Techno-economic assessment of grid-connected photovoltaic systems for peak shaving in arid areas based on experimental data

Bakhshi-Jafarabadi, Reza; Mousavi, Seyed Mahdi Seyed

DOI
10.1016/j.egyr.2024.05.035

Publication date
2024

Document Version
Final published version

Published in
Energy Reports

2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.00889

Taming Contrast Maximization for Learning Sequential, Low-latency, Event-based Optical Flow

Federico Paredes-Vallés[1,2] Kirk Y. W. Scheper[2] Christope De Wagter[1] Guido C. H. E. de Croon[1]

[1] Micro Air Vehicle Laboratory, Delft University of Technology
[2] Stuttgart Laboratory 1, Sony Semiconductor Solutions Europe, Sony Europe B.V.

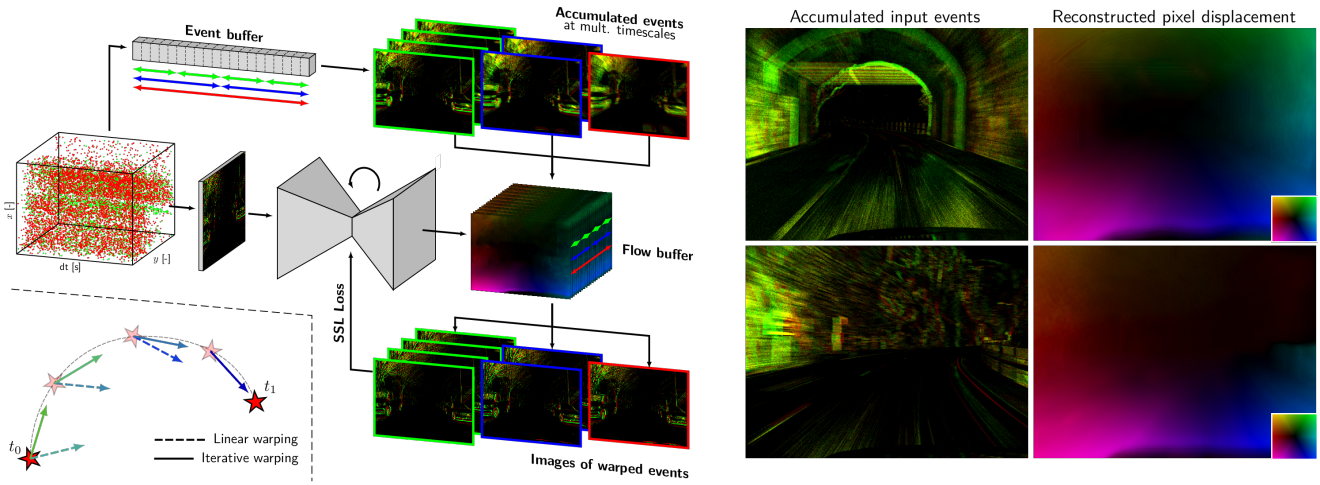

Figure 1: Our pipeline estimates event-based optical flow by sequentially processing small partitions of the event stream with a recurrent model. We propose a novel self-supervised learning framework (left) based on a multi-timescale contrast maximization formulation that is able to exploit the high temporal resolution of event cameras via iterative warping to produce accurate optical flow predictions (right).

Abstract

Event cameras have recently gained significant traction since they open up new avenues for low-latency and low-power solutions to complex computer vision problems. To unlock these solutions, it is necessary to develop algorithms that can leverage the unique nature of event data. However, the current state-of-the-art is still highly influenced by the frame-based literature, and usually fails to deliver on these promises. In this work, we take this into consideration and propose a novel self-supervised learning pipeline for the sequential estimation of event-based optical flow that allows for the scaling of the models to high inference frequencies. At its core, we have a continuously-running stateful neural model that is trained using a novel formulation of contrast maximization that makes it robust to nonlinearities and varying statistics in the input events. Results across multiple datasets confirm the effectiveness of our method, which establishes a new state of the art in terms of accuracy for approaches trained or optimized without ground truth.

1. Introduction

Event cameras capture per-pixel log-brightness changes at microsecond resolution [13]. This operating principle results in a sparse and asynchronous visual signal that, under constant illumination, directly encodes information about the apparent motion (i.e., optical flow) of contrast in the image space. These cameras offer several advantages, such as low latency and robustness to motion blur [13], and hence hold the potential of a high-bandwidth estimation of this optical flow information. However, the event-based nature of the generated visual signal poses a paradigm shift in the processing pipeline, and traditional, frame-based algorithms become suboptimal and often incompatible. Despite this, the majority of learning-based methods that have been proposed so far for event-based optical flow estimation are still highly influenced by frame-based approaches. This influence is normally reflected in two key aspects of their

The project's code and additional material can be found at https://mavlab.tudelft.nl/taming_event_flow/.

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.00889

pipelines: (i) the design of the network architecture, and (ii) the formulation of the loss function.

Regarding architecture design, most literature methods format subsets of the input events as dense volumetric representations [45] that are processed at once by stateless (i.e., non-recurrent) models [17, 29, 37, 45]. Similarly to their frame-based counterparts [12, 38, 39], these models estimate the per-pixel displacement over the time-length of the event volume using only the information contained within it. Consequently, these volumes need to encode enough spatiotemporal information for motion to be discernible. However, if done over relatively long time periods, the subsequent models suffer from limitations such as high latency or having to deal with large pixel displacements [17, 18, 42].

With respect to the loss function, multiple options have been explored due to the lack of real-world datasets with per-event ground truth. Pure supervised learning can be used with datasets such as MVSEC [43] or DSEC-Flow [17], but their ground truth only contains per-pixel displacement at low frequencies, which makes models difficult to train. On the other hand, a self-supervised learning (SSL) framework can be formulated using either the accompanying frames with the photometric error as a loss [11, 40, 44], or through an events-only contrast maximization [14, 15] proxy loss for motion compensation (i.e., event deblurring) [19, 26, 29, 35, 45]. However, despite not relying on ground truth, all the literature on SSL for optical flow assumes that the events move linearly within the time window of the loss, which ignores much of the potential of event cameras and their high temporal resolution (see Fig. 1, bottom left).

In this work, we focus on the estimation of high frequency event-based optical flow and how this can be learned in an SSL fashion using contrast maximization with a relaxed linear motion assumption. To achieve this, we build upon the continuous-operation pipeline from Hagenaars *et al*. [19], which retrieves optical flow by *sequentially* processing small partitions of the event stream with a stateful (i.e., recurrent) model over time, instead of dealing with large volumes of input events. We augment (and train) this pipeline with a novel contrast maximization formulation that performs event motion compensation in an iterative manner at multiple temporal scales, as shown in Fig. 1. Using this framework, we achieve the best accuracy of all contrast-maximization-based approaches on multiple datasets, only being outperformed by pure supervised learning methods trained with ground-truth data.

In summary, the extensions that we propose to the SSL framework in [19], i.e., our main contributions, are:

- The first iterative event warping module in the context of contrast maximization (see Section 3.2). This module unlocks a novel multi-reference loss function that better captures the trajectory of scene points over time, thus improving the accuracy of the predictions.
- The first multi-timescale approach to contrast maximization, which adds robustness, improves convergence, and reduces tuning requirements of loss-related hyperparameters (see Section 3.3).

As a result, we present the first self-supervised optical flow method for event cameras that relaxes the linear motion assumption, and hence that has the potential of exploiting the high temporal resolution of the sensor by producing estimates in a close-to continuous manner. We validate the proposed framework through extensive evaluations on multiple datasets. Additionally, we conduct ablation studies to show the effectiveness of each individual component.

2. Related work

Due to the aforementioned advantages of event cameras for optical flow estimation, extensive research has been carried out since these sensors were first introduced [1, 2, 4–8, 24, 27, 28, 35]. Regarding learning-based approaches, the first method was proposed by Zhu *et al*. in [44] with EV-FlowNet: a UNet-like [32] architecture trained with SSL, with the supervisory signal coming from the photometric error between subsequent frames captured with an accompanying camera. To avoid the need for a secondary vision sensor, in [45], Zhu *et al*. proposed an SSL framework around the contrast maximization for motion compensation idea from [14, 15], hence relying solely on event data. This pipeline was used and further improved in [19, 29, 35]. Other approaches trained EV-FlowNet in a supervised fashion with synthetic/real ground-truth data and showed higher accuracy levels through evaluations on public benchmarks [17, 37]. However, they also highlighted EV-FlowNet's inability to deal with large pixel displacements [17]. Because of this, inspired by the frame-based literature [39], Gehrig *et al*. proposed E-RAFT in [17]: the first architecture to introduce the use of correlation volumes in the event camera literature. This model, which was recently augmented with attention mechanisms [41], achieved state-of-the-art performance in multiple datasets.

A novel perspective to learning event-based optical flow is to move away from event accumulation over long timespans, and instead rely on continuously-running stateful models that integrate information over time. The first methods of this kind were proposed by Paredes-Vallés *et al*. [30] and Hagenaars *et al*. [19], but this idea has recently gained interest in the event camera literature [11, 31, 42]. The reason is that leveraging memory through sequential processing can potentially lead to lightweight and low-latency solutions that are also robust to large pixel displacements without the need for correlation volumes [42]. However, despite their potential of being scaled to high inference frequencies, all these solutions assume that events move linearly in the timespan of their loss function, and hence cannot capture the true trajectory of scene points over time.

When it comes to learning nonlinear pixel trajectories from event data, only the work of Gehrig *et al*. in [18] is to be highlighted. However, their approach uses multiple event and correlation volumes to fit Bézier curves to the trajectory of scene points, resulting in a high-latency solution. In contrast, in this work, we extend the continuous-operation pipeline from Hagenaars *et al*. [19] with an SSL framework in which discrete trajectories are regressed at high frequency by leveraging memory within the models. This approach allows, for the first time, to exploit the high temporal resolution of event cameras while capturing more accurately the trajectory of scene points over time thanks to the proposed iterative event warping mechanism.

Among non-learning-based approaches, the work of Shiba *et al*. in [35] bears particular relevance due to certain similarities with our framework. Specifically, Shiba *et al*. propose a tile-based method for event-based optical flow estimation that extends contrast maximization [14, 15] by incorporating (i) multiple spatial scales, (ii) a multi-reference focus loss, and (iii) a "time-aware" optical flow formulation. Similarly, our learning-based approach also employs a multi-scale strategy and utilizes a multi-reference focus loss. However, instead of making assumptions about the events' motion over extended time periods, we propose to learn their potentially nonlinear trajectories at high frequency by leveraging the iterative event warping module.

3. Method

The goal of this work is to learn to sequentially estimate optical flow at high frequencies from a continuous stream of events. In such a pipeline, if the inference frequency is sufficiently high, events need to be processed nearly as soon as they are triggered by the sensor, with no pre-processing in between. Because of this, the integration of temporal information needs to happen in the network itself. To accomplish this, we propose the framework in Fig. 1, in which a stateful model is trained using our novel formulation of contrast maximization for sequential processing. The components of this framework are described in the following sections.

3.1. Input format and contrast maximization

For an ideal camera, an event $\boldsymbol{e}_i = (\boldsymbol{x}_i, t_i, p_i)$ of polarity $p_i \in \{+, -\}$ is triggered at pixel $\boldsymbol{x}_i = (x_i, y_i)^T$ and time t_i whenever the change in log-brightness since the last event at that pixel location reaches the contrast sensitivity threshold for that polarity [13].

As in [19], we use a two-channel event count image as input representation, which gets populated with consecutive, non-overlapping, fine discrete partitions of the event stream $\boldsymbol{\varepsilon}_k^{\text{inp}} \doteq \{\boldsymbol{e}_i\}$ (further referred to as *input partitions*), each containing all the events in a time window of a certain duration, i.e., $t_i \in [t_k^{\text{begin}}, t_k^{\text{end}}]$. This representation does not contain temporal information by itself, and recurrent models are hence required for estimating of optical flow.

Regarding learning, we use the contrast maximization framework [14] to train, in an SSL fashion, to continuously estimate dense (i.e., per-pixel) optical flow from the event stream. Assuming brightness constancy, accurate flow information is encoded in the spatiotemporal misalignments (i.e., blur) among the events triggered by the same portion of a moving edge. To retrieve it, one has to compensate for this motion by geometrically transforming the events using a motion model. As in [19, 29, 35, 45], we transport each event to a reference time t_{ref} through:

$$\boldsymbol{x}_i' = \boldsymbol{x}_i + (t_{\text{ref}} - t_i)\boldsymbol{u}(\boldsymbol{x}_i) \tag{1}$$

where $\boldsymbol{u}(\boldsymbol{x}) = (u(\boldsymbol{x}), v(\boldsymbol{x}))^T$ denotes the optical flow map used to transport each event from t_i to t_{ref}. The result of aggregating the transformed events is further referred to as the image of warped events (IWE) at t_{ref}.

As the loss function of our SSL framework, we adopt the time-based focus objective function from [19]. Using the warped events at a given t_{ref}, we generate an image of the per-pixel average timestamps for each polarity p' via bilinear interpolation:

$$\begin{aligned} T_{p'}(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}}) &= \frac{\sum_j \kappa(x - x_j')\kappa(y - y_j')\bar{t}_j(t_{\text{ref}}, t_j)}{\sum_j \kappa(x - x_j')\kappa(y - y_j') + \epsilon} \\ \kappa(a) &= \max(0, 1 - |a|) \\ j = \{i \mid p_i = p'\}, &\qquad p' \in \{+, -\}, \qquad \epsilon \approx 0 \end{aligned} \tag{2}$$

where $\bar{t}_i$ denotes the normalized timestamp contribution of the ith event, according to Fig. 3 and Eq. 5.

Then, the contrast maximization loss function at t_{ref} is defined as the scaled sum of the squared temporal images:

$$\mathcal{L}_{\text{CM}}(t_{\text{ref}}) = \frac{\sum_{\boldsymbol{x}} T_+(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}})^2 + T_-(\boldsymbol{x};\boldsymbol{u}|t_{\text{ref}})^2}{\sum_{\boldsymbol{x}} [n(\boldsymbol{x}') > 0] + \epsilon} \tag{3}$$

where $n(\boldsymbol{x}')$ denotes a per-pixel event count of the IWE. The lower the $\mathcal{L}_{\text{CM}}$, the better the event deblurring at t_{ref}.

As discussed in [14, 19, 36], for any focus objective function to be a robust supervisory signal for contrast maximization, the event partition used in the optimization needs to contain enough motion information (i.e., blur) so it can be compensated for. However, as in [19], that is not the case in our input partitions due to the fine discretization of the event stream that we are targeting. Therefore, only at training time, we define the so-called *training partition* (or event buffer in Fig. 1) $\boldsymbol{\varepsilon}_{k\to k+R}^{\text{train}} \doteq \{\boldsymbol{\varepsilon}_i^{\text{inp}}\}_{i=k}^{k+R}$, which stacks together the events in the input partitions of R successive forward passes through the network. Once these are performed, we use this partition and the estimated optical flow maps to compute the loss, and use truncated backpropagation through time to update the model parameters. After

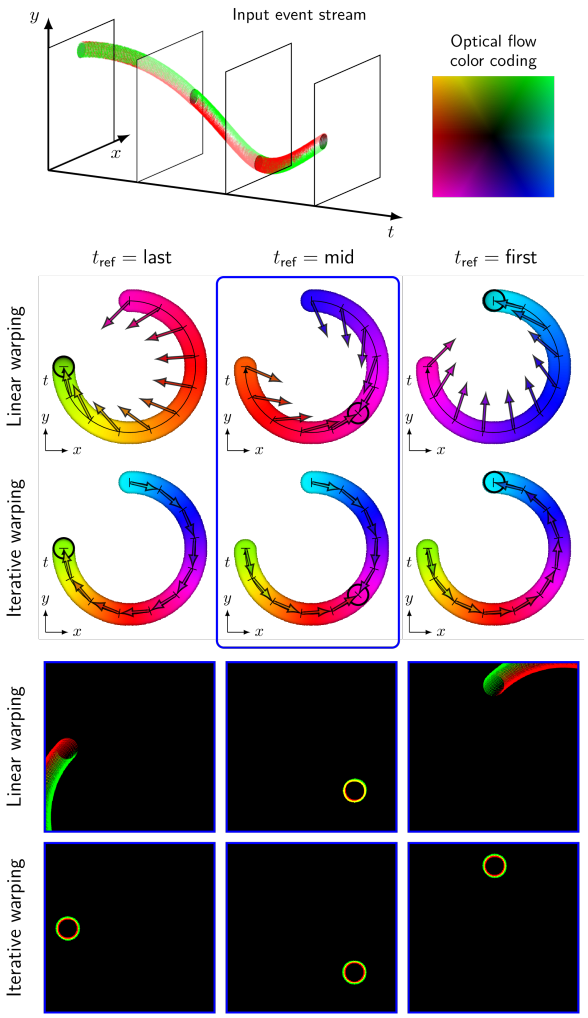

Figure 2: Incompatibility of multi-reference deblurring and linear event warping in the presence of nonlinearities in the pixel trajectories. *Top*: Events generated by a moving dot following circular motion (left), and optical flow color-coding scheme (right). *Middle*: Optical flow solutions required to produce sharp IWEs at different t_{ref} using linear and iterative warping. While the former requires a different solution for each t_{ref}, the proposed iterative warping can achieve this using a single solution. Arrows illustrate the direction of the required displacement $\Delta \boldsymbol{x}_i = (t_{\text{ref}} - t_i)\boldsymbol{u}(\boldsymbol{x}_i)$ at each (discretized) spatial location. *Bottom*: Resulting IWEs at different t_{ref} using the optimal optical flow map for $t_{\text{ref}} = \text{mid}$.

this update, we detach the states of the network from the computational graph and clear the training partition and optical flow buffer. The importance of sequential processing, as an alternative to training stateless models in short input partitions, is corroborated in the supplementary material.

3.2. Iterative event warping

In order to better approximate the trajectories of scene points over time, in this work we relax the linear motion assumption in the SSL framework for event-based optical flow by augmenting the sequential-estimation pipeline from [19] with *iterative event warping*. Instead of transporting events to a given t_{ref} assuming linear motion regardless of the length of the warping interval (as in [19, 29, 35, 45]), we perform a finer discretization of the event trajectories and assume that motion is only linear between optical flow estimates. Therefore, to express a group of events at a given t_{ref}, we geometrically transform them using all the intermediate optical flow estimates through multiple iterations of Eq. 1. Fig. 2 shows the differences between the linear event warping in [19] and the proposed iterative augmentation, as well as the limitations of the former. Note that our iterative warping is fundamentally different from that of the work of Wu *et al*. in [42], where input events are deblurred before being passed to the models using residual optical flow estimates until convergence.

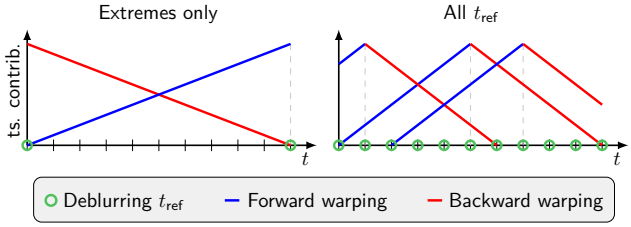

Figure 3: Timestamp normalization profiles for the per-event contributions to the images of average timestamps in Eq. 2, for R=10. *Left*: Deblurring only done at the extremes of the training partition, as in [19, 29, 45]. *Right*: Deblurring done at all reference times, thanks to the proposed iterative event warping. Normalization profiles only shown for three t_{ref} for a better visualization.

Literature methods on event-based optical flow with contrast maximization compute the focus objective function at multiple reference times in the training partition (usually at the extremes) to prevent overfitting and/or scaling issues during backpropagation [19, 29, 35, 45]. However, since these approaches assume optical flow constancy in the span of their loss function, they suffer with nonlinearities in the pixel trajectories (see the blurry IWEs in Fig. 2, bottom). On the contrary, because of the finer discretization of the event trajectories coming with our sequential processing pipeline and the proposed iterative warping, we can use any (combination of) reference time(s) for the computation of the focus objective function. In fact, as shown in Fig. 3 (right), we propose the use of *all* the discretization points as reference times for event deblurring. Apart from the aforementioned regularizing benefits, having to produce sharp IWEs at any t_{ref} forces the models to estimate a sequence of optical flow maps that is consistent with the velocity profile of the event stream. For a given training partition of

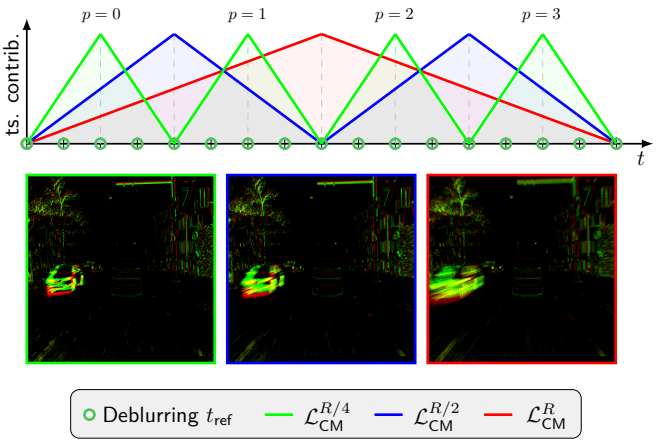

Figure 4: Multi-timescale approach to contrast maximization. For a given training partition of length R, we fit multiple sub-partitions of different lengths (in the figure: one of length R in red, two $R/2$ in blue, and four $R/4$ in green) and compute the loss in each of them according to Eq. 4. The global loss is computed as in Eq. 6. This figure only shows the timestamp normalization profiles of the central t_{ref} of each sub-partition, but the losses are still computed at all reference times. An image representation of the accumulated input events in a sub-partition of each timescale is also shown.

length R, the loss is computed as follows:

$$\mathcal{L}_{\text{CM}}^{R} = \frac{1}{R+1} \sum_{t_{\text{ref}}=0}^{R} \mathcal{L}_{\text{CM}}(t_{\text{ref}}) \tag{4}$$

which ensures that the IWEs (and the corresponding images of average timestamps in Eq. 2) at all reference times $t_{\text{ref}} \in [0, R]$ contribute equally to the loss.

As in [19, 29, 45], for $\mathcal{L}_{\text{CM}}(t_{\text{ref}})$ in Eq. 3 to be a valid supervisory signal, events that are temporally close to the reference time t_{ref} need to contribute to the temporal image in Eq. 2 with larger timestamp values than events that are far in time. Therefore, once the events have been transported to t_{ref}, their timestamp is normalized prior to the computation of Eq. 2 as follows:

$$\bar{t}_i(t_{\text{ref}}, t_i) = 1 - \frac{|t_{\text{ref}} - t_i|}{R}, \quad \text{with } t_i \in [0, R] \tag{5}$$

which results in the normalization profiles in Fig. 3, for a given training partition of length R.

Inspired by [25], we mask individual events from the computation of the loss whenever we detect that they are transported outside the image space through the event warping process in the span of a deblurring window defined at the reference time t_{ref}. This prevents our models from learning incorrect deformations at the image borders, and is motivated by the fact that the complete trajectory of the pixel is only partially observable in the training partition. An ablation study on the impact of this masking strategy can be found in the supplementary material.

Lastly, we do not augment the loss function in Eq. 4 with smoothing priors acting as regularization mechanisms. With iterative event warping, the error propagates through all the pixels covered in the warping process, regardless of whether they have input events or not. Therefore, the spatial coherence of the resulting optical flow maps is enhanced.

3.3. Deblurring at multiple timescales

Despite the addition of iterative event warping, the success of our SSL framework still heavily depends on the hyperparameters that control the amount of motion information perceived by the networks in the span of a deblurring window. In our pipeline, these are: dt_{input}, the timestep used to discretize the event stream; and R, the number of forward passes, and hence optical flow maps, per loss. Thus, the effective length (in units of time) of the event window used for motion compensation is given by $\text{dt}_{\text{input}} \times R$. We hypothesize that, for each training dataset, there is an optimal length for this window that depends on the statistics of the data (e.g., event density, distribution of optical flow magnitudes) and model architecture, and that deviations from this optimal length lead to the learning of suboptimal solutions. E.g., shorter windows may converge to solutions that are more selective to fast rather than slow moving objects, and vice versa. Note that not only our method is sensitive to the tuning of these parameters, but also previous approaches based on contrast maximization [19, 29, 35, 45].

To add a layer of robustness to the framework and relax its strong dependency on hyperparameter optimization, we propose the *multi-timescale approach* illustrated in Figs. 1 and 4. For a given training partition of length R, instead of computing a single focus loss through Eq. 4, we compute this loss at S temporal scales of length $R/2^s$, with $0 \leq s \leq S-1$, and combine them as follows:

$$\mathcal{L}_{\text{CM}}^{\text{multi}} = \frac{1}{S} \sum_{s=0}^{S-1} \frac{1}{2^s} \sum_{p=0}^{2^s-1} \mathcal{L}_{\text{CM},p}^{R/2^s} \tag{6}$$

As shown, we fit multiple non-overlapping sub-partitions in the training buffer if $s > 0$. The subscript p indicates their location in this buffer, starting from the earliest (see Fig. 4).

Note that, through this multi-timescale approach to contrast maximization, our models need to converge to a solution that is suitable for all the timescales in the optimization, regardless of their length. An alternative formulation would be to incorporate per-pixel learnable masks (i.e., an attention module in the loss space) so that, depending on the input statistics, learning only happens at the most adequate scale. However, for this to happen, the loss function would have to be augmented to stimulate this behavior, and it is unclear how that would be done in practice.

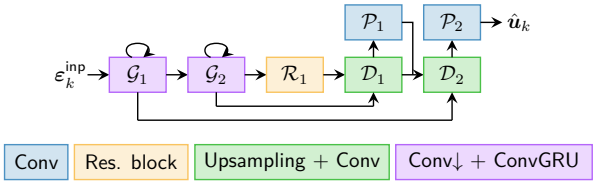

Figure 5: Schematic of the model architecture used in this work. It is characterized by $N_{\mathcal{G}}$ recurrent encoders, $N_{\mathcal{R}}$ residual blocks, and $N_{\mathcal{G}}$ decoder layers. Optical flow estimates are produced at all decoder levels. In this diagram, $N_{\mathcal{G}} = 2$ and $N_{\mathcal{R}} = 1$.

3.4. Network architecture

We use the recurrent version of EV-FlowNet [44] proposed in [19][1] (see Fig. 5). The events are represented as event count images (see Section 3.1), then passed through four encoders with strided convolutions followed by ConvGRUs [3] (channels doubling, starting from 64), two residual blocks [20], and then four decoders performing bilinear upsampling followed by convolution. After each decoder, there is a skip connection (using element-wise summation) from the corresponding encoder, as well as a depthwise convolution to produce estimates at lower scales, which are then concatenated with the activations of the previous decoder. Note that the proposed focus loss function (see Eq. 6) is applied to each intermediate optical flow estimate via upsampling. Lastly, all layers use 3×3 kernels and ReLU activations except for the prediction layers, which use TanH.

4. Experiments

We evaluate our method on the DSEC-Flow [16, 17] and MVSEC [43] datasets. We evaluate the accuracy of the predictions based on the following metrics: (i) EPE (lower is better, ↓), the endpoint error; (ii) $\%_{\text{3PE}}$ (↓), the percentage of points with EPE greater than 3 pixels; (iii) FWL (↑) [37], a deblurring quality metric based on the variance of the IWEs; and (iv) RSAT (↓) [19], a deblurring quality metric based on the per-pixel average timestamps of the IWEs. We compare our solution to the published baselines, which range from supervised learning (SL) methods trained with ground truth, to SSL methods trained with grayscale images (SSL_{F}) or events (SSL_{E}), and model-based approaches (MB).

We train all our models on a subset of sequences from the training dataset of DSEC-Flow (only daylight recordings, see supplementary material). This corresponds to 19 minutes of training data, which we split into 572 128×128 (randomly-cropped) sequences of 2 seconds each. We use a batch size of 8 and train until convergence with the Adam optimizer [21] and a learning rate of 1e-5. To keep memory usage within limits, we only propagate error gradients through up to 1e3 randomly-chosen events per millisecond of data. Despite this, note that we warp and use all the input events for the computation of the loss function.

[1] Our architecture is equivalent to ConvGRU-EV-FlowNet [19]. However, for the purpose of clarity within the rest of the paper, we will use this term to specifically refer to the original model from [19].

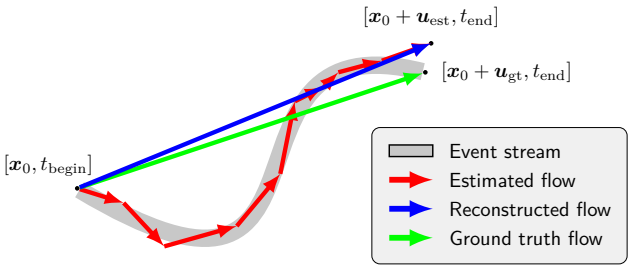

Figure 6: Reconstruction of the pixel displacement of a scene point in the time window of a ground-truth sample from the multiple optical flow maps estimated in this period, i.e., $\forall \boldsymbol{u}_k \in [t_{\text{begin}}, t_{\text{end}}]$. The error of the last optical flow estimate is magnified for clarity.

		EPE↓	$\%_{\text{3PE}}$↓	FWL↑	RSAT↓
SL	E-RAFT [17]	0.79	2.68	1.33	<u>0.87</u>
	EV-FlowNet, Gehrig *et al.* [17]	2.32	18.60	-	-
	Gehrig *et al.* [18]	0.75	2.44	-	-
	IDNet [42]	**0.72**	**2.04**	-	-
	TIDNet [42]	0.84	3.41	-	-
	TMA [23]	<u>0.74</u>	<u>2.30</u>	-	-
	Cuadrado *et al.* [9]	1.71	10.31	-	-
	E-Flowformer [22]	0.76	2.68	-	-
SSL_{E}	EV-FlowNet* [45]	3.86	31.45	1.30	**0.85**
	ConvGRU-EV-FlowNet* [19]	4.27	33.27	<u>1.55</u>	0.90
	dt $= 0.01$s, $R = 2$, $S = 1$ (Ours)	9.66	86.44	**1.91**	1.07
	dt $= 0.01$s, $R = 5$, $S = 1$ (Ours)	4.05	52.22	1.58	0.97
	dt $= 0.01$s, $R = 10$, $S = 1$ (Ours)	2.33	17.77	1.26	0.88
	dt $= 0.01$s, $R = 20$, $S = 1$ (Ours)	16.63	33.67	1.06	1.10
	dt $= 0.01$s, $R = 10$, $S = 3$ (Ours)	2.82	27.09	1.37	0.92
	dt $= 0.01$s, $R = 20$, $S = 4$ (Ours)	2.73	23.73	1.24	0.90
MB	Shiba *et al.* [35]	3.47	30.86	1.37	0.89

*Retrained by us on DSEC-Flow, linear warping.

Table 1: Quantitative evaluation on the DSEC-Flow dataset [17]. Best in bold, runner up underlined. A breakdown of the results is provided in the supplementary material.

4.1. Evaluation procedure

When evaluating our sequential models, if $\text{dt}_{\text{gt}} > \text{dt}_{\text{input}}$, we need to reconstruct the estimated per-pixel displacement in the ground-truth time window from the multiple optical flow maps estimated in this period. We do this by first averaging the (bilinearly interpolated) optical flow vectors that describe the trajectory of each scene point, and then by scaling the resulting optical flow vectors by $\text{dt}_{\text{gt}}/\text{dt}_{\text{input}}$. This converts them from units of pixels/input to units of pixel displacement. An illustration of this reconstruction is shown in Fig. 6 for a scene point following a nonlinear trajectory. Note that our solution is subject to cumulative errors when evaluated through this reconstruction on benchmarks with

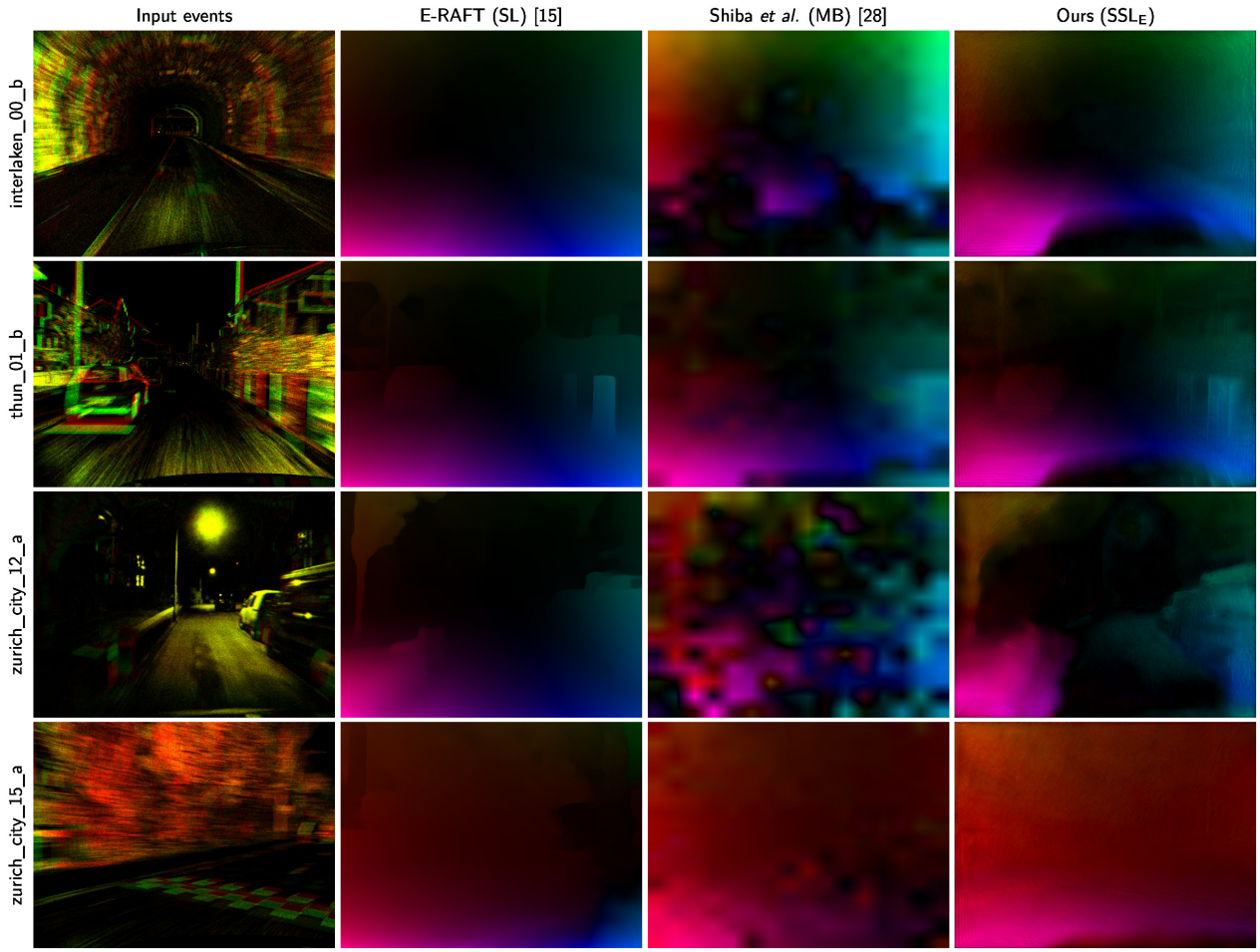

Figure 7: Qualitative comparison of our method with the state-of-the-art E-RAFT architecture [17] and the model-based approach from Shiba *et al*. [35] on sequences from the test partition of the DSEC-Flow dataset [17]. Ground truth not included due to unavailability. The optical flow color coding can be found in Fig. 2 (top), and the corresponding IWEs in the supplementary material.

ground truth provided at low rates (e.g., 10 Hz in DSEC-Flow [17]). Therefore, it will compare unfavorably to other, non-sequential, methods that only produce a single optical flow estimate in the timespan of a ground-truth sample.

4.2. Optical flow evaluation

Evaluation on DSEC-Flow: Quantitative results of our evaluation on DSEC-Flow are presented in Table 1, and are supported by the qualitative comparison in Fig. 7. For this experiment, we trained multiple models with the same $\text{dt}_{\text{input}} = 0.01$s (i.e., $\times 10$ faster than DSEC's ground truth) but different lengths of the training partition, and with and without the multi-timescale approach. Multiple conclusions can be derived from the reported results. Firstly, our best performing model (i.e., $R = 10$, $S = 1$) achieves the best accuracy of all contrast-maximization-based approaches on this dataset according to the EPE and the percentage of outliers. Specifically, it outperforms the baselines with an improvement in the EPE in the $33\% - 45\%$ range, only being outperformed by SL methods trained with ground truth on the same dataset [9, 17, 22, 23, 42]. This confirms that (i) the timestamp-based loss function in Section 3.1 allows us to learn accurate event-based optical flow (contrary to the findings of [34]); and that (ii) our augmentations to the sequential pipeline in [19] lead to a significant improvement in the accuracy of the model (i.e., 45% improvement in the EPE).

Secondly, these results also confirm our hypothesis that, for each training dataset, there is an optimal length for the training partition R in terms of the EPE. According to Table 1, the optimal R for this dataset, our model architecture, and our dt_{input} is 10 (i.e., 0.1s of event data), with the EPE increasing if the training partition is made shorter or longer. However, as also shown in this table, we can relax the strong dependency of the contrast maximization framework on this parameter through the proposed multi-timescale approach.

	EPE↓	$\%_{3PE}$↓	FWL↑	RSAT↓
dt = 0.1s*	3.48	34.72	0.98	**0.87**
dt = 0.05s	3.09	27.36	1.11	0.91
dt = 0.01s	**2.33**	**17.77**	1.26	0.88
dt = 0.005s	2.34	17.92	1.38	0.89
dt = 0.002s	2.66	21.83	**2.04**	0.90

*Non-recurrent, volum. event repr. with 10 bins.

Table 2: Impact of the input window length on the DSEC-Flow dataset [17]. Best in bold, runner up underlined.

		EPE↓	$\%_{3PE}$↓
SL	EV-FlowNet+ [37]	0.68	0.99
	E-RAFT [17]	**0.24**	1.70
	EV-FlowNet, Gehrig *et al*. [17]	0.31	**0.00**
	TMA [23]	0.25	0.07
	Cuadrado *et al*. [9]	0.85	-
SSL_F	EV-FlowNet, Zhu *et al*. [44]	0.49	0.20
	Ziluo *et al*. [11]	0.42	**0.00**
SSL_E	EV-FlowNet, Zhu *et al*. [45]	0.32	**0.00**
	EV-FlowNet, Paredes-Vallés *et al*. [29]	0.92	5.40
	EV-FlowNet, Shiba *et al*. [35]	0.36	0.09
	ConvGRU-EV-FlowNet [19]	0.47	0.25
	Ours	0.27	0.05
MB	Akolkar *et al*. [1]	2.75	-
	Brebion *et al*. [7]	0.53	0.20
	Shiba *et al*. [35]	0.30	0.11

Table 3: Quantitative evaluation on MVSEC's outdoor_1 sequence [43]. Best in bold, runner up underlined. Results on the indoor sequences can be found in the supplementary material.

Our $S > 1$ models converged to solutions that slightly underperform our best performing single-scale model (EPE went up by $17\% - 21\%$), but were trained without the need to fine-tune the length of the training partition. Note that, despite this slight drop in accuracy, these multi-timescale solutions still outperform the other non-SL baselines. To further support these results, a visualization of the distribution of the EPE of our models as a function of the ground truth magnitude is provided in the supplementary material.

Lastly, Table 1 also allow us to conclude that deblurring quality metrics FWL [37] and RSAT [19] are not reliable indicators of the quality of the estimated optical flow. The reason for this is their inability to capture "event collapse" issues (as described in [33]), and would give favorable scores to undesirable solutions that warp all the events into a few pixels. According to our results, the FWL metric, being the spatial variance of the IWE relative to that of the identity warp, suffers more from this issue: the best FWL value is obtained with a model with 9.66 EPE.

Regarding qualitative results, Fig. 7 shows a comparison of our best performing model with the state-of-the-art E-RAFT architecture [17] and the contrast-maximization-based approach from Shiba *et al*. [35] on multiple sequences from the test partition of DSEC-Flow (i.e., ground truth is unavailable). These results confirm that our models are able to estimate high quality event-based optical flow despite not having access to ground-truth data during training, and also show the superiority of our method over the current best contrast-maximization-based approach [35]. Two limit cases in which our models provide suboptimal solutions are also shown in this figure: (i) sequences recorded at night (e.g., zurich_city_12_a) due to the presence of large amounts of events triggered by flashing lights and not by motion; and (ii) the car hood, which is also problematic for E-RAFT (i.e., does not capture it) and for [35]. Note that, in our case, (ii) is an artifact of the pixel displacements reconstructed from multiple optical flow estimates, and could be mitigated by having an occlusion handling mechanism in this reconstruction process.

In addition to the evaluation in Table 1 and Fig. 7, we also conducted an experiment in which we trained multiple models with different dt_{input} (ranging from 0.1s to 0.002s) but with the same amount of information in the training partition: 0.1s of event data. Results in Table 2 show that our sequential pipeline leads to an improvement in the accuracy of the predicted optical flow maps with respect to the stateless EV-FlowNet, which processes the 0.1s of event data at once. This improvement is due to the fact that the complexity of dealing with large pixel displacements gets reduced when processing the input data sequentially using shorter input windows. In addition to this, Table 2 also shows that the accuracy of our models is not compromised when estimating optical flow at higher frequencies, despite the high sparsity levels in the input data at those rates.

Evaluation on MVSEC: Quantitative results of our evaluation on the oudoor_day1 sequence from MVSEC are presented in Table 3, and are supported by the qualitative comparison in the supplementary material. For this experiment, since (i) there is no consensus in the literature with respect to the training dataset [17, 19, 29, 35, 37, 44, 45], and (ii) the outdoor_day2 sequence (i.e., the other daylight, automotive sequence) is only 9 minutes of duration during which the event camera is subject to high frequency vibrations [43], we decided to transfer one of our models trained on DSEC-Flow to MVSEC. More specifically, we chose the model trained with $\text{dt}_{\text{input}} = 0.005$s and $R = 20$ from Table 2, as a model trained with a short input window on DSEC-Flow is expected to be robust to the slow motion statistics of MVSEC [17]. We deployed the model at the same frequency as the temporally-upsampled ground truth (i.e., 45 Hz). Results in Table 3 show that, by doing this, our model outperforms the great majority of methods in terms of the EPE (even some SL methods trained on this dataset), and is only surpassed by the current state-of-the-art E-RAFT architecture [17]. Besides reaffirming the high quality of the produced optical flow estimates, these results also confirm the generalizability of our method. For com-

pleteness, the results on the indoor evaluation sequences from MVSEC are provided in the supplementary material.

5. Limitations

The self-supervised method for event-based optical flow presented in this work, while demonstrating highly accurate and promising results, is not without limitations. Two critical challenges that need to be acknowledged are the brightness constancy assumption and the aperture problem. Firstly, the contrast maximization framework [14, 15] assumes constant illumination, leading our models to face difficulty in learning from events that are not due to motion in the image space but that arise from changes in illumination. Since this limitation is inherent to contrast maximization, it extends to other approaches based on the same principle [19, 29, 35]. Due to this assumption, we excluded sequences recorded at night from our training dataset (see supplementary material). Secondly, akin to many other optical flow methods, our approach is susceptible to the aperture problem. This indicates that only motion components normal to the orientation of an edge in the image space, also known as normal optical flow, can be reliably resolved [10]. Consequently, the proposed method might face challenges in accurately determining the true motion direction in certain ambiguous scenarios. The regularizing effect of the iterative event warping (see Section 3.2) and the multiple spatial scales at which dense optical flow is estimated in our architecture (see Section 3.4) are mechanisms in our proposed solution that collectively strive to counteract the aperture problem's influence.

6. Conclusion

In this paper, we presented the first learning-based approach to event-based optical flow estimation that is scalable to high inference frequencies while being able to accurately capture the true trajectory of scene points over time. The proposed pipeline is designed around a continuously-running stateful model that sequentially processes fine discrete partitions of the input event stream while integrating spatiotemporal information. We train this model through a novel, self-supervised, contrast maximization framework (i.e., event deblurring for supervision) that is characterized by an iterative event warping module and a multi-timescale loss function that add robustness and improve the accuracy of the predicted optical flow maps. We demonstrated the effectiveness of our approach on multiple datasets, where our models outperform the self-supervised and model-based baselines by large margins. Future research should look into how to learn to better combine the information from multiple timescales, as well as into the design of lightweight architectures that can keep up with real-time constraints.

We believe that the proposed approach opens up avenues for future research, especially in the field of neuromorphic computing. Spiking networks running on neuromorphic hardware have the potential of exploiting the main benefits of event cameras, but for that they need to process the input events shortly after they arrive. Our proposed framework is a step toward this objective, as it enables the estimation of optical flow in a close to continuous manner, with all the integration of information happening within the model itself.

References

[1] Himanshu Akolkar, Sio-Hoi Ieng, and Ryad Benosman. Real-time high speed motion prediction using fast aperture-robust event-driven visual flow. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):361–372, 2020. 2, 8

[2] Mohammed Almatrafi, Raymond Baldwin, Kiyoharu Aizawa, and Keigo Hirakawa. Distance surface for event-based optical flow. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 42(7):1547–1556, 2020. 2

[3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving deeper into convolutional networks for learning video representations. *Int. Conf. Learn. Representations*, 2015. 6

[4] 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.*, pages 884–892, 2016. 2

[5] 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, 2013. 2

[6] Ryad Benosman, Sio-Hoi Ieng, Charles Clercq, Chiara Bartolozzi, and Mandyam Srinivasan. Asynchronous frameless event-based optical flow. *Neural Networks*, 27:32–37, 2012. 2

[7] Vincent Brebion, Julien Moreau, and Franck Davoine. Real-time optical flow for vehicular perception with low-and high-resolution event cameras. *IEEE Trans. on Intell. Transp. Systems*, 2021. 2, 8

[8] Tobias Brosch, Stephan Tschechne, and Heiko Neumann. On event-based optical flow detection. *Frontiers in Neuroscience*, 9:137, 2015. 2

[9] Javier Cuadrado, Ulysse Rançon, Benoit R. Cottereau, Francisco Barranco, and Timothée Masquelier. Optical flow estimation from event-based cameras and spiking neural networks. *Frontiers in Neuroscience*, 17:1160034, 2023. 6, 7, 8

[10] David B. de Jong, Federico Paredes-Vallés, and Guido C. H. E. de Croon. How do neural networks estimate optical flow? A neuropsychology-inspired study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):8290–8305, 2021. 9

[11] Ziluo Ding, Rui Zhao, Jiyuan Zhang, Tianxiao Gao, Ruiqin Xiong, Zhaofei Yu, and Tiejun Huang. Spatio-temporal recurrent networks for event-based optical flow estimation. In *Proc. AAAI Conf. on Artificial Intell.*, volume 36, pages 525–533, 2022. 2, 8

[12] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2758–2766, 2015. 2

[13] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 2020. 1, 3

[14] Guillermo Gallego, Mathias Gehrig, and Davide Scaramuzza. Focus is all you need: Loss functions for event-based vis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12280–12289, 2019. 2, 3, 9

[15] 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.*, pages 3867–3876, 2018. 2, 3, 9

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

[17] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *Int. Conf. 3D Vis.*, pages 197–206. IEEE, 2021. 2, 6, 7, 8

[18] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from events and frames. *arXiv:2203.13674*, 2022. 2, 3, 6

[19] Jesse J. Hagenaars, Federico Paredes-Vallés, and Guido C. H. E. de Croon. Self-supervised learning of event-based optical flow with spiking neural networks. In *Advances in Neural Information Process. Syst.*, volume 34, pages 7167–7179, 2021. 2, 3, 4, 5, 6, 7, 8, 9

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 6

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Int. Conf. Learn. Representations*, 2014. 6

[22] 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. *arXiv:2303.07716*, 2023. 6, 7

[23] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. TMA: Temporal motion aggregation for event-based optical flow. *arXiv:2303.11629*, 2023. 6, 7, 8

[24] Min Liu and Tobi Delbruck. Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. In *Brit. Mach. Vis. Conf.*, 2018. 2

[25] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proc. AAAI Conf. on Artificial Intell.*, volume 32, 2018. 5

[26] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2018. 2

[27] Jun Nagata, Yusuke Sekikawa, and Yoshimitsu Aoki. Optical flow estimation by matching time surface with event-based cameras. *Sensors*, 21(4):1150, 2021. 2

[28] Garrick Orchard, Ryad Benosman, Ralph Etienne-Cummings, and Nitish V. Thakor. A spiking neural network architecture for visual motion estimation. In *IEEE Biomedical Circuits Syst. Conf.*, pages 298–301. IEEE, 2013. 2

[29] 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.*, 2021. 2, 3, 4, 5, 8, 9

[30] Federico Paredes-Vallés, Kirk Y. W. Scheper, and Guido C. H. E. de Croon. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 42(8):2051–2064, 2020. 2

[31] Wachirawit Ponghiran, Chamika Mihiranga Liyanagedera, and Kaushik Roy. Event-based temporally dense optical flow estimation with sequential neural networks. *arXiv:2210.01244*, 2022. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Computing and Comput.-Assisted Intervention*, pages 234–241. Springer, 2015. 2

[33] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Event collapse in contrast maximization frameworks. *Sensors*, 22(14):5190, 2022. 8

[34] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. A fast geometric regularizer to mitigate event collapse in the contrast maximization framework. *arXiv:2212.07350*, 2022. 7

[35] Shintaro Shiba, Yoshimitsu Aoki, and Guillermo Gallego. Secrets of event-based optical flow. In *Eur. Conf. Comput. Vis.*, 2022. 2, 3, 4, 5, 6, 7, 8, 9

[36] Timo Stoffregen and Lindsay Kleeman. Event cameras, contrast maximization and reward functions: An analysis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12300–12308, 2019. 3

[37] 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 *European Conf. Comput. Vis.*, 2020. 2, 6, 8

[38] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8934–8943, 2018. 2

[39] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419. Springer, 2020. 2

[40] Z Wan, Y Dai, and Y Mao. Learning dense and continuous optical flow from an event camera. *IEEE Trans. Image Process.*, 2022. 2

[41] Qimin Wang, Yongjun Zhang, Shaowu Yang, Zhe Liu, Lin Wang, and Luoxi Jing. E-HANet: Event-based hybrid attention network for optical flow estimation. In *IEEE Int. Conf. Syst., Man, Cybernetics*, pages 2189–2194. IEEE, 2022. 2

[42] Yilun Wu, Federico Paredes-Vallés, and Guido C. H. E. de Croon. Rethinking event-based optical flow: Iterative deblurring as an alternative to correlation volumes. *arXiv:2211.13726*, 2022. 2, 4, 6, 7

[43] Alex Z. Zhu, Dinesh Thakur, Tolga Özaslan, Bernd Pfrommer, Vijay Kumar, and Kostas Daniilidis. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception. *IEEE Robot. and Autom. Lett.*, 3(3):2032–2039, 2018. 2, 6, 8

[44] Alex Z. Zhu and Liangzhe Yuan. EV-FlowNet: Self-supervised optical flow estimation for event-based cameras. In *Robot.: Science and Syst.*, 2018. 2, 6, 8

[45] Alex Z. 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.*, pages 989–997, 2019. 2, 3, 4, 5, 6, 8