Joint Filtering of Intensity Images and Neuromorphic Events for High-Resolution Noise-Robust Imaging

Zihao W. Wang[†#§] Peiqi Duan[‡#] Oliver Cossairt[†] Aggelos Katsaggelos[†] Tiejun Huang[‡] Boxin Shi[‡*]
[†]Northwestern University [‡] Peking University
Project page: https://sites.google.com/view/guided-event-filtering

Abstract

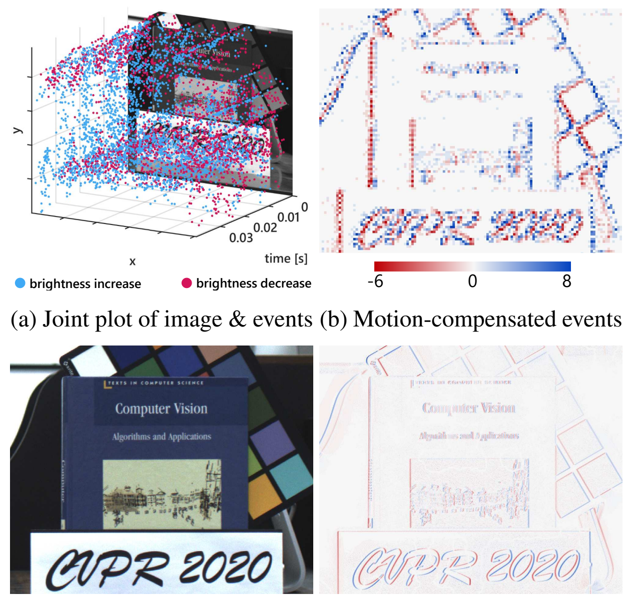
We present a novel computational imaging system with high resolution and low noise. Our system consists of a traditional video camera which captures high-resolution intensity images, and an event camera which encodes high-speed motion as a stream of asynchronous binary events. To process the hybrid input, we propose a unifying framework that first bridges the two sensing modalities via a noise-robust motion compensation model, and then performs joint image filtering. The filtered output represents the temporal gradient of the captured space-time volume, which can be viewed as motion-compensated event frames with high resolution and low noise. Therefore, the output can be widely applied to many existing event-based algorithms that are highly dependent on spatial resolution and noise robustness. In experimental results performed on both publicly available datasets as well as our new RGB-DAVIS dataset, we show systematic performance improvement in applications such as high frame-rate video synthesis, feature/corner detection and tracking, as well as high dynamic range image reconstruction.

1. Introduction

Recently, a new breed of bio-inspired sensors called event cameras, or Dynamic Vision Sensors (DVS), has gained growing attention with its distinctive advantages over traditional frame cameras such as high speed, high dynamic range (HDR) and low power consumption [22, 45]. Thus far, event cameras have shown promising capability in solving classical as well as new computer vision and robotics tasks, including optical flow and scene depth estimation [1, 31, 40, 49], high frame-rate HDR video synthesis [15, 30, 37, 38, 41, 43, 52, 55], 3D reconstruction and tracking [11, 19, 27, 36], visual SLAM [51], object/face detection [34, 35] and autonomous wheel steering [26].

Despite numerous advances in event-based vision [8],

[#] Equal contribution. [*] Corresponding author.
[§] Part of this work was finished while visiting Peking University.



(a) Joint plot of image & events (b) Motion-compensated events

(c) Captured image (d) Filtered events

Figure 1: Compared to traditional frame cameras, event cameras (*e.g.*, DAVIS240) can capture high-speed motion (a), but bear low resolution and severe noise (b). Our system jointly filters between a high-resolution image (c) and high-speed events to produce a high-resolution low-noise event frame (d), which can interface with downstream event-based algorithms with improved performance.

current event sensor prototypes, *e.g.*, DAVIS240, still bear low spatial resolution and severe noise (Fig. 1(a) & (b)). Moreover, the unique event sensing mechanism according to which each pixel individually responds to brightness changes and outputs a cloud of continuously timestamped address points (Fig. 1(a)) renders event-based super resolution and denoising elusively challenging. On the other hand, commercial frame sensors can easily acquire millions of pixels, and image-based super resolution and denoising algorithms are highly advanced after decades of development. These sensory and algorithmic imbalances motivate us to ask: Can we make complementary use of event and frame sensing? What is the unifying mechanism? How does their synergy benefit related visual tasks and applications?

To answer these questions, we build a hybrid camera system using a low-resolution event camera, *i.e.*, DAVIS240 and a high-resolution RGB camera. We establish a computational framework that bridges event sensing with frame sensing. Our system inherits the high-resolution property ($8\times$ higher than DAVIS) from the frame camera and is robust to event sensor noise.

Contributions:

• We propose a novel optimization framework, guided event filtering (GEF), which includes a novel motion compensation algorithm unifying event and frame sensing. By taking complimentary advantages from each end, GEF achieves high-resolution, noise-robust imaging.

• We build a prototype hybrid camera system and collect a novel dataset, *i.e.*, RGB-DAVIS. Validation experiments have been conducted on both publicly available datasets and RGB-DAVIS.

• We show broad applications of GEF to benefit optical flow estimation, high frame rate video synthesis, HDR image reconstuction, corner detection and tracking.

Limitations: Since our work is based on the assumption that frame sensing and event sensing have complementary advantages, one of the limitations is when one sensing mode under-performs significantly. For example, when the frame sensor suffers from significant blur or noise, our framework should only utilize event information, *i.e.*, to use events as both the guidance and the input. On the event side, events triggered from fast lighting variations are not modeled in our linear motion compensation model, and therefore may hinder the effectiveness of GEF due to incorrect flow estimation. Our hybrid camera does not preserve the low power consumption benefit of an event camera.

2. Related works

Event denoising. Event denoising is considered a preprocessing step in the literature [6, 7, 18, 24, 29]. Existing event denoising approaches exploit local spatial-temporal correlations, and label isolated events as noise to be canceled [53]. However, these denoisers face challenges when retrieving missing events for low contrast spatial texture. We address this issue by exploiting the correlation between events and an intensity image.

Event-based motion compensation. Motion compensation is an emerging technique to associate local events. It has shown benefits for downstream applications such as depth estimation [9], motion segmentation [48] and feature tracking [10]. The assumption is that local events are triggered by the same edge signal and should comply with the same motion flow [4]. The flow parameter can be estimated by maximizing the contrast of the histogram/image of the warped events [9]. Recent works have incorporated smooth constraints such as total variation [56].

Computational high speed cameras. The tradeoff between spatial resolution and temporal resolution in modern sensors introduces a fundamental performance gap between still cameras and video cameras. To address this issue, several methods [5, 13, 42] have emerged that utilize inter-frame correspondences via optical flow and/or space-time regularization. Hybrid cameras have been designed towards flexible [14], adaptive [59] sensing of high speed videos. Recently, a number of compressive video sensing prototypes [2, 17, 25, 39, 47] have been devised with additional spatio-temporal encoders and compressive sensing algorithms for data recovery and inference. Extensions of compressive sensing high-speed imaging have achieved single-shot 3D video recovery by incorporating active illumination [54].

Guided/joint image filters. The goal of guided/joint image filters is to transfer structural information from a reference image to a target image. The reference and the target can be identical, in which case the filtering process becomes an edge-preserving one [12, 16, 20, 46]. Although similar ideas of guided/joint image filtering (GIF) have been explored between RGB and near infrared (NIR) images [57], 3D-ToF [32], and hyperspectral data [33], the major challenge for applying GIF to event cameras is that events do not directly form an image and are spatio-temporally misaligned by scene motions or illumination variations.

3. Methods

In this section, we first briefly review the event sensing preliminaries in Sec. 3.1, and derive its relation to intensity/frame sensing in Sec. 3.2. Our framework guided event filtering (GEF) is then introduced in Sec. 3.3 (for the motion compensation step), Sec. 3.4 (for the joint filtering step) and Sec. 3.5 (for the implementation details).

3.1. Event sensing preliminaries

Consider a latent space-time volume ($\Omega \times T \in \mathbb{R}^2 \times \mathbb{R}$) in which an intensity field is sampled simultaneously by a frame-based camera which outputs intensity images $I(x, y; t)$ and an event camera which outputs a set of events, *i.e.*, $\mathcal{E} = \{e_{t_k}\}_{k=1}^{N_e}$, where N_e denotes the number of events. Each event is a four-attribute tuple $e_{t_k} = (x_k, y_k, t_k, p_k)$, where x_k, y_k denote the spatial coordinates, t_k the timestamp (monotonically increasing), p_k the polarity. $p_k \in \{-1, 1\}$ indicates the sign of the intensity variation in log space. *I.e.*, $p_k = 1$ if $\theta_t > \epsilon_p$ and $p_k = -1$ if $\theta_t < \epsilon_n$, where $\theta_t = \log(I_t + b) - \log(I_{t-\delta t} + b)$. b is an infinitesimal positive number to prevent $\log(0)$. I_t and $I_{t-\delta t}$ denote the intensity values at time t and $t - \delta t$, respectively, and ϵ_p and ϵ_n are contrast thresholds. We will use $\mathcal{L}_t$ to denote the log intensity at time t, *i.e.*, $\mathcal{L}_t \doteq \log(I_t + b)$. For now, we assume that I and $\mathcal{E}$ have the same spatial resolution.

3.2. Event-intensity relation

We show that the event and intensity/frame sensing are bridged via temporal gradients. On the intensity side, we employ the optical flow assumption for deriving the temporal gradient of the latent field $\mathcal{L}$. Assume that in a small vicinity, there exists a small flow vector $\delta\mathbf{u} = [\delta x, \delta y, \delta t]^\top$ under which the intensity is assumed to be constant. Mathematically, this assumption can be expressed as:

$$\mathcal{L}(x + \delta x, y + \delta y, t_{\text{ref}} + \delta t) = \mathcal{L}(x, y, t_{\text{ref}}). \tag{1}$$

The Taylor series expansion of the left side of Eq. (1) gives:

$$\mathcal{L}_{t_{\text{ref}}+\delta t} = \mathcal{L}_{t_{\text{ref}}} + \nabla_{\text{xyt}}\mathcal{L}_{t_{\text{ref}}}\delta\mathbf{u} + o(|\delta x| + |\delta y| + |\delta t|), \tag{2}$$

where $\nabla_{\text{xyt}}\mathcal{L}_{t_{\text{ref}}} = [\frac{\partial\mathcal{L}}{\partial x}, \frac{\partial\mathcal{L}}{\partial y}, \frac{\partial\mathcal{L}}{\partial t}]\big|_{t_{\text{ref}}}$ denotes the gradient operator evaluated at time t_{ref}. If we substitute only the zero- and first-order terms to approximate $\mathcal{L}_{t_{\text{ref}}+\delta t}$ and re-arrange Eq. (1), we can obtain the following relation:

$$\frac{\partial\mathcal{L}}{\partial t}\bigg|_{t_{\text{ref}}} \simeq -\nabla_{\text{xy}}\mathcal{L}_{t_{\text{ref}}}\mathbf{v} \doteq Q^l, \tag{3}$$

where $\nabla_{\text{xy}}\mathcal{L}_{t_{\text{ref}}} = [\frac{\partial\mathcal{L}_{t_{\text{ref}}}}{\partial x}, \frac{\partial\mathcal{L}_{t_{\text{ref}}}}{\partial y}]$ denotes the spatial gradient of $\mathcal{L}_{t_{\text{ref}}}$, and $\mathbf{v} = [\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}]^\top$ is the velocity vector. For future reference, we define the temporal gradient derived from intensity image as Q^l.

On the event side, the flow velocity $\mathbf{v}$ shall result in position shifts for local events. This is based on the assumption that local events are triggered by the same edge, as shown in Fig. 2(a). Therefore, the temporal gradient can be approximated by the tangent of a set of warped events in a local window:

$$\frac{\partial\mathcal{L}}{\partial t}\bigg|_{t_{\text{ref}}} \approx \frac{\sum_{(t_k - t_{\text{ref}})\in(0,\delta t)} \epsilon_k \hat{\delta}(\mathbf{x} - \mathbf{x}'_k)}{\delta t} \doteq Q^e, \tag{4}$$

where $\epsilon_k = \epsilon_p$, if $p_k = 1$; and $\epsilon_k = \epsilon_n$, if $p_k = -1$. $\hat{\delta}(\cdot)$ is the Dirac delta function. $\mathbf{x}'_k$ is the event location by warping (back propagating) measured events to time t_{ref} according to the flow velocity $\mathbf{v}$, *i.e.*, $\mathbf{x}'_k = \mathbf{x}_k - (t_k - t_{\text{ref}})\mathbf{v}$, where $\mathbf{x} = [x, y]^\top$, $\mathbf{x}_k = [x_k, y_k]^\top$ and $\mathbf{x}'_k = [x'_k, y'_k]^\top$. In the rest of the paper, we define the temporal gradient derived from events as Q^e.

From Eq. (4) and Eq. (3) we obtain,

$$Q^e \simeq Q^l. \tag{5}$$

The above equation establishes the relation between events and image spatial gradients. There are two unknowns, ϵ_k and $\mathbf{v}$ in the relation, where $\epsilon_k \in \{\epsilon_p, \epsilon_n\}$ can be obtained from the event camera configuration. Numerically, ϵ_k can be viewed as a constant scaling value to match Q^e with Q^l. The key unknown is the flow velocity $\mathbf{v}$.

Events generated by illumination variation are not considered here.

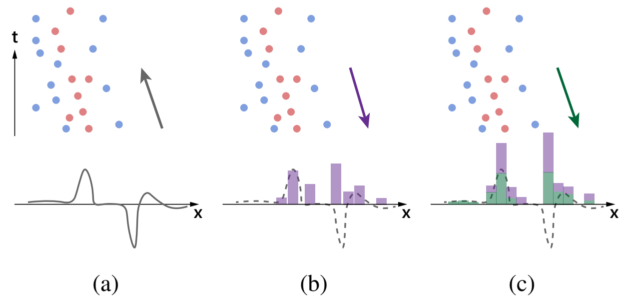


Figure 2: (a) A latent edge signal (gray curve) triggers a set of (noisy) events due to motion. (b) In contrast maximization (CM) [9], the events are warped back at t_{ref} to form a histogram (purple). (c) In our joint contrast maximization (JCM), an image is formed jointly by the events (purple) and the edge of the intensity image (green).

3.3. Joint contrast maximization

Previous work [9] proposed contrast maximization (CM) to optimize the flow parameter based on the contrast of the image (or histogram) formed only by the warped events, as shown in Fig. 2(b). However, CM is designed for event data alone. In the presence of an intensity image, we extend the framework of CM and propose joint contrast maximization (JCM) to estimate the flow vector based on intensity image and events. Particularly, we propose to maximize the contrast of an image/histogram jointly formed by the absolute edge of the intensity image and the warped events, as shown in Fig. 2(c). Mathematically, the image of warped events and intensity edge is expressed as:

$$J(\mathbf{x}; \mathbf{v}) = \sum_{k=1}^{N_e} \hat{\delta}(\mathbf{x} - \mathbf{x}'_k(\mathbf{v})) + \alpha S(\mathbf{x}), \tag{6}$$

where $S(\mathbf{x})$ is the edge image and can be defined as $S(\mathbf{x}) = \sqrt{|g_x I(\mathbf{x})|^2 + |g_y I(\mathbf{x})|^2}$. We use the Sobel edge (without thresholding) as a discrete approximation. The x-axis kernel can be defined as $g_x = [-1, 0, 1; -2, 0, 2; -1, 0, 1]$, $g_y = g_x^\top$, and $\alpha = \frac{N_e}{\sum_{i,j} S(i,j)}$ is a normalization coefficient to balance the energy of the two data.

The objective for estimating the flow velocity is:

$$\hat{\mathbf{v}} = \underset{\mathbf{v}}{\text{argmax}} \frac{1}{N_p} \sum_{ij} (J_{ij} - \bar{J})^2, \tag{7}$$

where N_p indicates the number of pixels in image patch J, while $\bar{J}$ denotes the mean value of J. Note that when no intensity image is available or it has low quality (*e.g.*, blurry), the Sobel term can be set to zero and the formulation degenerates to event-only contrast maximization [9]. With non-zero S, the maximal contrast corresponds to the flow velocity that transports events to the image edge. Non-optimal velocity will lead to a deterioration of the contrast.

Here, we perform a numerical comparison between CM and JCM, shown in Fig. 3. We follow the analysis in [22]

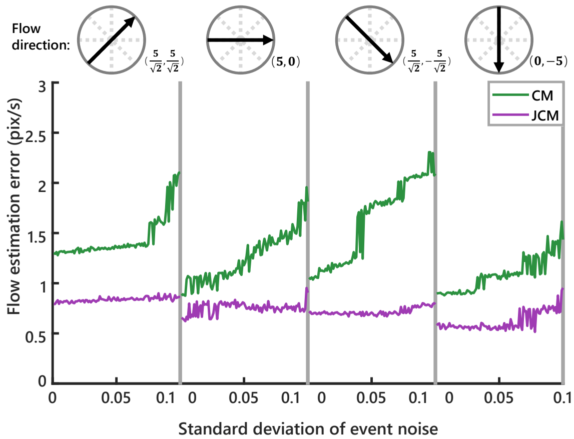


Figure 3: Comparison between CM and JCM [9] for flow estimation w.r.t. event noise.

and [28] for event simulation from images. *I.e.*, a thresholding operation ($\epsilon_p = 0.2$, $\epsilon_n = -0.2$) is applied on the difference image between the flow-shifted image and the original/last image. The event noise follows a Gaussian distribution around the per-pixel threhold values [22]. We consider a standard deviation range of $\sigma_e \in (0, 0.1)$, and compare the accuracy for flow estimation w.r.t. different flow directions with fixed flow radius of 5 pixels. We use the Euclidean distance to quantify the flow estimation error. The error is averaged over 18 images of size 30×30. Details of this experiment as well as visual examples can be found in the supplementary material. As shown in Fig. 3, both JCM and CM error increases as noise level increases. However, JCM maintains low error across all spectrum of the noise level, revealing a more noise-robust property than CM.

3.4. Joint filtering

The goal of joint/guided filtering is to construct an optimized output inheriting mutual structures from Q^e and Q^l. In guided image filtering, an output image patch Q^o is defined as an affine transformation of the guidance image patch Q^l:

$$Q^o = g_a Q^l + g_b. \tag{8}$$

By the above formulation, Q^o inherits the spatial structure of Q^l, *i.e.*, $\nabla Q^o = g_a \nabla Q^l$ in each local patch. The objective is generally defined as a data term and a regularization term:

$$\underset{g_a, g_b}{\operatorname{argmin}} ||Q^o - Q^e||_2^2 + \lambda \Phi, \tag{9}$$

where Φ is the regularization functional and λ the regularization parameter. In particular, we consider three popular as well as emerging filters, namely,

- Guided Image Filtering (GIF) [16]: In this case, $\Phi = g_a^2$. This regularization term is to prevent coefficient g_a from being too large.

Algorithm 1 Guided Event Filtering (GEF)

Input: Intensity image I, events $\mathcal{E}$.
Output: Filtered temporal gradient Q^o.
1: Estimate the flow field $\mathbf{v}$ using JCM in Eq. (7);
2: Compute Q^l in Eq. (3) and Q^e in Eq. (4);
3: Perform guided filtering according to Eq. (9).

- Side Window Guided Filtering (SW-GF) [58]: In this case, the regularization term is the same as the GIF, but the regression is computed on 8 (upper-half, lower-half, left-half, right-half, northwest, northeast, southwest, southeast) side windows instead of a single window centered around the target pixel. Compared to GIF, this filter has the property of better preserving the edges of the filter input image.
- Mutual-Structure for Joint Filtering (MS-JF) [44]: This filter emphasizes the mutual structure between the input and guidance images, and performs filtering in a bidirectional manner. The mutual structure is sought after by minimizing a similarity measure term, *i.e.*, $E_s = ||g_a Q^l + g_b - Q^e||_2^2 + ||g'_a Q^e + g'_b - Q^l||_2^2$, where g'_a and g'_b denotes the counterpart coefficients for using Q^e to represent Q^l. Additionally, the regularization term consists of the smoothness term, *i.e.*, $E_r = \lambda_1 g_a^2 + \lambda_2 g_a'^2$, as well as the deviation term which avoids filtered output deviating too far from the original images, *i.e.*, $E_d = \lambda_3 ||g_a Q^l + g_b - Q^l||_2^2 + \lambda_4 ||g'_a Q^e + g'_b - Q^e||_2^2$. The objective is to minimize the summed loss terms, *i.e.*, $E = E_s + E_r + E_d$, over g_a, g_b, g'_a, g'_b.

3.5. Implementation details

The steps of GEF is summarized in Algorithm 1.

In the JCM step, we use a local window with radius r_w to estimate pixel-wise flow. Areas with events fewer than 1 are skipped. r_w may vary due to the structure of the scene. A large r_w can be used when the scene has sparse and isolated objects, in exchange for more time to compute the flow field. The intensity image support is slightly larger (about several pixels on four sides) than the event window to prevent fallout of events due to large velocity.

Both the computation of flow velocity and Q^l use the spatial gradient. Therefore, the spatial gradient image can be computed once. Q^l is normalized to match the range of Q^e before the filtering step. This normalization step also functions as an estimation for the event threshold (ϵ_k). The pixel values of the output image Q^o are rounded to integers, which can be interpreted as the event counts.

In the filtering step, we set the window width to be 1 for all three filters. The filtering is switched between intensity-event joint guiding and event self-guiding. When a windowed image patch has low spatial contrast, and therefore large α values, we set $\alpha = 0$ in Eq. (6) and $Q^l = Q^e$. We run 20 iterations for MS-JF. For GIF and SW-GF, λ is set to 1×10^{-3}. For MS-JF, the same values are assigned for the parameter pairs, *i.e.*, λ_1 and λ_2 ($\sim 1 \times 10^{-2}$), as well as λ_3 and λ_4 (~ 3). This is to encourage equal weights between the input and guidance. Filtering is performed when Q^e and Q^l are at the same resolution and are both grayscale. Details for filtering color events are included in the supplementary material. The filtered output does not preserve the ternary representation as the original events. Our image-based event representation is better suited for downstream algorithms that process events in image-based fashion [55]. It is possible to warp the events back in the space-time volume to restore the ternary representation. One possible restoration approach is to evenly distribute events along the computed flow direction.

Similar to CM [9], the computational complexity of JCM is linear on the number of events to be warped. The additional computation of JCM contrast is typically negligible compared to CM. Both GIF and SW-GF have linear computation time w.r.t. patch pixel size. MS-JF is iteration-dependent.

4. Experiments

4.1. Numerical evaluation

Guided denoising. In this experiment, we compare GEF (considering all three filters) with two state-of-the-art event-based denoising approaches, *i.e.*, Liu *et al.* [24] and EV-gait [53]. To quantify the denoising performance, we use zero-noise event frame as the ground truth. The denoised images are compared against the ground truth images using the root mean squared error (RMSE) criterion. The smaller the RMSE values, the better denoising the performance. At each noise level, the RMSE values are averaged over 18 images. The results are plotted in Fig. 4. As can be seen, all three GEF methods have better denoising performance compared to non-guidance-based methods. Among the three guided filters, MS-JF [44] has the lowest RMSE values than the other two filters across the whole range. Therefore, we choose MS-JF as the filtering algorithm within GEF. We only show MS-JF results in the following experiments. Additional results using GIF and SW-GF are shown in the supplementary material.

Qualitatively, we compare the denoising performance on the captured real-world scenarios dataset (which will be introduced in Sec. 4.2). The results are shown in Fig. 5. Compared to existing approaches, GEF (MS-JF) is able to enhance the edge features as well as removing event noise.

Guided super resolution. Because it is challenging to obtain ground truth image and events at multiple scales, we perform quantitative evaluation for upsampling in simulation. We use 18 high resolution (HR) images to simulate the ground truth HR events. To simulate the low resolution (LR) events, the HR images are first downsized and used to generate zero-noise events using the same procedure described in Sec. 3.3. We consider three downsizing scales up to 8×. For future reference, we use 2×, 4×, and 8× to denote the upsampling factors. For 2× upsampling, we first bicubically upsample the low-resolution Q^e for 2×, and then perform same-resolution joint filtering with 2× Q^l (downsized from HR). The 2× upsampling procedure is iteratively applied for higher scales.

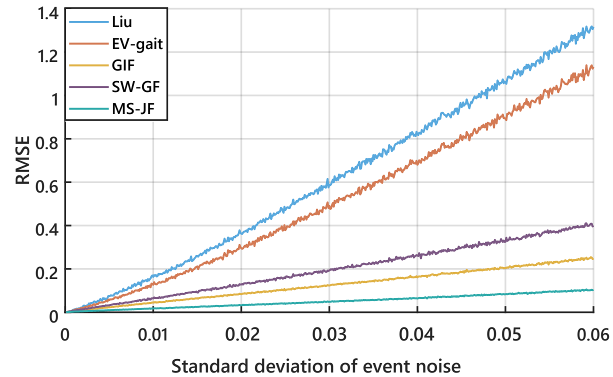


Figure 4: Comparison of event denoising performance. Intensity-guided filters (GIF [16], SW-GF [58] and MS-JF [44] unanimously outperform non-guidance-based methods (Liu *et al.* [24] and EV-gait [53]).

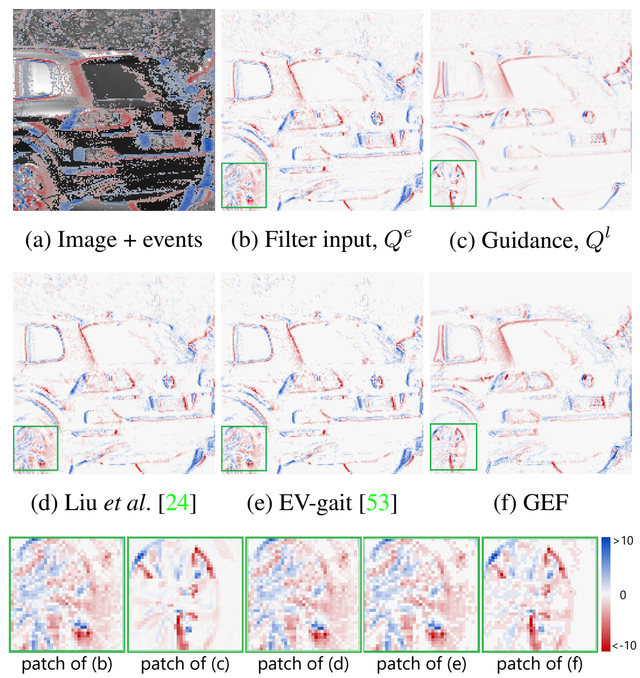
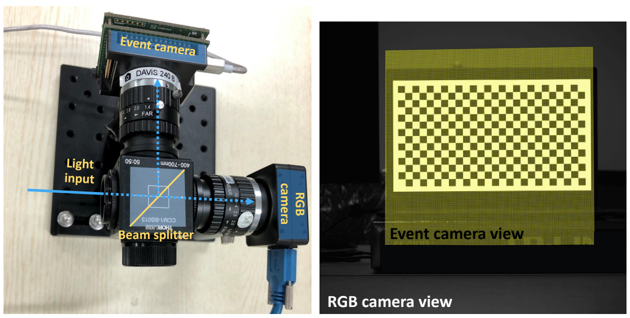


Figure 5: Comparison of denoising performance on our RGB-DAVIS dataset. (a) An image overlaid with events; (b) Q^l as filter guidance; (c) warped events, Q^e, as filter input; (d-f) denoising results using (d) Liu *et al.* [24], (e) EV-gait [53] and (f) our GEF (MS-JF). Additional results are presented in the supplementary material.

Table 1: PSNR comparison for super resolution

methods		2×	4×	8×
(1) no guidance SR	Bicubic	40.110	39.133	39.368
	EDSR [23]	39.976	39.363	39.319
	SRFBN [21]	40.572	39.937	40.152
	EDSR-ev	40.315	40.577	39.961
	SRFBN-ev	40.837	40.309	40.110
(2) guided SR, w/ SR image	Bicubic	42.591	42.612	44.144
	EDSR [23]	42.599	42.655	44.174
	SRFBN [21]	42.603	43.037	44.170
(3) GEF		**42.755**	**43.319**	**44.218**



(a) Experimental setup (b) Calibrated views

Figure 6: Our RGB-DAVIS imaging system.

We compare three super resolution (SR) schemes: (1) no guidance SR. The scheme refers to direct SR without guidance. Such methods include the baseline bicubic upsampling, and two state-of-the-art single image SR methods: EDSR [23] and SRFBN [21]. We apply both pre-trained models as well as re-trained ones. Re-trained models are denoted as EDSR-ev and SRFBN-ev, respectively. (2) guided SR, w/ SR image. In this case, the joint filtering is applied between the computed SR image and the event image. (3) GEF. GEF here is referred as joint filtering between the pristine HR image and the event image. The results are summarized in Table 1. We use Peak Signal to Noise Ratio (PSNR) as performance measurement. As can be seen, (2) and (3) both have higher PSNR than (1), which suggests the effectiveness of using image as guidance. In (1), re-training SR networks slightly improves the performance, but still underperforms (2) and (3). Another interesting effect in (2) and (3) is that PSNR values increase as scale factor increases. This is because the event image at high resolution has sparse non-zero signals representing thin edge. Examples and additional analysis are included in the supplementary material.

4.2. RGB-DAVIS camera system

To test GEF for real-world scenarios, we build a hybrid camera consisting of a high-resolution machine vision camera and a low-resolution event camera, *i.e.*, DAVIS. We refer to our camera prototype as RGB-DAVIS camera.

Setup and calibration. As shown in Fig. 6(a), we collocate an event camera (DAVIS240b, resolution of 180×190 pixels, with F/1.4 lens) and a machine vision camera (Point Grey Chameleon3, resolution of 2448×2048 pixels, 50 FPS, with F/1.4 lens). A beam splitter (Thorlabs CCM1-BS013) is mounted in front of the two cameras with 50% splitting. We use a 13.9" 60Hz monitor for offline geometric calibration for two signals. For geometric calibration, we consider homography and radial distortion between two camera views. In order to extract keypoints from event data, we display a blinking checkerboard pattern on the monitor and integrate the captured events over a time window to form a checkerboard image, as shown in Fig. 6(b). For temporal synchronization, we write a synchronization script to trigger the two cameras simultaneously. Details about the calibration procedure can be found in the supplementary material.

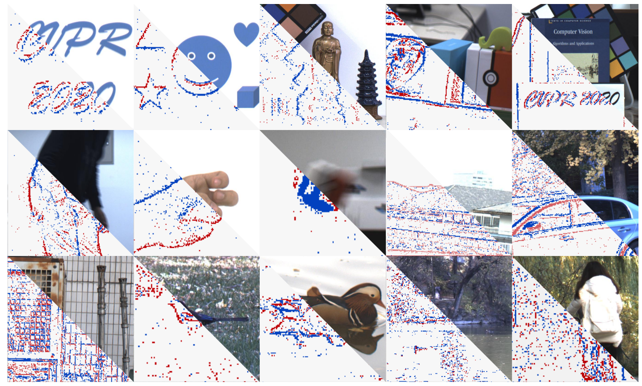


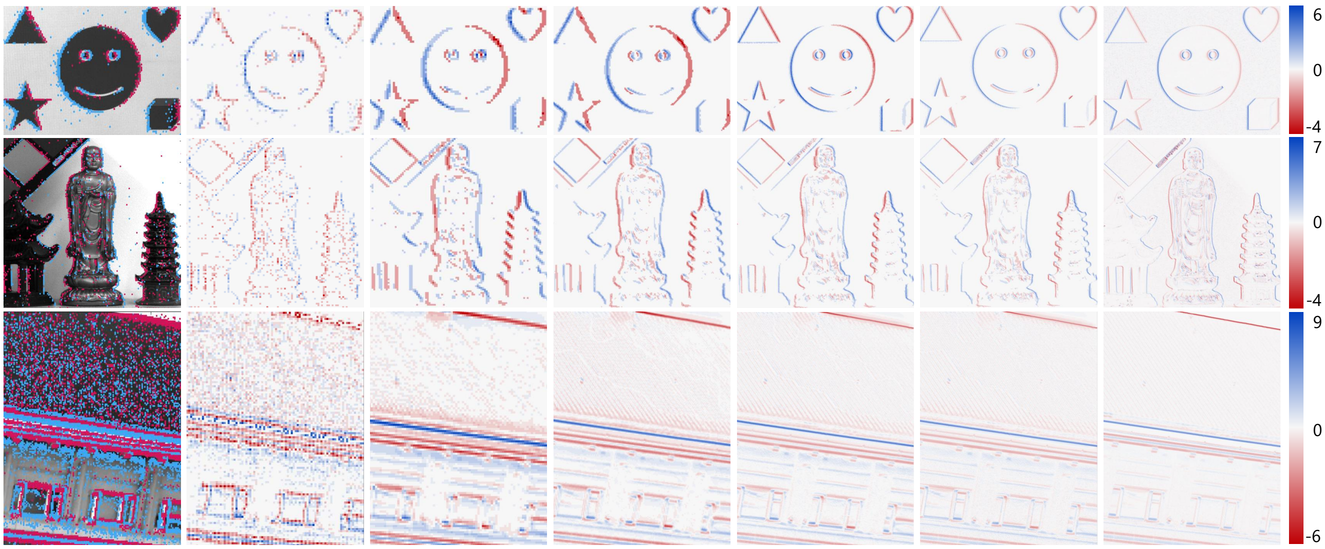
Figure 7: Examples of our proposed RGB-DAVIS dataset. In each square, lower-left is the converted event frame, and upper-right is the RGB image. Please find images of our complete dataset in the supplementary material.

Dataset collection. We use RGB-DAVIS to collect various sequences of event-RGB video clips. Examples are shown in Fig. 7. Both indoor and outdoor scenarios are captured. The scenes widely range from simple shapes to complex structures. All the clips involve camera motion and/or scene motion.

Results. After calibration, we perform guided filtering with three upsampling scales, *i.e.*, $2\times$, $4\times$, $8\times$. The flow is estimated at $1\times$. We show three upsampling examples corresponding to monitor, indoor and outdoor scenarios of our captured dataset in Fig. 8. The captured images as well as calibrated events are shown in Fig. 8(a), with the filtered output shown in Fig. 8 (c-f). As can be seen, the events are gradually and effectively upsampled and denoised. Please see additional results for scene motion as well as filtering results using other filters in the supplementary material.

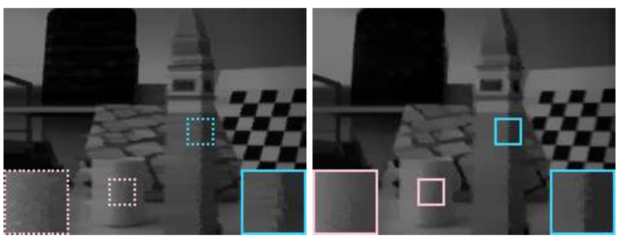
5. Applications

GEF has a variety of applications for event-based tasks. Here, we enumerate several example applications.



(a) Image + events (b) Q^e (c) $1\times$ (d) $2\times$ (e) $4\times$ (f) $8\times$ (g) Q^l

Figure 8: Guided upsampling results on our RGB-DAVIS data.



(a) w/o GEF (b) w/ GEF

Figure 9: Frame prediction using the DMR method in [55].

5.1. High frame-rate video frame synthesis

The task is to reconstruct high frame-rate video frames using a hybrid input of image(s) and events [30, 55].

Future frame prediction. In this case, we perform future frame prediction, *i.e.*, given a start intensity frame and the subsequent events to predict the future frame. We implement the differentiable model-based reconstruction (DMR) method in [55]. Without GEF, the reconstruction performance for the case of "slider_depth" is 25.10 (PSNR) and 0.8237 (SSIM). With GEF, the reconstruction performance improves to 26.63 (PSNR) and 0.8614 (SSIM). For a qualitative comparison, the #5 frame out of 12 reconstructed frames are shown in Fig. 9. The complete results can be found in the supplementary material.

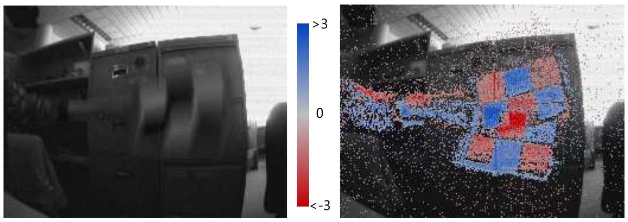
Motion deblur. GEF can be applied to improve event-based motion deblur [30]. Given a blurry image (Fig. 10(a)) and the events captured during the exposure time (Fig. 10(b)), Pan *et al.* [30] proposed an event-based double integral (EDI) approach to recover the underlying sharp image(s), as shown in Fig. 10(c). We employ the same formulation, but use our GEF to first filter the events. Note that in this case, the blurry image does not provide useful edge information, we therefore warp neighbor events to form the guidance images. The result is shown in Fig. 10(e). Even without the guidance of an intensity image, GEF can still reduce the event noise using neighbor events. We further compare the EDI result with denoised EDI output using bilateral filtering, as shown in Fig. 10(g). Compared to the post-denoising scheme, GEF (Fig. 10(f)) is more effective in eliminating the event noise.

5.2. HDR image reconstruction

GEF is able to improve HDR image reconstruction because of its effectiveness for motion compensation and denoising. As shown in Fig. 11(a) and (c), the intensity image contains over-exposed regions while the warped event image preserves structures in those regions. We follow a previous approach which employs Poisson reconstruction for HDR reconstruction [3]. The difference in our case is that the intensity image is used for reconstruction. In such case, GEF is applied by setting the warped event image Q^e as guidance and Q^l as filter input. The restored gradient field $\nabla_{xy} I'$ along with the estimated flow $\mathbf{v}$ and the intensity image are then used to reconstruct an HDR image. As can be seen in Fig. 11(c) and (d), the reconstructed HDR image w/ GEF has higher contrast and less artifacts than w/o GEF.

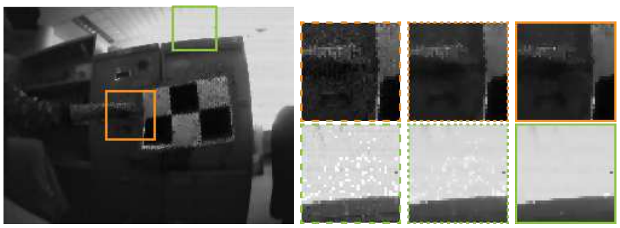
5.3. Corner detection and tracking

GEF can be applied on event-based feature/corner detection and tracking. To demonstrate the benefit of guided upsampling, we use RGB-DAVIS camera to capture a periodic circularly moving checkerboard pattern. We employ the event-based Harris corner detector (evHarris) [50] as the backbone corner detector. A slight difference between our implementation and the original evHarris is that we use the warped event image (motion compensated), instead of di-



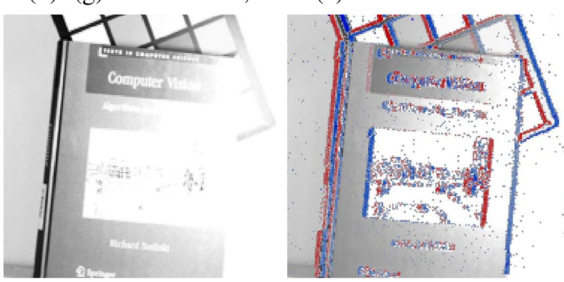
(a) Blurry image (b) w/ events in exposure

(c) EDI w/o GEF (d) Bilateral denoising of (c)

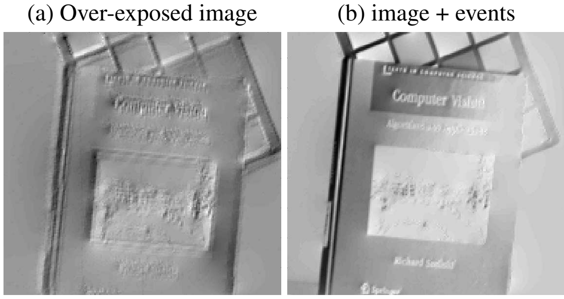


(e) EDI w/ GEF (f) (g) (h)

Figure 10: Motion deblur using EDI [30]. (f) EDI w/o GEF, from (c). (g) EDI result (w/o GEF) + bilateral denoising, from (d). (g) EDI w/ GEF, from (e).

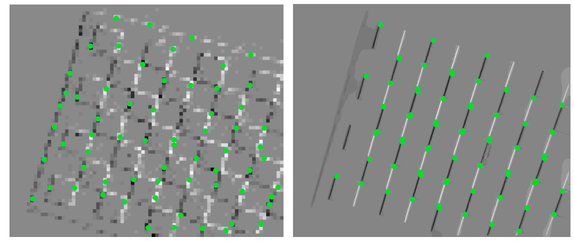


(a) Over-exposed image (b) image + events

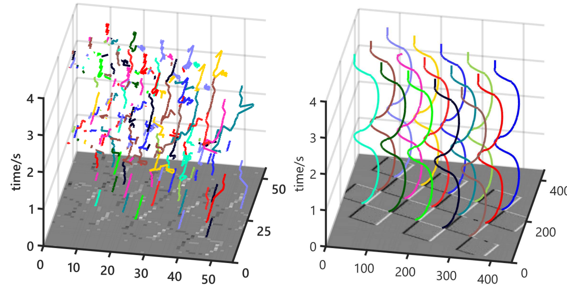


(c) w/o GEF (d) w/ GEF

Figure 11: HDR image reconstruction based on Poission method in [3]. (a) Low dynamic range image. (b) Overlaid with events. (c) Reconstructed HDR image w/o GEF. (f) Reconstructed HDR image w/ GEF.



(a) w/o GEF, 1× (b) w/ GEF, 8×



(c) w/o GEF, 1× (d) w/ GEF, 8×

Figure 12: Corner detection using evHarris [50].

rectly accumulating events in local windows. As shown in Fig. 12(a) and (b), with GEF (8× guided upsampling), the checkerboard corners are detected more accurately than w/o GEF. We also compare the corner tracks computed both *w/o* and *w/* GEF process. The results are shown in Fig. 12(c) and Fig. 12(d). As can be seen, the corner points that are upsampled by the GEF can be tracked more accurately than the original frames.

6. Concluding remarks

There are several interesting takeaways from our experimental study. First, our results showed that with the assistance of intensity images, performance improvement has been achieved for flow estimation, event denoising and event super resolution (SR). Second, for event SR, our results indicated that directly applying state-of-the-art CNN-based SR algorithms, w/ or w/o re-training, performs worse than first applying the same SR algorithms on intensity images and then performing joint filtering. Third, we have evaluated three joint filtering approaches with different properties. Our results concluded that finding the mutual structure (MS-JF) is better suited than the other two filters. Fourth, we have demonstrated the benefit of event denoising and SR by testing on a variety of downstream tasks.

Acknowledgment

This work is in part supported by National Natural Science Foundation of China under Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), Beijing Academy of Artificial Intelligence (BAAI), DARPA Contract No. HR0011-17-2-0044, and NSF CAREER IIS-1453192.

References

[1] Mohammed Mutlaq Almatrafi and Keigo Hirakawa. Davis camera optical flow. *IEEE Transactions on Computational Imaging*, 2019. 1

[2] Richard G Baraniuk, Thomas Goldstein, Aswin C Sankaranarayanan, Christoph Studer, Ashok Veeraraghavan, and Michael B Wakin. Compressive video sensing: algorithms, architectures, and applications. *Signal Processing Magazine*, 34(1):52–66, 2017. 2

[3] Souptik Barua, Yoshitaka Miyatani, and Ashok Veeraraghavan. Direct face detection and video reconstruction from event cameras. In *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 7, 8

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

[5] Pravin Bhat, C Lawrence Zitnick, Noah Snavely, Aseem Agarwala, Maneesh Agrawala, Michael Cohen, Brian Curless, and Sing Bing Kang. Using photographs to enhance videos of a static scene. In *Proc. of the 18th Eurographics conference on Rendering Techniques*, pages 327–338, 2007. 2

[6] Wensheng Cheng, Hao Luo, Wen Yang, Lei Yu, Shoushun Chen, and Wei Li. DET: A high-resolution DVS dataset for lane extraction. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, 2019. 2

[7] Daniel Czech and Garrick Orchard. Evaluating noise filtering for event-based asynchronous change detection image sensors. In *6th International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pages 19–24. IEEE, 2016. 2

[8] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew Davison, Joerg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *arXiv preprint arXiv:1904.08405*, 2019. 1

[9] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2018. 2, 3, 4, 5

[10] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Asynchronous, photometric feature tracking using events and frames. In *Proc. of European Conference on Computer Vision (ECCV)*, 2018. 2

[11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, pages 1–18, 2019. 1

[12] Xiaojie Guo, Yu Li, Jiayi Ma, and Haibin Ling. Mutually guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[13] Ankit Gupta, Pravin Bhat, Mira Dontcheva, Oliver Deussen, Brian Curless, and Michael Cohen. Enhancing and experiencing spacetime resolution with videos and stills. In *Proc. of International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2009. 2

[14] Mohit Gupta, Amit Agrawal, Ashok Veeraraghavan, and Srinivasa G Narasimhan. Flexible voxels for motion-aware videography. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 100–114. Springer, 2010. 2

[15] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[16] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2012. 2, 4, 5

[17] Michael Iliadis, Leonidas Spinoulas, and Aggelos K Katsaggelos. Deep fully-connected networks for video compressive sensing. *Digital Signal Processing*, 72:9–18, 2018. 2

[18] Alireza Khodamoradi and Ryan Kastner. O(N)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors. *IEEE Transactions on Emerging Topics in Computing*, 2018. 2

[19] Hanme Kim, Stefan Leutenegger, and Andrew J. Davison. Real-time 3D reconstruction and 6-DoF tracking with an event camera. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 2016. 1

[20] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 154–169. Springer, 2016. 2

[21] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2019. 6

[22] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor. *IEEE journal of solid-state circuits*, 43(2):566–576, 2008. 1, 3, 4

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 136–144, 2017. 6

[24] Hongjie Liu, Christian Brandli, Chenghan Li, Shih-Chii Liu, and Tobi Delbruck. Design of a spatiotemporal correlation filter for event-based sensors. In *Proc. of International Symposium on Circuits and Systems (ISCAS)*, pages 722–725, 2015. 2, 5

[25] Patrick Llull, Xuejun Liao, Xin Yuan, Jianbo Yang, David Kittle, Lawrence Carin, Guillermo Sapiro, and David J Brady. Coded aperture compressive temporal imaging. *Optics Express*, 21(9):10526–10545, 2013. 2

[26] 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 *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[27] Elias Mueggler, Basil Huber, and Davide Scaramuzza. Event-based, 6-DoF pose tracking for high-speed maneuvers. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2761–2768, 2014. 1

[28] 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. 4

[29] Vandana Padala, Arindam Basu, and Garrick Orchard. A noise filtering algorithm for event-based asynchronous change detection image sensors on truenorth and its implementation on truenorth. *Frontiers in Neuroscience*, 12:118, 2018. 2

[30] 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 *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 7, 8

[31] Federico Paredes-Vallés, Kirk Yannick Willehm Scheper, and Guido Cornelis Henricus Eugene 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*, 2019. 1

[32] Jaesik Park, Hyeongwoo Kim, Yu-Wing Tai, Michael S Brown, and Inso Kweon. High quality depth map upsampling for 3D-ToF cameras. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1623–1630, 2011. 2

[33] Jiahui Qu, Yunsong Li, and Wenqian Dong. Hyperspectral pansharpening with guided filter. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2152–2156, 2017. 2

[34] Bharath Ramesh, Andrés Ussa, Luca Della Vedova, Hong Yang, and Garrick Orchard. Low-power dynamic object detection and classification with freely moving event cameras. *Frontiers in Neuroscience*, 14:135, 2020. 1

[35] Bharath Ramesh and Hong Yang. Boosted kernelized correlation filters for event-based face detection. In *Proc. of Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 155–159, 2020. 1

[36] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. EMVS: Event-based multi-view stereo—3D reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018. 1

[37] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3857–3866, 2019. 1

[38] 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*, 2019. 1

[39] Dikpal Reddy, Ashok Veeraraghavan, and Rama Chellappa. P2C2: Programmable pixel compressive camera for high speed imaging. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 329–336, 2011. 2

[40] Bodo Rueckauer and Tobi Delbruck. Evaluation of event-based algorithms for optical flow with ground-truth from inertial measurement sensor. *Frontiers in Neuroscience*, 10:176, 2016. 1

[41] Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, and Davide Scaramuzza. Fast image reconstruction with an event camera. In *Proc. of Winter Conference on Applications of Computer Vision (WACV)*, pages 156–163, 2020. 1

[42] Falk Schubert and Krystian Mikolajczyk. Combining high-resolution images with low-quality videos. In *Proc. of British Machine Vision Conference (BMVC)*, pages 1–10, 2008. 2

[43] Prasan Shedligeri and Kaushik Mitra. Photorealistic image reconstruction from hybrid intensity and event-based sensor. *Journal of Electronic Imaging*, 28(6):063012, 2019. 1

[44] Xiaoyong Shen, Chao Zhou, Li Xu, and Jiaya Jia. Mutual-structure for joint filtering. In *Proc. of International Conference on Computer Vision (ICCV)*, 2015. 4, 5

[45] Bongki Son, Yunjae Suh, Sungho Kim, Heejae Jung, Jun-Seok Kim, Changwoo Shin, Keunju Park, Kyoobin Lee, Jinman Park, Jooyeon Woo, et al. A 640×480 dynamic vision sensor with a 9μm pixel and 300Meps address-event representation. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 66–67, 2017. 1

[46] Pingfan Song, Xin Deng, João FC Mota, Nikos Deligiannis, Pier-Luigi Dragotti, and Miguel Rodrigues. Multimodal image super-resolution via joint sparse representations induced by coupled dictionaries. *IEEE Transactions on Computational Imaging*, 2019. 2

[47] Vladimir Stanković, Lina Stanković, and Samuel Cheng. Compressive video sampling. In *16th European Signal Processing Conference*, pages 1–5. IEEE, 2008. 2

[48] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[49] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1527–1537, 2019. 1

[50] Valentina Vasco, Arren Glover, and Chiara Bartolozzi. Fast event-based harris corner detection exploiting the advantages of event-driven cameras. In *Proc. of International Conference on Intelligent Robots and Systems (IROS)*, pages 4144–4149, 2016. 7, 8

[51] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018. 1

[52] Lin Wang, S. Mohammad Mostafavi I. , Yo-Sung Ho, and Kuk-Jin Yoon. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[53] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. EV-Gait: Event-based robust gait recognition using dynamic vision sensors. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[54] Zihao Wang, Leonidas Spinoulas, Kuan He, Lei Tian, Oliver Cossairt, Aggelos K Katsaggelos, and Huaijin Chen. Compressive holographic video. *Optics Express*, 25(1):250–262,

2017. 2

[55] Zihao Winston Wang, Weixin Jiang, Kuan He, Boxin Shi, Aggelos Katsaggelos, and Oliver Cossairt. Event-driven video frame synthesis. In *Proc. of International Conference on Computer Vision (ICCV) Workshops*, 2019. 1, 5, 7

[56] Jie Xu, Meng Jiang, Lei Yu, Wen Yang, and Wenwei Wang. Robust motion compensation for event cameras with smooth constraint. *IEEE Transactions on Computational Imaging*, 6:604–614, 2020. 2

[57] Qiong Yan, Xiaoyong Shen, Li Xu, Shaojie Zhuo, Xiaopeng Zhang, Liang Shen, and Jiaya Jia. Cross-field joint image restoration via scale map. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1537–1544, 2013. 2

[58] Hui Yin, Yuanhao Gong, and Guoping Qiu. Side window guided filtering. *Signal Processing*, 165:315 – 330, 2019. 4, 5

[59] Hagit Zabrodsky and Shmuel Peleg. Attentive transmission. *Journal of Visual Communication and Image Representation*, 1(2):189–198, 1990. 2