Computer Vision and Image Understanding
journal homepage: www.elsevier.com

High Frame Rate Optical Flow Estimation from Event Sensors via Intensity Estimation

Prasan Shedligeri[a,**], Kaushik Mitra[a]

[a]*Department of Electrical Engineering, Indian Institute of Technology Madras, Chennai, 600036, India*

ABSTRACT

Optical flow estimation forms the core of several computer vision tasks and its estimation requires accurate spatial and temporal gradient information. However, if there are fast-moving objects in the scene or if the camera moves rapidly, then the acquired images will suffer from motion blur, which will lead to poor optical flow estimation. Such challenging cases can be handled by event sensors which are a novel generation of sensors that acquire pixel-level brightness changes as binary events at a very high temporal resolution. Brightness constancy constraint, which is the basis of several optical flow algorithms cannot be directly used on event sensors making it challenging to estimate optical flow. We overcome this challenge by imposing brightness constancy constraint on intensity images predicted from event sensor data. For this task, we design a recurrent neural network that jointly predicts a sparse optical flow and intensity images from the event data. While intensity estimation is supervised using ground truth frames, optical flow estimation is self-supervised using the predicted intensity frames. However, in our case the temporal resolution of the ground truth intensity frames is far lower than the temporal resolution of the predicted intensity frames, making it challenging to supervise. As we use recurrent neural network, such a challenge can be overcome by sharing the weights for each of the predicted intensity frames. Quantitatively our predicted optical flow is better than previously proposed algorithms for optical flow estimation from event sensors. We also show our algorithm's robustness against challenging cases of fast motion and high dynamic range scenes.

1. Introduction

Many of the modern computer vision applications rely on acquiring data from conventional image sensors. Optical flow forms basis for many of the computer vision tasks such as object-tracking, moving object segmentation, autonomous navigation, etc (Fortun et al., 2015). The dense texture rich information acquired from conventional image sensors, enable dense optical flow prediction. The brightness constancy based energy functional introuduced by Horn and Schunck (Horn and Schunck, 1981; Fortun et al., 2015) is the basis of many modern optical flow estimation algorithms. This energy functional relies on accurate sensing of image intensities between successive frames. This brightness constancy constraint fails to hold when the acquired images are degraded from motion blur due to fast-moving objects or due to the rapid camera motion as shown in Fig. 1. Again, due to low frame rate of image sensors, it becomes challenging to estimate optical flow for cases of large scene motion even without significant blur. This challenge can be overcome if we use an image sensor with a very high temporal resolution. Conventional image sensors, that acquire high temporal resolution video are significantly expensive and require large data bandwidth and hence event-based sensors can provide a viable alternative. Event-based sensors are a novel generation of neuromorphic sensors which asynchronously sense only the pixel-level brightness changes with a temporal resolution of the order of microseconds (Delbrück et al., 2010). At each pixel, the event sensor outputs a positive/negative event when it senses an increase/decrease in brightness over a specified threshold. Its extremely high temporal resolution has been demonstrated by reconstructing intensity frames at a frame rate of several thousand frames per second. These sensors also have a much higher dynamic range compared to conventional image sensors making them attractive for several computer vision applications(Gallego et al., 2019).

**Corresponding author:
e-mail: ee16d409@ee.iitm.ac.in (Prasan Shedligeri)

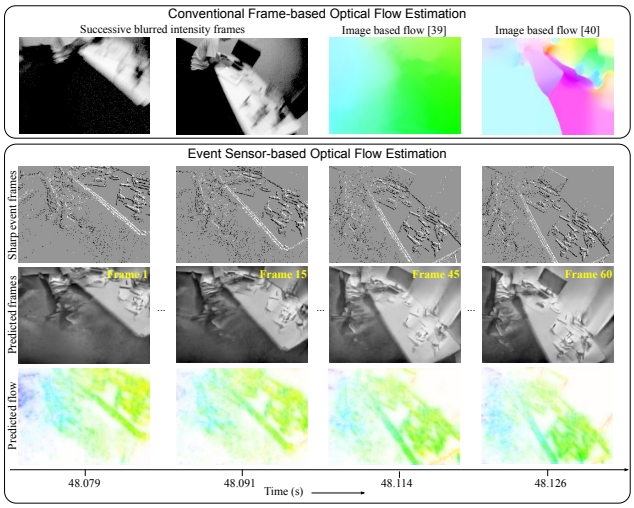

Fig. 1. Conventional frame-based optical flow algorithms suffer when the input images are degraded with motion blur as shown in the top row. Event sensors on the other hand operate at much higher temporal resolution and can sense much higher dynamic range than the frame-based sensors. We accumulate the events triggered between the two successive intensity images as event frames and show some of them in the second row. Our proposed algorithm takes these intermediate event frames as input and predicts corresponding intensity images and optical flow. In this example, optical flow and intensity images are predicted at 60 intermediate temporal locations corresponding to a 60x temporal super-resolution.

Optical flow estimation directly from event sensors is attractive but a challenging task as the brightness constancy based energy functional cannot be used directly. Despite this challenge, several algorithms have been proposed in the literature for event based optical flow estimation (Liu and Delbruck, 2018; Nagata et al., 2019; Paredes-Vallés et al., 2019; Khoei et al., 2019; Bardow et al., 2016; Zhu et al., 2018c; Haessig et al., 2018; Gallego et al., 2018; Almatrafi and Hirakawa, 2020). While learning based methods have shown significant improvement in optical flow prediction accuracy, they fail to exploit the advantages provided by the event sensor. EV-FlowNet(Zhu et al., 2018a), is one of the first learning-based algorithms proposed to predict optical flow from event sensor data. It used the low dynamic range and low frame rate intensity frames and use the brightness constancy as a supervisory signal, thus ignoring the high dynamic range and high temporal resolution offered by event sensors. In (Zhu et al., 2019), the authors propose to use a contrast maximization framework to estimate optical flow. This is an unsupervised algorithm, where the event sensor data alone is used to supervise optical flow prediction, thus fully utilizing the high dynamic range nature of event sensors. However, this algorithm requires an event volume of $30,000$ events as input and hence cannot predict optical flow at very high frame rates. This algorithm also makes a limiting assumption of linear object motion thus affecting the optical flow prediction accuracy. To make full use of the event sensor advantages, algorithms for optical flow estimation from event sensors should have the following desirable properties: (a) the optical flow should be predicted at high temporal resolution, (b) predicted optical flow should be reliable even for challenging high dynamic range scenes, (c) should not require difficult to acquire ground truth optical flow,

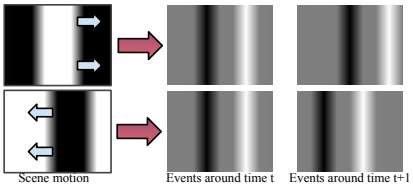

Fig. 2. Ambiguity in intensity image prediction from a single event frame. The first column shows two different scenes which have opposite motion with respect to the camera. These two scenes produce the same event frame at time t making it ambiguous to predict the corresponding scene intensity from the single event frame. However, when we consider the next event frame at time $t+1$, we clearly see the motion in the scene. Modeling this temporal information using recurrent neural network helps in predicting the intensity frames unambiguously from event data alone.

(d) should not make non-generalizable assumptions such as linear motion of the objects.

In our proposed method, intensity frames and a sparse optical flow are simultaneously predicted from the input event sensor data. The event sensor data is first converted to a series of event frames by stacking a fixed number of events per frame following the stacking by number (SBN) principle of (Wang et al., 2019). A sequence of event frames are given as input one-by-one to the neural network which predicts the corresponding intensity frame and optical flow. The intensity frame prediction is supervised using the temporally sparse ground truth intensity frames. While our proposed algorithm predicts intensity frame at a very high temporal resolution (at the rate of incoming events) the frames acquired from hybrid intensity and event based sensors (Brandli et al., 2014) are at a much lower temporal resolution. Thus, it is not possible for us to have a supervised loss for every predicted intensity frame. We overcome this challenge by using recurrent neural network architecture that makes it possible to use supervision only at a few time-steps by sharing weights across all the time-steps. Recurrent neural networks have already been used in (Rebecq et al., 2019b) to predict high frame rate intensity frames. We adapt this network to simultaneously predict intensity frames and optical flow. As demonstrated for optical flow prediction from conventional image sensors (Jason et al., 2016; Ren et al., 2017; Meister et al., 2018), we use the brightness constancy constraint as a supervisory signal for optical flow prediction from event sensors.

In summary, we make the following contributions:

- We propose a semi-supervised learning algorithm to predict high frame rate, sparse optical flow for high dynamic range scenes.
- Optical flow prediction is self-supervised using the high frame rate and high dynamic range intensity frames predicted directly from the event sensor data. Thus, ground truth optical flow is not necessary for training our proposed algorithm.
- We also demonstrate the generalizability of our proposed algorithm on a wide variety of open source event datasets captured with different sensors and in different environments.

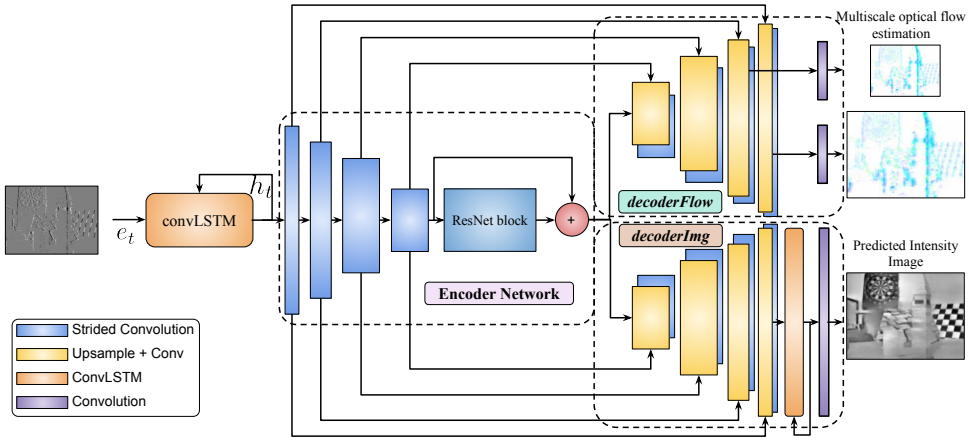

Fig. 3. Overall flow of our proposed method: Our proposed methods takes in a single event frame at each time-step, which is then input to a ConvLSTM (Convolutional Long-Short Term Memory) network. The updated hidden state from the convLSTM network is input to an encoder network consisting of four strided convolutional layers followed by a ResNet block. The hidden representation from the encoder network is then fed as input to two decoder networks, *decoderImg* and *decoderFlow*, which predict the intensity image and the optical flow, respectively.

2. Related Work

Motion estimation from event sensors: Although its a challenging task to estimate optical flow from event sensors, several algorithms have been proposed (Liu and Delbruck, 2018; Nagata et al., 2019; Paredes-Vallés et al., 2019; Khoei et al., 2019; Bardow et al., 2016; Zhu et al., 2018a, 2019, 2018c; Haessig et al., 2018; Gallego et al., 2018). Works such as (Gallego et al., 2018; Zhu et al., 2018c, 2019) use motion compensation on the space-time volume of events to estimate optical flow. In (Haessig et al., 2018), the authors design a spiking neural network to estimate optical flow and demonstrate their proposed algorithm on IBM's neuromorphic chip. A few learning based methods have also been proposed for estimating optical flow from event sensors(Zhu et al., 2019, 2018a).

Intensity image reconstruction: Previously researchers have attempted to estimate intensity frames from event sensor (Reinbacher et al., 2016; Scheerlinck et al., 2018; Bardow et al., 2016; Shedligeri and Mitra, 2019; Rebecq et al., 2019a; Wang et al., 2019), so that the intensity frames could be used as an input to off-the-shelf frame based computer vision algorithms. Recent learning based algorithms (Rebecq et al., 2019a; Wang et al., 2019) have shown a great improvement in intensity image quality compared to traditional methods. The closest work to ours is (Bardow et al., 2016), where the authors propose a framework to simultaneously estimate intensity and optical flow directly from the event sensor data.

3. Optical Flow Estimation from Event Sensors

3.1. Modeling events as sequential data

The output of an event sensor is a 4-tuple (x, y, t, p) where x and y represent the spatial location, t represents the time instant and p denotes the polarity ($+1$ or -1) of the triggered event. Following (Wang et al., 2019), we stack these events into a sequence of event frames to form the input to our algorithm. The temporal information is obviously lost due to this projection of spatio-temporal data as a spatial frame. In Fig. 2, we show a toy example where two different video sequences are used to generate an event frame at time t. Both the event frames look identical as they lack any temporal information about the events, leading to ambiguity in prediction of intensity frames.

To tackle this loss of temporal information we use a sequence of event frames akin to a sequence of image frames forming a temporal video. The effectiveness of this simple representation can be seen from Fig. 2 where a clear distinction emerges between the two cases of scene motion when considering a video sequence instead of looking at each frame independently. It's imperative for us to design a neural network that can effectively incorporate this temporal information so as to unambiguously predict the intensity images. LSTM (Long-Short Term Memory) (Gers et al., 1999) networks have been shown to be effective for such tasks and we use them to model the long-term temporal dependency in the sequence of event frames. Although the input to the algorithm at each timestep is a single event frame, the intensity frame is still unambiguously predicted, demonstrating the effectiveness of the proposed LSTM network to model sequential information.

3.2. Joint estimation of intensity image and optical flow

Fig. 3 shows our overall model to predict the intensity frames and optical flow from input event sensor data. The intensity frame prediction is supervised using temporally sparse raw intensity images acquired from the conventional image sensor present in DAVIS (Brandli et al., 2014). DAVIS is a hybrid sensor consisting of co-located intensity and event based sensors. The input frames are formed by accumulating events occurring in N non-overlapping sub-intervals between successive intensity frames. Each of these sub-intervals contain a fixed, predetermined number of events. These N event frames are given as input and at the output we obtain the N intensity frames and corresponding $N-1$ optical flow estimates. In the following sections we elaborate on the training algorithm for intensity and the optical flow estimation.

3.2.1. Intensity image prediction

We obtain the dataset to train our network from a hybrid intensity and event based sensor where the event data and intensity images are perfectly registered. Such hybrid sensors can acquire intensity frames at the rate of 25 − 30 frames per second and the event data at the temporal resolution of the order of microseconds. We first elaborate the process of predicting and supervising intensity image prediction considering two arbitrary intensity frames I_k and I_{k+1} and the N event frames between them. This process can be generalized to any number of successive raw intensity frames from a given video sequence.

For ease of training, we divide the interval between I_k and I_{k+1} into N sub-intervals based on equal time, instead of equal number of events in each interval. While training, we use the *stacking by time* (SBT) strategy and while testing, we use the *stacking by number* (SBN) (Wang et al., 2019) strategy for creating event frames. The events occurring in each of these sub-intervals are accumulated into separate event frames forming N event frames. At each time-step, the convLSTM network named *inLSTM* takes one event frame as input and updates its hidden state $\mathbf{h}_t$, as shown in Fig. 3. This hidden state $\mathbf{h}_t$ is then fed to an encoder network which outputs a hidden representation ϕ_e. The hidden representation is then fed to a decoder network, *decoderImg*, which outputs the intensity image corresponding to the event frame at time-step t. We denote the N intermediate frames predicted between raw frames I_k and I_{k+1} as $\hat{I}_k^1, \hat{I}_k^2, \hat{I}_k^3 \dots \hat{I}_k^N$. As we have obtained N event frames between two successive intensity frames I_k and I_{k+1}, we can have supervision for only one of those N predicted frames. Due to the way we have formed event frames only the N^{th} interval has the corresponding ground truth intensity frame, I_{k+1}, for supervision. Hence the network can be supervised for intensity image prediction at every N time-steps only. As the proposed recurrent network shares weights at each time-step, the network is able to predict intensity frames without being supervised at every time-step.

We supervise the intensity image predicted at N^{th} interval I_N with the loss $\mathcal{L}_{im}$ defined as,

$$\mathcal{L}_{im}(\hat{I}_k^N) = d(\hat{I}_k^N, I_{k+1}) \tag{1}$$

where $d(\cdot)$ is an appropriate distance metric. $L1$ distance metric has been popularly used in supervising learning based methods due to their ability to preserve edge sharpness. This distance metric is however unsuitable for our problem as the event sensor data has lost the absolute scene intensity information. So, by using a naive $L1$ metric, we are penalizing the network for not predicting something that it theoretically cannot predict with just events as input. To reflect this knowledge, we define our distance metric as,

$$d(\hat{I}, I) = \frac{1}{M} \sum \|(\nabla_x \hat{I} - \nabla_x I) \odot m\|_2 + \|(\nabla_y \hat{I} - \nabla_y I)) \odot m\|_2 \tag{2}$$

where M is the total number of pixels, ∇_x and ∇_y respectively are x and y-gradient operators. The gradient operator ∇ cancels out any absolute scene intensity information at each pixel of the image. We use a binary mask m which masks the saturated and low-intensity noisy image regions and is defined as,

$$m = \begin{cases} 1, 50 < I < 200 \\ 0, otherwise \end{cases}, \tag{3}$$

where the image intensity I varies between 0 and 255. We also do not penalize the network at saturated or the low-intensity noisy regions as the dynamic range of the intensity images is much lower than that of the event sensor data. We later show the effect of using the naive $L1$ loss as a distance metric on the performance of intensity frame and optical flow prediction.

3.2.2. Optical flow prediction

To predict the optical flow between the current and the previous time-steps, we feed the hidden representation ϕ_e obtained at the current time-step to the decoder network, *decoderFlow*. For image-based optical flow estimation, self-supervised learning based methods have been proposed in (Jason et al., 2016; Ren et al., 2017; Meister et al., 2018) as obtaining ground truth optical flow for a real dataset is a challenging task. We make use of these techniques to supervise optical flow prediction with the help of the intensity images predicted from the event sensor data. We define our self-supervised loss for optical flow as,

$$\mathcal{L}_{flow}(\hat{\mathbf{f}}_t^s) = \frac{1}{M} \sum_{t=2}^{N} \sum_{x,y} \|\hat{I}_k^t(x,y) - \hat{I}_k^{t-1}(x + \hat{u}_t^s, y + \hat{v}_t^s)\|_F \tag{4}$$

where $\hat{\mathbf{f}}_t^s = [\hat{u}_t^s, \hat{v}_t^s]^T$ is the predicted optical flow at time-step t and $\hat{I}_t, \hat{I}_{t-1}$ are respectively the predicted intensity images at timestep $t, t-1$. The superscript s in $\hat{\mathbf{f}}_t^s$ denotes the scale of the predicted optical flow. To overcome gradient locality (Godard et al., 2019; Zhou et al., 2017) of the bilinear sampler during image warping, optical flow is predicted at 2 different scales as can be seen in Fig. 3. Following (Godard et al., 2019), the optical flow at coarser scales is upsampled to the resolution of predicted intensity frame and the cost function in Eq. (4) is imposed. The final loss is the sum of costs at individual scales.

3.2.3. Overall cost function

Apart from $\mathcal{L}_{flow}$ and $\mathcal{L}_{im}$, we also impose the piece-wise smoothness constraint on the predicted intensity images and the optical flow as

$$\mathcal{L}_{im\_sm} = \frac{1}{M} \|\nabla_x \hat{I}_t\|_F + \|\nabla_y \hat{I}_t\|_F \tag{5}$$

$$\mathcal{L}_{flow\_sm}(\hat{\mathbf{f}}_t^s) = \frac{1}{M} \sum_{t=1}^{N} \|\nabla_x \hat{\mathbf{f}}_t^s\|_2 + \|\nabla_y \hat{\mathbf{f}}_t^s\|_2 \tag{6}$$

Overall, our training loss becomes,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{im} + \lambda_2 \sum_{s=1}^{2} \mathcal{L}_{flow}(\hat{\mathbf{f}}_t^s) + \lambda_3 \mathcal{L}_{im\_sm} + \lambda_4 \sum_{s=1}^{2} \mathcal{L}_{flow\_sm}(\hat{\mathbf{f}}_t^s) \tag{7}$$

where λ_i with $i = 1, 2, 3, 4$ are hyperparameters which weigh each of the loss terms for optimal performance. In the second

and fourth term $s = 1, 2$ represents the coarse and fine scale of the predicted optical flow. The optical flow at coarser scale is first upsampled to the resolution of the predicted intensity image before applying the loss function.

4. Experiments

4.1. Architectural details

As shown in Fig. 3 our proposed model consists of 4 major components, a LSTM network named *inLSTM*, an encoder network and two decoder networks named *decoderImg* and *decoderFlow*. The detailed description of architecture is shown in Table S.1 of the supplementary material. The convolutional LSTM network, *inLSTM*, consists of three 2D convolutional layers and has a hidden and cell state of size 32 channels. The convolutional LSTM network used at the output of the *decoderImg* has the same architecture as inLSTM. The inLSTM network is then followed by an encoder network and a ResNet block (He et al., 2016). The ResNet block is then followed by two decoder networks *decoderImg* and *decoderFlow*. Both the decoder networks mirror the encoder network with 4 convolutional layers. Each of the convolutional layers in the decoder block are preceded by a bilinear upsampling layer that upsamples the feature maps by a factor of 2. As shown in Fig. 3, the network also consists of skip connections between the encoder and the decoder networks, much like a U-Net (Ronneberger et al., 2015). The decoder network *decoderImg* outputs an intensity image at the same spatial resolution as the input event frame. We use the *decoderFlow* network to predict optical flow at 2 scales, as shown in Fig. 3. The feature maps from the final 2 layers of *decoderFlow* are input to separate 2D convolutional layers to predict the optical flow at 2 scales.

4.2. Implementation details:

To train our network we used the dataset proposed in (Mueggler et al., 2017). We provide the full architectural details of our proposed neural network in the supplementary material. Also, further information about the dataset used and the train-test split is provided in the supplementary material. For the quantitative evaluation of the predicted optical flow, we use the *MVSEC* dataset (Zhu et al., 2018b) which provides ground truth optical flow for event sensors. To further demonstrate the generalizability of our proposed algorithm, we also provide results on various event sensor datasets such as (Scheerlinck et al., 2018; Zhu et al., 2018b; Mueggler et al., 2017; Perot et al., 2020).

The dataset in (Mueggler et al., 2017), is acquired using DAVIS, a hybrid intensity and event sensor that captures raw image frames at a much lower temporal resolution than the event data. Hence, we divide the interval between each successive frame, I_k and I_{k+1} into N sub-intervals and generate N event frames, where the N^{th} event frame corresponds to the second raw intensity frame I_{k+1}. At each timestep, only one event frame is given as input to the proposed neural network model which then predicts the corresponding intensity frame and optical flow simultaneously. As we have the ground truth intensity frame for the N^{th} event frame, we supervise the intensity frame prediction with the loss metric in Eq. (1). Note that the $\mathcal{L}_{im}$ cost

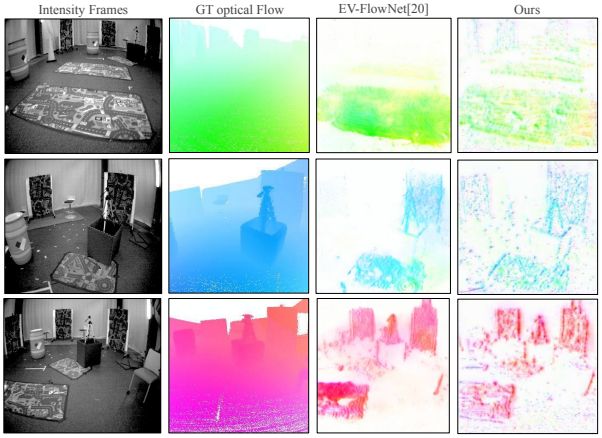

Fig. 4. We show some qualitative comparisons of the predicted optical flow on the *indoor_flying* sequence (Zhu et al., 2018b).

can be imposed for only 1 out of every N timesteps. However, optical flow is supervised using a self-supervised cost function in Eq. (4). This cost can be imposed for every timestep as our model predicts the intensity image at every timestep. However, in the initial phase of training, the output intensity frames are completely random. Hence, if we impose the optical flow loss in the initial stages, then the optical flow network will learn to match random images. To avoid this, we freeze the weights of the optical flow decoder for the first 1000 iterations and only supervise intensity image prediction. After the first 1000 iterations, network is trained to predict simultaneously the intensity frame and optical flow by minimizing the overall cost function, in Eq. (7).

We divide the time interval between successive raw frames into $N = 5$ uniform intervals and the corresponding events are accumulated into 5 event frames. We form our training set with such pairs of 5 event frames and the corresponding raw image frames. During training, we use 40 event frames and correspondingly 8 raw image frames, all in a sequence, of one video and input to our algorithm as one instance of the batch. The neural network is trained using our overall cost-function described in Eq. (7). The brightness constancy loss specified in Eq. (4) is applicable at all 40 time-steps. But, the intensity supervision specified in Eq. (1), is applicable only at 8 time-steps of the sequence.

For training our network, we use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1×10^{-4} which was decayed by a factor of 0.95 every $10k$ iterations. The hyperparamter in Eq. (7) were set to be $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.01$ and $\lambda_4 = 0.001$. The neural network is trained for $150k$ iterations with a batch size of 1. While testing, we accumulate a fixed number of events per event frame, which is akin to the Stacking By Number (SBN) framework proposed in (Wang et al., 2019). Accumulating the event frames using the SBN principle has the advantage of frame rate being adaptive to the event rate which corresponds to the amount of motion in the scene.

4.3. Optical Flow

In this section we evaluate the predicted optical flow, both qualtitatively and quantitatively, using the *indoor_flying* se-

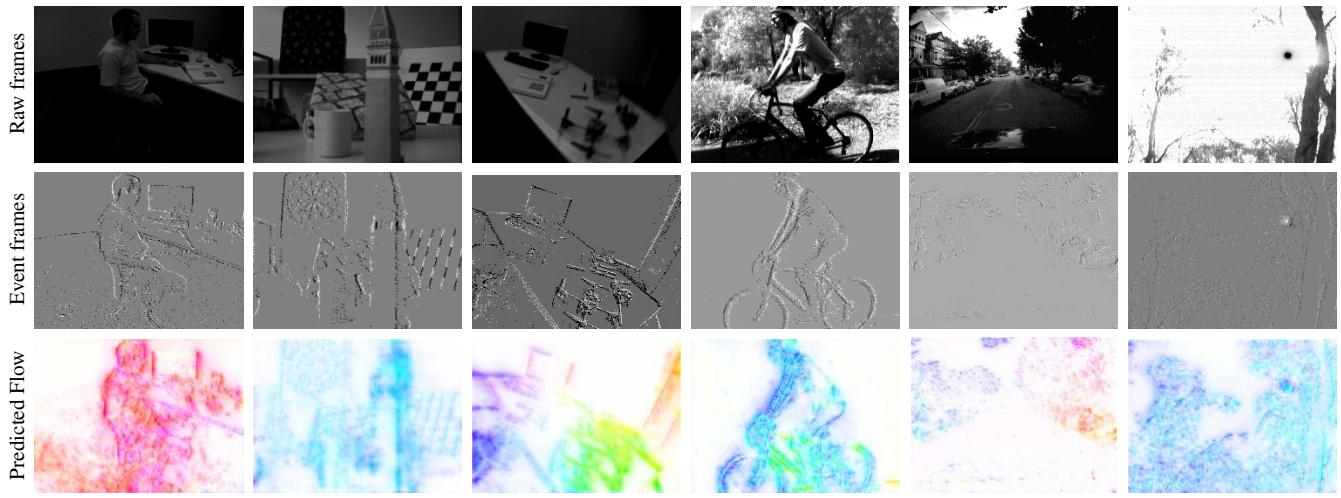

Fig. 5. We test our proposed optical flow model for its generalizability on various test sequences obtained from (Mueggler et al., 2017; Scheerlinck et al., 2018; Zhu et al., 2018b). *(If document is opened in Adobe Reader, videos can be viewed by clicking on the images).* We provide further results in the supplementary video (link to the video).

	indoor flying 1		indoor flying 2		indoor flying 3	
Method	AEE	% outliers	AEE	% outliers	AEE	% outliers
(Zhu et al., 2018a)	0.83	0.84	1.19	6.75	1.07	4.97
(Zhu et al., 2019)	0.58	0	1.02	4	0.87	3
Ours	**0.49**	**0.02**	**0.55**	**0.05**	**0.53**	**0.03**

Table 1. Quantitative comparison of the predicted optical flow on event sequences from (Zhu et al., 2018b).

quences from MVSEC dataset. Following (Zhu et al., 2018a), we choose the metrics (a) Average End-point Error (AEE) which measures the mean absolute error and (b) percentage outliers for quantitative comparison. Percentage outlier (% outlier) measures the percetange of pixels with end-point error above 3 pixels and 5% of the magnitude of the flow vector. For fair comparison, we select two state of the art *unsupervised* learning-based optical flow algorithms (Zhu et al., 2019, 2018a) to benchmark our proposed algorithm. In (Zhu et al., 2018a), all the events between two successive intensity frames are accumulated into a frame based representation and fed to the trained network. In (Zhu et al., 2019), a volume consisting of $30,000$ events divided over 10 event frames is fed into the optical flow network. Effectively, each of event frames in (Zhu et al., 2019) is formed by accumulating 3000 events from the event data. For a fair comparison, we too accumulate successive 3000 events into a single event frame which is then sequentially fed to our trained model.

In Table 1 we provide the quantitative metrics to compare our optical flow algorithms with the state of the art methods. We qualitatively compare the optical flow predicted from our model to that of the (Zhu et al., 2018a) in Fig. 4. We show optical flow predicted from various test sequences from datasets proposed by (Scheerlinck et al., 2018; Mueggler et al., 2017) in Fig. 5. Note that these test sequences do not have ground truth optical flow to be compared against. We also provide the video of the predicted optical flow for most of the sequences in the accompanying supplementary video.

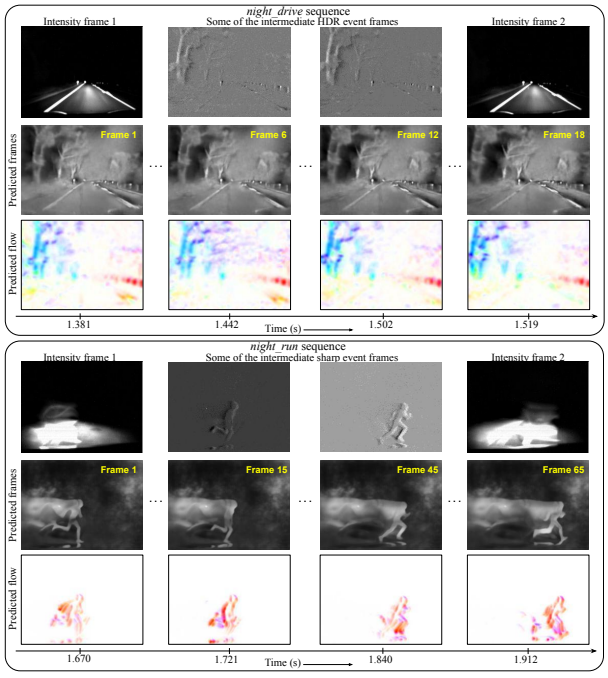

Fig. 6. The top figure shows the *night_drive* sequence shot in low-light conditions, demonstrating the ability of event sensors to sense objects at a high dynamic range, allowing the prediction of optical flow in extreme challenging cases. The *night_run* sequence combines two challenging scenarios, low-light and motion blur. With the help of event sensors we are able to predict the optical flow and intensity images at an effective rate of 1300 frames per second.

4.4. Advantages of event-based optical flow prediction

In this section, we demonstrate the advantages event sensors can provide over conventional image sensors for challenging scenes with fast motion and high dynamic range. In Fig. 1, we show an indoor scene with significant motion blur in the acquired image frames. A significant temporal information has also been lost between the two intensity frames. However, due to the high temporal resolution of the event sensors we are able

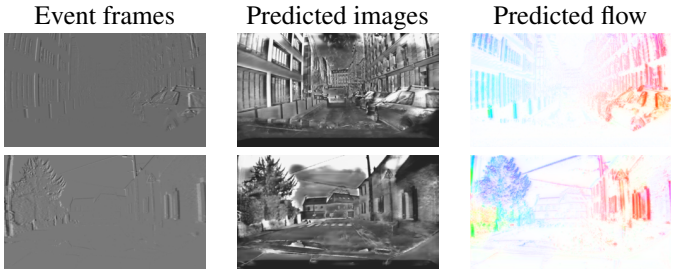

Fig. 7. Reconstruction results obtained from dataset proposed in (Perot et al., 2020). The dataset is collected using a 1MP resolution ATIS sensor which acquires only the event sensor data and no intensity frames. We observe that our proposed algorithm is able to generalize well to this new dataset.

to reconstruct multiple intensity frames, 60 in this case, between the successive intensity frames. Some of the 60 optical flow predictions have been shown in Fig. 1. Effectively, for this case, the intensity image and optical flow are being predicted at 1200 frames per second. This is a very high temporal resolution compared to many commercially available image sensors.

In Fig. 6, we consider two more cases. A *night_drive* sequence which is captured in extreme low-light conditions and a *night_run* sequence which combines both the extreme low-light and the fast scene motion cases. These two sequences are obtained from the dataset proposed in (Scheerlinck et al., 2018). In the *night_drive* sequence, the acquired intensity frames are under-saturated with most of the frame being dark. However, the intensity frames reconstructed from the event sensor reveals most of the details such as trees on the roadside. The *night_run* sequence reveals the high dynamic range and high temporal resolution nature of the event sensor. In this sequence, a person runs across the road in an extremely low-light scenario lit by only car headlamps. The acquired intensity frames are severely blurred along with parts of the image being saturated. Again, the intensity frames reconstructed from the event sensor data reveal the full details of the scene being captured. In this particular case, the intensity frames and optical flow are being reconstructed at an effective frame rate of 1300 frames per second. These examples clearly demonstrate the advantages of obtaining the optical flow directly from the event sensor data.

4.5. Generalization of the algorithm

4.5.1. Generalization to novel sensors:

The proposed algorithm is built assuming a specific category of event sensor where a positive or negative event is triggered when there is a change in the intensity. As long as this assumption is satisfied, we believe that the proposed algorithm should be able to predict the intensity image as well as the optical flow. To verify this, we considered a new dataset proposed in (Perot et al., 2020), collected using a 1 megapixel ATIS (Asynchronous Time-based Image Sensor) sensor (Posch et al., 2014). This dataset is sufficiently different from the one that we have used for training. The resolution of ATIS is far larger than the DAVIS sensor and the sensor technology is developed independently of the DVS/DAVIS sensor family. We provide the predicted optical flow and intensity images in Fig. 7 without training the proposed algorithm on this novel dataset. We observe that our proposed algorithm is able to generalize well showing that the algorithm works well with different types of sensors.

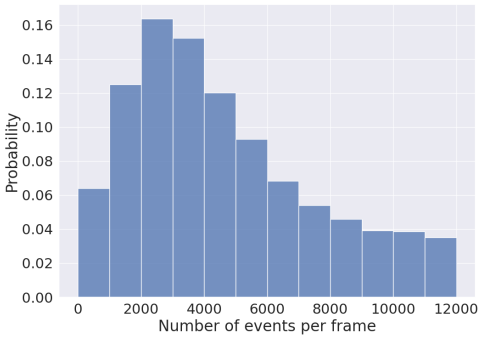

Fig. 8. Histogram of number of events per frame in the SBT strategy used to form event frames for training.

	indoor flying 1		indoor flying 2		indoor flying 3	
Events/frame	AEE	% outliers	AEE	% outliers	AEE	% outliers
1000	0.83	2.04	0.97	2.89	1.05	3.04
3000	0.49	0.02	0.55	0.05	0.53	0.03
5000	0.613	0.2	0.736	0.21	0.711	2.4
7000	0.842	1.05	1.04	2.4	1.02	2.27

Table 2. Quantitative optical flow comparison for different number of events per frame. Optical flow accuracy is highest for event frames with 3000 **events per frame and degrading gracefully for other values of the number of events.**

4.5.2. Generalization to new event rates:

We chose the *stacking by number* (SBN) strategy for event frame generation due to its property of being able to adapt to slow and fast motions. However, our proposed network was trained by generating frames with the *stacking by time* (SBT) strategy. In this strategy events from a fixed time interval are grouped into frames. We note that, the number of events in each fixed time interval can vary depending on the texture and relative camera motion. We provide the distribution of the number of events in a fixed time interval averaged across all sequences from the training set in Fig. 8. We observe that by using the SBT approach, we are training our algorithm for event frames containing different number of events per frame. However, a majority of the event frames contain number of events in the range $[2000, 4000]$. Hence, by using 3000 events per frame there's no major domain shift while testing. In Table 2, we show the quantitative results on optical flow accuracy for $1000, 3000, 5000$ and 7000 events per event frame.

4.6. Ablation studies

4.6.1. Choice of distance metric for intensity image supervision

In Eq. (2), we introduced a gradient-based L1 distance metric suitable for supervising intensity frame prediction from event sensors. Here, we evaluate the effectiveness of our proposed metric against other common metrics used for supervising image regression problems. We particularly consider two different cost functions, one based on pixel-wise error and the other based on perceptual similarity metric. For pixel-wise error we

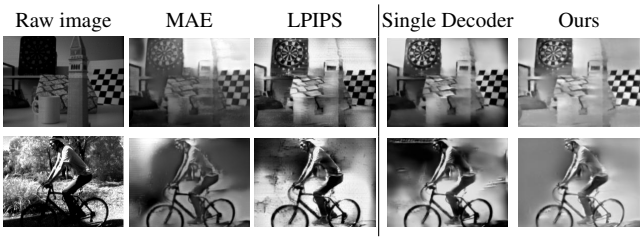

Fig. 9. We compare the effect of various architectural and supervision choices on intensity image estimation with respect to our proposed method. We show intensity image estimates for two different sequences obtained when using different two different cost functions, mean absolute error (MAE) and a perceptual metric LPIPS (Zhang et al., 2018). We also show intensity image estimates when using a single decoder to predict both the intensity frame and the optical flow.

consider the mean absolute error (MAE) defined as,

$$d(\hat{I}, I) = \frac{1}{M} \sum \|(\hat{I} - I) \odot m\|_1 \tag{8}$$

where I and $\hat{I}$ are respectively the ground truth and the predicted intensity images. The mask m is again used to mask the pixels which are saturated in the low dynamic range intensity images. In (Rebecq et al., 2019a), the authors use a learned perceptual similarity metric proposed in (Zhang et al., 2018) for supervising intensity image prediction from event data. We also use this perceptual metric called Learned Perceptual Image Patch Similarity(LPIPS)(Zhang et al., 2018) as the distance metric between our predicted and the ground truth intensity images. For a fair comparison, we retrain our proposed network on these two metrics with the same hyperparameters as used for the main experiment. In Fig. 9, we qualitatively compare the intensity images obtained by using the MAE and the LPIPS metrics. The MAE distance metric wrongly penalizes the neural network to predict the absolute intensity at each pixel which cannot be recovered from the event sensor data alone. As we use real data to train our proposed network, the mismatch in the dynamic range of the input event data and the ground truth intensity images make the LPIPS metric unsuitable. When using pixel-wise loss, the image regions which do not match the dynamic range can be masked. Such a flexibility is not provided by perceptual metrics such as LPIPS. Thus, we observe that the predicted images contain artifacts when using the MAE and the LPIPS metrics. From Table 3, we also see that the MAE and LPIPS metrics affect the accuracy of the predicted optical flow. Hence, our proposed gradient-based L1 metric performs better for the case of training with real data than other metrics for intensity image regression.

4.6.2. Single decoder network to predict intensity image and optical flow

Our network is trained in a multi-task learning fashion with a single encoder and two decoders for the two different tasks of intensity image and optical flow prediction. However, it is also possible to use only a single decoder to predict both the intensity image and optical flow. This leads to reduction in the number of parameters that need to be trained, hence reducing the amount of data required to train the network. We explored

	indoor flying 1		**indoor flying 2**		**indoor flying 3**	
	AEE	% outliers	AEE	% outliers	AEE	% outliers
MAE	0.57	0.04	0.63	0.75	0.61	0.07
LPIPS	0.53	0.1	0.58	0.5	0.58	0.1
Single Decoder	0.54	0.6	0.61	0.1	0.59	0.23
Ours	**0.49**	**0.02**	**0.55**	**0.05**	**0.53**	**0.03**

Table 3. We quantitatively compare the accuracy in optical flow estimation when the intensity image is supervised with mean absolute error (MAE) and LPIPS (Zhang et al., 2018). We also compare the optical flow accuracy for the case when a single decoder is used to predict both the intensity images and the optical flow.

Network	Number of parameters	Run time at resolution	
		180×240	256×256
Two decoders	2.4 M parameters	4.91 ms	5.89 ms
Single decoder	1.9 M parameters	3.9 ms	4.8 ms

Table 4. Runtime of different networks. Our proposed framework can process more than 150 frames per second at a resolution of 256×256.

this option of training a single decoder network to predict both the intensity image and the optical flow. For this experiment, we use our proposed decoder network *decoderImg* as our base network to predict the intensity images. To this network we augment two additional convolutional layers for optical flow prediction with 2 channels as output. These convolutional layers take as input the feature maps from the final 2 layers of the *decoderImg* network. Again, for a fair comparison we use the same hyperparamters to train this network as the ones used for our main experiment as described in Sec. 4. We provide qualitative results of the intensity images predicted from the single decoder network in Fig. 9. We also compare the optical flow estimation accuracy quantitatively for the different ablation experiments in Table 3. It can be observed that using a single decoder reduces the performance of the algorithm on both the intensity and optical flow prediction. However, use of two different decoder networks does not increase the runtime significantly as shown in Table 4 The inference time of the different networks is computed on a machine with Nvidia TitanX GPU with Intel Xeon processor. We can see that our proposed framework can process more than 150 frames per second at a resolution of 256×256.

5. Conclusion

In this work, we propose an algorithm to simultaneously predict the intensity and optical flow from event sensor data. The optical flow prediction is self-supervised and hence does not require difficult to acquire ground truth optical flow for event data. As our algorithm requires as few as 3000 events per time-step, the optical flow is predicted at a very high temporal resolution of more than 1000 frames per second for scenes with large motion. This high temporal resolution prediction also enables our algorithm to handle any non-linear relative motion of the scene. Due to the sparse nature of event sensor data, the predicted optical flow is sparse as well, and predicting a dense optical flow from event data alone can be an interesting future direction.

References

Almatrafi, M., Hirakawa, K., 2020. Davis camera optical flow. IEEE TCI 6, 396–407. doi:10.1109/TCI.2019.2948787.

Bardow, P., Davison, A.J., Leutenegger, S., 2016. Simultaneous optical flow and intensity estimation from an event camera, in: Proceedings of the IEEE Conference on CVPR, pp. 884–892.

Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T., 2014. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits 49, 2333–2341.

Delbrück, T., Linares-Barranco, B., Culurciello, E., Posch, C., 2010. Activity-driven, event-based vision sensors, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, IEEE. pp. 2426–2429.

Fortun, D., Bouthemy, P., Kervrann, C., 2015. Optical flow modeling and computation: a survey. Computer Vision and Image Understanding 134, 1–21.

Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A., Conradt, J., Daniilidis, K., et al., 2019. Event-based vision: A survey. arXiv preprint arXiv:1904.08405 .

Gallego, G., Rebecq, H., Scaramuzza, D., 2018. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation, in: Proceedings of the IEEE Conference on CVPR, pp. 3867–3876.

Gers, F.A., Schmidhuber, J., Cummins, F., 1999. Learning to forget: continual prediction with lstm, in: ICANN, pp. 850–855 vol.2. doi:10.1049/cp:19991218.

Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J., 2019. Digging into self-supervised monocular depth estimation, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3828–3838.

Haessig, G., Cassidy, A., Alvarez, R., Benosman, R., Orchard, G., 2018. Spiking optical flow for event-based sensors using ibm's truenorth neurosynaptic system. IEEE transactions on biomedical circuits and systems 12, 860–870.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on CVPR, pp. 770–778.

Horn, B.K., Schunck, B.G., 1981. Determining optical flow, in: Techniques and Applications of Image Understanding, International Society for Optics and Photonics. pp. 319–331.

Jason, J.Y., Harley, A.W., Derpanis, K.G., 2016. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness, in: European Conference on Computer Vision, Springer. pp. 3–10.

Khoei, M.A., Ieng, S.h., Benosman, R., 2019. Asynchronous event-based motion processing: From visual events to probabilistic sensory representation. Neural computation 31, 1114–1138.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Liu, M., Delbruck, T., 2018. Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. British Machine Vision Conference .

Meister, S., Hur, J., Roth, S., 2018. Unflow: Unsupervised learning of optical flow with a bidirectional census loss, in: Thirty-Second AAAI Conference on Artificial Intelligence.

Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D., 2017. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. The International Journal of Robotics Research 36, 142–149.

Nagata, J., Sekikawa, Y., Hara, K., Aoki, Y., 2019. Foe-based regularization for optical flow estimation from an in-vehicle event camera, in: International Workshop on Advanced Image Technology (IWAIT) 2019, International Society for Optics and Photonics. p. 110492V.

Paredes-Vallés, F., Scheper, K.Y.W., De Croon, G.C.H.E., 2019. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. IEEE TPAMI .

Perot, E., de Tournemire, P., Nitti, D., Masci, J., Sironi, A., 2020. Learning to detect objects with a 1 megapixel event camera. arXiv preprint arXiv:2009.13436 .

Posch, C., Serrano-Gotarredona, T., Linares-Barranco, B., Delbruck, T., 2014. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. Proceedings of the IEEE 102, 1470–1484. doi:10.1109/JPROC.2014.2346153.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019a. Events-to-video: Bringing modern computer vision to event cameras, in: Proceedings of the IEEE Conference on CVPR, pp. 3857–3866.

Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D., 2019b. High speed and high dynamic range video with an event camera. IEEET-PAMI .

Reinbacher, C., Graber, G., Pock, T., 2016. Real-time intensity-image reconstruction for event cameras using manifold regularisation. arXiv preprint arXiv:1607.06283 .

Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X., Zha, H., 2017. Unsupervised deep learning for optical flow estimation, in: Thirty-First AAAI Conference on Artificial Intelligence.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Scheerlinck, C., Barnes, N., Mahony, R., 2018. Continuous-time intensity estimation using event cameras. arXiv preprint arXiv:1811.00386 .

Shedligeri, P., Mitra, K., 2019. Photorealistic image reconstruction from hybrid intensity and event-based sensor. Journal of Electronic Imaging 28, 063012.

Wang, L., Ho, Y.S., Yoon, K.J., 2019. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks, in: Proceedings of the IEEE CVPR, pp. 10081–10090.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR.

Zhou, T., Brown, M., Snavely, N., Lowe, D.G., 2017. Unsupervised learning of depth and ego-motion from video, in: Proceedings of the IEEE Conference on CVPR, pp. 1851–1858.

Zhu, A., Yuan, L., Chaney, K., Daniilidis, K., 2018a. Ev-flownet: Self-supervised optical flow estimation for event-based cameras, in: Proceedings of Robotics: Science and Systems, Pittsburgh, Pennsylvania. doi:10.15607/RSS.2018.XIV.062.

Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K., 2018b. The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters 3, 2032–2039.

Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K., 2018c. Unsupervised event-based optical flow using motion compensation, in: European Conference on Computer Vision Workshop, Springer. pp. 711–714.

Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K., 2019. Unsupervised event-based learning of optical flow, depth, and egomotion, in: Proceedings of the IEEE CVPR, pp. 989–997.

Supplementary Material

A. Implementation details

In our proposed algorithm both the intensity frames and the optical flow are jointly predicted and the predicted optical flow is used to impose the temporal consistency in the predicted intensity frames. Another strategy would be to first learn a neural network to predict the intensity frames from event frames. These intensity frames will then be used to learn another neural network with self-supervised warping loss (Jason et al., 2016) to predict optical flow from the input event frames. Although, in theory, this strategy sounds better than our proposed method, it has a practical problem. The predicted intensity frames will not be temporally consistent and can lead to errors in the optical flow prediction. This has been shown in (Rebecq et al., 2019a), where the authors use pre-computed optical flow to explicitly impose temporal consistency loss between successive predicted intensity frames during training. Without explicitly imposing the temporal consistency during training, the predicted intensity frames are inconsistent temporally and will affect the optical flow prediction accuracy. In the initial stages of learning, the output of the image prediction network in our algorithm is completely random and this may affect our joint training strategy. Hence, we give a head-start to the image prediction network by computing only the supervised intensity loss. In practice, we found that a head-start of 1000 iterations for the image prediction network was enough.

Block	Layer	IC	OC	K	S	P	Remarks
inLSTM	conv2d	1	64	3	1	1	tanh
	conv2d	64	64	3	1	1	tanh
	conv2d	64	128	3	1	1	tanh; hidden size of 32
Encoder	conv2d	32	32	3	2	1	ReLU
	conv2d	32	64	3	2	1	ReLU
	conv2d	64	128	3	2	1	ReLU
	conv2d	128	256	3	2	1	ReLU
ResBlock	conv2d	256	128	3	1	1	ReLU
	conv2d	128	128	3	1	1	ReLU
	conv2d	128	256	3	1	1	ReLU
decoderFlow	conv2d	256	128	3	1	1	ReLU
	conv2d	256	64	3	1	1	skip connection; upsampling; ReLU
	conv2d	128	32	3	1	1	skip connection; upsampling; ReLU
	conv2d	64	32	3	1	1	skip connection; upsampling
Coarse flow	conv2d	32	2	3	1	1	tanh
Fine flow	conv2d	32	2	3	1	1	tanh
decoderImg	conv2d	256	128	3	1	1	ReLU
	conv2d	256	64	3	1	1	skip connection; upsampling; ReLU
	conv2d	128	32	3	1	1	skip connection; upsampling; ReLU
	conv2d	64	32	3	1	1	skip connection; upsampling; ReLU
outLSTM	conv2d	32	64	3	1	1	tanh
	conv2d	64	64	3	1	1	tanh
	conv2d	64	128	3	1	1	tanh; hidden size of 32
Out layer	conv2d	32	1	3	1	1	sigmoid; image output

Table S.1. Detailed description of the full network architecture. The number and type of layers constituting each block of the network is shown. We only use 2D convolutional layers with input channels *IC*, output channels *OC*, kernel size *K*, stride *S* and a padding of *P* pixels to the input. Non-linearity used for each layer has also been shown. The decoder blocks consist of bilinear upsampling layer and skip connections from the corresponding encoder block. More details are provided in Sec. 4.1 of the manuscript.

In the training phase, the time interval between the successive raw image frames is uniformly divided into 5 equal time intervals and the corresponding events are accumulated into 5 event frames. We form our training set with such pairs of 5 event frames and the corresponding raw image frames. During training, we use 40 event frames and correspondingly 8 raw image frames, all in a sequence, of one video and input to our algorithm as one instance of the batch. The neural network is trained using our overall cost-function described in Eq. (7). The brightness constancy loss specified in Eq. (4) is applicable at all 40 time-steps. But, the intensity supervision specified in Eq. (1), is applicable only at 8 time-steps of the sequence.

Hyperparameter selection: The hyperparameters in our training are chosen empirically to be $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.01$ and $\lambda_4 = 0.001$. The value of $\lambda_1 = 1.0$ was chosen as a reference value. The parameter $\lambda_2 = 0.1$, however, should be chosen carefully due to the nature of our proposed algorithm. Note that λ_1 weighs the importance of the intensity loss, $\mathcal{L}_{im}$ in Eq. (1) while λ_2 weighs the importance of the self-supervised optical flow loss in Eq. (4). A large value of λ_2 will try to make the predicted intensity frames as smooth as possible and sometimes predict an uniform intensity over the whole frame. However, a very small value of λ_2 will not learn to predict the optical flow accurately. Hence, to choose λ_2 we uniformly sampled values in $[0.05, 0.5]$ and found that the value of $\lambda_2 = 0.1$ worked best in terms of optical flow accuracy. The hyperparamters λ_3 and λ_4 impose spatial smoothness over the predicted intensity frame and optical flow respectively. The hyperparameters were chosen empirically to not impose too much or too little smoothness based on qualitative observations of the network predictions. Note that $\lambda_4 < \lambda_3$ as λ_3 imposes smoothness on the predicted intensity image and is imposed only once for every 5 timesteps.

B. Dataset

To train our proposed algorithm we obtain dataset from (Mueggler et al., 2017). This dataset consists of multiple video sequences collected from a commercially available hybrid intensity and event-based sensor named DAVIS240 (Brandli et al., 2014). Each of the video sequences are approximately 60s in length and contain the raw image sensor data acquired at about 20 frames per second and the event sensor data acquired from the event sensor. Both the intensity and the event data have a spatial resolution of 180×240. The video sequences are collected in variety of environments such as indoor, outdoor, planar scenes and with varying motion patterns such as pure rotation, pure translation, 6-DoF rapid motion, *etc*. We used 20 different sequences from the dataset where 13 sequences were

used for training and the rest for testing. The training sequences were *boxes_6dof, boxes_translation, boxes_rotation, office_spiral, office_zigzag, outdoors_running, outdoors_walking, poster_6dof, poster_rotation, poster_translation, shapes_6dof, shapes_translation, shapes_rotation*. The validation and testing sequences were *dynamic_6dof, dynamic_translation, dynamic_rotation, slider_depth, slider_close*. Video sequences that are captured in similar environments were put in either the training set or the test set, but not both. E.g. sequences such as *boxes* or *dynamic* appear either in the training data or in the test data, but not in both.

We quantitatively evaluate our proposed optical flow algorithm with the ground truth optical flow available in the dataset proposed by Zhu *et al.* (Zhu et al., 2018b). This dataset, also known as *MVSEC* dataset (Zhu et al., 2018b), consists of event sequences captured using a available hybrid sensor named DAVIS346. This sensor has a spatial resolution of 260×346. The dataset also contains ground truth depth maps captured using a LiDAR and the relative 6-DoF pose captured using a motion-capture system. For quantitative evaluation, the ground truth optical flow is also provided in the dataset. The ground truth optical flow is computed using the ground truth depth and the relative 6-DoF camera pose (Zhu et al., 2018a) under the assumption of a static scene. As the authors assume static scene to compute ground truth optical flow, we exclude the *outdoor_driving* sequence from optical flow evaluation and only use the *indoor_flying* sequences. This is because the *outdoor_driving* sequence has many moving objects such as cars, pedestrians, etc. where the static scene assumption does not hold and hence affects the optical flow evaluation. For a fair evaluation, we follow the procedure defined in (Zhu et al., 2018a) to compute metrics for optical flow evaluation by computing the error only at pixels where an event has fired.

C. More qualitative comparison on intensity image prediction

Intensity image prediction from event sensor data has been investigated by many researchers in the past few years. In this work we predict intensity images in order to facilitate the learning of optical flow directly from event sensors. However, in order to predict a good estimate of the optical flow, the predicted intensity frames should be temporally consistent, have high dynamic range and be free of any noise or other artifacts.

In Fig. S.1 of supplementary material we provide qualitative comparison of the intensity images from our proposed method with that of the other state-of-the-art intensity image-only estimation algorithms such as (Scheerlinck et al., 2018; Rebecq et al., 2019a; Reinbacher et al., 2016). Images predicted from (Reinbacher et al., 2016) contains various artifacts such as trailing edges as it relies on the hand-crafted image prior based on event manifolds. The algorithm proposed in (Scheerlinck et al., 2018) is sensitive to noisy events and can be seen producing noisy intensity estimates when a large number of events are being triggered in the scene. This sensitivity arises from the lack of any spatial regularization in the complementary filter model proposed in (Scheerlinck et al., 2018). In (Rebecq et al., 2019a), the authors propose a learning based method to predict intensity images from event sensor data. The method proposed in (Rebecq et al., 2019a) demonstrate high quality intensity image prediction from event sensor data. In contrast to MR (Reinbacher et al., 2016) and CF (Scheerlinck et al., 2018) which produce images with trailing edge artifacts, our reconstructions are smooth and are free of most of the artifacts. As seen in Fig. S.1, the predicted intensity frames from our model are comparable to the state of the art, learning-based, intensity only reconstruction method E2Vid (Rebecq et al., 2019a). The comparisons for all the methods in Fig. S.1 can be better visualized in the accompanying supplementary video (link to the video).

The learning based algorithm proposed in (Rebecq et al., 2019a) eliminates most of the artifacts dominant in the event based intensity reconstruction such as bleeding edges. However, in some cases the reconstructed intensity images start to show a dark region as shown in Fig. S.2. We also observe that the dark region on the reconstructed images grow and occupy larger area as more frames are reconstructed. The images from (Rebecq et al., 2019a) were reconstructed by accumulating N_e events into an event voxel-grid consisting of 5 frames. We consider $N_e = 0.35 \times H \times W$, where H and W represent the sensor resolution and this has shown to produce impressive results in most cases. The scenes shown in Fig. S.2 are from the *indoor_flying* and the *outdoor_driving* sequences from (Zhu et al., 2018b). These sequences are acquired with a hybrid sensor with a sensor resolution of $H = 260$, $W = 346$. As our algorithm requires only a single event frame as input, we accumulate $N_e/5$ events into each event frame to predict the intensity frames. In Fig. S.2, we provide the qualitative comparison for the predicted images from the two sequences *indoor_flying* and the *outdoor_driving*.

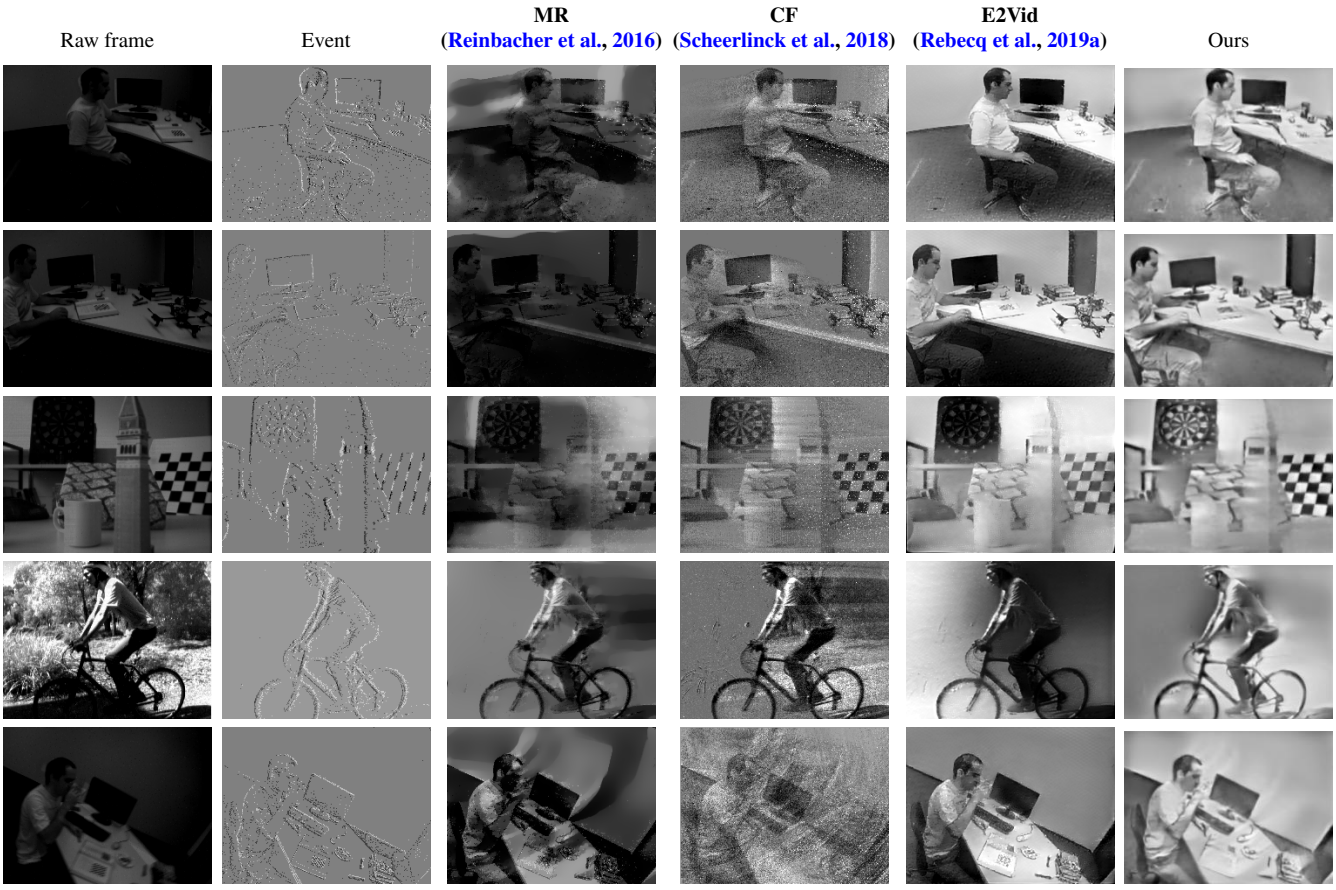

Fig. S.1. Qualitative comparison of intensity frame reconstruction on various event sensor sequences. We see that the reconstructed intensity images from our method do not have trailing edges and noisy regions as compared to (Reinbacher et al., 2016; Scheerlinck et al., 2018). Ours and (Rebecq et al., 2019a), both use learning based intensity reconstruction method and produce comparable results which can be better seen from the supplementary video material. (link to the video) *(If document is opened in Adobe Reader, videos can be viewed by clicking on the images).*

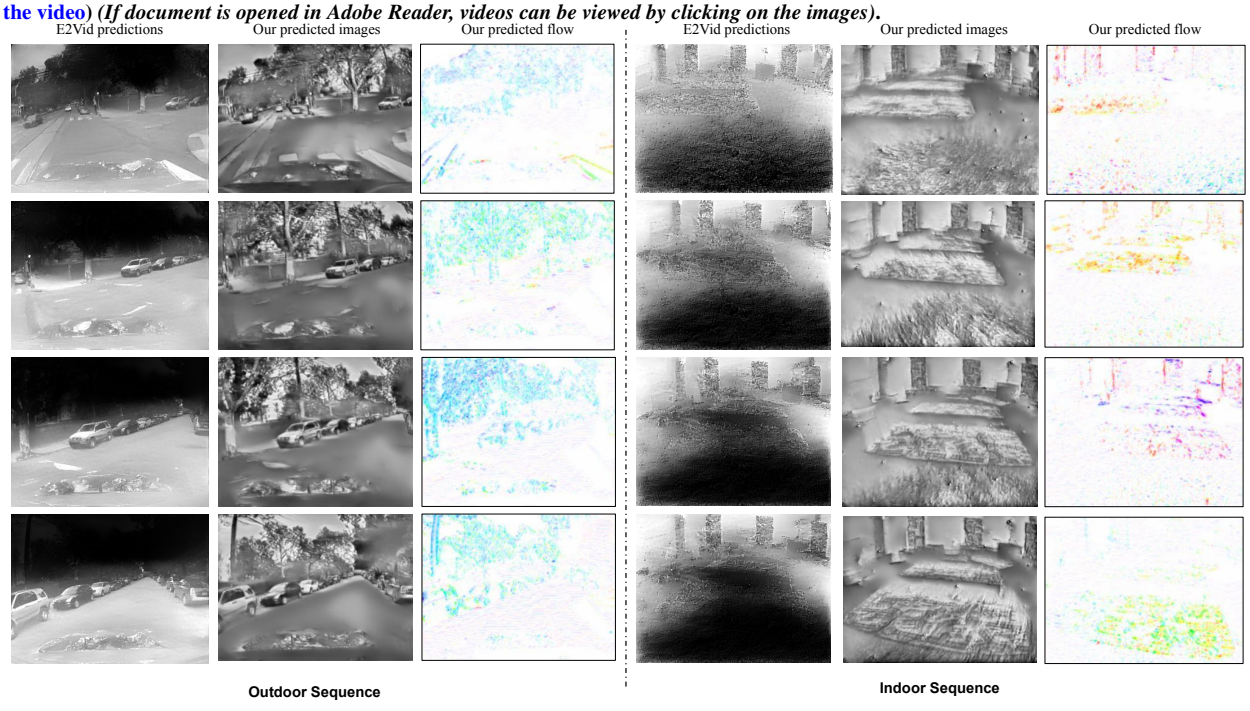

Fig. S.2. We show two sequences, one outdoor and another indoor, never seen by our method or the one proposed in (Rebecq et al., 2019a). We see that the images predicted by (Rebecq et al., 2019a) degrade with a growing dark region in the predicted intensity images in this particular case. Our proposed method generalizes enough to provide a reliable estimate of the intensity image and the optical flow.