Video synthesis from Intensity and Event Frames

Stefano Pini, Guido Borghi, Roberto Vezzani, and
Rita Cucchiara

University of Modena and Reggio Emilia, Italy
Department of Engineering "Enzo Ferrari"
{s.pini, name.surname}@unimore.it

Abstract. Event cameras, neuromorphic devices that naturally respond to brightness changes, have multiple advantages with respect to traditional cameras. However, the difficulty of applying traditional computer vision algorithms on event data limits their usability. Therefore, in this paper we investigate the use of a deep learning-based architecture that combines an initial grayscale frame and a series of event data to estimate the following intensity frames. In particular, a fully-convolutional encoder-decoder network is employed and evaluated for the frame synthesis task on an automotive event-based dataset. Performance obtained with pixel-wise metrics confirms the quality of the images synthesized by the proposed architecture.

Keywords: Video Synthesis · Event Camera · Event Frames · Automotive · Deep Learning

1 Introduction

Event cameras are optical sensors that asynchronously output events in case of brightness variations at pixel level. The major advantages of this type of neuromorphic sensors are the low power consumption, the low data rate, the high temporal resolution, and the high dynamic range [8]. On the other hand, despite exhibiting a higher power consumption and often a lower dynamic range, traditional cameras are able to record local information, like textures, and the majority of the computer vision algorithms are designed to work on this kind of data. Indeed, being able to apply existing algorithms to the output of event cameras could help the adoption of event-based sensors.

In this paper, aiming to conjugate the advantages of traditional and event cameras, we investigate the use of a deep learning-based method to interpolate intensity frames acquired by a low-rate camera with the support of the intermediate event data. Specifically, we exploit a fully-convolutional encoder-decoder architecture to predict intensity frames, relying on an initial or a periodic set of key-frames and a series of event frames, *i.e.* frames that collect the information captured by event cameras in a certain time interval.
Focusing on the automotive scenario, we employ a novel event-based dataset called DDD17 [4] (see Figure 1) and evaluate the feasibility of the proposed

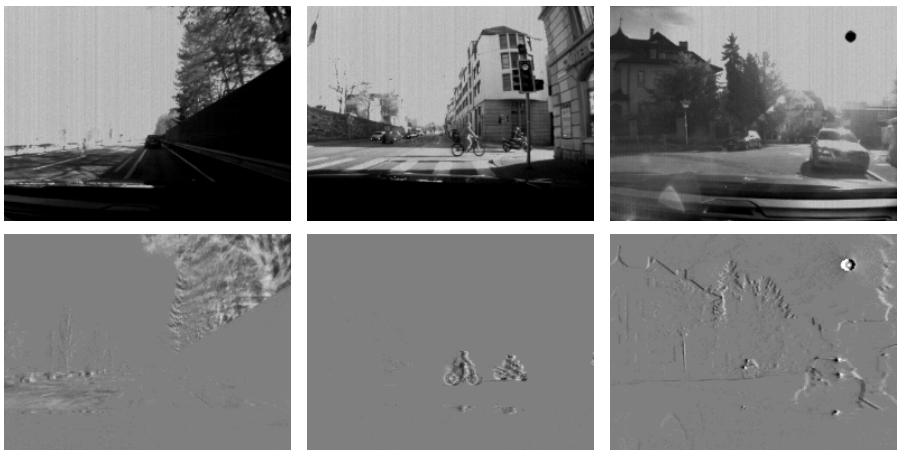

Fig. 1. Samples from the *DDD17* dataset. The first row contains the intensity grayscale images while the second one contains the event frames.

method with a wide set of pixel-level metrics. Quantitative and qualitative comparisons with a recent competitor [26] shows the superior quality of the images synthesized by the proposed model.

Summarizing, our contributions are twofold:

- We propose a fully-convolutional encoder-decoder architecture that combines traditional images and event data (as event frames) to interpolate consecutive intensity frames;
- We evaluate the effectiveness of the proposed approach on a public automotive dataset, assessing the ability to generate reasonable images and providing a fair comparison with a state-of-the-art approach.

2 Related Work

In the last years, event-based vision has increased its popularity in the computer vision community. Indeed, many novel algorithms have been proposed to deal with event-based data, produced by *Dynamic Vision Sensors* [11] (DVSs), like visual odometry [29], SLAM [17], optical flow estimation [8], and monocular [21] or stereo [1, 28] depth estimation.

Event cameras have also been exploited for the ego-motion estimation [7, 14], the real-time feature detection and tracking [15, 20], and the robot control in predator/prey scenarios [16]. Furthermore, it has been shown that event data can be employed to solve many classification tasks, such as the classification of characters [19], gestures [13], and faces [10]. Recently, an optimization-based algorithm that simultaneously estimates the optical flow and the brightness intensity was proposed in [3], while [18, 23] presented a manifold regularization method that

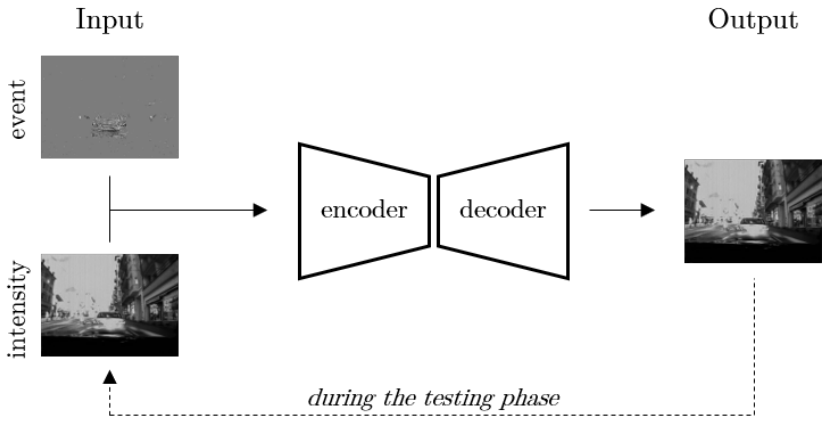

Fig. 2. Overview of the proposed method. The input of the encoder-decoder architecture is represented by the stack of an intensity and an event frame, while the output is the predicted intensity frame. During inference, the output at each step is used as the input intensity image in the following step.

reconstructs intensity frames from event data.
Lately, Scheerlinck *et al.* [26] proposed a complementary filter that combines image frames and events to estimate the scene intensity. The filter asynchronously updates the intensity estimation whenever new events or intensity frames are received. If the grayscale frames are missing, the estimation can be produced using events only.
This method is recent (at the time of writing) and outperforms previous existing works. Thus, we selected it as a baseline reference to evaluate our approach (see Section 4).

3 Proposed Method

In this Section, we formally define the event frame concept. Then, we present the investigated task from both a mathematical and an implementation point of view.

3.1 Event Frames

Following the notation of [14], the j-th event e_j provided by an event camera can be expressed as:

$$e_j = (x_j, y_j, t_j, p_j) \tag{1}$$

where x_j, y_j, and t_j are the spatio-temporal coordinates of a brightness change and $p_j \in \{-1, +1\}$ is the polarity of the brightness change (*i.e.* positive or negative variation).

An event frame can be defined as the pixel-wise integration of the events occurred in a time interval $[t, t+\tau]$:

$$\Psi_\tau(t) = \sum_{e_j \in [t,t+\tau]} p_j \tag{2}$$

where $e_j \in [t, t+\tau]$ means $\{e_j \,|\, t_j \in [t, t+\tau]\}$. In practice, an event frame can be formulated as a grayscale image that summarizes the events captured in a particular time interval. There is loss of information when the number of events exceeds the number of gray levels of the image.

3.2 Intensity Frame Estimation

We propose a method that corresponds to a learned parametric function F defined as:

$$F : \mathbb{R}^{2\times w\times h} \rightarrow \mathbb{R}^{w\times h} \tag{3}$$

that takes as input an intensity image $I_t \in \mathbb{R}^{w\times h}$ recorded at time t and an event frame $\Psi_\tau(t) \in \mathbb{R}^{w\times h}$ (which summarizes pixel-level brightness variations in the time interval $[t, t+\tau]$) in order to estimate the intensity image $\hat{I}(t+\tau) \in \mathbb{R}^{w\times h}$ at time $t+\tau$. w and h correspond to the width and the height of the event frames and the intensity images.
Formally, the synthesized image $\hat{I}(t+\tau)$ can be defined as:

$$\hat{I}(t+\tau) = F\left(I(t), \Psi_\tau(t), \theta\right) \tag{4}$$

where θ corresponds to the parameters of the function F.

3.3 Architecture

In practice, the parametric function F corresponds to an encoder-decoder architecture that predicts the intensity frame $\hat{I}(t+\tau)$ from the concatenation of an intensity frame $I(t)$ and an event frame $\Psi_\tau(t)$, as represented in Figure 2. In particular, the model is a fully-convolutional deep neural network with skip connections between layers i and $n-i$, with n corresponding to the total number of layers.

As in the *U-Net* architecture [24], the number of layers with skip connections is set to $n = 4$ with 128, 256, 512, 512 3×3 kernels in the encoder layers and with 256, 128, 64, 64 3×3 kernels in the decoder layers.
These skip-connected layers are preceded by two convolutional layers with 64 feature maps and followed by a convolutional layer with 1 feature map that projects the internal network representation to the final intensity estimation.

3.4 Training Procedure

The network is trained in a supervised manner using the *Mean Squared Error* (MSE) loss as objective function:

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 \tag{5}$$

Table 1. Pixel-wise metrics (lower is better) computed on the synthesized frames of *DDD17*.

Method	Norm ↓		Difference ↓		RMSE ↓		
	L_1	L_2	Abs	Sqr	Lin	Log	Scl
[26]	0.080	29.249	0.269	0.027	0.098	4.830	4.352
Ours	**0.027**	**8.916**	**0.179**	**0.007**	**0.040**	**4.048**	**3.571**

where y_i and $\hat{y}_i$ are respectively the i-th pixels of the ground truth and the generated image of the same size $N = w \cdot h$.
We optimize the network using the *Adam* optimizer [9] with learning rate $2 \cdot 10^{-4}$, $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a mini-batch size of 8.

During the training phase, two consecutive frames (one as input, one as ground-truth of the output) and the intermediate event frame (as input) are employed. During the testing phase, instead, in order to obtain a sequence of synthesized frames, the model iteratively receives the previously generated image as intensity input or a new key-frame after λ iterations.

4 Experimental Evaluation

In this section, we firstly present the dataset that has been employed to train and evaluate the proposed method. In the following, we report the procedure that we have adopted to evaluate the quality of the estimated intensity frames. Finally, we present and analyze the experimental results.

4.1 DDD17: End-to-end DAVIS Driving Dataset

Recently, Binas *et al.* [4] presented *DDD17: End-to-end DAVIS Driving Dataset*, the first open dataset of annotated DAVIS driving recordings. The dataset contains more than 12 hours of recordings captured with a *DAVIS* sensor [5] (some sample images are shown in Figure 1). Each recording includes both event data and grayscale frames along with vehicle information (*e.g.* vehicle speed, throttle, brake, steering angle). Recordings are captured in cities and highways, in dry and wet weather conditions, during day, evening, and night.
However, the quality of the gray-level images is low, the spatial resolution is limited to 346×260 pixels, and the framerate is variable (it depends on the brightness of the scene).

In our experiments, similar to [14], we use only the recordings acquired during the day. In contrast to Maqueda *et al.* [14], however, we create the train, validation, and test sets using different recordings.

4.2 Metrics

Inspired by [6], we employed a variety of metrics to check the quality of the generated images, being aware that evaluating synthesized images is in general

Table 2. Starting from the left, we report the percentage of pixels under three different thresholds, the *Peak Signal-to-Noise Ratio* (PSNR), and the *Structural Similarity* (SSIM) indexes, computed on the synthesized frames of *DDD17*. Higher is better.

Method	Threshold ↑			Indexes ↑	
	1.25	1.25^2	1.25^3	PSNR	SSIM
[26]	0.671	0.781	0.827	20.542	0.702
Ours	**0.775**	**0.848**	**0.875**	**29.176**	**0.864**

a difficult and still open problem [25].
In particular, we use distances (L_1 and L_2), differences (absolute and squared relative difference), the root mean squared error (in the linear, logarithmic, and scale-invariant version), and the percentage of pixels under a certain error threshold. Furthermore, with respect to [6], we introduce two additional metrics: the *Peak Signal-to-Noise Ratio* (PSNR) and the *Structural Similarity* index (SSIM) [27]. They are calculated to respectively evaluate the image noise level (in logarithmic scale) and the perceived image quality.
From a mathematical perspective, the PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{m \cdot |I|}{\sum_{y \in I} (\, y - \hat{y}\,)^2} \right) \tag{6}$$

where I is the ground truth image, $\hat{I}$ is the synthesized image, and m is the maximum possible value of I and $\hat{I}$. $\hat{y} \in \hat{I}$ corresponds to the element of the generated image at the same location of $y \in I$. In our experiments $m = 1$.
The SSIM is defined as:

$$\text{SSIM}(p, q) = \frac{(2\mu_p\mu_q + c_1)(2\sigma_{pq} + c_2)}{(\mu_p^2 + \mu_q^2 + c_1)(\sigma_p^2 + \sigma_q^2 + c_2)} \tag{7}$$

Given two windows $p \in I$, $q \in \hat{I}$ of equal size 11×11, $\mu_{p,q}$, $\sigma_{p,q}$ are the mean and variance of p and q while σ_{pq} is the covariance of p, q.
c_1 and c_2 are defined as $c_1 = (0.01 \cdot L)^2$ and $c_2 = (0.03 \cdot L)^2$ where L is the dynamic range (*i.e.* the difference between the maximum and the minimum theoretical value) of I and $\hat{I}$. In our experiments $L = 1$.

4.3 Experimental Results

We analyze the quality of the intensity estimations produced by our approach and by the method presented in [26] employing the pixel-wise metrics reported in Section 4.2.
In the experiments, we empirically set the number of consecutive synthesized frames (*i.e.* the sequence length) to $\lambda = 6$. It is worth noting that, within a sequence, the input intensity frame of the proposed method is the intensity estimation of the previous step except for the initial key-frame. We adapt the input images of *DDD17* to match the architecture requirements: the input data is resized to a spatial resolution of 256×192.

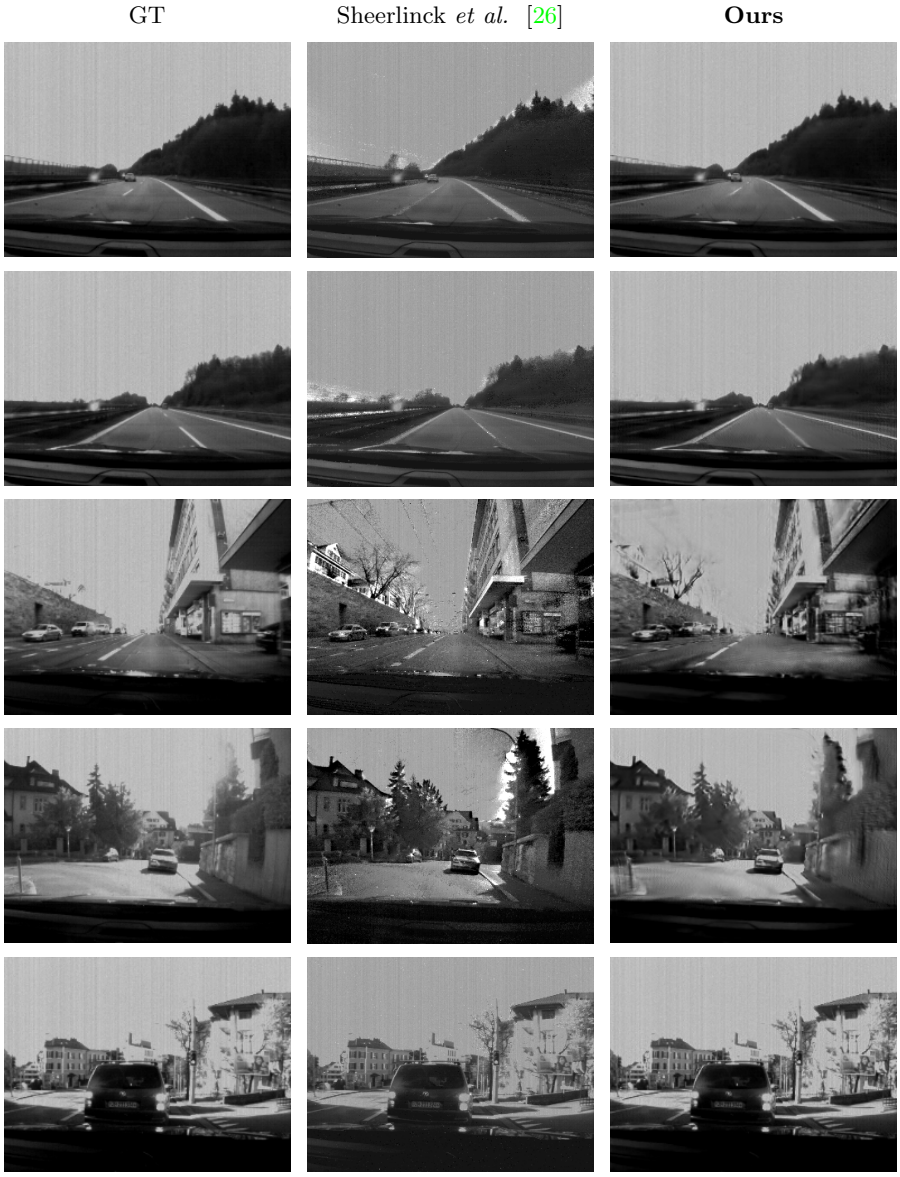

Fig. 3. Samples of synthesized frames produced by our method (last column) and the one of Scheerlinck *et al.* [26] (second column), while the first column contains ground truth images. As shown, the proposed method produces less artefacts, in the form of black or white spots, maintaining a good level of details, and it is able to preserve the overall structure and appearance of the original scene.

Quantitative results are reported in Table 1 and in Table 2. As it can be seen, the proposed method outperforms the competitor with a clear margin in every evaluation. In particular, PSNR and SSIM confirm the fidelity of the representation and the good level of perceived similarity between the generated and the ground truth images, respectively. Indeed, compared to the output of [26], frames synthesized by our method contain less artifacts and shadows and the overall structure of the scene is better preserved.
Visual examples, which are reported in Figure 3, highlight the ability of the proposed network to correctly handle the input event frames.

Finally, we investigate the performance of a traditional vision-based detection algorithm tested on the generated images. We adopt the well-known object detection network `Yolo-v3` [22], pre-trained on the *COCO* dataset [12], to investigate the ability of the proposed method to preserve the appearance of objects which are significant in the automotive context, like pedestrians, trucks, cars, and stop signals.
Since ground truth object annotations are not available in the dataset, we first run the object detector on the real images contained in *DDD17*, obtaining a sort of ground truth annotation. Then, we run `Yolo-v3` on the generated images and compare these detections with the produced annotations.
Results are expressed in terms of *Intersection-over-Union* (IoU) [2], which is defined as follows:

$$IoU(A, B) = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B| - |A \cap B|} \tag{8}$$

where A and B are the bounding boxes found in the original and the generated frames, respectively. A detection is valid if:

$$IoU(A, B) > \tau, \ \tau = 0.5 \tag{9}$$

A weighted object detection score is also employed: each class contributes to the final average according to its associated weight computed as the number of its occurrence on the total number present in the test sequences.
We obtained a mean Intersection-over-Union of 0.863 (the maximum reachable value is 1) with 61% of valid object detections. We believe that these results are remarkably promising because they show that the generated frames are semantically similar to the real ones. Therefore, the proposed method can be an effective way to apply traditional vision algorithms to the output of event cameras.

5 Conclusion

In this work, we have presented a deep learning-based method that performs intensity estimation given an initial or periodic collection of intensity key-frames and a group of events.
The model relies on a fully convolutional encoder-decoder architecture that

learns to combine intensity and event frames to produce updated intensity estimations. The experimental evaluation shows that the proposed method can be effectively employed to the intensity estimation task and that it is a valid alternative to current state-of-the-art methods.
As future work, we plan to test the framework on additional datasets as well as to take into account the long-term temporal evolution of the scene.

References

1. Andreopoulos, A., Kashyap, H.J., Nayak, T.K., Amir, A., Flickner, M.D.: A low power, high throughput, fully event-based stereo system. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 7532–7542 (2018) 2
2. Ballotta, D., Borghi, G., Vezzani, R., Cucchiara, R.: Fully convolutional network for head detection with depth images. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 752–757. IEEE (2018) 8
3. Bardow, P., Davison, A.J., Leutenegger, S.: Simultaneous optical flow and intensity estimation from an event camera. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 884–892 (2016) 2
4. Binas, J., Neil, D., Liu, S.C., Delbruck, T.: Ddd17: End-to-end davis driving dataset. Workshop on Machine Learning for Autonomous Vehicles (MLAV) in ICML 2017 (2017) 1, 5
5. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits **49**(10), 2333–2341 (2014) 5
6. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Neural Information Processing Systems (2014) 5, 6
7. Gallego, G., Lund, J.E., Mueggler, E., Rebecq, H., Delbruck, T., Scaramuzza, D.: Event-based, 6-dof camera tracking from photometric depth maps. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(10), 2402–2412 (2018) 2
8. Gallego, G., Rebecq, H., Scaramuzza, D.: A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition. vol. 1 (2018) 1, 2
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980 5
10. Lagorce, X., Orchard, G., Galluppi, F., Shi, B.E., Benosman, R.B.: Hots: a hierarchy of event-based time-surfaces for pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(7), 1346–1359 (2017) 2
11. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128 x 128 120db 30mw asynchronous vision sensor that responds to relative intensity change. In: 2006 IEEE International Solid State Circuits Conference-Digest of Technical Papers. pp. 2060–2069. IEEE (2006) 2
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision. Springer (2014) 8
13. Lungu, I.A., Corradi, F., Delbrück, T.: Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In: IEEE International Symposium on Circuits and Systems (ISCAS). pp. 1–1. IEEE (2017) 2

14. Maqueda, A.I., Loquercio, A., Gallego, G., Garcıa, N., Scaramuzza, D.: Event-based vision meets deep learning on steering prediction for self-driving cars. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 5419–5427 (2018) 2, 3, 5
15. Mitrokhin, A., Fermuller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. arXiv preprint arXiv:1803.04523 (2018) 2
16. Moeys, D.P., Corradi, F., Kerr, E., Vance, P., Das, G., Neil, D., Kerr, D., Delbrück, T.: Steering a predator robot using a mixed frame/event-driven convolutional neural network. In: 2016 Second International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP). pp. 1–8. IEEE (2016) 2
17. Mueggler, E., Rebecq, H., Gallego, G., Delbruck, T., Scaramuzza, D.: The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. The International Journal of Robotics Research **36**(2), 142–149 (2017) 2
18. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. International Journal of Computer Vision **126**(12), 1381–1393 (2018) 2
19. Orchard, G., Meyer, C., Etienne-Cummings, R., Posch, C., Thakor, N., Benosman, R.: Hfirst: a temporal approach to object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(10), 2028–2040 (2015) 2
20. Ramesh, B., Zhang, S., Lee, Z.W., Gao, Z., Orchard, G., Xiang, C.: Long-term object tracking with a moving event camera. In: British Machine Vision Conference (2018) 2
21. Rebecq, H., Gallego, G., Scaramuzza, D.: Emvs: Event-based multi-view stereo. In: British Machine Vision Conference (2016) 2
22. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 8
23. Reinbacher, C., Graber, G., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. In: British Machine Vision Conference (2016) 2
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 4
25. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Neural Information Processing Systems. pp. 2234–2242 (2016) 6
26. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. Asian Conference on Computer Vision (2018) 2, 3, 5, 6, 7, 8
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 6
28. Zhou, Y., Gallego, G., Rebecq, H., Kneip, L., Li, H., Scaramuzza, D.: Semi-dense 3d reconstruction with a stereo event camera. In: European Conference on Computer Vision (2018) 2
29. Zhu, A.Z., Atanasov, N., Daniilidis, K.: Event-based visual inertial odometry. In: IEEE International Conference on Computer Vision and Pattern Recognition. pp. 5816–5824 (2017) 2