CONVOLUTIONAL ISTA NETWORK WITH TEMPORAL CONSISTENCY CONSTRAINTS FOR VIDEO RECONSTRUCTION FROM EVENT CAMERAS

Siying Liu, Roxana Alexandru, Pier Luigi Dragotti

Department of Electrical and Electronic Engineering, Imperial College London, London, UK
{siying.liu20, roxana.alexandru12, p.dragotti}@imperial.ac.uk

ABSTRACT

Event cameras produce streams of events with high temporal resolution which do not suffer from motion blur. Current deep networks achieve high-quality video reconstruction from events, but most of them are large and difficult to interpret. In this work, we present a solution to this problem by systematically designing a deep network based on sparse representation. First, we investigate the relationship between events and intensity images. The reconstruction problem is then modelled as a sparse coding problem, which can be solved by the iterative shrinkage thresholding algorithm (ISTA). Second, we expand this into a convolutional ISTA network (CISTA) using algorithm unfolding. Finally, we introduce recurrent units and temporal similarity constraints to enhance the temporal consistency (TC) reconstruction of long videos. Results show that our CISTA-TC network achieves high-quality reconstruction compared with state-of-the-art methods, whilst leading to low memory consumption.

Index Terms— Event cameras, video reconstruction, sparse representation, deep learning

1. INTRODUCTION

Event cameras, also called neuromorphic cameras, are bio-inspired imaging systems whose pixels capture changes in brightness asynchronously [1]. In contrast to conventional cameras which capture videos at a fixed frame rate, event cameras generate events which encode the location, timestamp, and polarity of the brightness changes. They have many advantages, including high frame rate, low latency and high dynamic range. They thus have the potential to be used in many real-world applications, such as autonomous driving which relies on fast camera movements. However, events are asynchronous spikes and therefore, cannot be directly processed by main-stream computer vision algorithms. At the same time, spatio-temporal information of events can provide more details to reconstruct clear and sharp images. Thus a crucial task is to be able to reconstruct intensity images from event sequences.

Current methods which address this problem can be roughly categorised into model-based and learning-based methods. When casting intensity reconstruction problem as a model-based problem, some regularisation restrictions [2] are usually required, such as optical flow [3] or intensity information [4, 5, 6]. These approaches take advantage of the relationship between intensity frames and events, but hand-crafted priors are not always effective in challenging situations such as very fast motion or brightness changes.

In recent years, learning-based methods have achieved relatively high quality reconstruction. Barua et al. [7] introduced a dictionary learning approach, while in the context of image deblurring and super resolution using events, Wang et al. [8] proposed a network based on sparse learning, which is obtained by unfolding the iterations of the ISTA. It is similar to our idea, but corresponding intensity frames are required as input. Generative adversarial networks [9, 10] have also been used in this context but reconstructed images usually contain artifacts in regions where no events are generated. Some other networks rely on recurrent connections to enhance the performance. For example, E2VID [11, 12] achieves reconstruction of long videos from events only, using a recurrent network, however, it requires a significant memory consumption. Sheerlinck et al. [13] then turned E2VID into a smaller and faster network called FireNet without impairing the performance. SPADE-E2VID [14] feeds previous reconstructed frames into spatially-adaptive denormalization (SPADE) layers to achieve temporal coherence. However, these networks typically lack interpretability and have relatively high computational cost.

In this work, we aim to achieve video reconstruction from events by systematically designing a deep network based on sparse representation, which inherits the advantages of both model-based and learning-based methods. In Section 2, we model the relationship between intensity images and events by assuming they share the same sparse representation, then the iterative shrinkage thresholding algorithm (ISTA) [15] can be used to find the common sparse representation. We then transform the ISTA into a convolutional network by unfolding its iterations. Finally, a ConvLSTM unit and temporal similarity constraints are introduced to enhance the temporal consistency for long video reconstruction. As highlighted in the results in Section 3, our network achieves high quality reconstruction, and outperforms most state-of-the-art networks,

whilst being interpretable and relatively lightweight. Finally, we provide concluding remarks in Section 4.

2. METHODOLOGY

In Section 2.1, the reconstruction problem is modelled based on the working principle of event cameras and sparse representation. In Section 2.2, our network is designed from the ISTA algorithm using the unfolding approach, and we further introduce temporal constraints to enhance its performance.

2.1. Problem Formulation based on Sparse Representation

Events and intensity images. An event is a 4-element vector $\boldsymbol{e} = \{x, y, t, p\}$, where $(x, y)^{\mathrm{T}}$ is the location, t is the timestamp and $p = \pm 1$ is the polarity of the brightness change. Events can be modelled with the continuous-time function $\boldsymbol{e}(t) = p\delta(t - t_0)$, where t_0 denote the timestamp of the event and $\delta(\cdot)$ is the Dirac function. When the change in the logarithm of brightness exceeds a contrast threshold c, an event is triggered. Hence the intensity $\boldsymbol{I}_t$ of the whole image plane at instant t can be obtained from the intensity $\boldsymbol{I}_{t-1}$ at instant $t-1$ and the integral of events over the interval $[t-1, t]$.

However, the integral of events cannot be computed accurately because events are discrete. An alternative way of collecting events $\boldsymbol{E}_{t-1}^{t}$ is by using event voxel grids, as described in [16]. A stream of events is divided into B temporal bins to form a spatio-temporal voxel of size $B \times H \times W$, where $H \times W$ is the number of pixels in the camera. B is set to 5, and we set an upper limit to the number of events for each reconstruction, around 15000. Then, the intensity frame $\boldsymbol{I}_t$ can be potentially estimated from the event voxel grids $\boldsymbol{E}_{t-1}^{t}$ and the previous frame $\boldsymbol{I}_{t-1}$. If we denote with $\mathcal{F}(\cdot)$, the function mapping $\boldsymbol{I}_{t-1}$ and $\boldsymbol{E}_{t-1}^{t}$ to $\boldsymbol{I}_t$, then we aim to model it using a neural network designed by leveraging sparse representation.

Intensity estimation based on sparse representation. In the setting of sparse representation, the input signal can be represented as a linear combination of elementary atoms of a dictionary. We first merge $\boldsymbol{I}_{t-1}, \boldsymbol{E}_{t-1}^{t}$ using head layers $\boldsymbol{W}_I, \boldsymbol{W}_E$ and a fusion layer $\boldsymbol{W}$ (see Fig. 1(b)), where $\boldsymbol{W}$ performs downsampling in order to reduce the computational cost. This leads to a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ of the same size of the target intensity frame $\boldsymbol{I}$, where $n = H/2, m = W/2$. We assume that they can be represented by $\boldsymbol{X} = \boldsymbol{D}_X \boldsymbol{Z}_X$ and $\boldsymbol{I} = \boldsymbol{D}_I \boldsymbol{Z}_I$ respectively, where $\boldsymbol{D}_X, \boldsymbol{D}_I \in \mathbb{R}^{n \times d}$ are dictionaries, and $\boldsymbol{Z}_X, \boldsymbol{Z}_I \in \mathbb{R}^{d \times m}$ are sparse codes. We assume $\boldsymbol{X}$ and $\boldsymbol{I}$ share a common sparse representation $\boldsymbol{Z} = \boldsymbol{Z}_X = \boldsymbol{Z}_I$. Therefore, if proper dictionaries $\boldsymbol{D}_X$ and $\boldsymbol{D}_I$ are found, the reconstruction problem is modelled as a sparse coding problem as follows,

$$\min_{\boldsymbol{Z}} \|\boldsymbol{X} - \boldsymbol{D}_X \boldsymbol{Z}\|_2^2 + \lambda \|\boldsymbol{Z}\|_1. \tag{1}$$

where $\|\cdot\|_1$ denotes the l_1 norm. Once the sparse codes $\boldsymbol{Z}^*$ for input $\boldsymbol{X}$ is found, the target image can then be reconstructed using $\boldsymbol{I} = \boldsymbol{D}_I \boldsymbol{Z}^*$.

2.2. Convolutional ISTA Network with temporal consistency constraints (CISTA-TC)

ISTA network. Eq.(1) is a LASSO problem, which can be solved with the iterative shrinkage thresholding algorithm (ISTA) [15]. The ISTA updates the following equation to optimise the sparse codes $\boldsymbol{Z}^k$ iteratively,

$$\boldsymbol{Z}^k = h_{\boldsymbol{\theta}}(\boldsymbol{Z}^{k-1} + \boldsymbol{P}(\boldsymbol{X} - \boldsymbol{D}_X \boldsymbol{Z}^{k-1})), \tag{2}$$

where $k = 1, \cdots, K$ is the iteration number, $\boldsymbol{P} = \frac{1}{L}\boldsymbol{D}_X^T$, and $h_\theta(\boldsymbol{x}) = sign(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{\theta})_+$ is the soft thresholding function with a threshold $\boldsymbol{\theta}$. An ISTA network can then be constructed using the algorithm unfolding strategy [17] by cascading several iterations and transforming the dictionaries into trainable matrices.

CISTA-TC network. We then construct a convolutional ISTA (CISTA) network by replacing dictionaries with convolutional layers. We use a different ISTA block for each iteration, which consists of convolutional layers $\boldsymbol{D}_k$, $\boldsymbol{P}_k$ and a soft threshold vector $\boldsymbol{\theta}_k$. A $\boldsymbol{D}_I$ layer follows the ISTA blocks to achieve intensity reconstruction.

In order to reconstruct videos, the input $\boldsymbol{I}_{t-1}$ is replaced with the prediction $\hat{\boldsymbol{I}}_{t-1}$, as shown in Fig.1(a). The first input $\boldsymbol{I}_0$ of a sequence is initialised with zeros, following the strategy in [14].

We further adjust CISTA to enhance the temporal consistency (TC) and reduce the accumulated error for reconstruction of long videos. CISTA has two parts, ISTA blocks for sparse coding and $\boldsymbol{D}_I$ for intensity reconstruction. We can therefore improve our architecture for both stages. First, a convolutional long-short term memory (ConvLSTM) unit is added to $\boldsymbol{D}_I$ to achieve intensity coherence, where the state $\boldsymbol{a}_{t-1}$ keeps the memory of the previous reconstruction.

Second, a temporal-similarity (TS) constraint is introduced to preserve the coherence of sparse codes between two consecutive frames, which is inspired by [18]. We use the similarity $\boldsymbol{S}_{t-1,t}$ between the sparse codes $\boldsymbol{Z}_{t-1}$ and $\boldsymbol{Z}_t$ to calculate the constraints. $\boldsymbol{S}_{t-1,t}$ is a matrix with the same size of $\boldsymbol{Z}$. Specifically, the optimisation problem in Eq. (1) is extended as follows,

$$\min_{\boldsymbol{Z}_t} \|\boldsymbol{X}_t - \boldsymbol{D}_X \boldsymbol{Z}_t\|_2^2 + \lambda_1 \|\boldsymbol{Z}_t\|_1 + \lambda_2 \boldsymbol{S}_{t-1,t} \odot \|\boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}\|_2^2, \tag{3}$$

where λ_1, λ_2 are regularization parameters, which can be extended to vectors $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ for multiple channels. The update of the sparse codes at iteration k given in Eq. (2) then becomes,

$$\boldsymbol{Z}_t^k = h_{\boldsymbol{\theta}_k}\big[\boldsymbol{Z}_t^{k-1} + \boldsymbol{P}_k(\boldsymbol{X}_t - \boldsymbol{D}_k \boldsymbol{Z}_t^{k-1}) + \boldsymbol{S}_{t-1,t}^k \odot (\boldsymbol{Z}_{t-1}^K - \boldsymbol{Z}_t^{k-1})\big]. \tag{4}$$

Note that only the last $\boldsymbol{Z}_{t-1}^K$ in the previous reconstruction is used to compute $\boldsymbol{S}_{t-1,t}^k$, using the attention map between

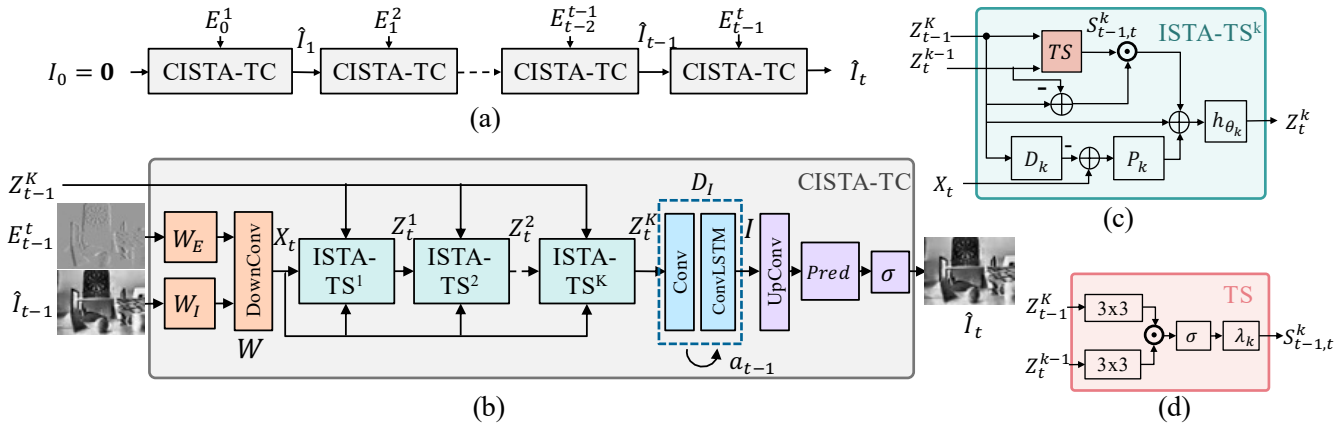

Fig. 1: CISTA with temporal consistency (TC) constraints – ConvLSTM for dictionary $\boldsymbol{D}_I$ and temporal similarity constraint for the sparse codes $\boldsymbol{Z}$. σ and $\odot$ denote the Sigmoid activation and the element-wise multiplication. (a) Flowchart of video reconstruction, (b) CISTA-TC network, (c) ISTA-TS block with temporal similarity (TS) constraint for $\boldsymbol{Z}$, (d) TS block.

$\boldsymbol{Z}_t^{k-1}$ and $\boldsymbol{Z}_{t-1}^K$, as shown in Fig.1(d). The ISTA block with TS constraint (ISTA-TS) is shown in Fig.1(c).

Finally, the overall CISTA-TC network is depicted in Fig 1(b). Specifically, all layers use 3×3 convolutions with a stride 1 except for $\boldsymbol{W}$ whose stride is 2 to perform downsampling. We choose 64 channels for $\boldsymbol{D}_k$, 128 channels for $\boldsymbol{P}_k$ and $K = 5$ in our experiments, which are chosen empirically to make a trade-off between memory consumption and performance. $\boldsymbol{D}_I$ consists of a ConvLSTM unit followed by an upsampling layer. At the end of the network, a prediction layer followed by a Sigmoid activation outputs the final predicted intensity frame $\hat{\boldsymbol{I}}_t$.

3. NUMERICAL RESULTS

Datasets. We use synthetic datasets for training, because paired events and high quality intensity images are difficult to obtain. We simulate events using ESIM [19], from images in MS-COCO datasets [20]. In our simulation, contrast threshold follows a normal distribution $C \sim \mathcal{N}(\mu, \sigma)$ where $\mu = \{0.2, 0.4, 0.6\}$ is chosen randomly and $\sigma = 0.05$. The refractory period is set to $1ms$. We simulate 1100 video sequences, containing around 60000 frames. We use 1000 of these sequences for training and 100 for testing. We also add random noise and hot pixels [21] to event voxel grids in training to enhance the resistance against noise of the network.

The Event Camera Dataset [22] is used for evaluation, which includes 7 video sequences and events captured by DAVIS240C. In testing, we use the first 550 frames of each sequence.

Training and testing settings. We use a combination of l_1 loss, structural similarity (SSIM) and perceptual loss (LPIPS) [23] between reconstructed and ground truth frames. LPIPS employs deep features as a similarity metric, where VGG is used in our experiment. We adopt the many-to-one scheme in training, which means that loss is only calculated once at the end of the sequence. For a video sequence of length L, the loss function is as follows,

$$\mathcal{L}_s = \|\boldsymbol{I}_L - \hat{\boldsymbol{I}}_L\|_1 + (1 - SSIM(\boldsymbol{I}_L, \hat{\boldsymbol{I}}_L)) + d(\boldsymbol{I}_L, \hat{\boldsymbol{I}}_L), \quad (5)$$

where $d(\cdot)$ denotes the perceptual loss. Other training details include the length of sequence is $L = 15$, batch size is 1, learning rate is initialised with 0.0001 and decays 10% every 10 epochs, and training is performed for 20 epochs.

In evaluation, we stretch the grayscale value to the range [0, 1] and implement histogram equalization for both ground truth and reconstructed images. Mean-square error (MSE), peak signal-to-noise ratio (PSNR), SSIM and LPIPS are used as image quality metrics.

Results. We compare our methods with three state-of-the-art networks. E2VID [11, 12] is a Unet architecture with ConvLSTMs and residual blocks. FireNet [13] simplifies E2VID since it uses fewer layers, smaller kernel size and has no downsampling. SPADE-E2VID [14] adds a SPADE layer in decoders of E2VID to fuse previous frames. These networks are retrained using our datasets and training strategies. SPADE-E2VID also requires much more memory as shown in Table 3.

Results are shown in Table 1 and Fig. 2. Our network outperforms E2VID and FireNet. FireNet has lower ability to recover intensity and reconstructed frames are more noisy. E2VID can recover more details but it performs worse than us and SPADE-E2VID in all metrics. SPADE-E2VID produces the best results on real datasets except for SSIM, while our model improves structural similarity and performs the best on simulated datasets. Also, the images of shapes_6dof reconstructed using SPADE-E2VID have smearing artefacts.

Ablation Study. We test on models without the ConvLSTM unit and temporal similarity constraints to verify their effectiveness. Table 2 and Fig. 3 show the results. CISTA is the

Table 1: Comparison of CISTA-TC and other networks, with corresponding citations E2VID [12], FireNet [13] and SPADE [14]. mean / mean_sim: average results on real / simulated datasets. Key: **Best** / *Second best*

	MSE				PSNR				SSIM				LPIPS			
Dataset	E2VID	FireNet	SPADE	Ours	E2VID	FireNet	SPADE	Ours	E2VID	FireNet	SPADE	Ours	E2VID	FireNet	SPADE	Ours
poster_6dof	0.048	0.057	**0.026**	*0.033*	13.29	12.54	**15.98**	*15.00*	0.502	0.478	**0.554**	*0.549*	0.330	0.376	**0.276**	*0.303*
boxes_6dof	**0.040**	0.057	0.046	*0.043*	**14.11**	12.52	13.59	*13.88*	0.521	0.474	*0.524*	**0.533**	0.298	0.364	**0.283**	*0.296*
calibration	0.034	0.060	**0.031**	*0.032*	14.82	12.58	**15.36**	*15.31*	0.486	0.437	*0.496*	**0.506**	**0.245**	0.318	0.249	*0.248*
shapes_6dof	0.035	0.033	*0.028*	**0.026**	14.73	14.95	*15.71*	**15.98**	0.620	**0.648**	0.581	*0.633*	0.447	0.430	*0.383*	**0.367**
slider_depth	0.055	0.064	*0.055*	**0.044**	12.85	12.11	*12.96*	**13.87**	0.443	0.441	*0.464*	**0.487**	0.357	0.400	**0.332**	*0.342*
dynamic_6dof	0.115	0.152	**0.072**	*0.081*	9.636	8.265	**11.65**	*11.29*	0.368	0.362	**0.421**	*0.415*	0.369	0.391	*0.322*	**0.313**
office_zigzag	**0.043**	0.112	0.051	*0.048*	*13.82*	9.621	13.17	**13.47**	*0.409*	0.353	*0.418*	**0.424**	0.303	0.393	**0.291**	*0.297*
mean	0.053	0.075	**0.042**	*0.043*	13.34	11.97	**14.31**	*14.21*	0.491	0.469	0.5059	**0.518**	0.336	0.378	**0.303**	*0.306*
mean_sim	0.036	0.054	*0.034*	**0.030**	15.17	13.32	*15.50*	**15.78**	*0.494*	0.463	0.510	**0.511**	*0.318*	0.394	**0.316**	0.330

Table 2: Comparison between variants of our models.

	MSE			PSNR			SSIM			LPIPS		
Dataset	CISTA	CISTA-D	CISTA-TC	CISTA	CISTA-D	CISTA-TC	CISTA	CISTA-D	CISTA-TC	CISTA	CISTA-D	CISTA-TC
mean_real	0.054	0.049	**0.043**	13.24	13.93	**14.21**	0.504	0.510	**0.518**	0.321	**0.306**	**0.306**
mean_sim	0.034	0.033	**0.030**	15.09	15.47	**15.78**	0.491	0.508	**0.511**	0.3490	0.336	**0.330**

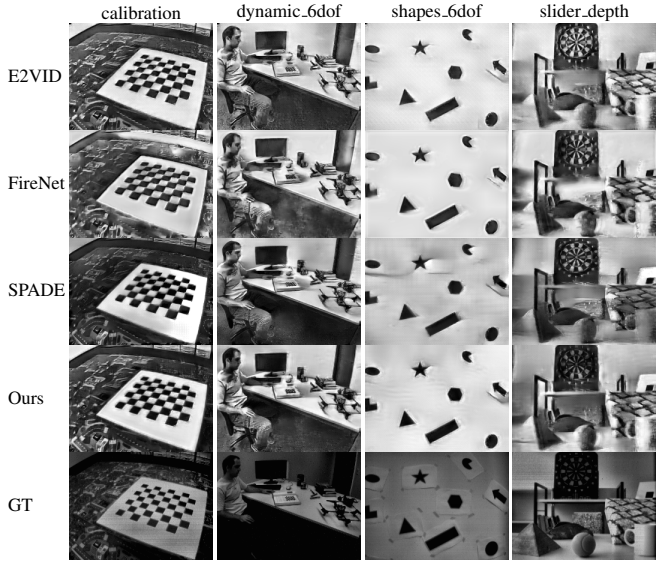

Fig. 2: Comparison of CISTA-TC and other networks

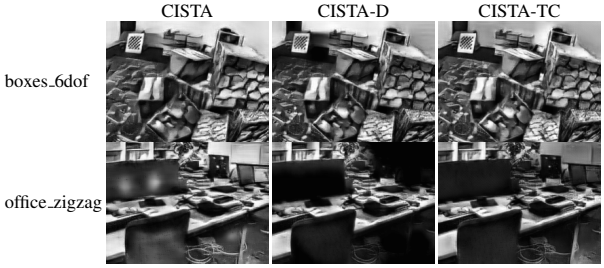

Fig. 3: Results for variants of CISTA-TC

model without any temporal consistency constraints. Fig. 1 especially office_zigzag shows that its ability of recovering intensity is affected. CISTA-D only contain ConvLSTM for $\boldsymbol{D}_I$, and its ability to recover intensity increases compared with CISTA, but predicted images may contain dark areas where details get lost. The reconstruction quality of CISTA-TC is significantly better overall. Images keep more intensity without losing details. Temporal similarity constraints help maintain temporal consistency.

Efficiency. Table 3 compares the memory use and the average reconstruction time of different models. The memory usage of E2VID and SPADE-E2VID is very large and FireNet is the fastest and smallest one. Our network is relatively light but the computational cost is close to SPADE-E2VID, because it has fewer downsampling layers so that calculation is executed on relatively large images and the computation of temporal similarity constraints is more complex. All tests are on an i7-8700 CPU @ 3.20GHz and a GeForce GTX 1080 Ti GPU.

Table 3: Comparison of memory usage and average reconstruction time per frame.

	E2VID	FireNet	SPADE-E2VID	Ours
Memory(Mb)	42.9	0.16	45.9	2.7
Time(ms)	2.0	1.2	2.9	3.0

4. CONCLUSION

We proposed a deep network for event-to-video reconstruction inspired by sparse representation. The reconstruction process is modelled as a sparse coding problem based on the relationship between events and intensity images. A convolutional ISTA network is then designed to solve this problem by unfolding ISTA iterations. The network is further improved using temporal consistency constraints. Our CISTA-TC network benefits from both model-based and learning-based approaches to achieve high-quality reconstruction, while being more interpretable than other networks and relatively lightweight.

5. REFERENCES

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based Vision: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[2] C. Reinbacher, G. Graber, and T. Pock, "Real-Time Intensity-Image Reconstruction for Event Cameras Using Manifold Regularisation," in *Proc. British Mach. Vis. Conf.*, 2016.

[3] P. Bardow, A. J. Davison, and S. Leutenegger, "Simultaneous Optical Flow and Intensity Estimation from an Event Camera," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 884–892, 2016.

[4] C. Brandli, L. Muller, and T. Delbrück, "Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor," in *IEEE Int. Symp. Circuits Syst. (ISCAS)*, pp. 686–689, 2014.

[5] C. Scheerlinck, N. Barnes, and R. Mahony, "Continuous-Time Intensity Estimation Using Event Cameras," in *Asian Conf. Comput. Vis. (ACCV)*, 2018.

[6] Z. W. Wang, P. Duan, O. Cossairt, A. Katsaggelos, T. Huang, and B. Shi, "Joint Filtering of Intensity Images and Neuromorphic Events for High-Resolution Noise-Robust Imaging," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1606–1616, 2020.

[7] S. Barua, Y. Miyatani, and A. Veeraraghavan, "Direct face detection and video reconstruction from event cameras," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016.

[8] B. Wang, J. He, L. Yu, G.-S. Xia, and W. Yang, "Event enhanced high-quality image recovery," in *Computer Vision–ECCV 2020: 16th European Conference*, pp. 155–171, Springer, 2020.

[9] L. Wang, I. M. Mostafavi, Y.-S. Ho, and K.-J. Yoon, "Event-Based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10073–10082, 2019.

[10] B. Su, L. Yu, and W. Yang, "Event-Based High Frame-Rate Video Reconstruction With A Novel Cycle-Event Network," in *IEEE Conf. Image Process. (ICIP)*, pp. 86–90, 2020.

[11] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-To-Video: Bringing Modern Computer Vision to Event Cameras," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3852–3861, 2019.

[12] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High Speed and High Dynamic Range Video with an Event Camera," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[13] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. E. Mahony, and D. Scaramuzza, "Fast Image Reconstruction with an Event Camera," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 156–163, 2020.

[14] P. R. G. Cadena, Y. Qian, C. Wang, and M. Yang, "SPADE-E2VID: Spatially-Adaptive Denormalization for Event-Based Video Reconstruction," *IEEE Trans. on Image Process.*, vol. 30, pp. 2488–2500, 2021.

[15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–1457, Nov. 2004.

[16] A. Z. Zhu, L. Yuan, K. Chaney, and K. Daniilidis, "EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras," in *Robot.: Sci. Syst.*, June 2018.

[17] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing," *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

[18] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, and S. Gao, "Video Anomaly Detection with Sparse Coding Inspired Deep Neural Networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, pp. 1070–1084, Mar. 2021.

[19] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an Open Event Camera Simulator," in *Conf. Robot. Learn. (CoRL)*, pp. 969–982, 2018.

[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *Comput. Vis. – ECCV 2014*, vol. 8693, pp. 740–755, Springer, 2014.

[21] T. Stoffregen, C. Scheerlinck, D. Scaramuzza, N. Barnes, L. Kleeman, and R. Mahony, "Reducing the Sim-to-Real Gap for Event Cameras," in *Eur. Conf. Comput. Vis. (ECCV)*, pp. 534–549, 2020.

[22] E. Mueggler, H. Rebecq, G. Gallego, T. Delbrück, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and SLAM," *Int. J. Robot. Res.*, vol. 36, pp. 142–149, Feb. 2017.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, pp. 586–595, 2018.