Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy

Federico Paredes-Vallés Guido C. H. E. de Croon
Micro Air Vehicle Laboratory, Delft University of Technology, The Netherlands

Abstract

Event cameras are novel vision sensors that sample, in an asynchronous fashion, brightness increments with low latency and high temporal resolution. The resulting streams of events are of high value by themselves, especially for high speed motion estimation. However, a growing body of work has also focused on the reconstruction of intensity frames from the events, as this allows bridging the gap with the existing literature on appearance- and frame-based computer vision. Recent work has mostly approached this problem using neural networks trained with synthetic, ground-truth data. In this work we approach, for the first time, the intensity reconstruction problem from a self-supervised learning perspective. Our method, which leverages the knowledge of the inner workings of event cameras, combines estimated optical flow and the event-based photometric constancy to train neural networks without the need for any ground-truth or synthetic data. Results across multiple datasets show that the performance of the proposed self-supervised approach is in line with the state-of-the-art. Additionally, we propose a novel, lightweight neural network for optical flow estimation that achieves high speed inference with only a minor drop in performance.

1. Introduction

Unlike conventional cameras recording intensity frames at fixed time intervals, event cameras sample light based on scene dynamics by asynchronously measuring per-pixel brightness[1] changes at the time they occur [1]. This results in streams of sparse events encoding the polarity of the perceived changes. Because of this paradigm shift, event cameras offer several advantages over their frame-based counterparts, namely low power consumption, high dynamic range (HDR), low latency and high temporal resolution.

Despite the advantages, the novel output format of event cameras poses new challenges in terms of algorithm design. Unless working with spiking neural networks [2], events are usually converted into intermediate representations that facilitate the extraction of information [1]. Among others, intensity frames are an example of a powerful representation since they allow the evaluation of the appearance of a visual scene, thus bridging the gap between event cameras and the existing frame-based computer vision literature [3, 4]. For this reason, there has been a significant research drive to develop new methods to reconstruct images from events with similar statistics to those captured by standard cameras.

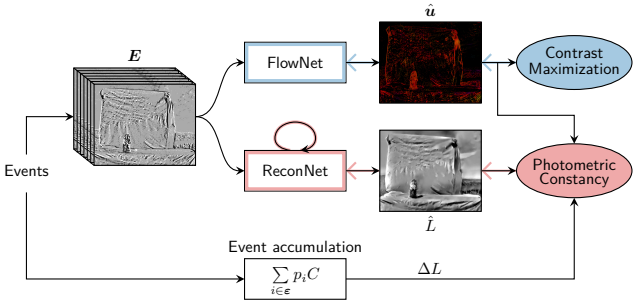

Figure 1: Overview of the proposed framework. Our model is trained in a self-supervised fashion to perform optical flow estimation and image reconstruction from event data using the contrast maximization proxy loss and the event-based photometric constancy, respectively. Colored reverse arrows indicate error propagation for each loss.

Recent work has mostly approached this problem from a machine learning perspective. With their E2VID artificial neural network, Rebecq *et al.* [3, 4] were the first to show that learning-based methods trained to maximize perceptual similarity via supervised learning outperform hand-crafted techniques by a large margin in terms of image quality. Later, Scheerlinck *et al.* [5] achieved high speed inference with FireNet, a simplified model of E2VID. Despite the high levels of accuracy reported, these architectures were trained with large sets of synthetic data from event camera simulators [6], which adds extra complexity to the reconstruction problem due to the *simulator-to-reality gap*. In fact, Stoffregen, Scheerlinck *et al.* [7] recently showed

[1] Defined as the logarithm of the pixel intensity, i.e., $L \doteq \log(I)$.

that if the statistics of the synthetic training datasets do not closely resemble those seen during inference, image quality degrades and the generalizability of these architectures remains limited.

In this work, we propose to come back to the theoretical basics of event cameras to relax the dependency of learning-based reconstruction methods on ground-truth and synthetic data. Specifically, we introduce the self-supervised learning (SSL) framework in Fig. 1, which consists of two artificial neural networks, *FlowNet* and *ReconNet*, for optical flow estimation and image reconstruction, respectively. FlowNet is trained through the contrast maximization proxy loss from Zhu *et al.* [8], while ReconNet makes use of the flow-intensity relation in the *event-based photometric constancy* [9] to reconstruct the frames that best satisfy the input events and the estimated flow. Using our method, we retrain several networks from the image reconstruction [3, 5] and optical flow [10] literature. In terms of accuracy, results show that the reconstructed images are in line with those generated by most learning-based approaches despite the lack of ground-truth data during training. Additionally, we propose *FireFlowNet*, a lightweight architecture for optical flow estimation that, inspired by [5], achieves high speed inference with only a minor drop in performance.

In summary, this paper contains *two main contributions*. First, a novel SSL framework to train artificial neural networks to perform event-based image reconstruction that, with the aid of optical flow, does not require ground truth of any kind and can learn directly on real event data. Second, we introduce FireFlowNet: a novel, lightweight neural network architecture that performs fast optical flow estimation from events. We validate our self-supervised method and optical flow network through extensive quantitative and qualitative evaluations on multiple datasets.

2. Related Work

Early methods to image reconstruction from event data approached the problem through the *photometric constancy*: each event provides one equation relating the intensity gradient and the optical flow [9]. Kim *et al.* [11] were the first in the field and developed an Extended Kalman Filter that, under rotational and static scene assumptions, reconstructs a gradient image that is later transformed into the intensity space via Poisson integration. They later extended this approach to 6 degrees-of-freedom camera motion [12]. Under the same assumptions, Cook *et al.* [13] simultaneously recovered intensity images, optical flow, and angular velocity through bio-inspired, interconnected network of interacting maps. Bardow *et al.* [14] developed a variational energy minimization framework to simultaneously estimate optical flow and intensity from sliding windows of events, relaxing for the first time the static scene assumption.

Instead of relying on the photometric constancy, several approaches based on direct event integration have been proposed, which do not assume scene structure or motion dynamics. Reinbacher *et al.* [15] formulated intensity reconstruction as an energy minimization problem via direct integration with periodic manifold regularization. Scheerlinck *et al.* [16] achieved computationally efficient reconstruction by filtering events with a high-pass filter prior to integration.

Several machine learning approaches have also been proposed. Training generative adversarial networks with real grayscale frames was proposed by Wang *et al.* [17] and Pini *et al.* [18]. However, Rebecq *et al.* [3, 4] showed that training in a supervised fashion with a large synthetic dataset allowed for higher quality reconstructions with their *E2VID* architecture. Focused on computational efficiency, Scheerlinck *et al.* [5] managed to significantly reduce E2VID complexity with *FireNet*, with only a minor drop in accuracy. Inspired by these works, Choi *et al.* [19] and Wang *et al.* [20] recently proposed hybrid approaches that incorporate super resolution aspects in the training process and architecture design to improve image quality. Lastly, Stoffregen, Scheerlinck *et al.* [7] recently highlighted that, when training with ground truth, the statistics of the training dataset play a major role in the reconstruction quality. They showed that a slight change in the training statistics of E2VID leads to significant improvements across multiple datasets.

Our proposed SSL framework (see Fig. 1) is based on the event-based photometric constancy used by early reconstruction methods. Similarly to Bardow *et al.* [14], we simultaneously estimate intensity and optical flow from the input events. However, instead of relying on a joint optimization scheme, we achieve it via two independent neural networks that only share information during training. Further, we reconstruct intensity directly from the photometric constancy, instead of from an oversimplified model of the event camera. This approach allows, for the first time, to relax the strong dependency of learning-based approaches on ground-truth and synthetic data.

3. Method

An event camera consist of an array of independent pixels that respond to changes in the brightness signal $L(t)$, and transmit these changes through streams of sparse and asynchronous events [21]. For an ideal camera, an event $\boldsymbol{e}_i = (\boldsymbol{x}_i, t_i, p_i)$ is triggered at pixel $\boldsymbol{x}_i = (x_i, y_i)^T$ and time t_i whenever the brightness change since the last event at that pixel reaches a contrast sensitivity threshold C. Therefore, the brightness increment occurred in a time window Δt_k is encoded in the event data via pixel-wise accumulation:

$$\Delta L_k(\boldsymbol{x}) = \sum_{\boldsymbol{e}_i \in \Delta t_k} p_i C \tag{1}$$

where $C > 0$, and the polarity $p_i \in \{+, -\}$ encodes the sign of the brightness change.

As in [9], under the assumptions of Lambertian surfaces, constant illumination and small Δt, we can linearize Eq. 1 to obtain the event-based photometric constancy:

$$\Delta L_k(\boldsymbol{x}) \approx -\nabla L_{k-1}(\boldsymbol{x}) \cdot \boldsymbol{u}_k(\boldsymbol{x}) \Delta t_k \tag{2}$$

which encodes that events are caused by the spatial gradients of the brightness signal, $\nabla L = (\delta_x L, \delta_y L)^T$, moving with optical flow $\boldsymbol{u} = (u, v)^T$. The dot product conveys that no events are generated if the flow vector is parallel to an edge ($\boldsymbol{u} \perp \nabla L$), while they are generated at the highest rate if perpendicular ($\boldsymbol{u} \parallel \nabla L$). Thus, events are caused by the projection of the optical flow vector in the ∇L direction.

3.1. Overview

Our goal is to learn, in an SSL fashion, to transform a continuous stream of events into a sequence of intensity images $\{\hat{I}_k\}$. To achieve this, we propose the pipeline in Fig. 1 in which two neural networks are jointly trained. On the one hand, *FlowNet* is a convolutional network that learns to estimate optical flow by compensating for the motion blur in the input events. On the other hand, *ReconNet* is a recurrent convolutional network that learns to perform image reconstruction through the event-based photometric constancy.

3.2. Input Event Representation

As proposed in [8], the input to both our networks is a voxel grid E_k with B temporal bins that gets populated with consecutive, non-overlapping partitions of the event stream $\boldsymbol{\varepsilon}_k \doteq \{\boldsymbol{e}_i\}_{i=0}^{N-1}$, each containing a fixed number of events, N. For each partition, every event (with index i) distributes its polarity p_i to the two closest bins according to:

$$E(\boldsymbol{x}_i, t_b) = \sum_i p_i \kappa(t_b - t_i^* \, (B-1)) \tag{3}$$

$$\kappa(a) = \max(0, 1 - |a|) \tag{4}$$

$$t_i^* = \frac{\left(t_i - t_0^k\right)}{\left(t_{N-1}^k - t_0^k\right)} \tag{5}$$

where b is the bin index, and $t_i^* \in [0, 1]$ denotes the normalized event timestamp. This representation adaptively normalizes the temporal dimension of the input depending on the timestamps of each partition of events.

3.3. Optical Flow via Contrast Maximization

We aim to learn to reconstruct L through the photometric constancy in Eq. 2, which, besides the spatial and temporal derivatives of the brightness itself, also depends on the optical flow $\boldsymbol{u}$. One could use ground-truth optical flow to solve for this ill-posed problem. However, due to the limited availability of event-camera datasets with accurate ground-truth data, we opt for training our FlowNet to perform flow estimation in a self-supervised manner, using the contrast maximization proxy loss for motion compensation [22].

A partition of events is said to be blurry whenever there is a spatiotemporal misalignment among its events, i.e., events generated by the same portion of a moving edge are captured with different timestamps and pixel locations. The idea behind the motion compensation framework [22] is that accurate optical flow can be retrieved by finding the motion model of each event that best deblurs $\boldsymbol{\varepsilon}_k$. Knowing the per-pixel optical flow, the events can be propagated to a reference time t_{ref} through:

$$\boldsymbol{x}_i' = \boldsymbol{x}_i + (t_{\text{ref}} - t_i)\boldsymbol{u}(\boldsymbol{x}_i) \tag{6}$$

In this work, we adopt the deblurring quality measure proposed by Mitrokhin *et al.* [23] and later refined by Zhu *et al.* [8]: the per-pixel and per-polarity average timestamp of the resulting image of warped events (IWE), H. The lower this metric, the better the deblurring. As in [8], we generate an image of the average (normalized) timestamp at each pixel for each polarity p' via bilinear interpolation:

$$T_{p'}(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}}^*) = \frac{\sum_j \kappa(x - x_j')\kappa(y - y_j')t_j^*}{\sum_j \kappa(x - x_j')\kappa(y - y_j') + \epsilon} \tag{7}$$
$$j = \{i \mid p_i = p'\}, \qquad p' \in \{+, -\}, \qquad \epsilon \approx 0$$

and minimize the sum of the squared images resulting from warping the events forward and backward to prevent scaling issues during backpropagation:

$$\mathcal{L}_{\text{contrast}}(t_{\text{ref}}^*) = \sum_{\boldsymbol{x}} T_+(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}}^*)^2 + T_-(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}}^*)^2 \tag{8}$$

$$\mathcal{L}_{\text{contrast}} = \mathcal{L}_{\text{contrast}}(1) + \mathcal{L}_{\text{contrast}}(0) \tag{9}$$

The total loss used to train FlowNet is then given by:

$$\mathcal{L}_{\text{FlowNet}} = \mathcal{L}_{\text{contrast}} + \lambda_1 \mathcal{L}_{\text{smooth}} \tag{10}$$

where $\mathcal{L}_{\text{smooth}}$ is a Charbonnier smoothness prior [24], and λ_1 is a scalar balancing the effect of the two losses. Note that, since $\mathcal{L}_{\text{contrast}}$ does no propagate the error back to pixels without events, we mask FlowNet's output so that null optical flow vectors are returned at these pixel locations.

3.4. Reconstruction via Photometric Constancy

We formulate the SSL reconstruction problem from an image registration perspective [25] via brightness increment images. Specifically, we propose to use the difference between the reference increment image ΔL (event integration, Eq. 1) and the predicted $\Delta \hat{L}$ (photometric constancy, Eq. 2) to reconstruct the brightness signal that best explains the input events, assuming known error-free optical flow. This reconstructed brightness is denoted by $\hat{L}$. FlowNet predictions are used in the computation of $\Delta \hat{L}$, and as registration parameters to warp both increment images to a common temporal frame (indicated by the superscript *). A schematic of the proposed formulation is shown in Fig. 2.

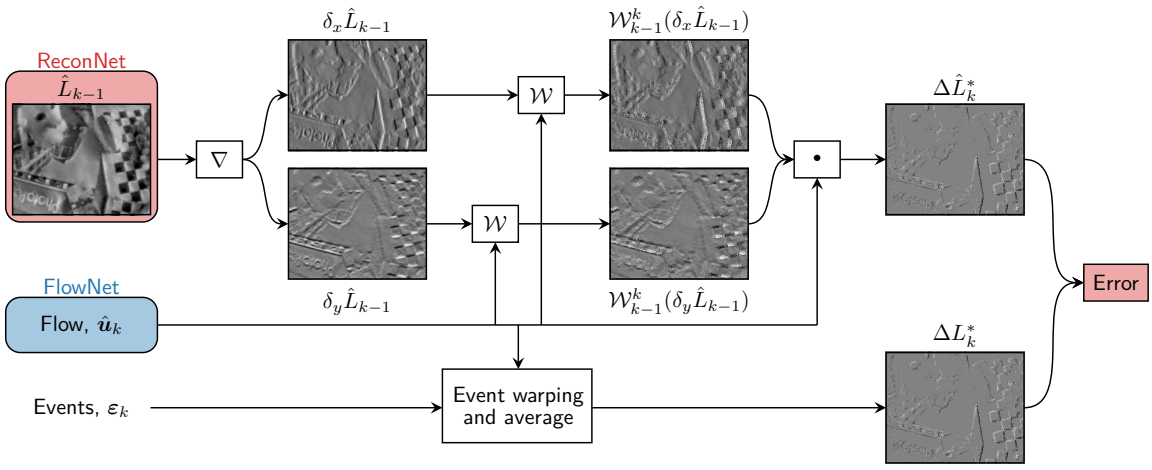

Figure 2: Brightness reconstruction via the event-based photometric constancy formulation proposed in this work. The most recent event-based optical flow estimate from FlowNet $\hat{\boldsymbol{u}}_k$ is used to (i) warp the input events, (ii) warp the spatial gradients of the last reconstructed image $\hat{L}_{k-1}$, and (iii) in the dot product with the warped gradients. The predicted brightness increment image $\Delta \hat{L}^*_k$ is compared to that obtained with the deblurred (and averaged) input events, ΔL^*_k, and the error is propagated backwards towards ReconNet to improve reconstruction accuracy.

To minimize motion blur in the reconstructed frames, instead of directly integrating the input events, we define the reference brightness increment ΔL^* via the per-pixel and per-polarity average number of warped events:

$$\Delta L^*(\boldsymbol{x};\boldsymbol{u}) \doteq C\left(G_+(\boldsymbol{x};\boldsymbol{u}|1) - G_-(\boldsymbol{x};\boldsymbol{u}|1)\right) \tag{11}$$

$$G_{p'}(\boldsymbol{x};\boldsymbol{u}|t^*_{\text{ref}}) = \frac{H_{p'}(\boldsymbol{x};\boldsymbol{u}|t^*_{\text{ref}})}{P_{p'}(\boldsymbol{x};\boldsymbol{u}|t^*_{\text{ref}}) + \epsilon} \tag{12}$$

where P is a two-channel image containing the number of pixel locations from where the IWE H receives events in the event warping process (see Section 3.3). Therefore, ΔL^* is a deblurred representation of the contrast change encoded in the input events. An ablation study on the impact of event deblurring prior to event integration can be found in the supplementary material.

On the other hand, we adapt the event-based photometric constancy in Eq. 2 and compute $\Delta \hat{L}$ by warping the spatial gradients of the last reconstructed image to the current time instance via spatial transformers [26]:

$$\Delta \hat{L}^*(\boldsymbol{x};\boldsymbol{u}) \doteq -\mathcal{W}^k_{k-1}(\nabla \hat{L}_{k-1}(\boldsymbol{x})) \cdot \hat{\boldsymbol{u}}_k(\boldsymbol{x}) \tag{13}$$

where $\mathcal{W}^k_{k-1}$ is the warping function of the optical flow $\hat{\boldsymbol{u}}_k$.

Following a maximum likelihood approach [21, 27], we define the photometric reconstruction loss as the squared L_2 norm of the difference of the warped brightness increments:

$$\mathcal{L}_{\text{PE}} = \left\| \Delta L^*(\boldsymbol{x};\boldsymbol{u}) - \Delta \hat{L}^*(\boldsymbol{x};\boldsymbol{u}) \right\|^2_2 \tag{14}$$

where, besides $\hat{L}$, the contrast threshold C is the only remaining unknown. To relax the dependency on this parameter, our ReconNet uses linear activation in its last layer instead of the frequently used sigmoid function [4, 5]. The resulting unbounded brightness estimate is first transformed into the intensity space through $\hat{I}_k = \exp(\hat{L}_k)$, and then linearly normalized to get the final reconstruction $\hat{I}^f_k$:

$$\hat{I}^f_k = \frac{\hat{I}_k - m}{M - m} \tag{15}$$

where m and M are the 1% and 99% percentiles of $\hat{I}_k$, and $\hat{I}^f_k$ is clipped to the range $[0, 1]$. This min/max normalization allows the use of any value of C for training as long as the ratio of positive and negative contrast thresholds resembles that of the evaluation sequences. We assume that most event-camera datasets were recorded with $C_+/C_- \approx 1$, and set both thresholds to 1.

On its own, Eq. 14 is not sufficient for the reconstruction of temporally consistent images. Because of the dot product in Eq. 13, the absence of input events can be ambiguously understood as lack of apparent motion, lack of spatial image gradients, or both. To solve for this issue, we introduce an explicit temporal consistency loss based on the frame-based formulation of the photometric constancy [28]. In essence, we define the temporal loss as the photometric error between two successive reconstructed frames:

$$\mathcal{L}_{\text{TC}} = \left\| \hat{L}_k - \mathcal{W}^k_{k-1}(\hat{L}_{k-1}) \right\|_1 \tag{16}$$

The total loss used to train ReconNet is then given by:

$$\mathcal{L}_{\text{ReconNet}} = \sum_{k=0}^{S} \mathcal{L}_{\text{PE}} + \lambda_2 \sum_{k=S_0}^{S} \mathcal{L}_{\text{TC}} + \lambda_3 \sum_{k=0}^{S} \mathcal{L}_{\text{TV}} \tag{17}$$

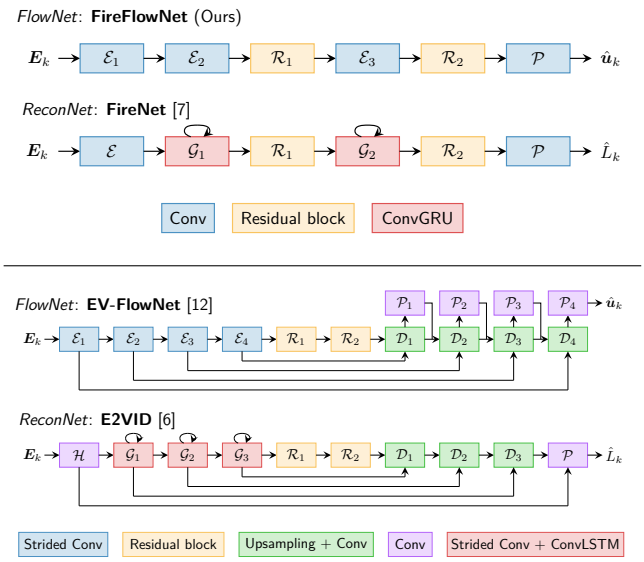

Figure 3: Neural networks evaluated in this work.

where S denotes the number of steps we unroll the recurrent network for during training, $\mathcal{L}_{\text{TV}}$ is a smoothness total-variation constraint [29], and λ_2 and λ_3 are scalars balancing the effect of the three losses.

3.5. Network Architectures

We evaluate the two trends on network design for event cameras when trained with our SSL framework. The evaluated architectures are shown in Fig. 3.

FlowNet: FireFlowNet. FireFlowNet is our proposed lightweight architecture for fast optical flow estimation. Inspired by FireNet [5], the network consists of three encoder layers that perform single-strided convolutions, two residual blocks [30], and a final prediction layer that performs depthwise (i.e., 1×1) convolutions with two output channels. All layers have 32 output channels and use 3×3 kernels and ReLU activations except for the final, which uses tanh activations. A comparison of the key architectural differences between our FireFlowNet and the current state-of-the-art is shown in Table 1.

FlowNet: EV-FlowNet [10]. The input voxel grid $\boldsymbol{E}_k$ is passed through four strided convolutional layers with output channels doubling after each layer starting from 64. The resulting activations are then passed through two residual blocks [30] and four decoder layers that perform bilinear upsampling followed by convolution. After each decoder, there is a (concatenated) skip connection from the corresponding encoder, as well as another depthwise convolution to produce a lower scale flow estimate, which is then concatenated with the activations of the previous decoder. The $\mathcal{L}_{\text{FlowNet}}$ loss (see Eq. 10) is applied to each intermediate flow estimate via flow upsampling. All layers use 3×3 convolutional kernels and ReLU activations except for the flow prediction layers, which use tanh activations.

Table 1: Main architectural differences between our FireFlowNet and EV-FlowNet [10]. FireFlowNet has $250\times$ fewer parameters, consuming only 0.41% of the memory.

	EV-FlowNet [10]	FireFlowNet (Ours)
No. params. (k)	14130.28	57.03
Memory (Mb)	53.90	0.22
Downsampling	Yes	No

ReconNet: FireNet [5]. Same architecture as FireFlowNet except for the second and third encoder, which are recurrent ConvGRU layers [32]. As in [5], each layer has 16 output channels, but we use linear activation in the final layer.

ReconNet: E2VID [4]. The input voxel grid $\boldsymbol{E}_k$ is passed through a convolutional head layer, three recurrent encoders performing strided convolution followed by ConvLSTM [33], two residual blocks [30], three decoder layers that perform bilinear upsampling followed by convolution, and a final depthwise convolutional prediction layer. There are (element-wise sum) skip connections between symmetric encoder and decoder layers, and the number of output channels in the head layer is 32 and doubles after each encoder. Head, encoder, and decoder layers use 5×5 kernels, while the rest uses 3×3. All layers use ReLU activations except for the final prediction layer which uses linear.

4. Experiments

We train our networks on the indoor forward facing sequences from the UZH-FPV Drone Racing Dataset (DR) [34], which is characterized by a much wider distribution of optical flow vectors than other datasets, such as MVSEC [31], the Event-Camera Dataset (ECD) [35], or the High Quality Frames (HQF) dataset [7]. Our training sequences consist of approximately 15 minutes of event data recorded with a racing quadrotor flying aggressive six-degree-of-freedom trajectories. We split these recordings and generate 440 128×128 (randomly cropped) sequences of 2 seconds each, and use them for training with $B = 5$. We further augment this data using random horizontal, vertical and polarity flips, besides with artificial pauses of the input event stream (i.e., forward-pass with null input voxel). For training, we fixed the number of input events per pixel to 0.3.

Our framework is implemented in PyTorch[2]. We use the Adam optimizer [36] and a learning rate of 0.0001 for both networks, and train with a batch size of 1 for 120 epochs. We empirically set the weights for each loss to $\{\lambda_1, \lambda_2, \lambda_3\} = \{1.0, 0.1, 0.05\}$, ReconNet's unrolling S to 20 steps, and S_0 to 10 steps.

[2]The project's code and additional qualitative results can be found at http://mavlab.tudelft.nl/ssl_e2v/.

Table 2: Quantitative evaluation of our FlowNet architectures on the MVSEC dataset [31]. For each sequence, we report the AEE (lower is better, ↓) in pixels and the percentage of points with endpoint error greater than 3 pixels, $\%_{\text{Outlier}}$ (↓). Best in bold, runner up underlined.

	outdoor_day1		indoor_flying1		indoor_flying2		indoor_flying3	
	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$
EV-FlowNet$_{\text{GT-SIM}}$ [7]	0.68	1.0	**0.56**	<u>1.0</u>	**0.66**	**1.0**	**0.59**	**1.0**
EV-FlowNet$_{\text{FW-MVSEC}}$ [10]	<u>0.49</u>	<u>0.2</u>	1.03	2.2	1.72	15.1	1.53	11.9
EV-FlowNet$_{\text{EW-MVSEC}}$ [8]	**0.32**	**0.0**	<u>0.58</u>	**0.0**	<u>1.02</u>	<u>4.0</u>	<u>0.87</u>	<u>3.0</u>
EV-FlowNet$_{\text{EW-DR}}$ (Ours)	0.92	5.4	0.79	1.2	1.40	10.9	1.18	7.4
FireFlowNet$_{\text{EW-DR}}$ (Ours)	1.06	6.6	0.97	2.6	1.67	15.3	1.43	11.0

Table 3: Computational cost evaluation of our FireFlowNet against EV-FlowNet [10]. We report inference time on GPU and the floating point operations (FLOPs) per forward-pass at common sensor resolutions. We used a single NVIDIA GeForce GTX 1080 Ti GPU for all experiments.

	GPU (ms)		FLOPs (G)	
	EV-FlowNet	FireFlowNet	EV-FlowNet	FireFlowNet
240×180	4.33	**1.97**	8.91	**2.47**
346×260	7.05	**3.81**	18.60	**5.14**
640×480	17.04	**12.55**	61.47	**17.59**
1280×720	49.32	**34.24**	184.41	**52.67**

Table 4: Quantitative evaluation of our FlowNet architectures on the ECD [35] and HQF [7] datasets. For each dataset, we report the mean FWL [7] (higher is better, ↑). Best in bold, runner up underlined.

	ECD*	HQF
EV-FlowNet$_{\text{FW-MVSEC}}$ [10]	1.36	1.25
EV-FlowNet$_{\text{GT-SIM}}$ [7]	**1.51**	1.39
EV-FlowNet$_{\text{EW-DR}}$ (Ours)	1.31	<u>1.51</u>
FireFlowNet$_{\text{EW-DR}}$ (Ours)	<u>1.39</u>	**1.58**

*Sequence cuts in the supplementary material.

4.1. Optical Flow Evaluation

To validate FireFlowNet as a lightweight alternative to the current state-of-the-art in event-based optical flow estimation, we evaluated both of our FlowNet architectures on the indoor_flying and outdoor_day sequences from the MVSEC dataset [31] with the ground-truth data provided by Zhu *et al.* [10]. Optical flow predictions were generated at each grayscale frame timestamp, and scaled to be the displacement between two successive frames.

Quantitative results are presented in Table 2. We use the average endpoint error (AEE) and the percentage of points with endpoint error greater than 3 pixels to compare our FlowNet architectures against three EV-FlowNet from literature; two of them trained with frame- (FW) [10] and event-warping (EW) [8] SSL proxy losses on MVSEC [31], and one trained with synthetic ground-truth data (GT) [7]. For our networks, the number of input events per pixel was set to 0.3. Error metrics were only acquired over pixels with valid ground-truth data and at least one event; and, for comparison, we used the quantitative results reported in [8, 7].

From Table 2, the first noticeable aspect is the accuracy gap between EV-FlowNet$_{\text{GT-SIM}}$ and the rest of networks. Training with ground-truth dense optical flow entails certain ability to resolve the aperture problem [37] that most SSL approaches lack. Regarding the latter, our EV-FlowNet performs consistently better than EV-FlowNet$_{\text{FW-MVSEC}}$ in all sequences except for outdoor_day1, but underperforms EV-FlowNet$_{\text{EW-MVSEC}}$ despite using the same architecture and training procedure. We believe this is mostly due to the different training datasets and the fact that we did not fine-tune the number of input events for this evaluation. Further, note that these literature architectures were trained on a very similar driving sequence from MVSEC, while our training data is much more diverse in terms of optical flow vectors [34].

Using our EV-FlowNet as reference, Table 2 shows that the proposed FireFlowNet is characterized by a comparable accuracy despite the significant reduction in model complexity. This performance drop is likely due to the narrow receptive field of the architecture, which entails limitations due to the aperture problem. Regarding computational cost, Table 3 shows that FireFlowNet runs $\sim$1.3-2.2 times faster than EV-FlowNet on GPU, requiring less than $\sim 30\%$ of FLOPS per forward-pass.

For completeness, we also evaluate our FlowNet architectures on the ECD [35] and HQF [7] datasets via the Flow Warp Loss (FWL) [7]. This metric, which does not require ground-truth data, measures the sharpness of the IWE in relation to that of the original partition of events. Similarly to [7], we set the number of input events to 50k for all sequences in this evaluation[3]. Table 4 shows that both our FlowNet architectures, which are specifically trained to perform event deblurring (see Section 3.3), are in line with or

[3]Note that the formulation of the FWL metric is sensitive to the number of input events [7].

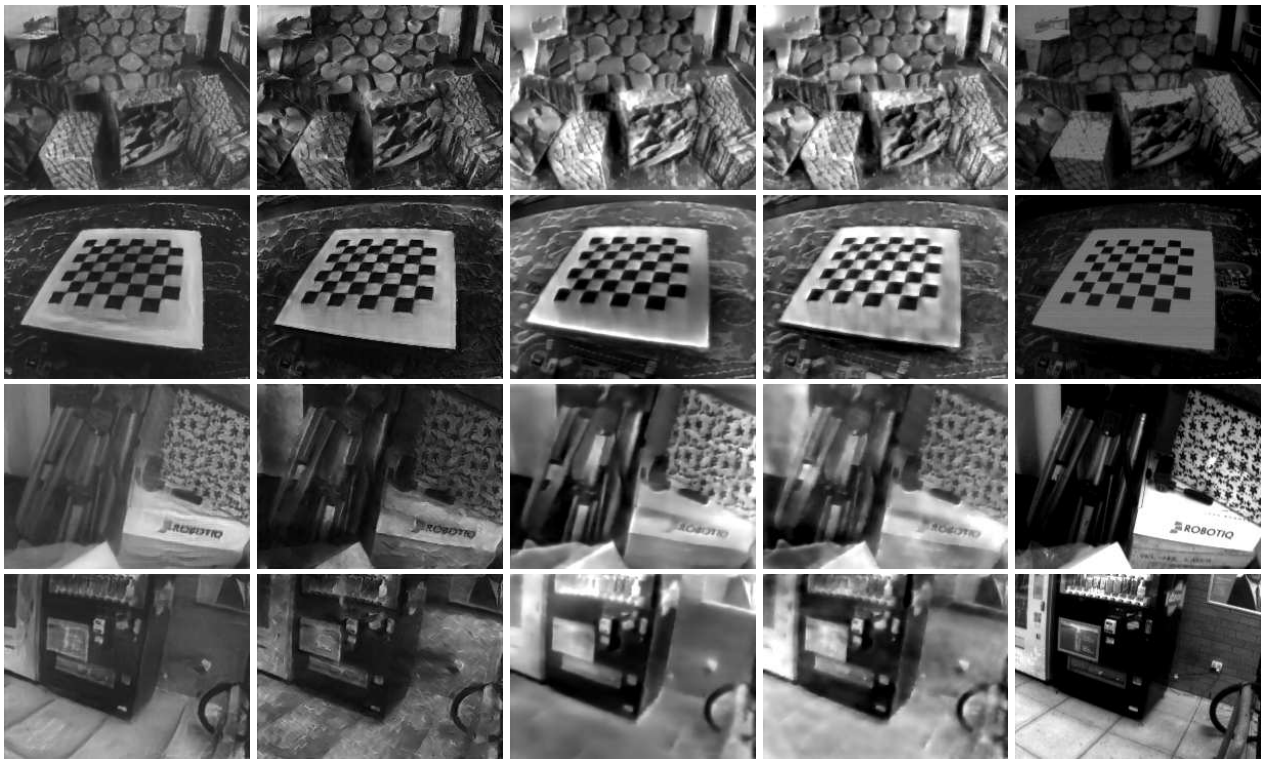

Figure 4: Qualitative comparison of our method with the state-of-the-art E2VID+ and FireNet+ architectures [7] on sequences from the ECD [35] and HQF [7] datasets. Local histogram equalization not used for this comparison.

outperform the state-of-the-art EV-FlowNet trained with either frames [10] or synthetic ground truth [7] according to this metric. More interestingly, FireFlowNet outperforms our EV-FlowNet in both datasets. A qualitative evaluation of our FlowNet architectures can be found in the supplementary material.

4.2. Reconstruction Evaluation

We evaluated the accuracy of our ReconNet architectures against the DAVIS240C [38] frames from the ECD [35] and HQF [7] datasets, and compared their performance to the state-of-the-art of image reconstruction networks trained with ground-truth supervision: E2VID [4], FireNet [5], E2VID+ [7], and FireNet+ [7]. Super resolution and adversarial methods are not considered in this comparison. We used the results and code provided by Stoffregen, Scheerlinck *et al.* [7] for the quantitative and qualitative evaluations. The subscripts F and E indicate whether our networks were trained together with FireFlowNet or EV-FlowNet.

For all methods, reconstructions were generated at each DAVIS frame timestamp. We first applied local histogram equalization [41] to both frames, and then computed mean squared error (MSE), structural similarity (SSIM) [39], and perceptual similarity (LPIPS) [40]. Only for this evaluation, instead of using a fixed number of input events, we used all the events *in between DAVIS frames*, thus generating image sets with the same number of frames as the ground truth. Quantitative results are presented in Table 5, and are supported by qualitative results in Figs. 4 and 5. Additional results can be found in the supplementary material.

Table 5: Quantitative evaluation of our ReconNet architectures on the ECD [35] and HQF [7] datasets. For each dataset, we report the mean MSE (↓), SSIM [39] (↑) and LPIPS [40] (↓). Best in bold; runner up underlined.

	ECD*			HQF		
	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS
E2VID [4]	0.08	0.54	0.37	0.14	0.46	0.45
FireNet [4]	0.06	<u>0.57</u>	<u>0.29</u>	0.07	<u>0.48</u>	0.42
E2VID+ [7]	**0.04**	**0.60**	**0.27**	**0.03**	**0.57**	**0.26**
FireNet+ [7]	<u>0.06</u>	0.51	0.32	<u>0.05</u>	0.47	<u>0.36</u>
E2VID$_F$ (Ours)	0.07	0.52	0.38	0.07	0.44	0.47
E2VID$_E$ (Ours)	0.06	0.55	0.37	0.06	0.48	0.47
FireNet$_F$ (Ours)	0.06	0.52	0.38	0.06	0.46	0.47
FireNet$_E$ (Ours)	0.06	0.51	0.41	0.06	0.46	0.51

*Sequence cuts in the supplementary material.

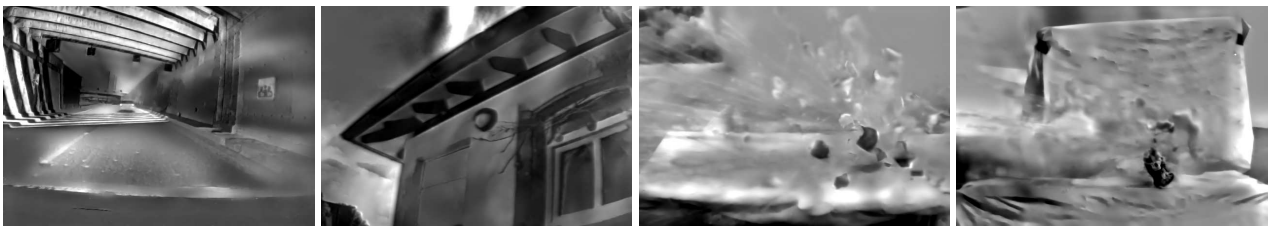

Figure 5: Qualitative results of our E2VID$_{\text{E}}$ on sequences from the High Speed and HDR Dataset [4].

Figure 6: Common failure cases of our SSL framework, namely motion blur in case of suboptimal optical flow estimation (left), ghosting artifacts in large texture-less regions (center), and inconsistent reconstructions due to the lack of information about the initial brightness L_0 (right).

Despite not using any ground-truth data during training, results show that our method is in line with the state-of-the-art in terms of reconstruction accuracy. Quantitatively, the error metrics of all our ReconNet architectures closely resemble the results obtained with the original E2VID and FireNet, but the accuracy gap increases if compared against these same networks trained with the refined data augmentation mechanisms from Stoffregen, Scheerlinck *et al.* [7]. This gap is particularly notable in the LPIPS loss because these literature networks are specifically trained to maximize perceptual similarity to ground-truth frames. On the other hand, there is no major quantitative difference between the evaluated versions of ReconNet, regardless of their architecture or the accompanying flow network.

Qualitative results confirm that our method reconstructs high quality HDR images. However, it is possible to identify several differences with respect to the state-of-the-art. Firstly, our images appear less sharp. Our architectures learn to correlate the spatial gradients of the estimated brightness $\hat{L}$ to the averaged IWE (see Section 3.4). This entails that the reconstructed images are affected by the accuracy of the optical flow. Suboptimal optical flow estimations lead to imperfect event deblurring during training, which in turn is reflected in the reconstructed images as motion blur. Note that this blur diminishes when using an appropriate fixed number of input events for each sequence. Secondly, the dynamic range of the images differs. State-of-the-art methods learn to map the input events into bounded estimates of $\hat{L}$ via supervised learning. On the contrary, our brightness estimate is unbounded, and normalization is used to encode this signal as bounded images. Besides this, there is no significant difference between the evaluated ReconNet versions, despite the limited smoothing capabilities of FireNet. Lastly, although our method does not suffer from the stretch marks mostly present in FireNet+ images, it is characterized by three common failure cases. As shown in Fig. 6, these are: (i) the aforementioned motion blur, (ii) "ghosting" artifacts in large texture-less regions due to limited extrapolation of edge information, and (iii) incoherent reconstructions due to the lack of information about the initial brightness L_0.

5. Conclusion

In this paper, we went back to the basics of event cameras and presented the first self-supervised learning-based approach to event-based image reconstruction, which does not rely on any ground-truth or synthetic data during training. Instead, our SSL method makes use of the flow-intensity relation used by early methods to reconstruct the frames that best satisfy the input events and the estimated optical flow. Results confirm that our method performs almost as well as the state-of-the-art, but that the reconstructed images are characterized by several artifacts that need to be addressed by future work. Additionally, we presented FireFlowNet: a fast, lightweight neural network that performs event-based optical flow estimation. We believe this work shows the exciting potential of SSL to take over the research on image reconstruction from event data, and it opens up avenues for further improvement by leveraging the great amount of unlabeled event data available. Moreover, we have proposed a general self-supervised learning framework that can be extended in multiple ways via more sophisticated reconstruction losses and other event-based optical flow algorithms.

References

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2020.

[2] F. Paredes-Vallés, K. Y. W. Scheper, and G. C. H. E. De Croon, "Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 42, no. 8, pp. 2051–2064, 2020.

[3] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 3857–3866.

[4] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2019.

[5] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza, "Fast image reconstruction with an event camera," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020, pp. 156–163.

[6] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Conf. Robot Learn.*, 2018, pp. 969–982.

[7] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, T. Drummond, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the sim-to-real gap for event cameras," in *European Conf. Comput. Vis. (ECCV)*, 2020.

[8] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "Unsupervised event-based learning of optical flow, depth, and egomotion," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 989–997.

[9] G. Gallego, C. Forster, E. Mueggler, and D. Scaramuzza, "Event-based camera pose tracking using a generative event model," *arXiv:1510.01972*, 2015.

[10] A. Z. Zhu and L. Yuan, "EV-FlowNet: Self-supervised optical flow estimation for event-based cameras," in *Robot.: Science and Systems (RSS)*, 2018.

[11] H. Kim, A. Handa, R. Benosman, S.-H. Ieng, and A. J. Davison, "Simultaneous mosaicing and tracking with an event camera," *J. Solid-State Circ.*, vol. 43, pp. 566–576, 2008.

[12] H. Kim, S. Leutenegger, and A. J. Davison, "Real-time 3d reconstruction and 6-dof tracking with an event camera," in *European Conf. Comput. Vis. (ECCV)*. Springer, 2016, pp. 349–364.

[13] M. Cook, L. Gugelmann, F. Jug, C. Krautz, and A. Steger, "Interacting maps for fast visual interpretation," in *Int. Joint Cong. Neural Networks (IJCNN)*. IEEE, 2011, pp. 770–776.

[14] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous optical flow and intensity estimation from an event camera," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 884–892.

[15] C. Reinbacher, G. Graber, and T. Pock, "Real-time intensity-image reconstruction for event cameras using manifold regularisation," *Int. J. Comput. Vis.*, vol. 126, no. 12, pp. 1381–1393, 2018.

[16] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-time intensity estimation using event cameras," in *Asian Conf. Comput. Vis. (ACCV)*, December 2018, pp. 308–324.

[17] 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 *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2019, pp. 10 081–10 090.

[18] S. Pini, G. Borghi, and R. Vezzani, "Learn to see by events: Color frame synthesis from event and rgb cameras," in *Int. Joint Conf. Comput. Vis., Imaging and Comput. Graphics Theory and Appl. (VISIGRAPP)*, 2019.

[19] J. Choi, K.-J. Yoon *et al.*, "Learning to super resolve intensity images from events," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2768–2776.

[20] L. Wang, T.-K. Kim, and K.-J. Yoon, "Eventsr: From asynchronous events to image reconstruction, restoration, and super-resolution via end-to-end adversarial learning," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 8315–8325.

[21] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 db 15μs latency asynchronous temporal contrast vis. sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[22] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 3867–3876.

[23] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2018, pp. 1–9.

[24] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *IEEE Int. Conf. Image Process. (ICIP)*, vol. 2, 1994, pp. 168–172.

[25] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vis." in *Int. Joint Conf. on Artificial Intell. (IJCAI)*, 1981.

[26] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Process. Systems (NeurIPS)*, 2015, pp. 2017–2025.

[27] D. Gehrig, H. Rebecq, G. Gallego, and D. Scaramuzza, "EKLT: Asynchronous photometric feature tracking using events and frames," *Int. J. Comput. Vis.*, vol. 128, no. 3, pp. 601–618, 2020.

[28] J. Y. Jason, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *European Conf. Comput. Vis. (ECCV)*. Springer, 2016, pp. 3–10.

[29] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778.

[31] 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 Robot. and Autom. Lett. (RA-L)*, vol. 3, no. 3, pp. 2032–2039, 2018.

[32] N. Ballas, L. Yao, C. Pal, and A. Courville, "Delving deeper into convolutional networks for learning video representations," *Int. Conf. Learn. Representations (ICLR)*, 2015.

[33] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Process. Systems (NeurIPS)*, 2015, pp. 802–810.

[34] J. Delmerico, T. Cieslewski, H. Rebecq, M. Faessler, and D. Scaramuzza, "Are we ready for autonomous drone racing? The UZH-FPV drone racing dataset," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019, pp. 6713–6719.

[35] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *Int. J. Robot. Research*, vol. 36, no. 2, pp. 142–149, 2017.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Int. Conf. Learn. Representations (ICLR)*, 2014.

[37] D. de Jong, F. Paredes-Vallés, and G. C. de Croon, "How do neural networks estimate optical flow? A neuropsychology-inspired study," *arXiv:2004.09317*, 2020.

[38] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240$\times$180 130 db 3μs latency global shutter spatiotemporal vision sensor," *IEEE J. Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *IEEE Conf. on Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 586–595.

[41] G. Yadav, S. Maheshwari, and A. Agarwal, "Contrast limited adaptive histogram equalization based enhancement for real time video system," in *Int. Conf. Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2014, pp. 2392–2397.