with the continuous temporal derivatives of logarithmic intensity:

$$\frac{\partial L((t_1 + t_2)/2)}{\partial t} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \sum_i e_i(t) \mathrm{d}t. \tag{4}$$

The left-hand side of Eq. 4 represents the partial derivative of logarithmic intensity with respect to time, while the right-hand side can be calculated by accumulating events. Intuitively, event-based video reconstruction can be formulated as solving Eq. 4, *i.e.*, finding a logarithmic intensity function that satisfies the equation. However, as stated earlier, solving Eq. 4 can be non-trivial due to the large amount, high noise rate of event data, and unknown boundary intensity values. To address these challenges, we employ INRs to solve Eq. 4 as they naturally scale up to large data and have high noise tolerance [20]. Boundary intensity values can also easily be regularized by injecting natural image priors as additional loss functions.

3.2 Learning INRs from Events

We aim to learn an INR $F_{\mathit{\Theta}}$ given an event stream $\{e_i\}_{i=0}^N$. Here, $\mathit{\Theta}$ represents the parameters of a fully connected MLP that predicts the logarithmic intensity value $\hat{L}$ at any spatiotemporal coordinate (x, y, t). We adapt the event frame representation [35] and stack the given events uniformly into T event frames.

Temporal Supervision The INR is optimized by directly minimizing the temporal loss between the predicted logarithmic intensity change $\mathit{\Delta}\hat{L}$ and the logarithmic intensity change estimated by event accumulation $\mathit{\Delta} L$. Here $\mathit{\Delta}\hat{L}$ can be derived by accumulating events as Eq. 2 and $\mathit{\Delta}\hat{L}$ is the change of $\hat{L}$ with respect to t obtained by performing double back-propagation [33] of $F_{\mathit{\Theta}}$:

$$\mathit{\Delta}\hat{L} = \frac{\partial F_{\mathit{\Theta}}((t_1 + t_2)/2)}{\partial t} \cdot (t_2 - t_1), \tag{5}$$

Here we adopt the mean squared error (MSE) loss for temporal supervision:

$$\mathcal{L}_{\text{temp}} = (\mathit{\Delta} L - \mathit{\Delta}\hat{L})^2, \tag{6}$$

where $\mathit{\Delta} L$ and $\mathit{\Delta}\hat{L}$ are given by Eq. 2 and Eq. 5, respectively.

Spatial Regularization Although the temporal supervision can estimate the intensity function, the results may still contain unnatural artifacts as the INR network F_Θ has no prior knowledge of the initial intensity values of each pixel. To address this issue, we introduce a spatial regularization term that encourages the solution to be in the space of natural-looking images. We adopt Tikhonov regularization [4] that penalizes the spatial gradients of logarithmic intensity:

$$\mathcal{L}_{\text{reg}} = (\frac{\partial F_\Theta}{\partial x})^2 + (\frac{\partial F_\Theta}{\partial y})^2. \tag{7}$$

It is worth noting that we do not employ more complex regularization methods, such as CNN denoisers [56] as used in [60], because they typically utilize large network models that can significantly slow down the training process.

Optimizing EvINR The overall objective is given by the equation:

$$\mathcal{L} = \mathcal{L}_{\text{temp}} + \lambda \mathcal{L}_{\text{reg}}, \tag{8}$$

where $\mathcal{L}_{\text{temp}}$ and $\mathcal{L}_{\text{reg}}$ were introduced in Sec. 3.2 and Sec. 3.2, respectively, as the temporal supervision and spatial regularization term. λ is a hyper-parameter used to adjust the weight of the spatial regularization term.

Tone-mapping The output of the EvINR is the logarithmic intensity of the reconstructed frames. We first use the exponential function to convert the predicted logarithmic intensity values to high dynamic range (HDR) intensity values $I \in [0, \infty)$:

$$I(x, y, t) = \exp(F_\Theta(x, y, t)). \tag{9}$$

Then, we adopt the Reinhard function [37] to map the HDR intensity values into the low dynamic range $[0, 1]$:

$$\Gamma(I) = (\frac{I}{I+1})^\gamma. \tag{10}$$

where γ is a hyper-parameter used to control the contrast of $\Gamma(I)$. We assess our reconstruction performance by applying $\Gamma(I)$ in all experiments.

3.3 Accelerating EvINR

While solving the event generation equation using the basic implementation of EvINR, as described in Sec. 3.3, yields satisfactory reconstruction results, it is not efficient enough for online tasks. Optimizing one INR network takes about 20 minutes, and a network can only represent approximately 1 second of a sequence. Reducing the training time or increasing the sequence time leads to severe performance degradation. To enable online usage for EvINR, we propose several techniques to reduce the training time and increase the representation capacity for EvINR, as illustrated in Fig. 3.

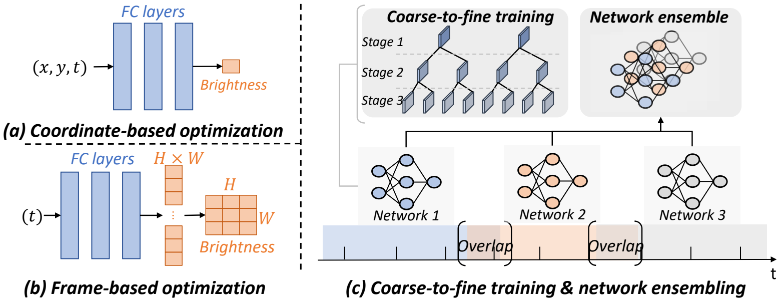


Fig. 3: Overview of acceleration techniques. (a) and (b) illustrate the difference between the basic coordinate-based and our frame-based optimization, respectively. (c) depicts our proposed coarse-to-fine training scheme and network ensembling technique.

Coordinate-based to frame-based optimization: The basic EvINR algorithm is designed to learn a mapping from 3D coordinates (x, y, t) to the logarithmic intensity L. However, this coordinate-based optimization requires the network to 'remember' $H \times W \times T$ coordinate-to-logarithmic intensity mappings, which can be highly complex and lead to slow convergence. To address this issue, we propose a frame-based optimization scheme that learns a mapping from timestamps t to the logarithmic intensity of all pixels at that timestamp. The frame-based optimization scheme can be represented as:

$$F_\Theta(t) = [L(i, j, t)]_{1 \le i \le H, 1 \le j \le W}, \tag{11}$$

where the right hand-side represent a matrix that contains all logarithmic intensity of all pixels at t. The frame-based optimization approach requires the network to remember only T coordinate-to-logarithmic intensity mappings, resulting in a significant reduction in complexity and faster convergence. Empirical results show that this approach reduces the convergence time by two orders of magnitude, from minutes to seconds. The comparison between coordinate-based and frame-based optimizations is illustrated in Fig. 3 (a) and (b).

Coarse-to-fine training: We adopt a coarse-to-fine training scheme, which enables the network to learn overall logarithmic intensity changes before focusing on finer details. We structure the training process into s distinct stages, at each of which we increase the temporal resolution by a factor of 2. Initially, we divide the event sequence into N segments of equal length and optimize the EvINR in accordance with Eq. 6 over a specified number of iterations. Subsequently, we proceed to bisect each of these N segments into two smaller ones, ensuring that they contain an equal number of events. For all the experiments, we set the number of stages s to 3 and scheduled the upsampling to occur after 100 and 200 iterations, respectively. This approach accelerates training by reducing the number of event frames needed in the early stages by approximately 2 times.

Network ensembling: As each EvINR network requires only about 1GB of GPU memory, we further leverage network ensemble techniques [16] to train N

EvINR networks simultaneously on a single GPU to achieve higher parallelism. This approach enables us to exploit the computational resources of modern GPUs more effectively and speed up the training process. We keep overlap periods between nearby networks to keep logarithmic intensity predictions constant among all networks. Fig. 3(c) depicts the coarse-to-fine training process and network ensembling techniques employed in our approach.

3.4 Dataset Collection

Most event-to-video reconstruction approaches are typically evaluated on event datasets collected using DAVIS sensors [5], such as IJRR [27], HQF [49], MVSEC [61], and CED [42]. Recently, other types of event cameras [1, 2, 39] have been developed, which share the same event generation model with DAVIS sensors but may differ in detailed configurations and settings. Therefore, it is essential to verify the generalization capability of existing methods and our EvINR on those new types of event sensors.

To fill the gap, we introduce a new real-world dataset, called the **A**LPIX **E**vent **D**ataset (AED), which is collected using an ALPIX-Eiger event camera [1], featuring static scenes accompanied by gradual camera motions. The camera provides well-aligned RGB frames and color events. The RGB frames have a resolution of 3264×2448 and the events have a resolution of 1632×1224. The AED dataset includes seven video sequences with diverse scenes, such as streets, buildings, indoor scenes, textures, and tools, and each approximately lasts for ten seconds. Note that, in this paper, we only focus on reconstructing grayscale frames to keep consistency with previous works [31, 60] and also for a fair comparison. Therefore, we first demosaic the RGB frames and events according to the Quad Bayer pattern, resulting in intensity frames with a resolution of 816×612 and event data with a resolution of 408×306. Details on the post-processing of AED dataset can be found in the supplementary material.

4 Experiments

4.1 Experiments Settings

Datasets: We conduct experiments to evaluate the effectiveness of our proposed method using three datasets: IJRR [27], HQF [49], and our AED dataset. The IJRR dataset consists of intensity frames and events in 25 real scenes and 2 synthetic scenes, captured by a DAVIS240C camera [5]. HQF dataset provides 14 event data sequences captured with two DAVIS240C cameras, delivering well-exposed and clear intensity frames. The spatial resolution of both the IJRR and HQF datasets is 240×180. The AED dataset contains 7 event sequences with a resolution of 408×306. More details of the exact time split can be found in the supplementary material.

Evaluation Metrics: We evaluate the efficacy of our method using several image quality metrics, including mean squared error (MSE), structural similarity (SSIM) [52], and learned perceptual image patch similarity (LPIPS) [57].

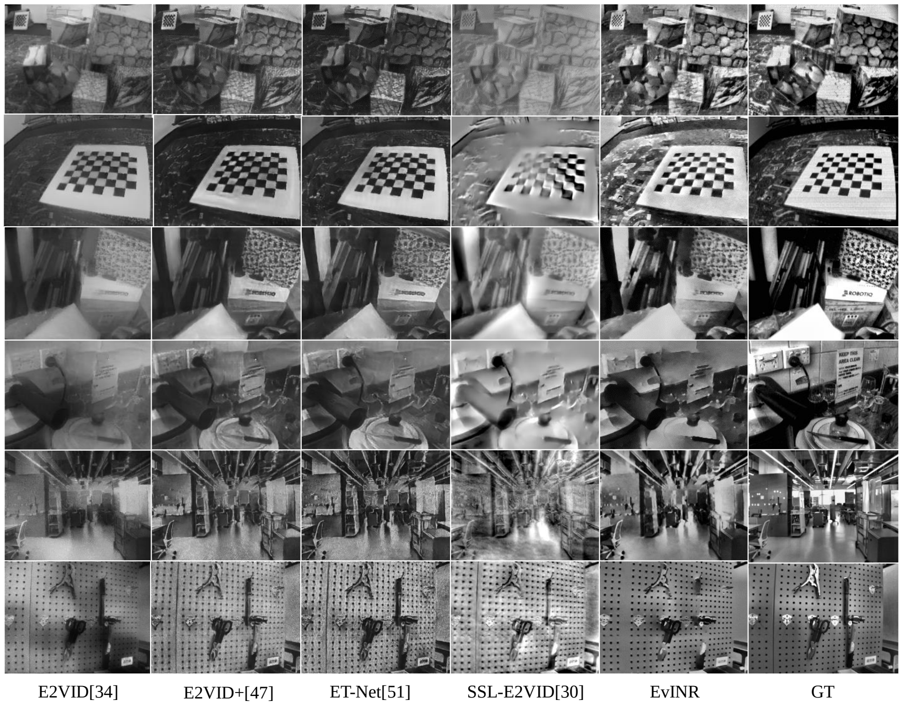


Fig. 4: Qualitative comparison with baseline methods on IJRR(Row 1&2), HQF (Row 3&4) and AED(Row 5&6).

Implementation Details: We use a SIREN network [46] with three hidden layers and 512 neurons per layer as our INR network. We partition each event sequence into sub-sequences lasting 5 seconds each and concurrently train all sub-sequences using model ensembling for 300 iterations. We first stack the events within $\frac{1}{32}$ second into an event frame and temporal upsample the event frames in 100 and 200 iterations, as described in Sec. 3.3. We adopt the Adam optimizer [23] with a learning rate of 1e−4 and exponentially reduce the learning rate every 10 iterations with a decay rate of 0.95. The weight of spatial regularization λ is set to 0.05, and γ in Eq. 10 is set to 0.6. The activated threshold C is set to 1 for the IJRR and HQF datasets and 0.25 for the AED dataset. The training process takes approximately 8 seconds on a single RTX3090 GPU.

4.2 Evaluation of Video Reconstruction

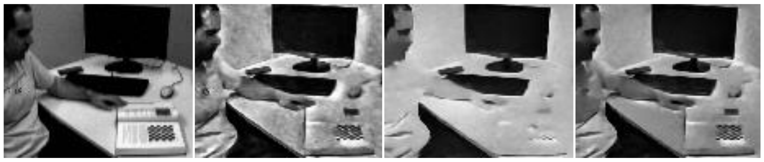
We assess the effectiveness of our approach by comparing it against eight state-of-the-art (SoTA) methods, classified based on the amount of data required for training. The methods are categorized into supervised learning (SL) methods that utilize synthetic ground-truth intensity frames for supervision and self-supervised learning (SSL) methods that rely solely on event data. Specifically, we compare against: **1)** Five SL methods: FireNet [41], E2VID [35], FireNet+ [49], E2VID+ [49], and ET-Net [53]. **2)** Three SSL method: SSL-E2VID [31], HF [40] and ELRP [60]. Reconstructed results for all methods were generated at each timestamp of the intensity frame. We use optical flow estimation from the FlowNet of SSL-E2VID [31] for ELRP [60]. We apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [54] to both ground truth and synthesized frames before evaluation, following [31]. Note that the quantitative results for FireNet [41], E2VID [35], FireNet+ [49], E2VID+ [49], and ET-Net [53] are referred from a recent benchmark paper [11]. For the visual comparison, we obtained the visualization results by downloading the publicly available checkpoints and test them in the local environment.

Table 1: Comparison of quantitative results on the IJRR, HQF, and AED datasets. Bold values indicate the best results among all methods, while underlined values indicate the best results among SSL methods.

	Methods	IJRR			HQF			AED		
		MSE	SSIM	LPIPS	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS
SL	FireNet [41]	0.131	0.502	0.320	0.094	0.533	0.441	0.074	0.298	0.579
	E2VID [35]	0.212	0.424	0.350	0.127	0.540	0.382	**0.056**	0.424	0.500
	FireNet+ [49]	0.063	0.555	0.290	0.040	0.614	0.314	0.094	0.232	0.566
	E2VID+ [49]	0.070	0.560	0.236	0.036	0.643	0.252	0.074	0.345	0.462
	ET-Net [53]	0.047	0.617	**0.224**	**0.032**	**0.658**	**0.260**	0.084	0.312	0.482
SSL	SSL_E2VID [31]	0.097	0.473	0.409	0.070	0.480	0.464	0.094	0.316	0.453
	HF [40]	0.164	0.334	0.658	0.133	0.232	0.670	0.080	0.240	0.943
	ELRP [60]	0.080	0.437	0.485	0.074	0.450	0.474	0.084	0.305	0.473
	Ours	**0.047**	**0.628**	0.251	0.048	0.531	0.333	0.067	**0.458**	**0.366**

The quantitative results are presented in Table 1. Our method demonstrates superior performance compared to the best SL methods, with improvements of 7%, 13%, and 4% in terms of MSE, SSIM, and LPIPS respectively on the IJRR dataset [27]. Although the gap between our method and SL methods widens on the HQF dataset [49] due to the lower event density, our method remains comparable. On the AED dataset, our method outperforms the state-of-the-art SL methods, E2VID+ and ET-Net, by a clear margin for all three metrics. Notably, our method shows a significant improvement over previous SSL methods. In particular, compared with SSL-E2VID, it improves MSE, SSIM, and LPIPS by 35%, 25%, and 21%, respectively.

A qualitative comparison of our method with the baseline methods is depicted in Fig. 4. Our method produces intensity frames with better contrast and overall visual quality, compared with other methods that suffer from issues such as foggy effects, artifacts, and loss of detailed structures. It is worth noting that the performance of E2VID+ and ET-Net degrades significantly on the AED dataset, *which suggests that their training strategy may overfit the DAVIS sensor setting.* Additional results can be found in the supplementary material.



Ground Truth No *Reg.* No *C2T.* Full Model

Fig. 5: The impact of removing the spatial regularization and coarse-to-fine training.

Table 2: Ablation of the spatial regularization term (*Reg.*) and coarse-to-fine training (*C2F.*) on the IJRR [27] dataset.

	MSE	SSIM	LPIPS	Time(s)	FPS
Base	0.063	0.542	0.357	17.29	13.77
Base+Reg.	0.044	0.612	0.275	17.44	13.64
Base+C2F.	0.061	0.566	0.343	**7.94**	**31.33**
Full Model	**0.047**	**0.658**	**0.251**	8.08	29.45

4.3 Ablation Study

We conduct ablation experiments on the IJRR dataset to assess the significance of each individual component of our method, and the results are presented in Tab. 2. Our method's basic implementation, which doesn't include spatial regularization term as described in Sec. 3.2 or use coarse-to-fine training as described in Sec. 3.3, is called *Base*. We also evaluate the performance of two modified versions of the full model: *Base + Reg.*, which adds a spatial regularization term (Eq. 7) with $\lambda = 0.05$, and *Base + C2F*, which uses coarse-to-fine training. Experimental results show that both spatial regularization and coarse-to-fine training contribute to the improved performance of our full model. Fig. 5 confirms the effectiveness of these two modules, showing that removing the spatial regularization term introduces noticeable noise (e.g., in the background), while removing the coarse-to-fine training scheme leads to an increase in training time and missing details (e.g., in the people's clothes and the book).

4.4 Discussions

Training speed: To assess the training speed of EvINR, we conducted an analysis using various model ensembling configurations on the whole *Calibration* sequence from the IJRR dataset [27]. We divide the event sequence, which had a total duration of 50 seconds, into N partitions, with each partition having a duration of τ seconds. Subsequently, we trained N EvINR models in parallel, incorporating the model ensembling techniques outlined in Sec. 3.4. The experimental findings are presented in Table 3. By training 10 EvINR models, each corresponding to 5 seconds of the event sequence, we achieved a performance gain of over 26%, while only requiring an additional 14% of training time compared to training a single EvINR model on the entire sequence. However, we observed a significant drop in performance when increasing the partition size to

Table 3: Impact of hyper-parameters of model ensembling on the training speed and reconstruction performance.

N	τ	MSE	SSIM	LPIPS	Time(s)	FPS
1	50	0.084	0.46	0.380	**15.71**	**75.66**
5	10	0.0688	0.600	0.253	16.91	70.31
10	5	**0.0625**	**0.612**	**0.244**	18.32	64.90
50	1	0.0875	0.488	0.343	29.43	40.39

50, suggesting that a 1-second sequence is insufficient for EvINR to converge to a stable solution of the event generation equation. We also note that the optimal choice of hyper-parameters may differ for different event sequences due to the per-scene optimization nature of our INR approach.

Event enhancement: Our approach provides a smooth and continuous representation of event data by representing events triggered within a small time window Δt as $\frac{\partial F_\Theta}{\partial t}\Delta t$. This event representation automatically reduces noise and preserves critical information through INR optimization, making it highly robust to noise. Fig. 6 compares the denoising results of our INR representation with several SoTA denoising methods [12, 51].

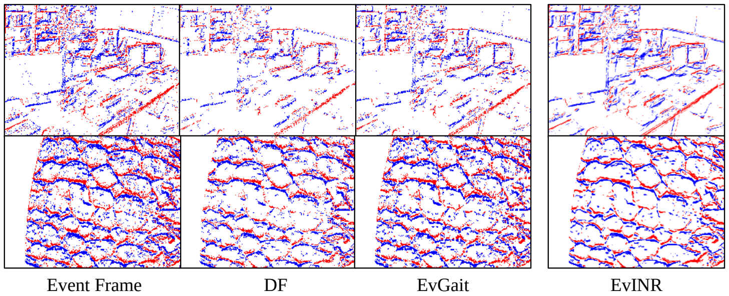
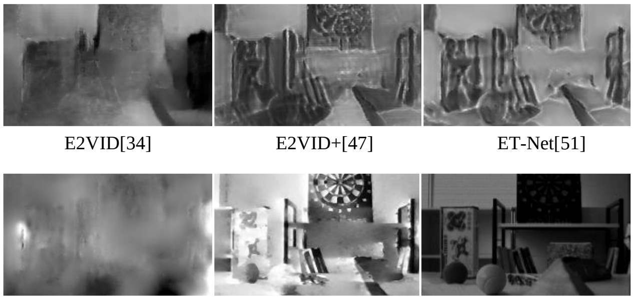


Fig. 6: Qualitative comparison with baseline methods of event stream denoising.

Early frames reconstruction: [6] suggested that RNN-based techniques for video reconstruction, *e.g.* E2VID [35], require an initialization period to achieve satisfactory results. Consequently, the initial frames generated by these methods may be of poor quality, restricting their usefulness to short event sequences. Conversely, our proposed approach is capable of producing realistic outcomes using a minimal number of events. Fig. 7 illustrates a comparison of the first frame generated by our method and RNN-based techniques, demonstrating our ability to rapidly create high-quality outputs with minimal input.

Inference speed: Compared to other SoTA approaches, our method offers a significant improvement in terms of inference speed, as demonstrated in Tab. 4. The main reason for this improvement is that other methods require converting event data into event frames or voxel grids during the inference period, which



E2VID[34] E2VID+[47] ET-Net[51]

SSL-E2VID[30] EvINR GT

Fig. 7: The reconstructed initial frame from different compared methods.

Table 4: Comparison of inference time in terms of frames-per-second (fps) at two resolutions from IJRR [27] and AED dataset respectively.

Resolution	FireNet	E2VID	ET-Net	Ours
(240, 180)	118.24	100.96	36.1	178.19
(408, 306)	28.83	28.56	18.27	55.37

adds significant overhead. However, our method only requires the spatiotemporal coordinates as input, thus avoiding this issue.

5 Conclusion

This paper introduced EvINR, a SSL method for event-to-video reconstruction that relieves the need for synthetic data or optical flow estimation. We, for the first time, show that high-quality videos can be reconstructed in a self-supervised and interpretable way without time-consuming end-to-end training. Our method is based on directly solving the event generation model via optimizing an INR whose temporal derivative is self-supervised by events and spatial derivative is regularized to reduce artifacts. Additionally, we propose several acceleration techniques that significantly reduce the training time of EvINR, making it applicable for online tasks. Experiments show that our approach significantly outperforms the previous SSL methods, and is competitive with the SoTA supervised methods. Our approach also demonstrates superior interpretability and robustness to various event dataset. Overall, our work contributes to the advancement of event-to-video reconstruction and offers a promising direction for future research that combines INRs with event data.

Limitations and Future Work: The current parameter size of EvINR takes up approximately the same amount of storage as the original event data. In future work, we plan to explore network pruning and quantization techniques to further reduce the parameter size.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (NSF) under Grant No. NSFC22FYT45, the Guangzhou City, University and Enterprise Joint Fund under Grant No.SL2022A03J01278, and Guangzhou Fundamental and Applied Basic Research (**Grant Number**: 2024A04J4072)

References

1. Alpsentek products. `https://www.alpsentek.com/product`, [Accessed: Feb. 17, 2023]
2. Pepperl fuchs event camera. `https://www.pepperl-fuchs.com/global/en/classid_11544.htm?view=productgroupliterature`, accessed on 17 Feb 2023
3. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 884–892 (2016)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
5. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014)
6. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. IEEE Transactions on Image Processing **30**, 2488–2500 (2021)
7. Chen, H., He, B., Wang, H., Ren, Y., Lim, S.N., Shrivastava, A.: Nerv: Neural representations for videos. Advances in Neural Information Processing Systems **34**, 21557–21568 (2021)
8. Chen, Y., Liu, S., Wang, X.: Learning continuous image representation with local implicit image function. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8628–8638 (2021)
9. Cook, M., Gugelmann, L., Jug, F., Krautz, C., Steger, A.: Interacting maps for fast visual interpretation. In: The 2011 International Joint Conference on Neural Networks. pp. 770–776. IEEE (2011)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Ercan, B., Eker, O., Erdem, A., Erdem, E.: Evreal: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3942–3951 (2023)
12. Feng, Y., Lv, H., Liu, H., Zhang, Y., Xiao, Y., Han, C.: Event density based denoising method for dynamic vision sensor. Applied Sciences **10**(6), 2024 (2020)
13. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., et al.: Event-based vision: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(1), 154–180 (2020)
14. Gehrig, D., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Video to events: Recycling video datasets for event cameras. In: IEEE Conf. Comput. Vis. Pattern Recog. (CVPR) (June 2020)

15. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
16. Hansen, L.K., Salamon, P.: Neural network ensembles. IEEE transactions on pattern analysis and machine intelligence **12**(10), 993–1001 (1990)
17. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1312–1321 (2021)
20. Hwang, I., Kim, J., Kim, Y.M.: Ev-nerf: Event based neural radiance field. arXiv preprint arXiv:2206.12455 (2022)
21. Kim, H., Handa, A., Benosman, R., Ieng, S.H., Davison, A.J.: Simultaneous mosaicing and tracking with an event camera. J. Solid State Circ **43**, 566–576 (2008)
22. Kim, H., Leutenegger, S., Davison, A.J.: Real-time 3d reconstruction and 6-dof tracking with an event camera. In: European conference on computer vision. pp. 349–364. Springer (2016)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Klenk, S., Koestler, L., Scaramuzza, D., Cremers, D.: E-nerf: Neural radiance fields from a moving event camera. IEEE Robotics and Automation Letters (2023)
25. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4460–4470 (2019)
26. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
27. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. The International Journal of Robotics Research **36**(2), 142–149 (2017)
28. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. International Journal of Computer Vision **126**(12), 1381–1393 (2018)
29. Niemeyer, M., Geiger, A.: Giraffe: Representing scenes as compositional generative neural feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11453–11464 (2021)
30. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5589–5599 (2021)
31. Paredes-Vallés, F., de Croon, G.C.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3446–3455 (2021)
32. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 165–174 (2019)

33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
34. Pini, S., Borghi, G., Vezzani, R.: Learn to see by events: Color frame synthesis from event and rgb cameras. arXiv preprint arXiv:1812.02041 (2018)
35. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3857–3866 (2019)
36. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
37. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. ACM Transactions on Graphics **21**(3), 267–276 (2002). `https://doi.org/10.1145/566654.566576`
38. Rudnev, V., Elgharib, M., Theobalt, C., Golyanik, V.: Eventnerf: Neural radiance fields from a single colour event camera. arXiv preprint arXiv:2206.11896 (2022)
39. Ryu, H.E.: Industrial dvs design; key features and applications. In: Conf. on Computer Vision and Pattern Recognition. vol. 3 (2019)
40. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: Asian Conference on Computer Vision. pp. 308–324. Springer (2018)
41. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D.: Fast image reconstruction with an event camera. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 156–163 (2020)
42. Scheerlinck, C., Rebecq, H., Stoffregen, T., Barnes, N., Mahony, R., Scaramuzza, D.: Ced: Color event camera dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
43. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: Generative radiance fields for 3d-aware image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
44. Shiba, S., Aoki, Y., Gallego, G.: Event collapse in contrast maximization frameworks. Sensors **22**(14), 5190 (2022)
45. Shiba, S., Aoki, Y., Gallego, G.: Secrets of event-based optical flow. In: European Conference on Computer Vision. pp. 628–645. Springer (2022)
46. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems **33**, 7462–7473 (2020)
47. Skorokhodov, I., Ignatyev, S., Elhoseiny, M.: Adversarial generation of continuous images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10753–10764 (2021)
48. Song, C., Huang, Q., Bajaj, C.: E-cir: Event-enhanced continuous intensity recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7803–7812 (2022)
49. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: European Conference on Computer Vision. pp. 534–549. Springer (2020)
50. Strümpler, Y., Postels, J., Yang, R., Gool, L.V., Tombari, F.: Implicit neural representations for image compression. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI. pp. 74–91. Springer (2022)

51. Wang, Y., Du, B., Shen, Y., Wu, K., Zhao, G., Sun, J., Wen, H.: Ev-gait: Event-based robust gait recognition using dynamic vision sensors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6358–6367 (2019)
52. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
53. Weng, W., Zhang, Y., Xiong, Z.: Event-based video reconstruction using transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2563–2572 (2021)
54. Yadav, G., Maheshwari, S., Agarwal, A.: Contrast limited adaptive histogram equalization based enhancement for real time video system. In: 2014 international conference on advances in computing, communications and informatics (ICACCI). pp. 2392–2397. IEEE (2014)
55. Yang, M., Liu, S.C., Delbruck, T.: A dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding. IEEE Journal of Solid-State Circuits **50**(9), 2149–2160 (2015)
56. Zhang, K., Li, Y., Zuo, W., Zhang, L., Van Gool, L., Timofte, R.: Plug-and-play image restoration with deep denoiser prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6360–6376 (2021)
57. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
58. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. ACM Transactions on Graphics (TOG) **40**(6), 1–18 (2021)
59. Zhang, Z., Yezzi, A., Gallego, G.: Image reconstruction from events. why learn it? arXiv preprint arXiv:2112.06242 (2021)
60. Zhang, Z., Yezzi, A.J., Gallego, G.: Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. IEEE Transactions on Pattern Analysis & Machine Intelligence **45**(07), 8372–8389 (2023)
61. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
62. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018)
63. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Unsupervised event-based learning of optical flow, depth, and egomotion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 989–997 (2019)
64. Zhu, L., Wang, X., Chang, Y., Li, J., Huang, T., Tian, Y.: Event-based video reconstruction via potential-assisted spiking neural network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3594–3604 (2022)