This paper has been accepted for publication at the
IEEE International Conference on Computer Vision (ICCV), Honolulu, 2025. ©IEEE

Unsupervised Joint Learning of Optical Flow and Intensity with Event Cameras

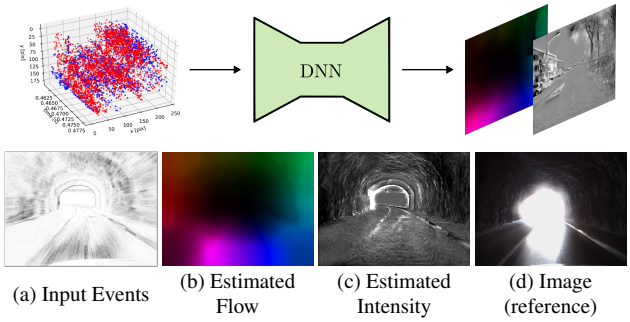
Shuang Guo[1], Friedhelm Hamann[1,2] and Guillermo Gallego[1,2,3].
[1] TU Berlin and Robotics Institute Germany,
[2] Science of Intelligence Excellence Cluster, [3] Einstein Center for Digital Future.

Abstract

Event cameras rely on motion to obtain information about scene appearance. This means that appearance and motion are inherently linked: either both are present and recorded in the event data, or neither is captured. Previous works treat the recovery of these two visual quantities as separate tasks, which does not fit with the above-mentioned nature of event cameras and overlooks the inherent relations between them. We propose an unsupervised learning framework that jointly estimates optical flow (motion) and image intensity (appearance) using a single network. From the data generation model, we newly derive the event-based photometric error as a function of optical flow and image intensity. This error is further combined with the contrast maximization framework to form a comprehensive loss function that provides proper constraints for both flow and intensity estimation. Exhaustive experiments show our method's state-of-the-art performance: in optical flow estimation, it reduces EPE by 20% and AE by 25% compared to unsupervised approaches, while delivering competitive intensity estimation results, particularly in high dynamic range scenarios. Our method also achieves shorter inference time than all other optical flow methods and many of the image reconstruction methods, while they output only one quantity. Project page: https://github.com/tub-rip/E2FAI

1. Introduction

Event cameras [18, 30] are novel bio-inspired vision sensors that offer attractive properties compared to traditional cameras: high temporal resolution, very high dynamic range (HDR), low power consumption and high pixel bandwidth, resulting in reduced motion blur. Hence, event cameras have a large potential for computer vision and robotics applications in challenging scenarios for traditional cameras, such as high speed motion and HDR illumination. However, novel methods are required to process the unconventional output of these sensors (a stream of asynchronous per-pixel brightness changes instead of conventional images) in order to unlock their potential [9].



(a) Input Events (b) Estimated Flow (c) Estimated Intensity (d) Image (reference)

Figure 1. *Our method* computes accurate optical flow and intensity images from event-camera data despite complex scenarios, fast motion and high dynamic range. The above result is obtained using data from interlaken_00_b sequence of DSEC [13].

In the past decade, a variety of computer vision algorithms have been developed to recover fundamental visual quantities from event streams [4, 9, 14, 50], such as EV-FlowNet [52] for optical flow estimation, and E2VID [31] for image intensity reconstruction. Despite achieving good performance, most of these methods are designed to estimate a single visual quantity: optical flow *or* image intensity. This philosophy does not fit well with the fact that, under constant illumination, motion and appearance are inherently entangled in the input data, since events are produced by *moving intensity patterns* on the image plane (e.g., Fig. 1a). More accurate and robust estimation becomes possible when the synergies between both visual quantities are properly leveraged.

To achieve this goal, we newly derive the event-based photometric error (PhE) as a function of optical flow and image intensity, and combine it with the state-of-the-art contrast maximization (CMax) framework [7], yielding a comprehensive loss function. Both PhE and CMax provide constraints on scene appearance and motion. The former focuses more on appearance (i.e., intensity) while the latter focuses more on motion (i.e., optical flow). Their complementary properties give rise to a well-behaved loss in both aspects. In addition, our loss function also includes a full consideration of the internal synergies between esti-

mated visual quantities: during training, the predicted flow is leveraged to warp the predicted image intensity to the adjacent interval so that a temporal consistency (TC) loss can be calculated, which significantly encourages consistency and robustness. We train a deep neural network (DNN) with this loss function in an unsupervised manner. The resulting model can predict precise optical flow and image intensity simultaneously from event data (see Fig. 1).

To the best of our knowledge, this paper presents the first unsupervised learning-based approach that jointly recovers optical flow and image intensity from event data, with a single network (Tab. 1). In the experiments, we evaluate our method in terms of optical flow and intensity on a variety of public datasets. For optical flow, our method achieves the best accuracy in the unsupervised learning category in the DSEC benchmark [13], with 20% and 25% improvements in terms of EPE and AE, respectively. For intensity, our method reports competitive results with respect to other unsupervised methods and even some supervised ones, especially in HDR scenarios. In terms of speed, our method reports minimal inference time to obtain both visual quantities. Our method robustly predicts precise flow and intensity on unseen data recorded in different scenarios and with different event cameras, thus demonstrating generalization.

Our contributions can be summarized as follows:

1. We propose the first unsupervised learning framework for the joint estimation, with a single network, of event-based optical flow and image intensity. Its working principle fits naturally with the characteristics of event data.
2. We derive event-based PhE, and combine it with CMax, yielding a comprehensive and well-behaved loss function for the estimation of motion and scene appearance. It can be directly adapted to various optimization and learning-based solutions for similar problems.
3. We conduct comprehensive experiments on public datasets, where our method shows state-of-the-art performance on optical flow, image intensity and inference time. Furthermore, our method shows excellent generalization on unseen data recorded in different scenarios (HDR and fast motion) and with different event cameras.

We hope the clear advantages of our approach and the source code provided will make its adoption appealing, thus bringing the tasks of optical flow estimation and intensity reconstruction closer than they currently are (until now, they have mostly been treated as independent problems).

2. Related Work

Given the high-speed and HDR properties of event cameras, extensive research has been carried out to utilize them for the estimation of optical flow and image intensity. Let us review the literature on each of these tasks in Sec. 2.1 and Sec. 2.2, respectively, and summarize the approaches that solve for both quantities in Sec. 2.3.

Method	Year	Type	DOFs	Joint	Loss function
IVM [5]	2011	MB	3	✓	"Consistency" of the data (Max. Likelihood)
SOFIE [1]	2016	MB	6	✓	Pixel-wise brightness change + TC
E-cGAN [25]	2021	SL	6	✗	Supervised: Error w.r.t. ground truth
BTEB [28]	2021	USL	6	✗	FlowNet: IWE sharpness; ReconNet: LEGM
Ours	2025	USL	6	✓	CMax + PhE of Flow and Intensity + TC

Table 1. *Event-based optical flow and intensity estimation methods.* The columns indicate: the method type, the number of degrees of freedom (DOFs) of the camera motion that the method can handle (3≡rotation, 6≡free motion), whether the method estimates flow and intensity jointly or separately, and the loss used.

2.1. Event-based Optical Flow Estimation

Previous works can be categorized into model-based (MB), supervised learning (SL) or unsupervised learning (USL). State-of-the-art MB methods [3, 39] formulate the problem using an objective function, and solve for optical flow through optimization. The most commonly-used objective is the sharpness of images of warped events (IWEs), such as the CMax loss [7, 8] used in [39], and the flow warping loss (FWL) [41] used in [3]. The optimization usually takes a number of iterations to converge, and it repeatedly warps all the involved events in every iteration, which makes these methods costly. Furthermore, the objectives derived from IWE sharpness may suffer from the issue of event collapse [36, 38], which affects flow accuracy and requires strong regularization to mitigate it.

Due to the better fitting capabilities and short inference times of DNNs, researchers have adopted them to compute optical flow. SL approaches [11, 13, 17, 19, 21, 45] train DNNs to learn the mapping from event data to the ground truth (GT) optical flow. However, the acquisition of GT relies on either simulation [17, 21] or calculation from external depth sensors using the motion field equation [13, 51]. Both are expensive. The former suffers from the sim-to-real gap, that is, the trained model may not perform well on real-world data, while the latter suffers from sensor inaccuracy and data sparsity [16, 39]. Although SL methods are generally ahead in optical flow benchmarks, the above issues about GT data are inherited by the trained model.

In contrast, USL methods [16, 29, 47, 52] forego costly data labeling, shifting the focus to the data and the loss function to learn optical flow patterns. Similarly to MB methods, current mainstream loss functions are still based on IWE sharpness [8]. For instance, EV-FlowNet [52] quantifies sharpness in terms of average timestamp images [22]. More recently, [16, 29, 39] adopted the CMax loss to train DNNs, which achieved obvious improvements with respect to MB methods that used the same loss functions. In addition, USL performs expensive event warping only during training, hence it has much shorter and more constant inference time than equivalent MB methods.

2.2. Event-based Image Intensity Reconstruction

Similarly, intensity estimation methods can be classified into MB, SL and USL categories. For the MB category, early works recovered brightness from events by means of temporal filtering [34] or temporal integration with manifold denoising [27]. A recent work [49] formulated this task as a linear inverse problem with image regularization and solved for intensity using ADMM [2]. However, it required accurate optical flow as input, which limits its applicability.

SL approaches [10, 31, 35, 44] trained DNNs using synthetic data to learn the mapping from event data to image intensity. In this way, the models were prevented from learning motion blur and under/over-exposure that happens in traditional cameras. Besides, these methods were aimed at video reconstruction and therefore adopted recurrent network architectures while introducing temporal consistency loss to guarantee the continuity of sequential images. However, they suffer from the sim-to-real gap and sometimes perform poorly in HDR scenarios (see, e.g., Fig. 4).

USL methods for intensity reconstruction are underexplored. Recent work [28] proposed to supervise the training with event-based photometric consistency. The loss was defined using the linearized event generation model (LEGM), as the error between brightness increments measured by event data and those obtained from the spatial gradient of the predicted image. However, this method requires accurate optical flow provided externally for both training and inference, for which the authors had to train a separate network to estimate optical flow (see details of the flow part in Sec. 2.3). The need to infer sequentially with two separate models takes more time and amplifies error propagation, thus significantly limiting practicality.

2.3. Event-based Flow-Intensity Estimation

Table 1 summarizes the main works that estimate both optical flow and image intensity with event cameras. The earliest work [5] proposed a joint estimation approach recovering optical flow, image intensity and camera ego-motion. However, it was limited to pure rotational motion. Bardow et al. [1] developed a variational algorithm that simultaneously optimized optical flow and image intensity. However, optical flow was only constrained by the brightness constancy between estimated intensity images, while the motion information in event data was not utilized. Hence, the resulting intensity and optical flow showed poor accuracy.

In the learning-based category, [25] presented an SL method for image reconstruction from events using conditional generative adversarial networks (cGANs), and also showed the applicability to predict depth and optical flow. The networks were trained and performed inference for each quantity separately, so the bonds between them were ignored. More recently, [28] proposed training an intensity reconstruction DNN (same architecture as [31]), but in an unsupervised manner. To do so, they separately trained a network for optical flow estimation using the same loss as [52], and subsequently fed its output flow into a second network for intensity reconstruction (see Sec. 2.2).

Our method overcomes and simplifies previous designs: it jointly estimates intensity and flow in an unsupervised manner by leveraging their synergies via the proposed loss function and architecture (i.e., using a single DNN).

3. Method

In this section, we first explain the preliminaries (Sec. 3.1), then introduce our proposed comprehensive loss function, including newly derived event-based PhE (Sec. 3.2), CMax loss (Sec. 3.3) and regularization terms (Sec. 3.4). Finally, we present the training pipeline in Sec. 3.5.

3.1. Preliminaries

Event Generation Model (EGM). Every pixel of an event camera independently measures brightness changes, generating an event $e_k \doteq (\mathbf{x}_k, t_k, p_k)$ when the logarithmic brightness change ΔL reaches a contrast threshold C [9]:

$$\Delta L \doteq L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t_k - \Delta t_k) = p_k C, \qquad (1)$$

where $p_k \in \{+1, -1\}$ indicates the polarity of the intensity increment, and Δt_k is the time elapsed since the previous event at the same pixel $\mathbf{x}_k$.

Event warping. Given a set of events $\mathcal{E} \doteq \{e_k\}_{k=1}^{N_e}$, we can warp them to a reference time t_{ref} with a motion model $\mathbf{W}$, yielding a set of warped events $\mathcal{E}'_{t_{\text{ref}}} \doteq \{e'_k\}_{k=1}^{N_e}$, where: $e_k \doteq (\mathbf{x}_k, t_k, p_k) \rightarrow e'_k \doteq (\mathbf{x}'_k, t_{\text{ref}}, p_k)$. Provided that the time span of $\mathcal{E}$ is small, we can assume all pixels move with constant but different velocities on the image plane, that is, the motion model $\mathbf{W}$ is given by:

$$\mathbf{x}'_k = \mathbf{x}_k + (t_{\text{ref}} - t_k) F(\mathbf{x}_k), \qquad (2)$$

where $F(\mathbf{x}_k)$ is the optical flow at $\mathbf{x}_k$ [39].

3.2. Event-based Photometric Error (PhE)

We leverage the original EGM to derive the event-based PhE. Starting from (1), we warp an event $e_k \doteq (\mathbf{x}_k, t_k, p_k)$ and its predecessor (at the same pixel) $e_{k-1} \doteq (\mathbf{x}_k, t_k - \Delta t_k, p_{k-1})$ to a reference time t_{ref}. According to (2), their warped locations on the camera plane at t_{ref} are:

$$\begin{aligned} \mathbf{x}'_k &= \mathbf{x}_k + (t_{\text{ref}} - t_k) F(\mathbf{x}_k), \\ \mathbf{x}'_{k-1} &= \mathbf{x}_k + (t_{\text{ref}} - (t_k - \Delta t_k)) F(\mathbf{x}_k). \end{aligned} \qquad (3)$$

Let the image intensity at t_{ref} be[1] $L(\mathbf{x})$, then we substitute (3) into (1) to obtain the PhE for the event e_k:

$$\epsilon_k \doteq (L(\mathbf{x}'_k) - L(\mathbf{x}'_{k-1})) - p_k C, \qquad (4)$$

[1] We omit "logarithmic" in the following text for simplification.

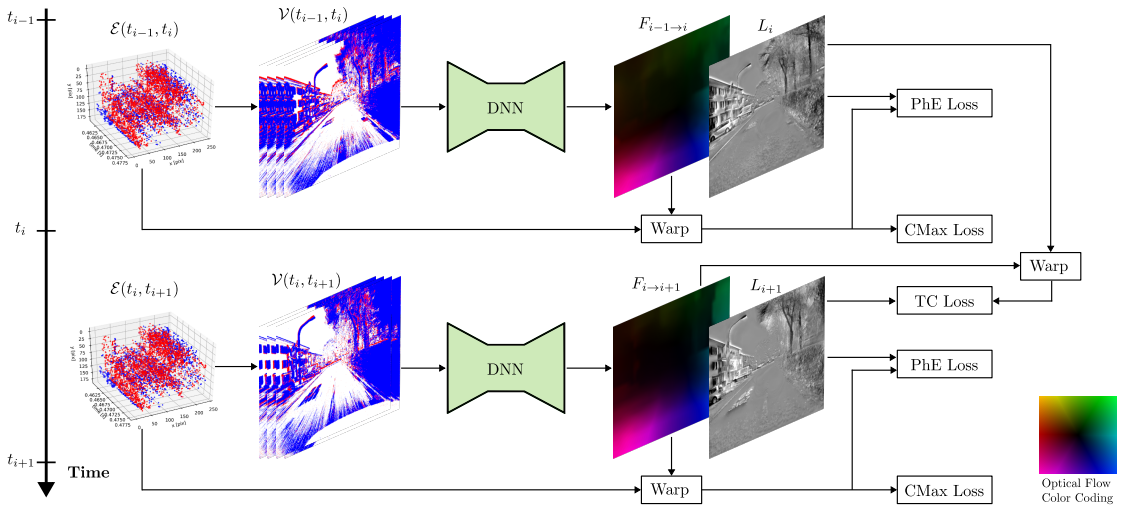


Figure 2. *The training pipeline of our proposed flow-intensity joint estimation method.* In every training step, two consecutive event data samples are input to the network, respectively. The CMax and PhE losses for each sample are calculated with the output optical flow and image intensity. The TC loss is defined by the photometric error between the predicted L_{i+1} and the one warped from L_i through $F_{i\to i+1}$.

which is the difference between the predicted intensity increment and the measured one. In practice, $\mathbf{x}'_k$ and $\mathbf{x}'_{k-1}$ have sub-pixel precision, so we use bilinear interpolation to compute $L(\mathbf{x}'_k)$ and $L(\mathbf{x}'_{k-1})$. Therefore, every PhE term (4) provides constraints for the intensity values at up to eight pixels around $\mathbf{x}'_k$ and $\mathbf{x}'_{k-1}$, as well as the optical flow value at one pixel $\mathbf{x}_k$. Finally, we sum the photometric error terms of all events in $\mathcal{E}$, obtaining the PhE loss:

$$\mathcal{L}_{\text{PhE}}(L, F) \doteq \frac{1}{N_e} \sum_{k=1}^{N_e} |\epsilon_k| \,. \tag{5}$$

It is an objective function of *both* intensity and flow, which opens the door to the joint estimation of both quantities. It is worth emphasizing that the PhE loss does not have the event collapse problem [36, 38] that CMax suffers from.

3.3. Contrast Maximization (CMax)

The CMax framework [7, 8] assumes events are triggered by moving edges, so that optical flow estimation can be determined by seeking the best motion compensation. Specifically, warped events $\mathcal{E}'_{t_{\text{ref}}}$ are aggregated on the image plane at t_{ref}, forming an image of warped events:

$$\text{IWE}(\mathbf{x}; \mathcal{E}'_{t_{\text{ref}}}, F) \doteq \sum_{k=1}^{N_e} \mathcal{N}\big(\mathbf{x};\, \mathbf{x}'_k,\, \sigma^2 = 1\text{px}\big), \tag{6}$$

where every pixel $\mathbf{x}$ counts its number of warped events.

In this work, we adopt the inverse of the L^1 magnitude of the gradient to quantify the CMax loss:

$$\mathcal{L}_{\text{CMax}}(F) \doteq 1 \Big/ \left(\frac{1}{|\Omega|} \int_\Omega \big\| \nabla \text{IWE}(\mathbf{x}) \big\|_1 \, d\mathbf{x} \right). \tag{7}$$

In contrast to PhE, the CMax loss recovers scene appearance in the form of a sharp edge map, which is the objective instead of the variable. The only optimizable variable is optical flow, which reflects that CMax loss focuses more on motion parameters, as mentioned in Sec. 1.

3.4. Regularization

Smoothness of Flow and Intensity. As stated in Sec. 3.2, the PhE loss only provides supervisory constraints on pixels that contain events (called valid pixels). To infer the values of the remaining pixels, regularization is needed; we use the total variation (TV) [33] to encourage smoothness of optical flow and intensity predictions: $\mathcal{L}_{\text{FTV}}(F) \doteq \frac{1}{|\Omega|} \int_\Omega \|\nabla F(\mathbf{x})\|_1 \, d\mathbf{x}$, and $\mathcal{L}_{\text{ITV}}(L) \doteq \frac{1}{|\Omega|} \int_\Omega \|\nabla L(\mathbf{x})\|_1 \, d\mathbf{x}$. Additionally, the smoothness of the flow mitigates the event collapse caused by the CMax loss.

Temporal Consistency (TC). A key advantage of joint estimation (over a separate one) is being able to leverage the synergies between the quantities for better consistency. We achieve this by establishing associations between temporally consecutive predictions. As depicted in Fig. 2, in each training step, the network takes as input two consecutive data samples to predict the corresponding optical flow and image intensity. We leverage flow $F_{i\to i+1}$ to transport L_i to t_{i+1}, yielding L'_{i+1}, whose photometric error with respect to L_{i+1} (predicted from the other data sample) is calculated to encourage temporal consistency (TC):

$$\mathcal{L}_{\text{TC}} \doteq \frac{1}{|\Omega|} \int_\Omega \big| L_{i+1}(\mathbf{x}) - \mathcal{W}(\mathbf{x}; L_i, F_{i\to i+1}) \big| \, d\mathbf{x}, \tag{8}$$

where $\mathcal{W}$ warps an image according to the optical flow. The introduction of the TC loss greatly improves the prediction quality, especially in image intensity (see Sec. 4.4).

3.5. Training Pipeline

An overview of the training pipeline is displayed in Fig. 2. During training, the loss terms (5), (7), etc. are evaluated with the predictions of each sample, while the TC loss (8), between consecutive samples. Consequently, the total loss is the weighted sum of all these terms:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{PhE}} + \lambda_2 \mathcal{L}_{\text{CMax}} + \lambda_3 \mathcal{L}_{\text{FTV}} + \lambda_4 \mathcal{L}_{\text{ITV}} + \lambda_5 \mathcal{L}_{\text{TC}},$$

where the first four terms are the sum for the two data samples (Fig. 2). For every sample, the reference time of CMax is set to a random number within the time span of the event set [16], that is, keeping the IWE sharp at any time, which helps reduce the possibility of event collapse, while that of PhE is always set to the end time of the sample. For inference, users just need to input one event voxel grid $\mathcal{V}(t_{i-1}, t_i)$ to predict flow $F_{i-1\to i}$ and intensity L_i.

4. Experiments

We begin by describing the experimental setup (Sec. 4.1), then present the results of optical flow evaluation (Sec. 4.2) and of image intensity evaluation (Sec. 4.3), before finally introducing the ablation study (Sec. 4.4).

4.1. Experimental setup

Datasets. Table 2 summarizes the datasets used in our experiments. The ECD dataset [26] is collected with a handheld DAVIS240C camera in indoor scenarios, where the motion speed varies from slow to fast. The DSEC dataset [13] features urban and highway driving scenarios in daylight and night. It is recorded with an on-board Prophesee Gen3 camera, and the GT optical flow is computed from the LiDAR disparity data. The HDR dataset [31] contains event data captured in HDR illumination conditions, such as the Sun and a car driving out of a tunnel. It highlights the HDR property of event cameras. The BS-ERGB dataset [42] provides high-resolution event data recorded by a handheld co-capture system composed of a Prophesee Gen4 camera (1 megapixel) and a FLIR RGB camera, in complex outdoor scenes. The high quality of the GT frames makes them suitable for the evaluation of event-based intensity reconstruction. For intensity evaluation, we use the EVREAL tool [6], and adopt its sequence selection. The sole change made is the removal of the last seconds of the may29_rooftop sequences from the evaluation, where the camera does not move and event data consists of pure noise.

Metrics. For optical flow, we adopt standard metrics: end-point error (EPE), angular error (AE) and %Out (the percentage of flow estimates whose EPE>3px). We also report the flow warp loss (FWL) proposed in [41]. For image intensity, we report full-reference metrics when GT images are available (e.g., BS-ERGB dataset): mean square error (MSE), structure similarity index (SSIM) [43] and perceptual similarity (LPIPS) [48]. Otherwise, we use no-reference metrics: BRISQUE [23], NIQE [24], MANIQA [46] (e.g., HDR dataset).

Dataset	Camera	Pixels	Scenarios & Features
ECD [26]	DAVIS240C	240×180	Indoor handheld, varying speed
DSEC [12]	Prophesee Gen3	640×480	Outdoor driving, daylight & night
HDR [31]	Samsung DVS Gen3	640×480	Indoor & outdoor, handheld & driving, HDR
BS-ERGB [42]	PSEE Gen4 & FLIR	970×625	Outdoor handheld motion, HDR

Table 2. *Configurations of the datasets used in the experiments*. The BS-ERGB data is cropped; its resolution is a bit smaller than the event camera resolution.

Implementation Details. We train our model only on the DSEC training split for 130 epochs with an AdamW optimizer [20], whose learning rate is 10^{-3} for the first 100 epochs and then decays to 10^{-4}. The training is carried out on four NVIDIA RTX A6000 GPUs with a total batch size of 24. This trained model is used for the evaluation of optical flow on the test split of DSEC and those of image intensity on BS-ERGB and HDR datasets, *without fine-tuning*. The weights of the loss terms are: $\lambda_1 = 30, \lambda_2 = 1, \lambda_3 = 10, \lambda_4 = 0.001, \lambda_5 = 1$. The contrast threshold C is set to 0.2. We adopt the classical U-Net architecture [32], which has 15 input channels (number of time bins in the event voxel grid [52]) and three output channels (first two: optical flow, third one: intensity).

During training and inference, we first apply average pooling with a kernel size of 16 to the raw output flow, and then perform bilinear interpolation to recover the original resolution for the output flow, to further guarantee flow smoothness. For inference, like [28], we convert the predicted logarithmic intensity into the linear scale through $I = \exp(L)$, then perform the robust min/max normalization before evaluation and visualization: $\hat{I} = (I - m)/(M - m)$, where m and M are the 1% and 99% percentiles of I, respectively. The inference time reported in Tabs. 3 and 5 is measured on a single GPU of the same type (NVIDIA RTX A6000).

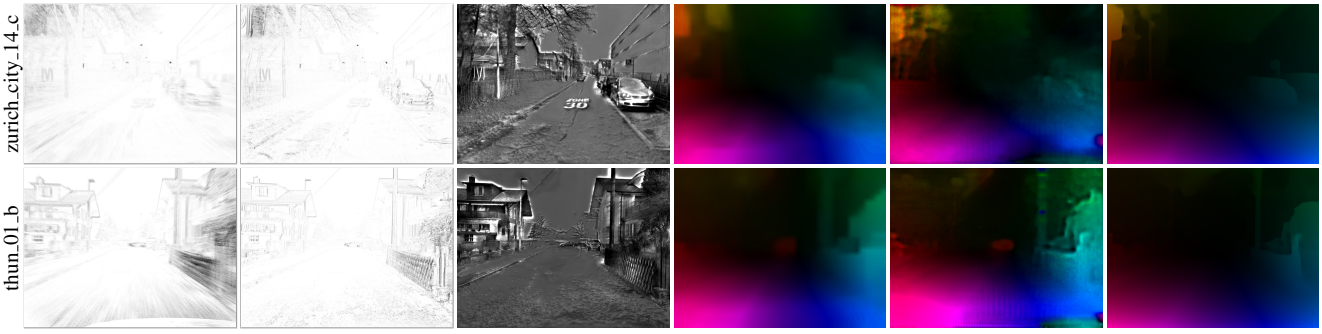
4.2. Optical Flow Evaluation

Accuracy. Table 3 presents a comprehensive comparison between our method and baseline approaches, where the methods that require GT optical flow (SL) and those that do not (MB/USL) are compared respectively in two categories. As mentioned in Sec. 2.1, SL methods achieve smaller errors because they are not only trained with GT, but also the GT of training and test sequences are from the same sensor (no distribution gap). In the MB/USL category, our method significantly outperforms all others in the summarized metrics ("All" columns) and the per-sequence values. The only exception is that our value of %Out on interlaken_00_b is slightly bigger than that of MotionPriorCM. Overall, our

Type	Method	t_{inf}[ms]	All				interlaken_00_b				interlaken_01_a				thun_01_a			
			EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑
SL	E-RAFT [13]	46.33	0.79	10.56	2.68	1.29	1.39	6.22	6.19	1.32	0.90	6.88	3.91	1.42	0.65	9.75	1.87	1.20
	IDNet [45]		**0.72**	**2.72**	**2.04**	–	**1.25**	**2.11**	**4.35**	–	**0.77**	**2.25**	**2.60**	–	**0.57**	**2.66**	**1.47**	–
MB/ USL	RTEF [3]	$9.9 \cdot 10^3$	4.88	–	41.95	**2.51**	8.59	–	59.84	**2.89**	5.94	–	47.33	**2.92**	3.01	–	29.70	**2.39**
	MultiCM [37]		3.47	13.98	30.86	1.37	5.74	9.19	38.93	1.50	3.74	9.77	31.37	1.51	2.12	11.06	17.68	1.24
	BTEB [28]		3.86	–	31.45	1.30	6.32	–	47.95	1.46	4.91	–	36.07	1.42	2.33	–	20.92	1.32
	Paredes et al. [29]	40.1	2.33	10.56	17.77	–	3.34	6.22	25.72	–	2.49	6.88	19.15	–	1.73	9.75	10.39	–
	EV-FlowNet [52]		3.86	–	31.45	1.30	6.32	–	47.95	1.46	4.91	–	36.07	1.42	2.33	–	20.92	1.32
	MotionPriorCM [16]	17.86	3.20	8.53	15.21	1.46	3.21	4.89	**20.45**	1.58	2.38	5.46	17.40	1.70	1.39	6.99	7.36	1.30
	VSA-SM [47]		2.22	8.86	16.83	–	3.20	6.23	24.61	–	2.46	7.00	20.23	–	1.55	6.63	10.67	–
	Ours	**15.12**	**1.78**	**6.44**	**11.24**	1.79	**3.08**	**3.87**	20.76	1.92	**1.90**	**4.11**	**12.62**	2.06	**1.26**	**5.69**	**6.61**	1.56

Type	Method	thun_01_b				zurich_city_12_a				zurich_city_14_c				zurich_city_15_a			
		EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑	EPE ↓	AE ↓	%Out ↓	FWL ↑
SL	E-RAFT [13]	0.58	8.41	1.52	1.18	0.61	23.16	**1.06**	1.12	**0.71**	10.23	**1.91**	1.47	0.59	8.88	1.30	1.34
	IDNet [45]	**0.55**	**2.07**	**1.35**	–	**0.60**	**4.56**	1.16	–	0.76	**3.74**	2.74	–	**0.55**	**2.55**	**1.02**	–
MB/ USL	RTEF [3]	3.91	–	34.69	**2.48**	3.14	–	34.08	**1.42**	4.00	–	45.67	**2.67**	3.78	–	37.99	**2.82**
	MultiCM [37]	2.48	12.05	23.56	1.24	3.86	28.61	43.96	1.14	2.72	12.62	30.53	1.50	2.35	11.82	20.99	1.41
	BTEB [28]	3.04	–	25.41	1.33	2.62	–	25.80	1.03	3.36	–	36.34	1.24	2.97	–	25.53	1.33
	Paredes et al. [29]	1.66	8.41	9.34	–	2.72	23.16	26.65	–	2.64	10.23	23.01	–	1.69	8.88	9.98	–
	EV-FlowNet [52]	3.04	–	25.41	1.33	2.62	–	25.80	1.03	3.36	–	36.34	1.24	2.97	–	25.53	1.33
	MotionPriorCM [16]	1.54	6.55	9.69	1.33	8.33	20.16	22.39	1.13	1.78	8.79	12.99	1.56	1.45	6.27	8.34	1.51
	VSA-SM [47]	1.74	6.76	13.07	–	2.19	17.13	15.24	–	1.69	7.57	11.02	–	1.85	8.06	13.55	–
	Ours	**1.15**	**4.89**	**5.81**	1.63	**1.92**	**14.35**	**13.31**	1.40	**1.50**	**6.93**	**10.51**	1.92	**1.26**	**5.46**	**6.41**	1.89

Table 3. *Optical flow evaluation*. Results on the DSEC optical flow benchmark [13]. Bold is the best in each category, except the column of t_{inf}, where only the shortest t_{inf} is marked. Note that our model predicts both flow and intensity, while others only predicts the former.



(a) Events (b) IWEs (c) Intensity (d) Flow (Ours) (e) Flow ECCV'24 [16] (f) Flow E-RAFT [13]

Figure 3. *Qualitative comparisons on DSEC*. From left to right: (a) input events; (b) image of warped events (IWE) with our predicted flow; (c) our predicted intensity; (d) our predicted flow; (e)-(f) two baseline methods that predict optical flow (USL and SL, respectively).

method achieves great improvements of 20%, 25% and 26% in terms of EPE, AE and %Out, respectively. Note that our method even outperforms E-RAFT (SL) in terms of AE, with a reduction of 39%. RTEF directly adopts FWL as the objective function to optimize, so it has the highest FWL scores on all sequences, followed by our method in the second place. Note that a high FWL value may not always be good, as it can be indicative of event collapse [39]. In addition to the overall metrics, our method shows the best performance on all individual sequences that are recorded in various scenarios (urban and highway) and various illumination conditions (daylight and night). This demonstrates the robustness and versatility of the proposed approach.

Figure 3 presents a qualitative comparison on the DSEC dataset. It is clear that the output flow maps (column d) of our method are more precise than those of the most recent USL method (MotionPriorCM, column e). Regarding scene appearance, our method not only generates sharp edge maps using the predicted flow (column b) –like flow-based methods–, but also simultaneously produces detailed intensity images (column c).

Runtime. We also compare the inference time of some methods in Tab. 3 ("t_{inf}" column). Our method shows the best efficiency despite predicting two quantities (flow and intensity). Note that MotionPriorCM also uses a U-Net architecture, where the only difference from ours is the way of up-sampling features in the decoder (we use bilinear interpolation whereas [16] uses 2D transposed convolutions).

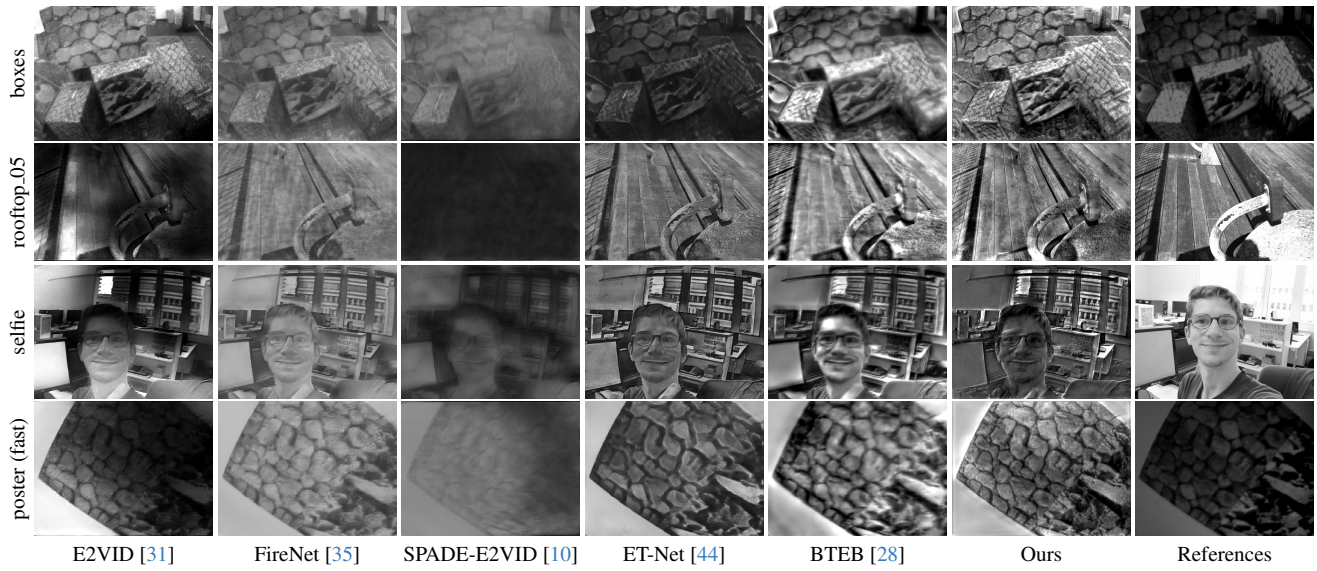


Figure 4. *Qualitative image-intensity comparisons* on ECD, BS-ERGB, HDR and ECD-fast data. Best viewed when zoomed in.

4.3. Evaluation of Image Intensity Reconstruction

The results of the benchmarks with and without GT reference images on the BS-ERGB and HDR datasets are presented in Tab. 4. In addition, the inference times of all methods are reported in Tab. 5.

Among supervised-learning (SL) approaches, ET-Net achieves the best image quality in five out of six metrics, but it is the slowest by a margin. Conversely, FireNet reports the shortest inference time, but its quality metrics are not as good as other methods.

In the unsupervised-learning (USL) category, our method reports comparable results in full-reference metrics as those of BTEB [28]. However, our method significantly outperforms BTEB in all three metrics on the HDR dataset, with an improvement of around 50%, where many SL methods are also surpassed. In particular, we achieve the best MANIQA value compared to all baselines. This implies that our method better unlocks the HDR properties of event cameras. In terms of inference time, our model stands in the middle of the ranking; however, our model outputs both flow and intensity, while the others only output intensity.

An important remark to make is that the above metrics have limitations: as analyzed in [49], different metrics often lead to divergent conclusions, where discrepancies occur. For example, some methods achieve better MSE just by darkening the predictions (as GT images are dark). Our method reports better LPIPS, but worse MSE and SSIM values than BTEB; E2VID reports better BRISQUE than ET-Net, but it performs worse in terms of NIQE and MANIQA. The numbers do not seem to entirely reflect visual quality from a human perspective, as we present in the qualitative comparisons of Fig. 4. From the images, it is clear that our approach is able to recover fine details (sharp image) of scene appearance despite the HDR illumination (selfie) and fast motion (poster fast). Figure 4 also illustrates the shortcomings of the above quantitative metrics. All images from SPADE-E2VID are visually blurred, but it reports better MSE and SSIM than FireNet and our method, whose images are markedly sharper and cleaner.

		BS-ERGB			HDR		
Type	Method	MSE↓	SSIM↑	LPIPS↓	BRISQUE↓	NIQE↓	MANIQA↑
SL	E2VID [31]	0.14	0.33	0.56	**12.63**	4.27	0.30
	FireNet [35]	0.10	0.34	0.53	18.57	3.85	0.30
	SPADE-E2VID [10]	0.09	0.35	0.63	24.51	7.17	0.28
	ET-Net [44]	**0.07**	**0.37**	**0.44**	19.20	**3.45**	**0.32**
USL	BTEB [28]	**0.09**	**0.36**	0.62	51.47	6.24	0.18
	Ours	0.10	0.31	**0.56**	**25.03**	**3.78**	**0.40**

Table 4. *Image-intensity quality assessment*. Full-reference evaluation results on the BS-ERGB [42] dataset (left), and non-reference evaluation results on the HDR [31] dataset (right). Bold is the best in each category. The between-frame event slicing is used for BS-ERGB while the fixed-duration slicing ($\Delta t = 100$ms) is used for HDR, where no frame is available.

Resolution	E2VID (2019)	FireNet (2020)	SPADE-E2VID (2021)	ET-Net (2021)	BTEB (2021)	Ours (2024)
640×480	10.95	4.94	36.07	173.56	10.59	15.11
1280×720	31.04	14.67	105.87	1606.33	29.89	40.78

Table 5. *Runtime evaluation* [ms] of event-based intensity estimation methods at VGA and HD resolutions. Note that our model infers both flow and intensity, while others only infer intensity.

4.4. Ablation Study

We conduct an ablation study to illustrate the effects of some loss terms and the superiority of joint flow-intensity

estimation over flow-only estimation. The quantitative and qualitative results are presented in Tab. 6 and Fig. 5. Please refer to the supplementary for the ablation studies on loss weights and camera contrast threshold C.

TC Loss. First, we disable the TC loss ($\lambda_5 = 0$). Although flow accuracy does not decrease dramatically (around 10%), intensity accuracy is reduced significantly (Tab. 6, 1st row). This is also revealed in Fig. 5 (column a), where texture at valid pixels is recovered, while invalid pixels are filled with markedly incorrect values due to the lack of constraints. This confirms our claim in Sec. 1, that better estimation for both visual quantities is achieved by leveraging their synergies.

TV regularization. Next, we disable the TV regularization of both flow and intensity ($\lambda_3 = \lambda_4 = 0$). Contrary to the previous test, intensity accuracy remains while flow accuracy drops considerably (Tab. 6, 2nd row). The same conclusion can be drawn from Fig. 5 (column b), where the estimated flow shows artifacts, mostly at invalid pixels. As a consequence, the output image intensity at such pixels is not as smooth as that of our main model (column c).

Flow-only Estimation. Finally, we train the network to predict only optical flow ($\lambda_1 = \lambda_4 = \lambda_5 = 0$, thus changing the number of output channels from three to two). In this case, the loss function basically reduces to the one used by recent purely CMax-based optical flow estimation methods [16, 37]. As reported in the third row of Tab. 6, the flow accuracy is slightly better than that of w/o TV reg., but still falls far behind our main model, with respective gaps of 43%, 33%, 53% in terms of EPE, AE and %Out. This again confirms the advantages of our proposed joint estimation method over the separate ones.

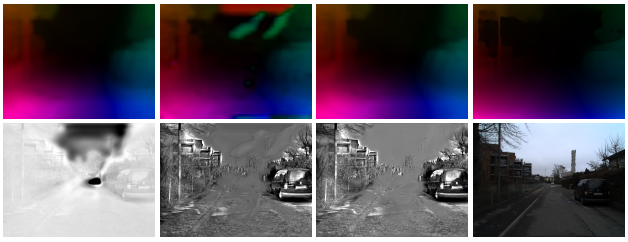
4.5. Single Network vs. Dual Network

To illustrate the superiority of joint learning with a single network, we compare it to a dual network (two U-Nets, one for optical flow prediction and the other for intensity reconstruction), with configuration identical to ours (loss function, loss weights and training settings). Table 6 show that the estimation results of the dual network (4th row) are considerably worse than those of the single U-Net model (last row). Our joint learning scheme enables one single network to learn the motion and appearance, as well as their synergies, thus achieving better performance than separately.

5. Limitations

The event camera output depends on scene texture and camera motion. In regions where no events are produced, it is thus difficult to recover motion parameters and/or scene appearance. In this case, regularization is used to fill in those pixels by encouraging spatial smoothness. However, this may cause inaccuracies. This issue could be overcome by combining two visual modalities (events and frames), not without considering data fusion challenges, and/or by adding recurrent connections (at the expense of increasing the initialization time and inertia of the system [31]).

Events triggered by flickering lights or hot pixels do not satisfy the brightness constancy assumption, which is the basis of all involved MB/USL methods. As outliers, they can undermine estimation accuracy. Learning-based methods have some capacity to deal with such outliers. However outliers would be more sensibly treated in pre-processing by some denoising method, or more recently, jointly [40].



(a) w/o TC loss (b) w/o TV reg. (c) Ours (d) References

Figure 5. *Qualitative results of ablation study.* The above are results from the thun_01_b sequence of DSEC. The reference optical flow is generated using E-RAFT [13].

	Flow			Intensity		
Ablation	EPE↓	AE↓	%Out↓	MSE↓	SSIM↑	LPIPS↓
w/o TC loss	2.01	7.92	13.96	0.16	0.23	0.64
w/o TV reg.	3.30	14.93	24.95	**0.10**	0.30	**0.55**
Flow-only	3.13	9.61	23.56	–	–	–
Dual network	3.30	10.15	22.85	0.11	0.29	0.59
Ours	**1.78**	**6.44**	**11.24**	**0.10**	**0.31**	0.56

Table 6. *Quantitative results of an ablation study.* Flow evaluation is performed on the DSEC dataset and intensity evaluation is done on the BS-ERGB dataset.

6. Conclusion

We have presented the first unsupervised joint learning framework for optical flow and image intensity estimation with event cameras. This enabled us to train a single and lightweight network for predicting both visual quantities simultaneously, exploiting their synergies. It is designed to match the natural sensing principle of event cameras, that is, that motion and appearance are intertwined and, therefore, shall be jointly estimated. A comprehensive and well-behaved loss function is proposed by combining the event generation model, photometric consistency, event alignment and regularization. The experiments demonstrate that our method is accurate and efficient: (*i*) it achieves the highest accuracy among MB/USL optical flow methods; (*ii*) it shows competitive intensity quality, especially in HDR conditions; and (*iii*) it reports very short inference time while predicting both flow and image intensity.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

Supplementary Material

7. Additional Ablation Study

As an expansion of Sec. 4.4, we present the results of the ablation studies on loss weights and contrast threshold C.

7.1. Loss Weights

In addition to disabling some loss terms to show their effects (Sec. 4.4), an ablation study on the weight of CMax loss λ_2 is performed to show how the ratio between CMax and PhE influences model performance. The results are reported in the upper half of Tab. 7. It turns out that increasing or decreasing λ_2 does not lead to a further improvement in performance. Therefore, the choice of λ_1 and λ_2 to train our main model yields a sensible combination of the CMax and PhE loss terms.

7.2. Contrast Threshold

The contrast threshold C can vary across event cameras and change even within the same dataset [41]. Hence, it is worth analyzing the influence of C on model performance.

The works of [14, 15] have shown that the PhE and its linearized version are insensitive to the value of C, due to the PhE being calculated using thousands or millions of events instead of a few. The user just needs to set a mean value for C, and then the optimizer seeks a balanced motion and brightness to best explain the events. This finding has also been verified by the results in Fig. 4 in our main paper and Figs. 6 to 8 in this supplementary: our model was trained only on the DSEC dataset (Prophesee Gen3) with $C = 0.2$, but nevertheless it is able to predict precise optical flow and image intensity on several other datasets, such as ECD (DAVIS240C), HDR (Samsung DVS Gen3) and BS-ERGB (Prophesee Gen4).

Furthermore, we also train models with different C values, and report the results in the lower half of Tab. 7. The results agree with the statements above; the accuracy remains approximately constant for different C values.

8. DSEC: Training Sequence Selection

As mentioned in Sec. 5, all MB/USL methods rely on the brightness constancy assumption to estimate optical flow or image intensity. Events triggered by flickering lights or by hot pixels would undermine the estimation accuracy. To this end, the sequences in the training split of the DSEC dataset [13] are screened before being used for training, according to the data quality. Here we present the list of the selected training sequences in Tab. 8.

		Flow			Intensity		
Ablation	Value	EPE↓	AE↓	%Out↓	MSE↓	SSIM↑	LPIPS↓
λ_2	0.2	2.78	8.12	18.93	0.10	0.31	0.57
	5.0	2.34	9.73	16.42	0.10	0.31	0.56
C	0.1	1.89	6.80	12.34	0.11	0.30	**0.55**
	0.4	1.88	**6.42**	12.32	0.10	**0.32**	0.56
Main ($\lambda_2 = 1.0, C = 0.2$)		**1.78**	6.44	**11.24**	**0.10**	0.31	0.56

Table 7. Results of ablation studies on λ_2 and C.

zurich_city_02_a	zurich_city_02_b	zurich_city_02_c	zurich_city_02_d
zurich_city_02_e	zurich_city_03_a	zurich_city_04_a	zurich_city_04_b
zurich_city_04_c	zurich_city_04_d	zurich_city_04_e	zurich_city_04_f
zurich_city_05_a	zurich_city_06_a	zurich_city_07_a	zurich_city_08_a
zurich_city_11_a	zurich_city_11_b	interlaken_00_c	interlaken_00_d
interlaken_00_e	interlaken_00_f	interlaken_00_g	thun_00_a

Table 8. Sequences from the DSEC training split that were used for training our model.

9. BS-ERGB: Evaluation Sequence Cropping

Here we describe the removal of the last seconds of the may29_rooftop sequences, for the intensity evaluation on the BS-ERGB [42] dataset (970 × 625 px resolution), as presented in Tab. 9. We perform the cropping because the camera is not moving in the last seconds (mentioned in Sec. 4.1) of those sequences, thus the recorded event data is pure noise.

Sequence	Start Time [s]	End Time [s]
may29_rooftop_handheld_01	0.0	24.0
may29_rooftop_handheld_02	0.0	17.0
may29_rooftop_handheld_03	0.0	14.0
may29_rooftop_handheld_05	0.0	9.5

Table 9. Details of the cropping of the may29_rooftop sequences in the BS-ERGB dataset.

10. Additional Qualitative Results

In this section, we present additional qualitative results of our model on the high-resolution BS-ERGB [42] datasets in Fig. 6, including (a) input events, (b) image of warped events (IWE) with the predicted flow, (c) optical flow, (d) image intensity and (e) reference images. A qualitative comparison between our method and other baselines on the same dataset is also presented in Fig. 7, where some regions of interest (ROIs) are highlighted for further zoomed-in visualization in Fig. 8.

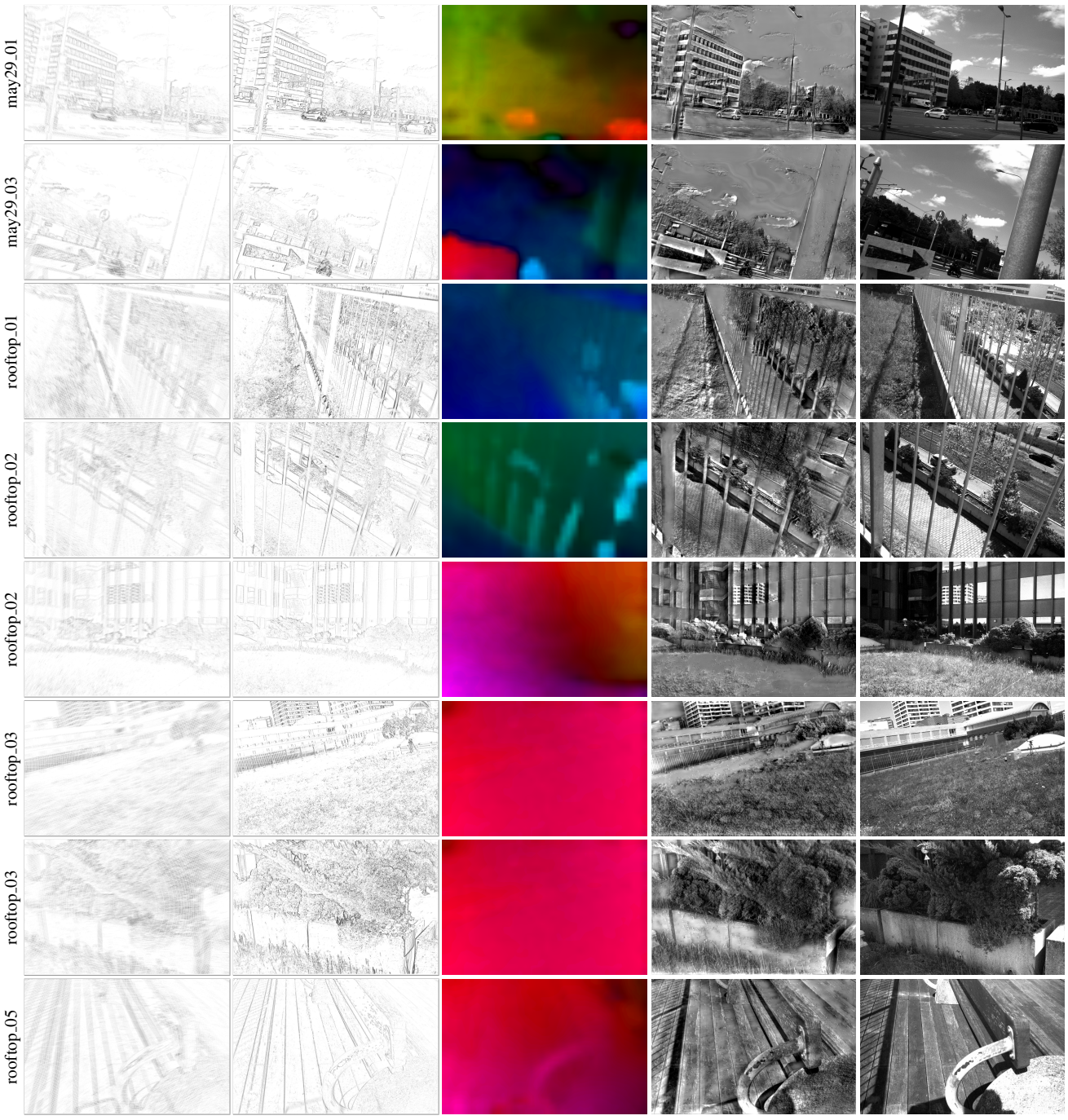


Figure 6. *Additional qualitative results on the BS-ERGB dataset*. From left to right: (a) input events; (b) image of warped events (IWE) with our predicted flow; (c) our predicted flow; (d) our predicted intensity; (e) reference image.

Figure 6 demonstrates that our model recovers precise optical flow and image intensity on unseen data (i.e., not used for training). In column c (optical flow), independent moving objects (IMOs) are clearly identified with respect to the background (e.g., the cars in the first row and the motorbike in the second row). Besides, flow discontinuities agree with the contours of different objects at different depths (e.g., the fence in the third and fourth rows and the bench in the last row). For image intensity, our model reconstructs fine details at the valid pixels (e.g., the contours

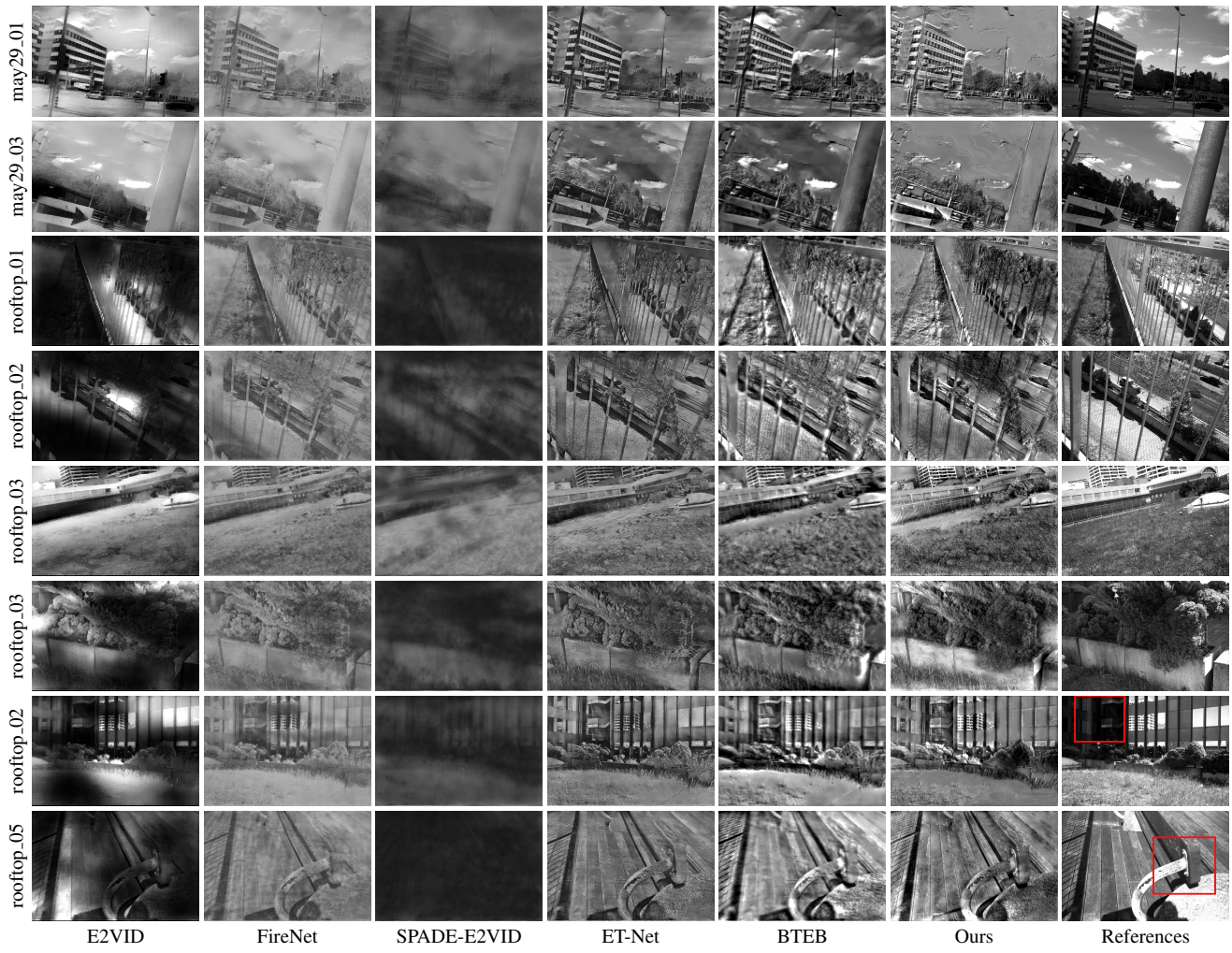


Figure 7. *Additional qualitative comparison of image intensity reconstruction on the BS-ERGB dataset*. For the last two rows, the regions marked by red boxes are enlarged in Fig. 8.

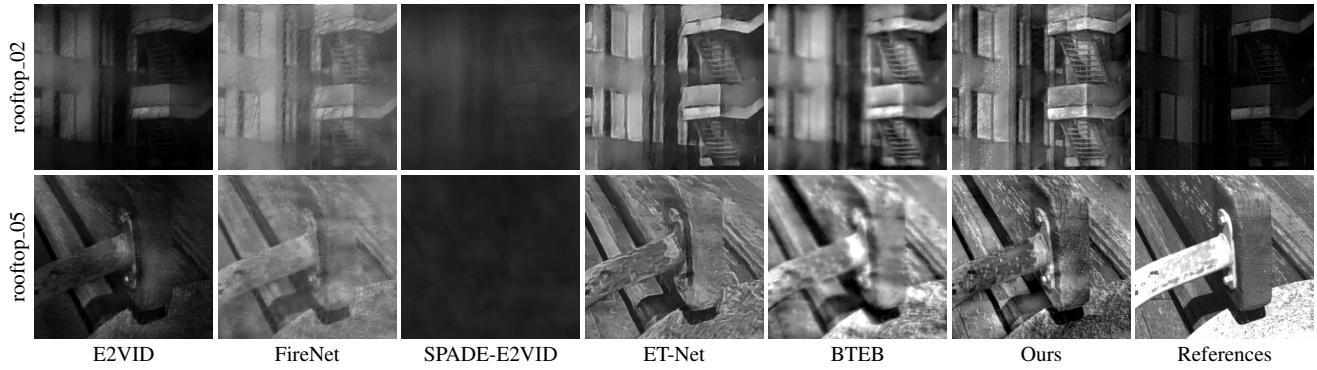


Figure 8. *Additional qualitative comparison of image intensity reconstruction*. Enlarged regions indicated by red boxes in Fig. 7.

of objects), while it leverages the total variation regularization to partially fill in the regions that lack texture and rarely trigger events.

Figure 7 confirms that our model produces competitive results compared to baseline methods. Our reconstructed images are overall sharper, and are more precise in HDR

conditions. To highlight this, we select two HDR regions (i.e., the stairs of the building and the bench on the rooftop) and present their zoomed-in versions in Fig. 8. It can be clearly seen that: E2VID and SPADE-E2VID report poor HDR performance; FireNet produces intensity images with low contrast, where strong artifacts (like spider webs) caused by the rectification of event data using the camera's intrinsic parameters are clearly observed; BTEB's intensity images are blurred; ET-Net oversmooths the fine textures on the building wall, and shows strange wrong textures on the bench (especially the stone part in the bottom right). In contrast, our reconstructed intensity reveals sharp edges and fine details in HDR illumination, where frame-based cameras suffer from under/over exposure problems.

References

[1] 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.

[2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[3] Vincent Brebion, Julien Moreau, and Franck Davoine. Real-time optical flow for vehicular perception with low- and high-resolution event cameras. *IEEE Trans. Intell. Transport. Syst.*, pages 1–13, 2021.

[4] Guang Chen, Hu Cao, Jorg Conradt, Huajin Tang, Florian Rohrbein, and Alois Knoll. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Process. Mag.*, 37(4):34–49, 2020.

[5] Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *Int. Joint Conf. Neural Netw. (IJCNN)*, pages 770–776, 2011.

[6] Burak Ercan, Onur Eker, Aykut Erdem, and Erkut Erdem. EVREAL: Towards a comprehensive benchmark and analysis suite for event-based video reconstruction. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 2023.

[7] 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.

[8] 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)*, pages 12272–12281, 2019.

[9] 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. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2022.

[10] Pablo Rodrigo Gantier Cadena, Yeqiang Qian, Chunxiang Wang, and Ming Yang. SPADE-E2VID: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Trans. Image Process.*, 30:2488–2500, 2021.

[11] 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)*, pages 5632–5642, 2019.

[12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robot. Autom. Lett.*, 6(3):4947–4954, 2021.

[13] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *Int. Conf. 3D Vision (3DV)*, pages 197–206, 2021.

[14] Shuang Guo and Guillermo Gallego. Event-based mosaicing bundle adjustment. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 479–496, 2024.

[15] Shuang Guo and Guillermo Gallego. Event-based photometric bundle adjustment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[16] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 18–37, 2024.

[17] Yijin Li, Zhaoyang Huang, Shuo Chen, Xiaoyu Shi, Hongsheng Li, Hujun Bao, Zhaopeng Cui, and Guofeng Zhang. BlinkFlow: A dataset to push the limits of event-based optical flow estimation. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 3881–3888, 2023.

[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] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. TMA: Temporal motion aggregation for event-based optical flow. In *Int. Conf. Comput. Vis. (ICCV)*, pages 9651–9660, 2023.

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Representations (ICLR)*, 2019.

[21] Xinglong Luo, Kunming Luo, Ao Luo, Zhengning Wang, Ping Tan, and Shuaicheng Liu. Learning optical flow from event camera with rendered dataset. In *Int. Conf. Comput. Vis. (ICCV)*, pages 9847–9857, 2023.

[22] Anton Mitrokhin, Cornelia Fermuller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, pages 1–9, 2018.

[23] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

[25] S.M. Mostafavi I., Lin Wang, and Kuk-Jin Yoon Yoon. Learning to reconstruct HDR images from events, with applications to depth and flow prediction. *Int. J. Comput. Vis.*, 129(4):900–920, 2021.

[26] 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.

[27] Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *Int. J. Comput. Vis.*, 126(12): 1381–1393, 2018.

[28] Federico Paredes-Vallés and Guido C. H. E. de Croon. Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3445–3454, 2021.

[29] Federico Paredes-Vallés, Kirk YW Scheper, Christophe De Wagter, and Guido CHE de Croon. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *Int. Conf. Comput. Vis. (ICCV)*, pages 9661–9671, 2023.

[30] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proc. IEEE*, 102(10):1470–1484, 2014.

[31] 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.*, 43 (6):1964–1980, 2021.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[33] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992.

[34] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conf. Comput. Vis. (ACCV)*, pages 308–324, 2018.

[35] 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)*, pages 156–163, 2020.

[36] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. A fast geometric regularizer to mitigate event collapse in the contrast maximization framework. *Adv. Intell. Syst.*, page 2200251, 2022.

[37] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 628–645, 2022.

[38] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Event collapse in contrast maximization frameworks. *Sensors*, 22(14):1–20, 2022.

[39] Shintaro Shiba, Yannick Klose, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow, depth, and ego-motion by contrast maximization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):7742–7759, 2024.

[40] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Simultaneous motion and noise estimation with event cameras. In *Int. Conf. Comput. Vis. (ICCV)*, 2025.

[41] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 534–549, 2020.

[42] Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time Lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 17755–17764, 2022.

[43] 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, 2004.

[44] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Int. Conf. Comput. Vis. (ICCV)*, pages 2543–2552, 2021.

[45] Yilun Wu, Federico Paredes-Vallés, and Guido C. H. E. de Croon. Lightweight event-based optical flow estimation via iterative deblurring. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 14708–14715, 2024.

[46] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pages 1191–1200, 2022.

[47] Hongzhi You, Yijun Cao, Wei Yuan, Fanjun Wang, Ning Qiao, and Yongjie Li. Vector-symbolic architecture for event-based optical flow. *arXiv e-prints*, 2024.

[48] 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.

[49] Zelin Zhang, Anthony Yezzi, and Guillermo Gallego. Formulating event-based image reconstruction as a linear inverse problem with deep regularization using optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8372–8389, 2023.

[50] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv e-prints*, 2023.

[51] 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)*, pages 1–9, 2018.

[52] 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)*, pages 989–997, 2019.