Reducing the Sim-to-Real Gap for Event Cameras

Timo Stoffregen[⋆,1,3], Cedric Scheerlinck[⋆,2,3], Davide Scaramuzza[4], Tom Drummond[1,3], Nick Barnes[2,3], Lindsay Kleeman[1], and Robert Mahony[2,3]

[1] Dept. of ECSE, Monash University, Melbourne, Australia
[2] Australian National University, Canberra, Australia
[3] Australian Centre for Robotic Vision, Australia
[4] University of Zurich, Zurich, Switzerland

Abstract. Event cameras are paradigm-shifting novel sensors that report asynchronous, per-pixel brightness changes called 'events' with unparalleled low latency. This makes them ideal for high speed, high dynamic range scenes where conventional cameras would fail. Recent work has demonstrated impressive results using Convolutional Neural Networks (CNNs) for video reconstruction and optic flow with events. We present strategies for improving training data for event based CNNs that result in 20-40 % boost in performance of existing state-of-the-art (SOTA) video reconstruction networks retrained with our method, and up to 15 % for optic flow networks. A challenge in evaluating event based video reconstruction is lack of quality ground truth images in existing datasets. To address this, we present a new **High Quality Frames (HQF)** dataset, containing events and ground truth frames from a DAVIS240C that are well-exposed and minimally motion-blurred. We evaluate our method on HQF + several existing major event camera datasets.

Video, code and datasets: https://timostoff.github.io/20ecnn

1 Introduction

Event-based cameras such as the Dynamic Vision Sensor (DVS) [18] are novel, bio-inspired visual sensors. Presenting a paradigm-shift in visual data acquisition, pixels in an event camera operate by asynchronously and independently reporting intensity changes in the form of events, represented as a tuple of x, y location, timestamp t and polarity of the intensity change s. By moving away from fixed frame-rate sampling of conventional cameras, event cameras deliver several key advantages in terms of low power usage (in the region of 5 mW), high dynamic range (140 dB), low latency and timestamps with resolution on the order of µs.

⋆ Equal contribution.

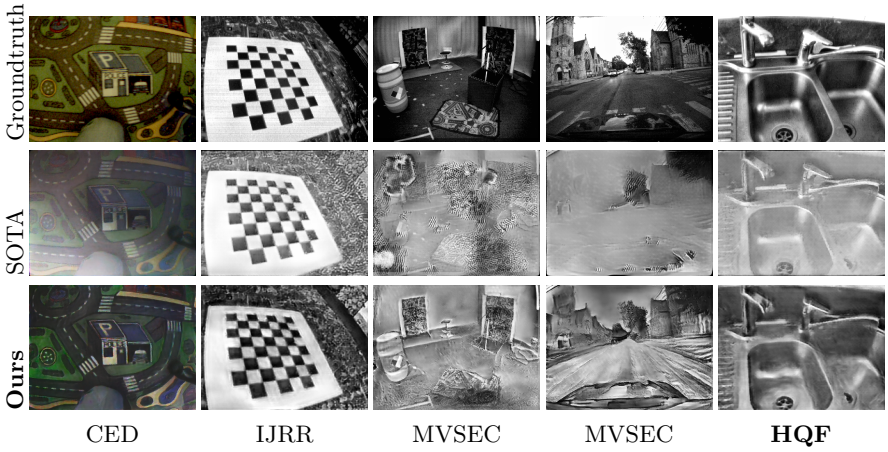


Fig. 1: Top: ground truth reference image. Middle/bottom: state-of-the-art E2VID [28] vs. our reconstructed images from events only. Challenging scenes from event camera datasets: CED [33], IJRR [23], MVSEC [42] and our **HQF** dataset.

With the recent preponderance of deep learning techniques in computer vision, the question of how to apply this technology to event data has been the subject of several recent works. Zhu *et al.* [43] propose an unsupervised network able to learn optic flow from real event data, while Rebecq *et al.* [27,28] showed that supervised networks trained on synthetic events transferred well to real event data. Simulation shows promise since data acquisition and ground truth are easily obtainable, in contrast to using real data. However, mismatch between synthetic and real data degrades performance, so a key challenge is simulating realistic data.

We generate training data that better matches real event camera data by analyzing the statistics of existing datasets to inform our choice of simulation parameters. A major finding is that the contrast threshold (CT) - the minimum change in brightness required to trigger an event - is a key simulation parameter that impacts performance of supervised CNNs. Further, we observe that the apparent contrast threshold of real event cameras varies greatly, even within one dataset. Previous works such as event based video reconstruction [28] choose contrast thresholds that work well for some datasets, but fail on others. Unsupervised networks trained on real data such as event based optic flow [43] may be retrained to match any real event camera - at the cost of new data collection and training. We show that using CT values for synthetic training data that are correctly matched to CTs of real datasets is a key driver in improving performance of retrained event based video reconstruction and optic flow networks across multiple datasets. We also propose a simple noise model which yields up to 10 % improvement when added during training.

A challenge in evaluating image and video reconstruction from events is lack of quality ground truth images registered and time-synchronized to events, because most existing datasets focus on scenarios where event cameras excel (high speed, HDR) and conventional cameras fail. To address this limitation, we introduce a new High Quality Frames (HQF) dataset that provides several sequences in well lit environments with minimal motion blur. These sequences are recorded with a DAVIS240C event camera that provides perfectly aligned frames from an integrated Active Pixel Sensor (APS). HQF also contains a diverse range of motions and scene types, including slow motion and pauses that are challenging for event based video reconstruction. We quantitatively evaluate our method on two major event camera datasets: IJRR [23] and MVSEC [42], in addition to our HQF, demonstrating gains of 20-40 % for video reconstruction and up to 15 % for optic flow when we retrain existing SOTA networks.

Contribution We present a method to generate synthetic training data that improves generalizability to real event data, guided by statistical analysis of existing datasets. We additionally propose a simple method for dynamic train-time noise augmentation that yields up to 10 % improvement for video reconstruction. Using our method, we retrain several network architectures from previously published works on video reconstruction [28, 32] and optic flow [43, 44] from events. We are able to show significant improvements that persist over architectures and tasks. Thus, we believe our findings will provide invaluable insight for others who wish to train models on synthetic events for a variety of tasks. We provide a new comprehensive High Quality Frames dataset targeting ground truth image frames for video reconstruction evaluation. Finally, we provide our data generation code, training set, training code and our pretrained models, together with dozens of useful helper scripts for the analysis of event-based datasets to make this task easier for fellow researchers.

In summary, our major contributions are:

- A method for simulating training data that yields 20 %-40 and up to 15 % improvement for event based video reconstruction and optic flow CNNs.
- Dynamic train-time event noise augmentation.
- A novel High Quality Frames dataset.
- Extensive analysis and evaluation of our method.
- An optic flow evaluation metric *Flow Warp Loss (FWL)*, tailored to event data, that does not require ground truth flow.
- Open-source code, training data and pretrained models.

The remainder of the paper is as follows. Section 2 reviews related works. Section 3 outlines our method for generating training data, training and evaluation, and introduces our HQF dataset. Section 4 presents experimental results on video reconstruction and optic flow. Section 5 discusses our major findings and concludes the paper.

2 Related Works

2.1 Video Reconstruction

Video and image reconstruction from events has been a popular topic in the event based vision literature. Several approaches have been proposed in recent years; Kim *et al.* [14] used an EKF to reconstruct images from a rotating event camera, later extending this approach to full 6-DOF camera motions [15]. Bardow *et al.* [2] used a sliding spatiotemporal window of events to simultaneously optimize both optic flow and intensity estimates using the primal-dual algorithm, although this method remains sensitive to hyperparameters. Reinbacher *et al.* [29] proposed direct integration with periodic manifold regularization on the Surface of Active Events (SAE [22]) to reconstruct video from events. Scheerlinck *et al.* [30,31] achieved computationally efficient, continuous-time video reconstruction via complementary and high-pass filtering. This approach can be combined with conventional frames, if available, to provide low frequency components of the image. However, if taken alone, this approach suffers from artifacts such as ghosting effects and bleeding edges.

Recently, convolutional neural networks (CNNs) have been brought to bear on the task of video reconstruction. Rebecq *et al.* [27,28] presented E2VID, a recurrent network that converts events (discretized into a voxel grid) to video. A temporal consistency loss based on [16] was introduced to reduce flickering artifacts in the video, due to small differences in the reconstruction of subsequent frames. E2VID is current state-of-the-art. Scheerlinck *et al.* were able to reduce model complexity by 99 % with the *FireNet* architecture [32], with only minor trade-offs in reconstruction quality, enabling high frequency inference.

2.2 Optic Flow

Since event based cameras are considered a good fit for applications involving motion [8], much work has been done on estimating optic flow with event cameras [1–4, 6, 10, 20, 35, 36]. Recently, Zhu *et al.* proposed a CNN (EV-FlowNet) for estimating optic flow from events [43], together with the Multi-Vehicle Stereo Event Camera (MVSEC) dataset [42] that contains ground truth optic flow estimated from depth and ego-motion sensors. The input to EV-FlowNet is a 4-channel image formed by recording event counts and the most recent timestamps for negative and positive events. The loss imposed on EV-FlowNet was an image-warping loss [13] that took photometric error between subsequent APS frames registered using the predicted flow. A similar approach was taken by Ye *et al.* [39], in a network that estimated depth and camera pose to calculate optic flow. In [44], Zhu *et al.* improved on prior work by replacing the image-warping loss with an event-warping loss that directly transports events to a reference time using the predicted flow. We use a similar method to evaluate optic flow performance of several networks (see Section 4.1). Zhu *et al.* [44] also introduced a novel input representation based on event discretization that places events into bins with temporal bilinear interpolation to produce a voxel grid. EV-FlowNet was

trained on data from MVSEC [42] and Ye *et al.* [39] even trained, then validated on the same sequences; our results (Section 4.1) indicate that these networks suffer from overfitting.

2.3 Input Representations

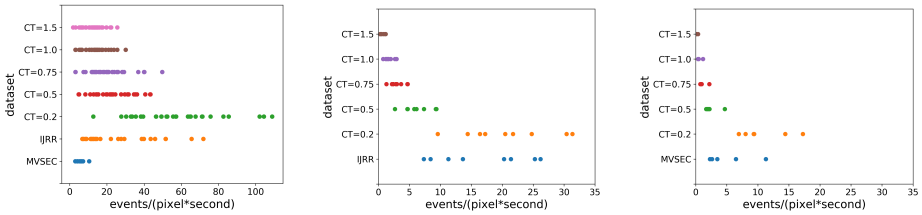
To use conventional CNNs, events must first be transformed into an amenable grid-based representation. While asynchronous spiking neural networks can process raw events and have been used for object recognition [17, 24, 25] and optic flow [3, 4], lack of appropriate hardware or effective error backpropagation techniques renders them yet uncompetitive with state-of-the-art CNNs. Several grid-based input representations for CNNs have been proposed: simple event images [21, 43] (events are accumulated to form an image), Surface of Active Events (SAE) [43] (latest timestamp recorded at each pixel), Histogram of Averaged Time Surfaces (HATS) [34] and even learned input representations, where events are sampled into a grid using convolutional kernels [12]. Zhu *et al.* [44] and Rebecq *et al.* [28] found best results using a voxel grid representation of events, where the temporal dimension is essentially discretized and subsequently binned into an n dimensional grid (eq. 1).

3 Method

3.1 Event Camera Contrast Threshold

In an ideal event camera, a pixel at (x, y) triggers an event e_i at time t_i when the brightness since the last event e_{i-1} at that pixel changes by a threshold C, given $t - t_{i-1} > r$, the refractory period of that pixel. C is referred to as the contrast threshold (CT) and can be typically adjusted in modern event cameras. In reality, the values for C are not constant in time nor homogeneous over the image plane nor is the positive threshold C_p necessarily equal to the negative threshold C_n. In simulation (e.g. using ESIM [26]), CTs are typically sampled from $\mathcal{N}(\mu{=}0.18, \sigma{=}0.03)$ to model this variation [12, 27, 28]. The CT is an important simulator parameter since it determines the number and distribution of events generated from a given scene.

While the real CTs of previously published datasets are unknown, one method to estimate CTs is via the proxy measurement of average events per pixel per second ($\frac{\text{events}}{\text{pix}\cdot\text{s}}$). Intuitively, higher CTs tend to reduce the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ for a given scene. While other methods of CT estimation exist (see supp. material), we found that tuning the simulator CTs to match $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of real data worked well. Since this measure is affected by scene dynamics (i.e. faster motions increase $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ independently of CT), we generated a diverse variety of realistic scene dynamics. The result of this experiment (Figure 2a) indicates that a contrast threshold setting of between 0.2 and 0.5 would be more appropriate for sequences from the IJRR dataset [23]. The larger diversity of motions is also apparent in the large spread of the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ compared to MVSEC [42] whose sequences are tightly clustered.



(a) IJRR/MVSEC vs. ESIM (b) IJRR vs ESIM (c) MVSEC vs ESIM

Fig. 2: Each dot represents a sequence from the given dataset (y-axis). (a) $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of IJRR and MVSEC vs. ESIM training datasets (CT 0.2-1.5) described in Section 3.2. (b) $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of IJRR vs. ESIM events simulated from IJRR *APS frames*. (c) $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of MVSEC vs. ESIM events simulated from MVSEC *APS frames*.

(a) IJRR (b) IJRR (c) MVSEC (d) MVSEC (e) HQF

Fig. 3: Note that in many sequences from the commonly used IJRR and MVSEC datasets, the accompanying APS frames are of low quality. The top row shows the APS frames, the bottom row overlays the events. As can be seen, many features are not visible in the APS frames, making quantitative evaluation difficult. This motivates our own High Quality Frames dataset (HQF).

As an alternative experiment to determine CTs of existing datasets, we measured the $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ of events simulated using the actual APS (ground truth) frames of IJRR and MVSEC sequences. Given high quality images with minimal motion blur and little displacement, events can be simulated through image interpolation and subtraction. Given an ideal image sequence, the simulator settings should be tunable to get the exact same $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ from simulation as from the real sensor. Unfortunately APS frames are not usually of a very high quality (Figure 3), so we were limited to using this approach on carefully curated snippets (Figure 4). The results of this experiment in Figure 2b and 2c indicate similar results of lower contrast thresholds for IJRR and higher for MVSEC, although accuracy is limited by the poor quality APS frames.

(a) Poorly exposed from IJRR and MVSEC (b) Well exposed from IJRR and MVSEC

Fig. 4: Examples of frames from IJRR and MVSEC after local histogram equalization, with poorly exposed sequences in 4a, and better exposed images in 4b.

3.2 Training Data

We used an event camera simulator, ESIM [26] to generate training sequences for our network. There are several modes of simulation available, of which we used "Multi-Object-2D" that facilitates moving images in simple 2D motions, restricted to translations, rotations and dilations over a planar background. This generates sequences reminiscent of Flying Chairs [7], where objects move across the screen at varying velocities. In our generation scheme, we randomly selected images from COCO [19] and gave them random trajectories over the image plane. Our dataset contains 280 sequences, 10 s in length. Sequences alternate between four archetypal scenes; slow motion with 0-5 foreground objects, medium speed motion with 5-10 foreground objects, fast speed with 5-20 foreground objects and finally, full variety of motions with 10-30 foreground objects. This variety encourages networks to generalize to arbitrary real world camera motions, since a wide range of scene dynamics are presented during training. Sequences were generated with contrast thresholds (CTs) between 0.1 and 1.5 in ascending order. Since real event cameras do not usually have perfectly balanced positive and negative thresholds, the positive threshold $\mathrm{C_p} = \mathrm{C_n} \cdot x, x \in \mathcal{N}(\mu{=}1.0, \sigma{=}0.1)$.

The events thus generated were discretized into a voxel grid representation. In order to ensure synchronicity with the ground truth frames of our training set and later with the ground truth frames of our validation set, we always take all events between two frames to generate a voxel grid. Given N events $e_i = \{x_i, y_i, t_i, s_i\}_{i=0,\dots,N}$ spanning $\Delta_T = t_N - t_0$ seconds, a voxel grid V with B bins can be formed through temporal bilinear interpolation via

$$V_{k\in[0,B-1]} = \sum_{i=0}^{N} s_i \max(0, 1 - |t_i^* - k|) \tag{1}$$

where t_i^* is the timestamp normalized to the range $[0, B-1]$ via $t_i^* = \frac{t_i - t_0}{\Delta_T}(B-1)$ and the bins are evenly spaced over the range $[t_0, t_N]$. This method of forming voxels has some limitations; it is easy to see that the density of the voxels can vary greatly, depending on the camera motion and frame rate of the camera. Thus, it is important to train the network on a large range of event rates $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ and voxel densities. During inference, other strategies of voxel generation can be employed, as further discussed in the supplementary materials. We used $B = 5$

throughout the experiments in this paper. In earlier experiments we found values of $B = 2, 5, 15$ produced no significant differences.

3.3 Sequence Length

To train recurrent networks, we sequentially passed L inputs to the network and computed the loss for each output. Finally, the losses were summed and a backpropagation update was performed based on the gradient of the final loss with respect to the network weights. Since recurrent units in the network are initialized to zero, lower values of L restrict the temporal support that the recurrent units see at train time. To investigate the impact of sequence length L, we retrain our networks using $L = 40$ (as in E2VID [28]) and $L = 120$. In the case of non-recurrent networks such as EV-FlowNet [43, 44], we ignore the sequence length parameter.

3.4 Loss

For our primary video reconstruction loss function we used "learned perceptual image patch similarity" (LPIPS) [41]. LPIPS is a fully differentiable similarity metric between two images that compares hidden layer activations of a pre-trained network (e.g. Alex-Net or VGG), and was shown to better match human judgment of image similarity than photometric error or SSIM [38]. Since our event tensors were synchronized to the ground truth image frames by design (the final event in the tensor matches the frame timestamp), we computed the LPIPS distance between our reconstruction and the corresponding ground truth frame. As recommended by the authors [41], we used the Alex-Net variant of LPIPS. We additionally imposed a temporal consistency loss [16] that measures photometric error between consecutive images after registration based on optic flow, subject to an occlusion mask. For optic flow, we used the L1 distance between our prediction and ground truth as the training loss.

3.5 Data Augmentation

During training, Rebecq *et al.* [28] occasionally set the input events to zero and performed a forward-pass step within a sequence, using the previous ground truth image frame to compute the loss. The probability of initiating a pause when the sequence is running $P(p|r) = 0.05$, while the probability of maintaining the paused state when the sequence is already paused $P(p|p) = 0.9$ to encourage occasional long pauses. This encourages the recurrent units of the network to learn to 'preserve' the output image in absence of new events. We used pause augmentation to train all recurrent networks.

Event cameras provide a noisy measurement of brightness change, subject to background noise, refractory period after an event and hot pixels that fire many spurious events. To simulate real event data, we applied a refractory period of 1ms. At train time, for each sequence of L input event tensors we optionally

added zero-mean Gaussian noise ($\mathcal{N}(\mu$=0, σ=0.1)) to the event tensor to simulate uncorrelated background noise, and randomly elected a few 'hot' pixels. The number of hot pixels was drawn from a uniform distribution from 0 to 0.0001, multiplied by the total number of pixels. Hot pixels have a random value ($\mathcal{N}(\mu$=0, σ=0.1)) added to every temporal bin in each event tensor within a sequence. To determine whether augmenting the training data with noise benefits performance on real data, we retrained several models with and without noise (Table 5).

3.6 Architecture

To isolate the impact of our method from choice of network architecture, we retrained state-of-the-art (SOTA) video reconstruction network E2VID [28] and SOTA optic flow network EV-FlowNet described in [43,44]. Thus, differences in performance for each task are not due to architecture. Additionally, we aim to show that our method generalizes to multiple architectures. While we believe architecture search may further improve results, it is outside the scope of this paper.

3.7 High Quality Frames Dataset

To evaluate event camera image reconstruction methods, we compared reconstructed images to temporally synchronized, registered ground truth reference images. Event cameras such as the DAVIS [5] can capture image frames (in addition to events) that are timestamped and registered to the events, that may serve as ground truth. Previous event camera datasets such as IJRR [23] and MVSEC [42] contain limited high quality DAVIS frames, while many frames are motion-blurred and or under/overexposed (Figure 3). As a result, Rebecq *et al.* [28] manually rejected poor quality frames, evaluating on a smaller subset of IJRR.

We present a new High Quality Frames dataset (HQF) aimed at providing ground truth DAVIS frames that are minimally motion-blurred and well exposed. In addition, our HQF covers a wider range of motions and scene types than the evaluation dataset used for E2VID, including: static/dynamic camera motion vs. dynamic camera only, very slow to fast vs. medium to fast and indoor/outdoor vs. indoor only. To record HQF, we used two different DAVIS240C sensors to capture data with different noise/CT characteristics. We used default bias settings loaded by the RPG DVS ROS driver[5], and set exposure to either auto or fixed to maximize frame quality. Our HQF provides temporally synchronized, registered events and DAVIS frames (further details in supplementaries, Table 6).

[5] https://github.com/uzh-rpg/rpg_dvs_ros

4 Experiments

4.1 Evaluation

We evaluated our method by retraining two state-of-the-art event camera neural networks: E2VID [27, 28], and EV-FlowNet [43, 44]. Our method outperforms previous state-of-the-art in image reconstruction and optic flow on several publicly available event camera datasets including IJRR [23] and MVSEC [42], and our new High Quality Frames dataset (HQF, Section 3.7).

For video reconstruction on the datasets HQF, IJRR and MVSEC (Table 1) we obtained a 40 %, 20 % and 28 % improvement over E2VID [28] respectively, using LPIPS. For optic flow we obtained a 12.5 %, 10 % and 16 % improvement over EV-FlowNet [43] on flow warp loss (FWL, eq. 3). Notably, EV-FlowNet was trained on MVSEC data (`outdoor_day2` sequence), while ours was trained entirely on synthetic data, demonstrating the ability of our method to generalize to real event data.

Image As in [28] we compared our reconstructed images to ground truth (DAVIS frames) on three metrics; mean squared error (MSE), structural similarity [38] (SSIM) and perceptual loss [40] (LPIPS) that uses distance in the latent space of a pretrained deep network to quantify image similarity.

Since many of these datasets show scenes that are challenging for conventional cameras, we carefully selected sections of those sequences where frames appeared to be of higher quality (less blurred, better exposure etc.). The exact cut times of the IJRR and MVSEC sequences can be found in the supplementary materials. However, we were also ultimately motivated to record our own dataset of high quality frames (HQF, Section 3.7) of which we evaluated the entire sequence.

Flow A warping loss (similar to [11]) was used as a proxy measure of accuracy as it doesn't require ground truth flow. Events $E = (x_i, y_i, t_i, s_i)_{i=1,\dots,N}$ are warped by per-pixel optical flow $\phi = (u(x, y), v(x, y))^T$ to a reference time t' via

$$I(E, \phi) = \begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (t' - t_i) \begin{pmatrix} u(x_i, y_i) \\ v(x_i, y_i) \end{pmatrix}. \tag{2}$$

The resulting image I becomes sharper if the flow is correct, as events are motion compensated. Sharpness can be evaluated using the variance of the image $\sigma^2(I)$ [9,37], where a higher value indicates a better flow estimate. Since image variance $\sigma^2(I)$ depends on scene structure and camera parameters, we normalize by the variance of the unwarped event image $I(E, 0)$ to obtain the *Flow Warp Loss (FWL)*:

$$\text{FWL} := \frac{\sigma^2(I(E, \phi))}{\sigma^2(I(E, 0))}. \tag{3}$$

FWL < 1 implies the flow is worse than a baseline of zero flow. FWL enables evaluation on datasets without ground truth optic flow. While we used ground

Table 1: Comparison of state-of-the-art methods of video reconstruction and optic flow to networks trained using our dataset on HQF, IJRR and MVSEC. Best in bold.

Sequence	MSE		SSIM		LPIPS		FWL	
	E2VID	Ours	E2VID	Ours	E2VID	Ours	EVFlow	Ours
HQF								
bike_bay_hdr	0.16	**0.03**	0.41	**0.52**	0.51	**0.30**	1.22	**1.23**
boxes	0.11	**0.03**	0.50	**0.59**	0.38	**0.26**	1.75	**1.80**
desk_6k	0.15	**0.03**	0.51	**0.60**	0.39	**0.22**	1.23	**1.35**
desk_fast	0.12	**0.04**	0.54	**0.61**	0.40	**0.25**	1.43	**1.50**
desk_hand_only	0.12	**0.05**	0.53	**0.57**	0.63	**0.39**	**0.95**	0.85
desk_slow	0.16	**0.04**	0.53	**0.62**	0.47	**0.25**	1.01	**1.08**
engineering_posters	0.13	**0.03**	0.42	**0.57**	0.47	**0.26**	1.50	**1.65**
high_texture_plants	0.16	**0.03**	0.37	**0.65**	0.38	**0.14**	0.13	**1.68**
poster_pillar_1	0.14	**0.03**	0.38	**0.50**	0.54	**0.27**	1.20	**1.24**
poster_pillar_2	0.15	**0.04**	0.40	**0.47**	0.56	**0.26**	**1.16**	0.96
reflective_materials	0.13	**0.03**	0.44	**0.55**	0.44	**0.28**	1.45	**1.57**
slow_and_fast_desk	0.16	**0.03**	0.48	**0.62**	0.45	**0.25**	0.93	**0.99**
slow_hand	0.18	**0.04**	0.41	**0.57**	0.57	**0.30**	**1.64**	1.56
still_life	0.09	**0.03**	0.51	**0.63**	0.35	**0.22**	1.93	**1.98**
Mean	0.14	**0.03**	0.46	**0.58**	0.46	**0.26**	1.20	**1.35**
IJRR								
boxes_6dof_cut	**0.04**	0.04	0.63	**0.64**	0.29	**0.25**	1.42	**1.46**
calibration_cut	0.07	**0.03**	0.61	**0.62**	0.22	**0.18**	1.20	**1.31**
dynamic_6dof_cut	0.17	**0.05**	0.45	**0.53**	0.38	**0.27**	1.37	**1.39**
office_zigzag_cut	0.07	**0.04**	0.49	**0.51**	0.31	**0.26**	**1.13**	1.11
poster_6dof_cut	0.07	**0.03**	0.60	**0.66**	0.26	**0.19**	1.50	**1.56**
shapes_6dof_cut	0.03	**0.02**	**0.80**	0.77	0.26	**0.22**	1.15	**1.57**
slider_depth_cut	0.08	**0.03**	0.54	**0.62**	0.35	**0.24**	1.73	**2.17**
Mean	0.07	**0.03**	0.61	**0.64**	0.28	**0.22**	1.32	**1.45**
MVSEC								
indoor_flying1_data_cut	0.25	**0.08**	0.19	**0.36**	0.72	**0.45**	1.02	**1.14**
indoor_flying2_data_cut	0.23	**0.09**	0.18	**0.36**	0.71	**0.45**	1.13	**1.36**
indoor_flying3_data_cut	0.25	**0.09**	0.18	**0.37**	0.73	**0.44**	1.06	**1.23**
indoor_flying4_data_cut	0.21	**0.08**	0.23	**0.36**	0.72	**0.45**	1.24	**1.50**
outdoor_day1_data_cut	0.32	**0.13**	0.31	**0.34**	0.66	**0.52**	1.15	**1.27**
outdoor_day2_data_cut*	0.30	**0.10**	0.29	**0.34**	0.57	**0.43**	**1.21**	1.20
Mean	0.29	**0.11**	0.27	**0.35**	0.65	**0.47**	1.12	**1.30**

*Removed from mean tally for EV-FlowNet, as this sequence is part of the training set.

Table 2: Comparison of various methods to optic flow estimated from Lidar depth and ego-motion sensors [42]. The average-endpoint-error to the Lidar estimate (AEE) and the percentage of pixels with AEE above 3 and greater than 5 % of the magnitude of the flow vector ($\%_{\text{Outlier}}$) are presented for each method (lower is better, best in bold). Zeros shows the baseline error of zero flow. Additional works are compared in Table 9 which can be found in the supplementary materials.

Dataset	outdoor_day1		outdoor_day2		indoor_flying1		indoor_flying2		indoor_flying3	
	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$	AEE	$\%_{\text{Outlier}}$
Zeros	4.31	0.39	1.07	**0.91**	1.10	**1.00**	1.74	**0.89**	1.50	**0.94**
EVFlow [43]	**0.49**	**0.20**	-	-	1.03	2.20	1.72	15.10	1.53	11.90
Ours	0.68	0.99	**0.82**	0.96	**0.56**	**1.00**	**0.66**	1.00	**0.59**	1.00

truth from the simulator during training, we evaluated on real data using FWL (Table 1). We believe training on ground truth (L1 loss) rather than FWL encourages dense flow predictions.

Table 2 shows average endpoint error (AEE) of optic flow on MVSEC [42]. MVSEC provides optic flow estimates computed from lidar depth and ego motion sensors as 'ground truth', allowing us to evaluate average endpoint error (AEE) using code provided in [43]. However, lidar + ego motion derived ground truth is subject to sensor noise, thus, AEE may be an unreliable metric on MVSEC. For example, predicting zero flow achieves near state-of-the-art in some cases on MVSEC using AEE, though not with our proposed metric FWL (by construction, predicting zero flow yields FWL = 1.0).

4.2 Contrast Thresholds

We investigated the impact of simulator contrast threshold (CT, see Section 3.1) by retraining several networks on simulated datasets with CTs ranging from 0.2 to 1.5. Each dataset contained the same sequences, differing only in CT. Table 3 shows that for reconstruction (evaluated on LPIPS), IJRR is best on a lower CT ≈ 0.2, while MVSEC is best on high CT ≈ 1.0. Best or runner up performance was achieved when a wide range of CTs was used, indicating that exposing a network to additional event statistics outside the inference domain is not harmful, and may be beneficial. We believe training with low CTs (thus higher $\frac{\text{events}}{\text{pix}\cdot\text{s}}$) reduces dynamic range in the output images (Table 4), perhaps because the network becomes accustomed to a high density of events during training but is presented with lower $\frac{\text{events}}{\text{pix}\cdot\text{s}}$ data at inference. When retraining the original E2VID network, dynamic range increases with CTs (Table 4).

4.3 Training Noise and Sequence Length

To determine the impact of sequence length and noise augmentation during training, we retrained E2VID architecture using sequence length 40 (L40) and

Table 3: Evaluation of image reconstruction and optic flow networks trained on simulated datasets with a variety of contrast thresholds (CTs) from 0.2 to 1.5. 'All' is a dataset containing the full range of CTs from 0.2 to 1.5. All networks are trained for 200 epochs and evaluated on datasets HQF (excluding `desk_hand_only` on FWL), IJRR [23], MVSEC [42]. We report mean squared error (MSE), structural similarity (SSIM) [38] and perceptual loss (LPIPS) [41] for reconstruction and FWL for optic flow. Key: **best** | *second best*.

Contrast threshold	HQF				IJRR				MVSEC			
	MSE	SSIM	LPIPS	*FWL*	MSE	SSIM	LPIPS	*FWL*	MSE	SSIM	LPIPS	FWL
0.20	0.05	0.50	0.38	1.93	**0.04**	**0.60**	**0.25**	*1.45*	0.10	**0.35**	0.55	1.15
0.50	**0.04**	*0.51*	0.36	1.90	0.04	0.57	0.27	1.42	*0.10*	0.31	0.52	1.19
0.75	0.05	**0.51**	**0.36**	1.90	0.05	0.56	0.28	1.44	0.11	0.29	0.53	*1.22*
1.00	0.05	0.48	0.36	1.91	0.05	0.53	0.29	1.42	0.12	0.27	*0.51*	1.18
1.50	0.05	0.47	0.38	*1.93*	0.06	0.52	0.30	1.44	0.09	0.30	0.52	1.14
All	*0.05*	0.50	**0.36**	**1.96**	**0.04**	*0.59*	*0.27*	**1.46**	**0.08**	*0.34*	**0.51**	**1.24**

Table 4: Dynamic range of reconstructed images from IJRR [23]: original E2VID [28] versus E2VID retrained on simulated datasets covering a range of contrast thresholds CTs. We report the mean dynamic range of the 10th-90th percentile of pixel values.

	Original [28]	Retrained					
Contrast threshold	∼0.18	0.2	0.5	0.75	1.0	1.5	All
Dynamic range	77.3	89.2	103.7	105.9	104.8	100.0	103.3

120 (L120), with and without noise augmentation (N) (see Table 5). Increasing sequence length from 40 to 120 didn't impact results significantly. Noise augmentation during training improved performance of L40 models by ∼ 5-10 %, while giving mixed results on different datasets for L120 models. Qualitatively, adding more noise encourages networks to smooth outputs, while less noise may encourage the network to 'reconstruct' noise events, resulting in artifacts (Figure 1) observed in E2VID [28] (trained without noise).

5 Discussion

The significant improvements gained by training models on our synthetic dataset exemplify the importance of reducing the sim-to-real gap for event cameras in both the event rate induced by varying the contrast thresholds and the dynamics of the simulation scenes. Our results are quite clear on this, with consistent improvements across tasks (reconstruction and optic flow) and architectures (recurrent networks like E2VID, and U-Net based flow estimators) of up to 40 %.

We believe this highlights the importance for researchers to pay attention to the properties of the events they are training on; are the settings of the camera

Table 5: Mean LPIPS [41] on our HQF dataset, IJRR [23] and MVSEC [42], for various training hyperparameter configurations. E2VID architecture re-trained from scratch in all experiments. Key: L40/L120=sequence length 40/120, N=noise augmentation during training.

Model	HQF			IJRR			MVSEC		
	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS	MSE	SSIM	LPIPS
L40	0.044	**0.583**	0.296	0.042	**0.650**	0.229	0.151	0.330	0.526
L40N	**0.033**	0.579	**0.256**	**0.034**	0.636	**0.224**	0.105	**0.346**	**0.467**
L120	0.040	0.544	0.279	0.038	0.619	0.237	0.132	0.311	0.478
L120N	0.036	0.547	0.290	0.040	0.608	0.241	**0.099**	0.344	0.498

or simulator such that they are generating more or less events? Are the scenes they are recording representative of the wide range of scenes that are likely to be encountered during inference?

In particular, it seems that previous works have inadvertently overfit their models to the events found in the chosen target dataset. EV-FlowNet performs better on sequences whose dynamics are similar to the slow, steady scenes in MVSEC used for training, examples being `poster_pillar_2` or `desk_slow` from HQF that feature long pauses and slow motions, where EV-FlowNet is on par or better than ours. For researchers looking to use an off-the-shelf pretrained network, our model may be a better fit, since it targets a greater variety of sensors and scenes. A further advantage of our model that is not reflected in the FWL metric, is that training in simulation allows our model to predict *dense* flow (see supp. material), a challenge for prior self-supervised methods.

Similarly, our results speak for themselves on image reconstruction. While we outperform E2VID [28] on all datasets, the smallest gap is on IJRR, the dataset we found to have lower CTs. E2VID performs worst on MVSEC that contains higher CTs, consistent with our finding that performance is driven by similarity between training and evaluation event data.

In conclusion, future networks trained with synthetic data from ESIM or other simulators should take care to ensure the statistics of their synthetic data match the final use-case, using large ranges of CT values and appropriate noise and pause augmentation in order to ensure generalized models.

Acknowledgments

This work was supported by the Australian Government Research Training Program Scholarship and the Australian Research Council through the "Australian Centre of Excellence for Robotic Vision" under Grant CE140100016.

References

1. Mohammed Mutlaq Almatrafi and Keigo Hirakawa. DAViS camera optical flow. IEEE Trans. Comput. Imaging, pages 1–11, 2019.
2. Patrick Bardow, Andrew J. Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), pages 884–892, 2016.
3. Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. IEEE Trans. Neural Netw. Learn. Syst., 25(2):407–417, 2014.
4. Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi, and Mandyam Srinivasan. Asynchronous frameless event-based optical flow. Neural Netw., 27:32–37, 2012.
5. Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240x180 130dB 3us latency global shutter spatiotemporal vision sensor. IEEE J. Solid-State Circuits, 49(10):2333–2341, 2014.
6. Tobias Brosch, Stephan Tschechne, and Heiko Neumann. On event-based optical flow detection. Front. Neurosci., 9, Apr. 2015.
7. Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In Int. Conf. Comput. Vis. (ICCV), pages 2758–2766, 2015.
8. Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. arXiv e-prints, abs/1904.08405, 2019.
9. Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vision. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.
10. Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), pages 3867–3876, 2018.
11. Guillermo Gallego and Davide Scaramuzza. Accurate angular velocity estimation with an event camera. IEEE Robot. Autom. Lett., 2(2):632–639, 2017.
12. Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In Int. Conf. Comput. Vis. (ICCV), 2019.
13. Yu Jason, Harley Adam, and Derpanis Konstantinos. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. 2016.
14. Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew J. Davison. Simultaneous mosaicing and tracking with an event camera. In British Mach. Vis. Conf. (BMVC), 2014.
15. Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In Eur. Conf. Comput. Vis. (ECCV), pages 349–364, 2016.
16. Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In Eur. Conf. Comput. Vis. (ECCV), 2018.
17. Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. Front. Neurosci., 10:508, 2016.

18. Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15 μs latency asynchronous temporal contrast vision sensor. IEEE J. Solid-State Circuits, 43(2):566–576, 2008.
19. Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In Eur. Conf. Comput. Vis. (ECCV), 2014.
20. Min Liu and Tobi Delbruck. Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. In British Mach. Vis. Conf. (BMVC), 2018.
21. Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), pages 5419–5427, 2018.
22. Elias Mueggler, Christian Forster, Nathan Baumli, Guillermo Gallego, and Davide Scaramuzza. Lifetime estimation of events from dynamic vision sensors. In IEEE Int. Conf. Robot. Autom. (ICRA), pages 4874–4881, 2015.
23. Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM. Int. J. Robot. Research, 36(2):142–149, 2017.
24. Garrick Orchard, Cedric Meyer, Ralph Etienne-Cummings, Christoph Posch, Nitish Thakor, and Ryad Benosman. HFirst: A temporal approach to object recognition. IEEE Trans. Pattern Anal. Mach. Intell., 37(10):2028–2040, 2015.
25. José A. Perez-Carrasco, Bo Zhao, Carmen Serrano, Begoña Acha, Teresa Serrano-Gotarredona, Shouchun Chen, and Bernabé Linares-Barranco. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward ConvNets. IEEE Trans. Pattern Anal. Mach. Intell., 35(11):2706–2719, Nov. 2013.
26. Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. ESIM: an open event camera simulator. In Conf. on Robot. Learning (CoRL), 2018.
27. Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.
28. Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. IEEE Trans. Pattern Anal. Mach. Intell., 2020.
29. Christian Reinbacher, Gottfried Graber, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. In British Mach. Vis. Conf. (BMVC), 2016.
30. Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In Asian Conf. Comput. Vis. (ACCV), 2018.
31. Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Asynchronous spatial image convolutions for event cameras. IEEE Robot. Autom. Lett., 4(2):816–822, Apr. 2019.
32. Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In IEEE Winter Conf. Appl. Comput. Vis. (WACV), 2020.
33. Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. CED: color event camera dataset. In IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW), 2019.
34. Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, and Ryad Benosman. HATS: Histograms of averaged time surfaces for robust event-based

object classification. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), pages 1731–1740, 2018.
35. Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In Int. Conf. Comput. Vis. (ICCV), 2019.
36. Timo Stoffregen and Lindsay Kleeman. Simultaneous optical flow and segmentation (SOFAS) using Dynamic Vision Sensor. In Australasian Conf. Robot. Autom. (ACRA), 2017.
37. Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: an analysis. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.
38. Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process., 13(4):600–612, Apr. 2004.
39. Chengxi Ye, Anton Mitrokhin, Chethan Parameshwara, Cornelia Fermüller, James A. Yorke, and Yiannis Aloimonos. Unsupervised learning of dense optical flow and depth from sparse event data. arXiv e-prints, 2019.
40. Linguang Zhang and Szymon Rusinkiewicz. Learning to detect features in texture images. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2018.
41. Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2018.
42. Alex Zihao Zhu, Dinesh Thakur, Tolga Ozaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. IEEE Robot. Autom. Lett., 3(3):2032–2039, July 2018.
43. Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In Robotics: Science and Systems (RSS), 2018.
44. Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In IEEE Conf. Comput. Vis. Pattern Recog. (CVPR), 2019.