IMAGE RECONSTRUCTION FROM EVENT CAMERAS FOR AUTONOMOUS DRIVING

Daniel Dauner
University of Tübingen
daniel.dauner@gmail.com

ABSTRACT

Event cameras are novel sensors that output a stream of asynchronous per-pixel brightness changes called 'events' rather than capturing brightness images. They offer outstanding performance in capturing high-speed motion and high dynamic range scenarios where traditional cameras are prone to fail. In particular, autonomous systems can benefit from event cameras by acquiring more robust visual information. Although the events theoretically encode a complete visual signal, event streams are incompatible with conventional computer vision techniques. Recent work has demonstrated the qualitative reconstruction of intensity images from event streams. This approach acts as a bridge between event-based vision and conventional computer vision. This report aims to introduce the field of event vision, present state-of-the-art image reconstruction techniques and examine their application for autonomous driving.

1 INTRODUCTION

Event cameras are bio-inspired vision sensors that operate fundamentally different to conventional cameras. Instead of acquiring a sequence of images at a fixed frame rate, event cameras record pixel-wise *intensity changes* (called *"events"*) asynchronously at the time they occur (Gallego et al., 2020). The output is a stream of events, each encoding the time, location, and polarity of the brightness change (as depicted in Figure 1). Event cameras offer several advantages over traditional cameras: high temporal resolution, high dynamic range, and low power and bandwidth requirements.

Event sensors are fast (in the order of µs), lightweight, and robust alternatives for acquiring visual information. Event cameras are advantageous where traditional cameras have shortcomings, e.g., in fast-motion scenarios and under challenging illumination conditions. Conventional cameras constitute a primary sensor for many self-driving cars, together with LiDAR and radar sensors (Chen et al., 2020). However, advanced autonomous systems require reliable vision to operate safely in their environment. Thus, event cameras could serve as an alternative or complementary vision sensor for perception tasks in self-driving.

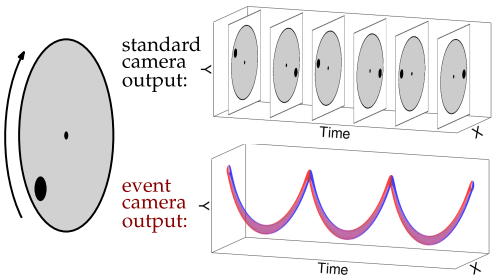


Figure 1: Comparison of standard and event cameras when recording a rotating disk with a black dot. While the conventional camera captures discrete frames, the event camera continuously reports pixel-wise brightness changes. Figure from Rebecq et al. (2019b).

Recent work shows that event cameras improve end-to-end steering prediction of autonomous systems, such as self-driving cars (Binas et al., 2017; Maqueda et al., 2018; Hu et al., 2020) and robots (Moeys et al., 2016). When using specialized algorithms, event cameras have shown remarkable performance in tasks such as optical flow (Benosman et al., 2013; Bardow et al., 2016; Zhu et al., 2018b), feature tracking (Kueng et al., 2016; Zhu et al., 2017), and visual odometry (Kim et al., 2016; Rebecq et al., 2016; 2017). On the other hand, such methods are highly task-specific and cannot offer a broadly applicable framework for processing event data in diverse tasks. While the results

Figure 2: Image reconstructions of E2VID from Rebecq et al. (2019b) in driving scenarios under challenging lighting conditions. The first row shows event frames, while the second row shows images from a regular camera. The reconstructions of E2VID in the third row can capture the scene in a high dynamic range and without motion blur.

indicate the value of event cameras, more advanced paradigms for autonomous cars and robots depend on various perception tasks from computer vision (Janai et al., 2020; Yurtsever et al., 2020). Since the output of event cameras is an asynchronous stream of events (instead of conventional image frames), established methods from computer vision are not directly applicable.

An approach in the literature to overcome this limitation is the reconstruction of conventional intensity images from event cameras, which allows the visualization of events or the application of off-the-shelf computer vision algorithms. Early approaches focused on reconstructing images by simultaneously estimating the event camera movement and temporal brightness gradients (Cook et al., 2011; Kim et al., 2014). Bardow et al. (2016) formulated reconstruction as an optimization problem to estimate intensity images and optical flow based on an energy minimization problem with a sliding window. Another approach is to filter the event stream, e.g., spatially via time-surfaces (Munda et al., 2018), or temporally (Scheerlinck et al., 2018), and directly integrating the events in a pixel-wise manner. This research insight focuses on the work of Rebecq et al. (2019a;b), which proposed E2VID, a neural network trained to reconstruct images from events in a supervised fashion. E2VID achieved a significant leap in reconstruction quality compared to previous work and still constitutes a state-of-the-art method in the field. Section 3 analyzes E2VID and presents follow-up research of the approach (Scheerlinck et al., 2020; Stoffregen et al., 2020). Before that, Section 2 provides a general background in event vision and examines several approaches to represent event data. Event cameras are a radically different way to perceive visual information and promise to be particularly impactful in autonomous driving. Section 4 discusses the challenges and opportunities of event cameras for autonomous driving and the role of image reconstruction.

2 BACKGROUND

Event sensors are inspired by the functionality of the human retina. Increments and decrements in light stimulate photoreceptor cells that subsequently send signals (called *"spikes"*) to the brain (Posch et al., 2014). This section formally presents how event cameras operate and introduces established data representations of events.

2.1 EVENT GENERATION MODEL

Event sensors report an *asynchronous* stream of events that are *independently* triggered by pixels for logarithmic changes in brightness $L = \log I$. In a noise-free scenario, this can be formalized by the ideal event generation model (Mueggler et al., 2018; Gallego et al., 2020). Here an event is formulated as a tuple $e_k = (\mathbf{x}_k, t_k, p_k)$ that occurs at time t_k and at a pixel $\mathbf{x}_k = (x_k, y_k)^\top$, if the brightness increment exceeds a contrast threshold C. The condition can be formulated as

$$p_k(L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t - \Delta t_k)) > C, \tag{1}$$

where Δt_k is the elapsed time since the last event at pixel $\mathbf{x}_k$, and $p_k \in \{-1, 1\}$ is the polarity (or sign) of the brightness change. The model can be extended for colored events (Li et al., 2015; Scheerlinck et al., 2019). The threshold C depends on bias currents on the sensor and can be fixed by the user for the given conditions (Nozaki & Delbruck, 2017).

An event vision sensor has several benefits in design. The pixel circuit is fast in detecting an event, and the camera can timestamp the event with microsecond resolution. Therefore, event cameras have a *high temporal resolution* without motion blur like conventional cameras. The pixels operate independently without waiting for global exposures, resulting in a *low latency* (sub-millisecond). Third, each pixel can adapt to bright and dark stimuli due to its independence and logarithmic operation. Thus, the event sensor can acquire visual information with a *high dynamic range*.

Under ideal circumstances (noise-free, ideal sensor response), event sensors can capture real brightness images when integrating the events over time. The pixel log-intensity $\hat{L}(\mathbf{x}; t)$ is reconstructed by accumulating the events $e_k = (\mathbf{x}_k, t_k, p_k)$ as follows,

$$\hat{L}(\mathbf{x}; t) = L(\mathbf{x}; 0) + \sum_{0 < t_k \leq t} p_k C \delta_{\mathrm{K}}(\mathbf{x} - \mathbf{x}_k) \delta_{\mathrm{D}}(t - t_k), \tag{2}$$

where $L(\mathbf{x}; 0)$ is the true log-intensity offset at $t = 0$, δ_{K} and δ_{D} are the Kronecker and Dirac delta functions, respectively, which select the pixel to update (Mueggler et al., 2017). A step function approximates the continuous change of intensity. However, in practice, even a fixed threshold C highly varies depending on factors such as the temperature (Nozaki & Delbruck, 2017), manufacturing imperfections, and circuit noise (Wang et al., 2020), but also the polarity of the events (Stoffregen et al., 2020). Therefore, it must be assumed that C is neither constant nor uniform for each pixel. Furthermore, the log-intensity offset is unknown. Thus the reconstructed intensity is relative to $t = 0$. Scheerlinck et al. (2018) address the limitations by applying a pixel-wise filter on the event stream and merging the pixel intensity of a conventional camera. Nonetheless, filtering methods still suffer from artifacts caused by the noisy event stream while requiring a regular camera for static information.

2.2 EVENT REPRESENTATION & PROCESSING

Due to the novelty of event cameras, there is no consensus collection of algorithms to extract features from events. The following introduces common data representations of events and corresponding processing methods.

Individual Events $e_k = (\mathbf{x}_k, t_k, p_k)$ are used for processing methods that are applied for each incoming event. This allows minimal latency but requires heavy processing, especially at high event rates. Methods include Spiking Neural Networks (SNNs), along with probabilistic and deterministic filters. Examples in literature include: (Kim et al., 2014; Scheerlinck et al., 2018; Paredes-Vallés et al., 2019; Gehrig et al., 2020).

Event Packets: Packets $\varepsilon_k = \{e_k\}_{k=1}^{N}$ are aggregated groups of subsequent events that are processed together. The packet size N is either fixed or varies to capture a constant time interval Δt. Packets introduce latency but require fewer processing steps which can be crucial for real-time applications. Possible processing methods depend on the downstream representation of the package, such as event frames or voxel grids. Examples in literature: (Rogister et al., 2011; Mueggler et al., 2018; Rebecq et al., 2019a;b).

Event Frames: Event packets are converted into an image structure (2D grid) by a simple method, e.g., by pixel-wise summation of the events or accumulating the polarity. Thereby, the images

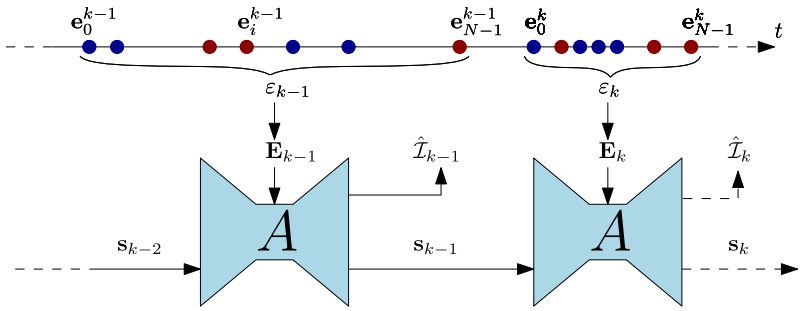


Figure 3: Overview of E2VID from Rebecq et al. (2019b). The incoming events (visualized as red/blue dots) are collected in packets ε_k with a fixed number of N events. The packets are converted into a 3D spatiotemporal tensor $\mathbf{E}_k$ that forms the input together with the previous state $\mathbf{s}_{k-1}$. The recurrent network outputs a new image reconstruction $\hat{\mathcal{I}}_k$ and the updated state $\mathbf{s}_k$ for each event packet.

represent a 2D pixel histogram of the events. This approach removes a large portion of temporal information. Nevertheless, event frames are often used in literature because they allow the application of well-studied computer vision algorithms, even though event frames do not share the statistics of natural images. It enables the usage of successful deep learning architectures of vision tasks, such as Convolutional Neural Networks (CNNs). Event frames are depicted in Figure 2 and 4. Examples in literature are: (Maqueda et al., 2018; Hu et al., 2020; Rebecq et al., 2016; Cook et al., 2011).

Voxel Grids: Similarly to frames, an event package is converted into a 3D grid of voxels representing a spatial-temporal histogram. Events are accumulated in voxels, where the voxel represents a pixel in a defined time interval. Although time discretization is needed, the voxel grid preserves more temporal information than event frames. Voxel grids are compatible with CNNs and other common vision techniques. However, the 3D grid comes with greater memory and computation requirements. Examples in literature: (Bardow et al., 2016; Wang et al., 2019; Rebecq et al., 2019a;b).

Reconstructed Images are regular brightness images reconstructed from events. In contrast to event frames, reconstructed images share the characteristics of natural images, thus becoming easy to interpret for the user. In addition, reconstructed images ideally capture the benefits of event cameras, such as a high dynamic range or no motion blur. Therefore, the images act as an intermediate representation for event data while being compatible with conventional vision algorithms and applicable to any downstream task (Rebecq et al., 2019a;b).

3 Image Reconstruction

In recent years, learning-based methods have been applied to image reconstruction (Barua et al., 2016; Wang et al., 2019). Rebecq et al. (2019a;b) introduced E2VID, a recurrent neural network to reconstruct video frames from an event stream. The network is trained using supervised learning on simulated data.

The event stream is represented as event packets $\varepsilon_k = \{e_k\}_{k=1}^N$ with a fixed number of N events. The goal is to learn a mapping to reconstruct an image $\hat{\mathcal{I}}_k \in [0,1]^{W \times H}$ for every incoming event package. Here, the mapping is a recurrent neural network based on the convolutional auto-encoder architecture of U-Net (Ronneberger et al., 2015). Before applying convolutions, each package ε_k is converted into a tensor $\mathbf{E}_k$ representing a spatiotemporal voxel grid. The events are discretized into B temporal bins, resulting in the input tensor $\mathbf{E}_k$ of shape $B \times W \times H$. The input is encoded by applying several downsampling layers, consisting of regular 2D convolutions and ConvLSTM layers (Shi et al., 2015). The network maintains a state $\mathbf{s}_k$ at each time step k, corresponding to a set of hidden states from the ConvLSTM layers. Thereby, the network becomes recurrent where at each step k, the tensor $\mathbf{E}_k$ is encoded based on the past state $\mathbf{s}_{k-1}$ (see Fig. 3). After the encoding stage, the network applies intermediate residual blocks and the decoder stage, which utilize bilinear upsampling and convolutions, together with skip connections of symmetric encoder layers. After a final prediction layer, the network outputs the reconstructed image $\hat{\mathcal{I}}_k$.

(a) Frame (b) E2VID (c) ECNN

Figure 4: Challenging scenes from the MVSEC dataset (Zhu et al., 2018a). Depicted are (a) the conventional frame images, (b) failed predictions of E2VID (Rebecq et al., 2019b), and (c) the predictions of ECNN (Stoffregen et al., 2020). Since ECNN shares the architecture with E2VID but was trained on augmented data to reduce the sim-to-real gap, the predictions are notably more robust.

Given the complexity of the task, a large dataset of event streams with corresponding ground truth images is needed to successfully apply supervised training. Furthermore, the purpose of reconstructing images is to incorporate the superior aspects of event cameras, such as high dynamic range or the absence of motion blur. Thus, images of conventional cameras would provide insufficient ground truth data. Therefore, E2VID is trained exclusively on synthetic data generated with the event simulator ESIM (Rebecq et al., 2018) and tested on actual event data.

The network is trained using combined loss with a reconstruction and temporal consistency component. The reconstruction term is a calibrated perceptual loss (LPIPS) (Zhang et al., 2018), where the target image $\mathcal{I}_k$ and reconstructed image $\hat{\mathcal{I}}_k$ are passed into a VGG-Net (Simonyan & Zisserman, 2014), pre-trained on Image-Net (Russakovsky et al., 2015). The loss corresponds to the distance of the VGG feature maps on multiple layers. The network learns to reconstruct images based on the statistics of natural images when minimizing the perceptual loss. On the other hand, the temporal consistency term is added to penalize temporal artifacts between successive frames. The loss is a weighted L_1-distance between subsequent reconstructions, where image $\hat{\mathcal{I}}_{k-1}$ is warped forward to time k with optical flow maps that are available during training. The term is weighted to reduce the penalty for occlusions between frames.

Compared to filtering-based methods, E2VID suffers less from artifacts and outperforms previous methods visually and quantitatively (see Rebecq et al. (2019b)). Moreover, the reconstructions surpass conventional cameras in low-light and high-speed scenarios, as shown in Figure 2. Due to the temporal resolution of the events, high frame-rate videos can be reconstructed (in the range of thousands of FPS). Furthermore, the reconstructed videos allow for processing event data with well-studied algorithms for conventional cameras directly. The authors demonstrate this by applying standard vision methods to reconstructions in tasks such as object classification and visual-inertial odometry, thereby outperforming previous methods specifically designed for event data.

Nevertheless, E2VID comes with limitations that are partially addressed by further research. Firstly, the reconstructions are computationally expensive, which limits the real-time applicability. For example, a forward pass of E2VID takes 93 ms for a 1280x720 image on a GPU, which would result in a 10 FPS video. Therefore, Scheerlinck et al. (2020) propose a smaller network, called FireNet, which reduces the number of parameters by 99% with minor trade-offs regarding the reconstruction quality. The reduction is achieved by having convolutions without down-sampling and with smaller kernels, as well as GRUs as recurrent units (Chung et al., 2014). Consequently, FireNet runs three times faster than E2VID on the same GPU (31.01 ms vs. 93.34 ms).

A second limitation of E2VID originates from training solely on simulated data. Thereby, inference with actual events is carried out exclusively on out-of-distribution data. Stoffregen et al. (2020) propose a new strategy for generating training data by emphasizing the contrast threshold C when

synthesizing data. The E2VID network is retrained on data generated with a wide range of contrast thresholds and noise augmentation to resemble actual event data closer. The training approach improves generalizability (see Fig. 4) and outperforms the conventionally trained E2VID across several benchmarks. The results highlight the vital role of synthetic data in learning-based reconstruction. The training data must share the statistics of the actual use case.

4 DISCUSSION

Autonomous systems can become more robust by introducing event cameras as visual sensors. With their temporal resolution, event cameras are ideal for tracking and detection in high-speed scenes. Their HDR capabilities enable excellent vision at night, even with glaring light sources (e.g., the sun or oncoming headlights). However, due to their novelty relative to traditional sensors, there are no established methods to process or fuse event data in existing self-driving systems.

While image reconstructions from events offer a familiar representation, they also introduce flaws. Learning-based approaches have to be trained exclusively on simulated data. Thus, the simulation must be tailored to the use case and the camera parameters. The estimation is prone to errors if the training distribution mismatches the real-world application despite precautions. Dissimilarities of simulated and real-world data are especially concerns for functional safety in autonomous driving. Furthermore, image reconstruction with learned models is computationally demanding, resulting in slow processing even with advanced hardware. Smaller networks can reduce computational cost (Scheerlinck et al., 2020) but still introduce latency and require hardware accelerators for processing (i.e., GPUs). Moreover, downstream applications require time to process the images for the actual task. The application in an autonomous system becomes infeasible when real-time requirements cannot be satisfied.

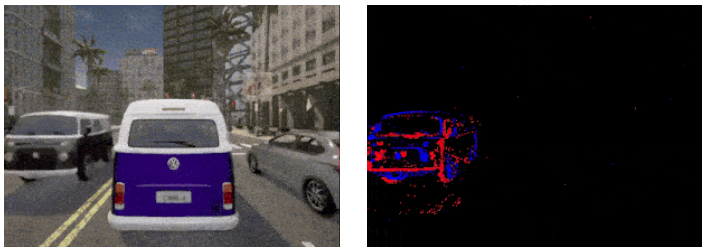
Most importantly, image reconstruction is an expensive intermediate step when applied to a downstream target task. Reconstruction techniques only transform the provided input and also discard positive aspects of event cameras. Theoretically, the event stream contains the same visual information but in a fast and highly compressed format. Event vision needs a generally applicable framework to process event streams directly. Spiking Neural Networks (SNNs) seem promising due to fitting input modality and the same bio-inspired nature. SNNs have shown promising results for specific tasks, such as optical flow estimation (Paredes-Vallés et al., 2019) or angular velocity regression (Gehrig et al., 2020). A combination of SNNs with neuromorphic processors would offer a low memory and low power vision approach that is desirable for autonomous systems (Galluppi et al., 2014).

Furthermore, conventional cameras and event cameras are not competing but complementary technologies. Autonomous systems also need visual information in static or slowly varying scenes. Traditional cameras are ideally suited for such scenarios. It is most likely beneficial to consider both cameras as input data, either separately or merged. In fact, event and frame-based cameras have been fused in literature to remove motion blur (Jiang et al., 2020; Lin et al., 2020; Pan et al., 2019), generate high-speed videos (Tulyakov et al., 2021), and to increase dynamic range (Scheerlinck et al., 2018). Enhancing videos with event data has potential in numerous areas, such as film-making or smartphone applications.

For autonomous driving, simulators offer an ideal environment to investigate how event cameras can be effectively integrated into self-driving cars. Particularly the CARLA simulator provides an implementation of event cameras[1] (see Fig. 5) and offers a customizable platform to test various sensor, lighting, and weather configurations (Dosovitskiy et al., 2017). Therefore, benchmarks can be generated in CARLA under challenging conditions, e.g., at high-speed scenes, at night-time, or with injected motion blur of the regular camera recordings. The benchmarks would allow the evaluation of existing approaches in challenging scenarios and examine how event cameras can be utilized to increase robustness.

Overall, image reconstruction remains an important task in current event vision research. Image reconstruction functions as a data representation of events. The representation allows for establishing a baseline for a task with conventional frame-based methods. Using off-the-shelf algorithms enables one to validate that event data offers a valuable contribution before creating more sophisti-

[1] https://carla.readthedocs.io/en/latest/ref_sensors/#dvs-camera



(a) RBG Camera (b) Event Camera

Figure 5: Sensor comparison in CARLA Dosovitskiy et al. (2017). The RBG camera (a) captures static information, while the event camera (b) retrieves dynamic information by only sensing pixel-wise brightness changes. The blue and red pixels in (b) show the polarity of the events.

cated techniques tailored to an approach. Furthermore, the reconstructions remain the most familiar and interpretable visualization for humans of event data.

5 Conclusion

Event cameras introduce a novel approach to how machines perceive and represent visual information. They offer several advantages compared to conventional cameras, such as high-speed capabilities, high dynamic range, low power, and low latency. Consequently, event cameras show convincing potential in autonomous driving, mainly to increase robustness in challenging scenarios. The field of event-based vision has several challenges ahead, especially the development of algorithms that unlock the unique properties of events. Intensity image reconstruction can reduce the development gap by making traditional computer vision accessible to event cameras. This research insight analyzed a line of methods for learning reconstructions as intermediate event camera representation. Learning-based approaches define current state-of-the-art by training neural networks to reconstruct images. However, such networks have high computational costs and can only be trained in simulation, restricting their application for autonomous driving. Overall, image reconstruction from events remains a vital discipline in the current state of research. The reconstructions enable a bridge to traditional computer vision, are valuable for prototyping, and visualize the outstanding capabilities of event cameras.

References

Patrick Bardow, Andrew J Davison, and Stefan Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 884–892, 2016.

Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9. IEEE, 2016.

Ryad Benosman, Charles Clercq, Xavier Lagorce, Sio-Hoi Ieng, and Chiara Bartolozzi. Event-based visual flow. *IEEE transactions on neural networks and learning systems*, 25(2):407–417, 2013.

Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd17: End-to-end davis driving dataset. *arXiv preprint arXiv:1711.01458*, 2017.

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

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Matthew Cook, Luca Gugelmann, Florian Jug, Christoph Krautz, and Angelika Steger. Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pp. 770–776. IEEE, 2011.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Francesco Galluppi, Christian Denk, Matthias C Meiner, Terrence C Stewart, Luis A Plana, Chris Eliasmith, Steve Furber, and Jörg Conradt. Event-based neural computing on an autonomous mobile platform. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2862–2867. IEEE, 2014.

Mathias Gehrig, Sumit Bam Shrestha, Daniel Mouritzen, and Davide Scaramuzza. Event-based angular velocity regression with spiking networks. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4195–4202. IEEE, 2020.

Yuhuang Hu, Jonathan Binas, Daniel Neil, Shih-Chii Liu, and Tobi Delbruck. Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6. IEEE, 2020.

Joel Janai, Fatma Güney, Aseem Behl, Andreas Geiger, et al. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020.

Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, and Yebin Liu. Learning event-based motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3320–3329, 2020.

Hanme Kim, Ankur Handa, Ryad Benosman, Sio-Hoi Ieng, and Andrew Davison. Simultaneous mosaicing and tracking with an event camera. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. doi: http://dx.doi.org/10.5244/C.28.26.

Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *European Conference on Computer Vision*, pp. 349–364. Springer, 2016.

Beat Kueng, Elias Mueggler, Guillermo Gallego, and Davide Scaramuzza. Low-latency visual odometry using event-based feature tracks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 16–23. IEEE, 2016.

Chenghan Li, Christian Brandli, Raphael Berner, Hongjie Liu, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. Design of an rgbw color vga rolling and global shutter dynamic and active-pixel vision sensor. In *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 718–721. IEEE, 2015.

Songnan Lin, Jiawei Zhang, Jinshan Pan, Zhe Jiang, Dongqing Zou, Yongtian Wang, Jing Chen, and Jimmy Ren. Learning event-driven video deblurring and interpolation. In *European Conference on Computer Vision*, pp. 695–710. Springer, 2020.

Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5419–5427, 2018.

Diederik Paul Moeys, Federico Corradi, Emmett Kerr, Philip Vance, Gautham Das, Daniel Neil, Dermot Kerr, and Tobi Delbrück. 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.

Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International Journal of Robotics Research*, 36(2):142–149, 2017. doi: 10.1177/0278364917691115. URL https://doi.org/10.1177/0278364917691115.

Elias Mueggler, Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. Continuous-time visual-inertial odometry for event cameras. *IEEE Transactions on Robotics*, 34(6):1425–1440, 2018.

Gottfried Munda, Christian Reinbacher, and Thomas Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision*, 126:1381–1393, 2018.

Yuji Nozaki and Tobi Delbruck. Temperature and parasitic photocurrent effects in dynamic vision sensors. *IEEE Transactions on Electron Devices*, 64(8):3239–3245, 2017. doi: 10.1109/TED.2017.2717848.

Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6820–6829, 2019.

Federico Paredes-Vallés, Kirk YW Scheper, and Guido CHE De Croon. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):2051–2064, 2019.

Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Linares-Barranco, and Tobi Delbruck. Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output. *Proceedings of the IEEE*, 102(10):1470–1484, 2014. doi: 10.1109/JPROC.2014.2346153.

Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. Evo: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2(2):593–600, 2016.

Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. 2017.

Henri Rebecq, Daniel Gehrig, and Davide Scaramuzza. Esim: an open event camera simulator. In *Conference on Robot Learning*, pp. 969–982. PMLR, 2018.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3857–3866, 2019a.

Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019b.

Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbruck. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 23(2):347–353, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Cedric Scheerlinck, Nick Barnes, and Robert Mahony. Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pp. 308–324. Springer, 2018.

Cedric Scheerlinck, Henri Rebecq, Timo Stoffregen, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Ced: Color event camera dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 156–163, 2020.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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 Conference on Computer Vision*, pp. 534–549. Springer, 2020.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16155–16164, 2021.

Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10081–10090, 2019.

Ziwei Wang, Yonhon Ng, Pieter van Goor, and Robert E. Mahony. Event camera calibration of per-pixel biased contrast threshold. *CoRR*, abs/2012.09378, 2020. URL `https://arxiv.org/abs/2012.09378`.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based feature tracking with probabilistic data association. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4465–4470. IEEE, 2017.

Alex Zihao 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 Robotics and Automation Letters*, 3(3):2032–2039, 2018a.

Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *arXiv preprint arXiv:1802.06898*, 2018b.