

Temporal-Mapping Photography for Event Cameras

Yuhan Bao^{1*}, Lei Sun^{1,2*}, Yuqin Ma¹, and Kaiwei Wang¹

¹ National Research Center for Optical Instrumentation, Zhejiang University

² INSAIT, Sofia University

wangkaiwei@zju.edu.cn, leo_sun@zju.edu.cn

Abstract. Event cameras, or Dynamic Vision Sensors (DVS) are novel neuromorphic sensors that capture brightness changes as a continuous stream of “events” rather than traditional intensity frames. Converting sparse events to dense intensity frames faithfully has long been an ill-posed problem. Previous methods have primarily focused on converting events to video in dynamic scenes or with a moving camera. In this paper, for the first time, we realize events to dense intensity image conversion using a stationary event camera in static scenes with a transmittance adjustment device for brightness modulation. Different from traditional methods that mainly rely on event integration, the proposed Event-Based Temporal Mapping Photography (EvTemMap) measures the time of event emitting for each pixel. Then, the resulting Temporal Matrix is converted to an intensity frame with a temporal mapping neural network. At the hardware level, the proposed EvTemMap is implemented by combining a transmittance adjustment device with a DVS, named Adjustable Transmittance Dynamic Vision Sensor (AT-DVS). Additionally, we collected TemMat dataset under various conditions including low-light and high dynamic range scenes. The experimental results showcase the high dynamic range, fine-grained details, and high-grayscale resolution of the proposed EvTemMap. The code and dataset are available in <https://github.com/YuHanBaozju/EvTemMap>.

Keywords: Dynamic Vision Sensor · Image reconstruction · High dynamic range imaging · Low-light photography

1 Introduction

Event cameras offer high temporal resolution and dynamic range, swiftly responding to brightness changes through positive and negative events. While the distinctive asynchronous data output from event cameras diminishes data redundancy, it concurrently discards the grayscale information of the image. To capture grayscale images, a dynamic and active pixel vision sensor (DAVIS) [5] integrates an active pixel sensor (APS) and a DVS sensor on a single chip. This integration allows the sensor to output both grayscale frames and events. The

* Authors contributed equally to this work.

fusion of grayscale images with events facilitates various computer vision tasks. Many event-based image/video enhancement tasks rely on such devices, such as motion deblur [17, 26, 40, 52], video interpolation [41, 43], high dynamic range (HDR) imaging [30], auto-focusing [3] *etc.* These tasks leverage the high temporal resolution and dynamic range of events to enhance grayscale images.

However, due to the integration of multiple imaging sensors on a single chip, DAVISs have lower fill factors, and smaller pixel resolutions [13]. Additionally, the grayscale images produced by the DAVIS device are consistent with traditional images and lack the high temporal resolution and dynamic range of DVS. Thus, to leverage the advantages of event cameras, some works aim to utilize events from pure DVS to reconstruct videos. Previous reconstruction approaches are based on hand-crafted priors and strong assumptions about the imaging process as well as the statistics of natural images [31]. Recent work, such as E2VID [35, 39], trains a recurrent neural network to reconstruct videos from a stream of events. However, these methods suffer from grayscale distortion, lack of texture detail, and failure on static scenes. In contrast to E2VID, tailored for dynamic scenes, our objective is to capture diverse scenes, encompassing both static and dynamic ones, akin to traditional photography.

Since the temporal dimension of events exhibits theoretically infinite description levels, with $1\ \mu\text{s}$ representing one level, can we utilize time to measure grayscale? Imagine a scene: the event camera is initially in complete darkness, and when the lighting gradually increases, the event camera captures positive events as each pixel gets brighter. Each pixel lights up at a different rate, depending on its corresponding point’s light intensity in the scene. As the event sensor initiates from darkness and gradually intensifies, the time each pixel takes to reach a specified threshold (i.e., the timestamp of the initial positive event, IPE) indicates its brightening speed. Hence, through the measurement of the timestamp associated with each IPE, we can derive the light intensity of the corresponding point within the scene.

Building upon the mentioned basic concept, we introduce our EvTemMap. Initially, we block all incoming light to the event camera. Then, as we progressively raise the transmittance of the optical system, the brightness of each pixel increases at different rates. Throughout this process, we record the timestamp of IPE for each pixel, forming a Temporal Matrix. Subsequently, the Temporal Matrix undergoes time-intensity mapping and finally produces a grayscale image. However, due to hardware constraints, the time-intensity mapping is not as ideal as we derive. Various factors such as uncertain contrast threshold, timestamp error, *etc.* degrade the final image. To address this issue, we design a temporal and spatial degradation space, especially for event-to-image conversion. Trained on the designed degradation model, our temporal-mapping network converts Temporal Matrix to high-quality, high-grayscale-resolution, and high-dynamic-range image faithfully. Additionally, we record Temporal Matrix across diverse scenes and compile a dataset named TemMat dataset. Through extensive comparisons with existing event-to-video methods, we demonstrate the superiority of the proposed EvTemMap.

In summary, we make the following main contributions:

- We revisit the mechanism of the event camera, and propose event-based temporal-mapping photography (EvTemMap) to convert events from a stationary AT-DVS in static scenes to an image, which is the first work in this area to the best of our knowledge.
- We further explore the temporal and spatial degradation, design a degradation model and train a temporal mapping network for our EvTemMap.
- We present TemMat dataset consisting of diverse scenarios, varying illuminance levels, and high dynamic range scenes. Experiments on the dataset underscore the superiority of the proposed method.

2 Related Work

2.1 Event to Video Reconstruction

Non-learning-based methods. The pioneering work [9] demonstrated the feasibility of extracting intensity information from event data and reconstructing images from events captured by an event camera traversing a stationary scene. Their approach mainly depends on the inherent connection between intensity gradients and optical flow, utilizing each event to formulate equations for brightness constancy. Subsequently, event integration was proposed to be independent of assumptions regarding scene structure or motion dynamics [31] by Munda *et al.* Scheerlinck *et al.* [37] defined the reconstruction as an energy minimization problem and applied a high-pass filter to events before integration. However, these methods frequently encounter "bleeding edges" caused by spatially variant contrast thresholds.

Learning-based Methods. With the advancement of deep learning, Barua *et al.* [4] utilized K-SVD [2] on simulated data to train a dictionary mapping small patches of integrated events to image gradients. Subsequently, they employed Poisson integration for intensity image reconstruction. Rebecq *et al.* [36] designed an end-to-end recurrent model trained with synthetic events, resulting in the high-speed video reconstruction method E2VID. Subsequent works [6, 38, 39, 51] maintained this paradigm while improving efficiency [38], fine-grained details [6], and performance in low-light conditions [51]. More recently, Paredes-Vall'es *et al.* [33] extended E2VID to self-supervised learning by introducing optical flow as a constraint. However, all these methods only recover intensity up to an unknown initial image, leading to "ghosting" effects and non-uniform grayscale distribution. Moreover, these methods are primarily applicable in dynamic scenes, where either the camera or the scene exhibits motion.

2.2 High Dynamic Range Imaging

Bracket exposure HDR. The most common method for achieving HDR photography is to use different exposure parameters to expose the same scene multiple times, obtain multiple low dynamic range (LDR) images, and merge them

to reconstruct an HDR image [19]. This technique, commonly referred to as bracket exposure, poses a crucial and challenging step in aligning multiple LDR images [44]. Among the representative alignment methods are: Optical Flow-based Image Alignment [10, 19, 49], Direct Feature Concatenation [7, 32, 47], Correlation-Guided Feature Alignment [8, 28, 48] and Image Translation-Based Alignment [22, 23]. After alignment, the process of fusing LDR images has similarities to many image fusion methods, as summarized in [20]. Motion within the multi-exposure is the key factor that influences the method. Messikommer *et al.* [30] utilize the inter-frame motion information provided by the events to assist in the alignment of LDR stacks. While such an approach results in high-confidence light and dark details, multi-frame alignment still produces excessive alignment artifacts for large-scale motion.

Deep single-exposure HDR. Single-exposure HDR image reconstruction aims to recover missing information in the saturated areas of an LDR image and reconstruct an HDR image using a neural network [44]. Representative methods in this domain include HDRCNN [11] and deep reverse tone mapping [12]. While the single-exposure HDR approach avoids the alignment problem in bracketing, its recovery of overexposed and underexposed areas is an ill-posed problem, and synthetic artifacts may compromise the credibility of the result.

3 Event-Based Temporal-Mapping Photography

The independent pixels on the event cameras can respond to changes in the brightness $L(x, y, t)$. Eq. (1) shows how the event camera works:

$$p(x, y, t_i) = \begin{cases} +1, & \text{if } L(x, y, t_i) - L(x, y, t_{i-1}) > C, \\ -1, & \text{if } L(x, y, t_i) - L(x, y, t_{i-1}) < -C, \end{cases} \quad (1)$$

where $L(x, y, t_i)$ denotes the brightness on the pixel coordinates (x, y) at the current moment t_i , $L(x, y, t_{i-1})$ denotes the brightness when the event is last triggered, and C denotes the contrast threshold. When the brightness change exceeds the threshold, events with different polarities are emitted. Brightness is a logarithmic map of light intensity [13]. Intensity images cannot be obtained solely through DVS. Previous work [36] recovered grayscale frames/videos from moving cameras or in dynamic scenes by training an event-to-video network. However, events cannot be directly integrated to recover accurate grayscale images if both the camera and the scene are static.

3.1 Principles of Event-Based Temporal-Mapping Photography

Returning to the fundamental principle introduced in Eq. (1), the core of the event camera is the change in brightness. Various factors contribute to changes in brightness, with motion in a textured scene being one of them. E2VID [36] leverages such events to recover its grayscale images. Rather than confining our thinking to motion events, can we focus on the essence of event cameras—detecting changes in brightness?

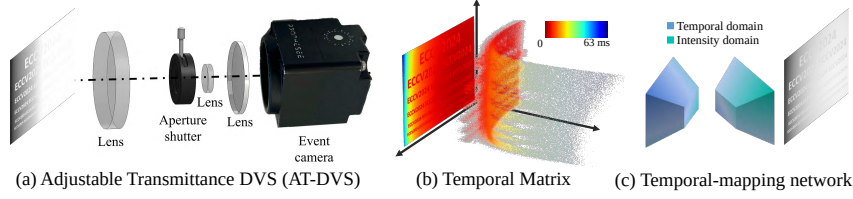


Fig. 1: Pipeline of EvTemMap. The actual scenes are initially recorded by AT-DVS, generating a Temporal Matrix. Following this, the Temporal Matrix undergoes time-intensity mapping and denoising through a Temporal-mapping network, producing a grayscale image.

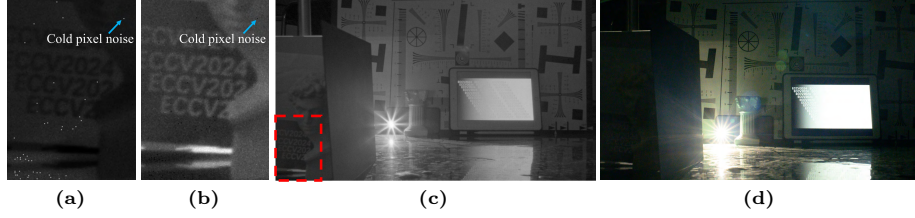


Fig. 2: An implementation example of EvTemMap. (a) Temporal Matrix. (b) The intensity image obtained by substituting Temporal Matrix into Eq. (4). (c) The adaptive dynamic range image ensures optimal grayscale resolution without underexposure or overexposure. (d) Restricted dynamic range image shows overexposing and underexposing.

Suppose that the initial brightness on the DVS sensor plane is 0, namely $L(x, y, t_{i-1})$ is 0 in Eq. (1). As we gradually increase the transmittance of the DVS, the brightness on the sensor plane increases. The IPE of each pixel triggers when $L(x, y, t_i)$ reaches C . Since the corresponding objects in the scene have different intensities, different pixels exhibit distinct rates of brightness growth. Consequently, the brightness $L(x, y, t_i)$ reaches the threshold C at different times for different pixels. Unlike conventional cameras that accumulate photons over a fixed period to measure light intensity, our method incrementally elevates the brightness everywhere on the sensor and records the time reaching a predefined brightness level, *i.e.*, the IPE timestamp. IPE timestamps are negatively correlated with the absolute light intensity in the scene.

3.2 Details of Implementation

The transmittance adjustment device (TA) is the key to our method, which needs to be able to adjust the transmittance of the incident light **starting from 0**. We adopt an aperture shutter as the TA, as shown in Fig. 1a. Other alternative TAs such as rotary polarizer attenuators and liquid crystal shutters are discussed in the supplementary material. TA makes the transmittance rate (TR) of the

optical system change from 0 to 1. In general, TR changes over time according to the following function:

$$TR(t) = \begin{cases} f(t), & 0 \leq t \leq t_{end}, \\ 1, & t > t_{end}, \end{cases} \quad (2)$$

where $f(t)$ is a monotonically increasing function with a value range between 0 and 1, defined by users. t_{end} denotes the point in time when the transmittance reaches 1. By adjusting t_{end} , we can control the speed of change of the transmittance, thus affecting the acquisition time of EvTemMap.

In EvTemMap, the accuracy of the temporal measurement of the DVS affects the accuracy of the grayscale recovery. Event cameras exhibit significant event delay in low light, where the timestamp of an event lags behind its actual time of occurrence. This delay is due to the decrease in the bandwidth of the photodiode in low light conditions [34]. In low-light situations, due to the presence of event delays, the $L(x, y, t_i)$ in Eq. (1) is no longer the current instantaneous value but should be considered as an integral value from 0 up to the present moment. We interpret Eq. (1) in the EvTemMap process at the circuit level:

$$V_{diff} = \frac{1}{C_{PD}} \log(1 + \int_0^{t^*(x,y)} I_{max}(x, y) \cdot TR(t) dt) - V_{ref} \equiv V_{thd}, \quad (3)$$

where $t^*(x, y)$ is the timestamp when the IPE is generated for a given pixel (x, y) . C_{PD} is the capacitance of the photodiode. V_{ref} is the reference voltage, and since transmittance increases from 0, V_{ref} is ideally 0. V_{thd} is the voltage threshold, quantitatively related to the event trigger threshold C in Eq. (1). To conform to the dimension, $I_{max}(x, y)$ can be considered here as the photocurrent when the transmittance is 1.

As $TR(t)$ increases from 0 to 1, the brightness on the sensor plane increases, and the photocurrent increases accordingly. During this process, the photodiode continues to accumulate photons until V_{diff} exceeds V_{thd} . There is a linear mapping relationship between the photocurrent, light intensity, and the grayscale, so $I_{max}(x, y)$ is the target value. We perform a simple transformation of Eq. (3) to reflect the mapping from $t^*(x, y)$ to $I_{max}(x, y)$ more clearly:

$$I_{max}(x, y) = \frac{\exp((V_{ref} + V_{thd}) \cdot C_{PD}) - 1}{h(t^*(x, y))}, h(t^*) = \int_0^{t^*(x,y)} TR(t) dt. \quad (4)$$

For grayscale images, we care about the relative values among $I_{max}(x, y)$ rather than their exact magnitudes. So the numerator term, as a coordinate-independent constant, can be eliminated by normalization operations. By substituting each $t^*(x, y)$ in the Temporal Matrix into Eq. (4) and normalizing $I_{max}(x, y)$ by the maximum and minimum values, we derive the normalized grayscale image $\overline{I_{max}}$, shown in Fig. 2b, where $\overline{I_{max}} \in [0, 1]^{W \times H}$.

In Eq. (4), we define the definite integral over $TR(t)$ as another function $h(t^*(x, y))$ concerning $t^*(x, y)$, which is also a monotonically increasing function. While the true $TR(t)$ may deviate from the predefined function (see Eq. (2)), this disparity influences the mapping curve of $\overline{I_{max}}$, akin to the Look Up Table (LUT) mapping curve in traditional images.

Overview. Figure 1 illustrates the pipeline of our work. First, we integrate the transmittance adjustment device and the event camera, forming AT-DVS. Then, we record the events generated when the transmittance increases from 0 and record the timestamp when each pixel triggers the IPE, forming a Temporal Matrix. Subsequently, through a Temporal-mapping network, we perform time-intensity mapping and denoising to the Temporal Matrix. And finally, we get the result of EvTemMap.

Adaptive Dynamic Range. For conventional cameras, the dynamic range of is a fixed value (typically 70 dB) when the exposure parameters are fixed. In high dynamic range scenes (103 dB in the scene of Fig. 2c), single-exposure images from conventional cameras are often overexposed and underexposed, as presented in Fig. 2d. However, EvTemMap enables adaptive dynamic range photography, effectively preserving both dark and light details within the image. This capability stems from the fact that the Temporal Matrix operates within the range spanning from the brightest pixel t^* to the darkest pixel t^* .

High grayscale resolution. Another benefit of EvTemMap is its higher grayscale resolution. For conventional cameras, a 12-bit raw format file is capable of storing up to 4096 gray levels. In contrast, our EvTemMap utilizes the high temporal resolution of the event camera to achieve extremely high grayscale resolution. Taking the scene in Fig. 2c as an example, the difference in t^* between the brightest and darkest is 292111 μs , and the temporal resolution of the event camera is 1 μs . Theoretically, the effective gray resolution level of Fig. 2c is 292111 levels, which is **71.3 times** that of a 12-bit raw format file.

3.3 Degradation Model for EvTemMap

Equation (4) delineates the ideal formulation for the conversion of events to intensity image. However, for event cameras in the real world, there are many practical factors that make it impossible to derive high-quality intensity images directly from the events: (1) uncertain threshold C , (2) event timestamp error, (3) noise brought by cold and hot pixels, and (4) other degradations. Previous work [45, 50] designed a practical degradation model to restore low-quality images. Inspired by this, we analyze each factor for our temporal-mapping imaging in both **temporal domain** and **spatial domain** and design a degradation space as large and realistic as possible for our EvTemMap.

Spatially variant contrast threshold. The threshold values C for a given event camera vary both over time and across the image instead of remaining constant [39], which results from the variant V_{thd} in Eq. (4), as shown in Fig. 3a. In addition, V_{ref} is not ideally 0 due to dark current or TA light leakage. We consider the term $((V_{ref} + V_{thd}) \cdot C_{PD})$ as the threshold C for the event camera and summarize the above sources of noise, embodied as random perturbations of

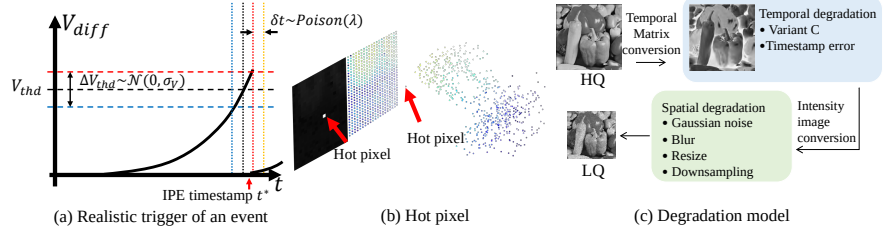


Fig. 3: Degradation paradigm. (a) The real IPE timestamp is affected by the threshold random perturbation ΔV_{thd} and the time measurement error δt . (b) The hot pixel is reflected as an unusually early value on the Temporal Matrix and as the brightest white dot in the grayscale image. (c) The process of generating training data for the degradation model.

C. To cope with different threshold conditions, we choose C randomly from 0.1 to 0.4 for each sample. Then we add a ΔC to the set threshold C for different pixels, where the ΔC obeys the zero-mean Gaussian noise model $\mathcal{N}(0, \sigma)$. σ is uniformly sampled from 0.02 to 0.08.

Event timestamp error comes from the time uncertainty of DVS output events. This time uncertainty is manifested as a time delay that approximates a Poisson distribution [15]. This temporal measurement noise is also spatially random on EvTemMap images. In our degradation model, we add δt to $t^*(x, y)$ in Eq. (4), and t is sampled from $\text{Poisson}(\lambda)$, where λ is uniformly sampled from 10 to 1000 μs . Extreme conditions are discussed in supplementary materials.

Cold and hot pixels. Because of the limitation of the circuit, there are cold pixels, which are less prone to generating events, and hot pixels, which constantly generate events in DVS [15]. In the Temporal Matrix, the cold pixel is the location where t^* is null, which is artificially set to the maximum value of t^* , as shown in Figs. 2a and 2b. In contrast, the hot pixel exhibits an unusually early t^* , resulting in the brightest white dots in the grayscale image, as shown in Fig. 3b. Experimentally, it is found that hot pixels consistently occupy specific locations. We locate the position of hot pixels with pixel calibration [46]. We then remove the hot pixels by defining the t^* of them as the minimum of the other positions in Temporal Matrix, *i.e.* temporal pre-processing.

Other degradation. Apart from the above degradation, there is still some unpredictable degradation such as out-of-focus blur, noise from dark current, *etc.* Thus we choose Gaussian noise model $\mathcal{N}(\mu, \sigma)$, where σ is uniformly sampled from $1/255$ to $25/255$. For the out-of-focus blur, we add the same blur kernel as BSRGAN [50]. The current event cameras suffer from relatively low resolution due to limitations in transmission bandwidth [14]. To manifest the need for high resolution, we super-resolve the final result of EvTemMap with a scale factor of 4. Thus we also incorporate downsampling degradation into our degradation space following [50].

Up to this point, we have thoroughly analyzed various types of degradation within the framework of EvTemMap. To enlarge the degradation space, we use the random shuffle strategy for all the degradation processes [50]. Fig. 3c shows our degradation model. In the training stage, given a high-quality (HQ) image, we first convert the image from the spatial domain to the temporal domain with Eq. (4) and apply degradation for variant contrast threshold, event timestamp error, and cold pixels. After the temporal degradation, we convert the degraded Temporal Matrix to intensity image and apply the rest spatial degradation. Finally, the resulting low-quality image (LQ) is used as the input of the Temporal-mapping network for training. In contrast to E2VID [36], our approach does not require synthetic events for training, thereby obviating the need to consider the domain gap issue arising from it.

4 Experiments

4.1 TemMat Dataset

For comparison, we collect TemMat dataset. The dataset features (1) diverse scenarios including both outdoor and indoor scenes, (2) a significant luminance range from 0.02 lux to 20000 lux, and (3) high dynamic range scenes, reaching up to 114 dB. We collect the Temporal Matrix with AT-DVS. To compare with other E2VID [36, 39] methods, for each scene, we also collect motion events by slowly and gently shaking the DVS, without changing the transmittance of the optical system. For capturing reference images, a Nikon D780 SLR camera equipped with an AF-S NIKKOR 50mm f/1.4G lens, configured in bracket-exposure HDR mode, is employed. To achieve a depth of field comparable to that of the event camera, an aperture of F16 is utilized in aperture priority mode. Because the Temporal Matrix, motion events for E2VID, and HDR images by SLR camera are obtained separately, the results of EvTemMap, E2VID, and reference images are not pixel-aligned.

4.2 Settings

AT-DVS details. We adopt prophesee EVK4 as DVS sensor in our experiments, with a spatial resolution of 1280×720 and a temporal resolution of $1 \mu s$. Different from DAVIS, which outputs spatially aligned intensity images and events, EVK4 only outputs events. For the transmittance function $TR(t)$, we place an aperture shutter at the aperture diaphragm to control the transmittance of the optical system by adjusting the aperture. Please refer to the supplement for more details on our AT-DVS.

Training details. The novelty of the paper lies in the brand new event-to-image conversion method, *i.e.* temporal mapping and corresponding degradation model. To demonstrate the advantage of the proposed method, we choose widely-used Swinir [24] as our image restoration model for the temporal-mapping network. The model is trained on a composition comprising DIV2K [1], Flick2K [25],

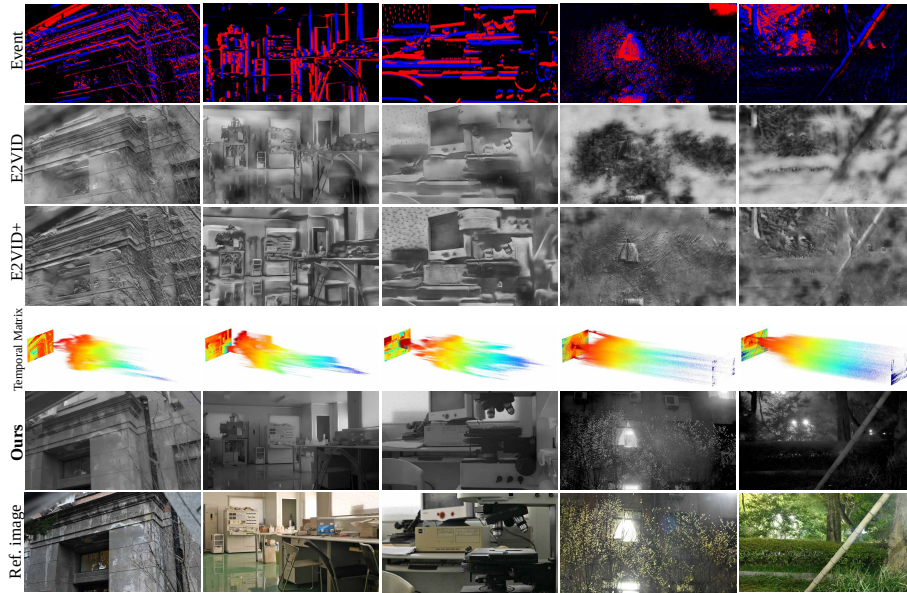


Fig. 4: Qualitative results on TemMat dataset. Rows from top to bottom: Visualized event frame, results of E2VID [35] and E2VID+ [39], visualized Temporal Matrix, results of EvTemMap, reference image. Our method achieves the most faithful and finest result. Best viewed on a screen and zoomed in.

42], WED [29]. With the HQ images in the dataset, we produce the LQ images with the proposed temporal and spatial degradation model on the fly. We trained model with 72×72 LQ patch size. Following BSRGAN [50], we train the model with a weighted combination of L1 loss, perceptual loss [18], and spectral norm-based least square PatchGAN loss [16] with weights 1, 1, and 0.1, respectively. We train the model using Adam [21] with an initial learning rate of 0.0002 and cosine learning rate decay with a minimum learning rate of 10^{-7} . We train the model with a batch size of 48 for 1000k iterations.

Compared methods. We are the first method that is able to convert a static scene from a stationary event camera to the best of our knowledge, the most similar work is E2VID, which translates motion events to video. Thus, we compare our method with E2VID [36] and E2VID+ [39].

4.3 Comparisons with State-of-the-Art Methods

Qualitative results. Qualitative results on the TemMat dataset are presented in Fig. 4, comparing our approach with two event-to-video methods: E2VID [35] and E2VID+ [39]. Our approach achieves grayscale image recovery closest to the reference, showcasing the realism in our grayscale restoration in both indoor and outdoor scenes. Specifically, in the leftmost column, our method effectively

captures the grayscale transition from outdoors to indoors, and even reflections from glass are discernible. In addition, in the night scenes presented in the two right columns, the trees in the shadows and the lights in the bright areas are correctly grayscale mapped. However, while the E2VID and E2VID+ methods roughly depict the outlines of the scene in the left three columns, their grayscales are significantly distorted. In the last two columns, they recover a grayscale that is even opposed to the real situation.

For high-frequency textures, our EvTemMap also performs far better than other event-based methods, such as the tree branches and blooming flowers in the fourth column. E2VID and E2VID+, on the other hand, only recover the general outline of the scene but blur the details, such as tree branches. Intriguingly, our results showcase the capability to discern the approximate texture of illuminated indoor environments from outdoor settings at night, a challenge even for HDR images from SLR cameras.

Table 1: Quantitative comparison on TemMat dataset.

Methods	EvTemMap	E2VID [35]	E2VID+ [39]	SPADE [6]
PSNR	15.75	10.48	11.21	10.20
SSIM	0.69	0.58	0.55	0.52

Quantitative results. Table 1 presents the quantitative comparison results with the E2VID-based methods [6, 35, 39]. We take the bracketed HDR grayscale images captured by the SLR camera as the ground truth, and use LightGlue [27] to align the results of various methods with the ground truth, and then perform PSNR and SSIM evaluations. In terms of quantitative results, EvTemMap performs the best, while the three E2VID-based methods do not differ much. More detailed quantitative and qualitative results can be explored in the supplementary material.

4.4 High Dynamic Range Photography

To test the dynamic range of our EvTemMap, we construct an HDR scene with a range of 114dB in a dark room, featuring 20430 lux at the brightest point and 0.04 lux at the darkest point measured by an illuminance meter. Figure 5a shows that EvTemMap records details in both shadow and highlight areas without overexposure or underexposure. Around the brightest point, we can observe the gradation of brightness within the elliptical light source as well as the diffracted asterisms around it. Near the darkest point, the result of EvTemMap effectively captures intricate details in the hair of David’s sculpture and natural variations in light and shadow on the shoulder. In the rest of the image, the “ECCV2024” texts on the screen and in the shadows, as well as the background resolution board, are clearly visible. Comparing Fig. 5a and Fig. 5b, it can be found that EvTemMap can truly depict various details in the scene, while E2VID can only

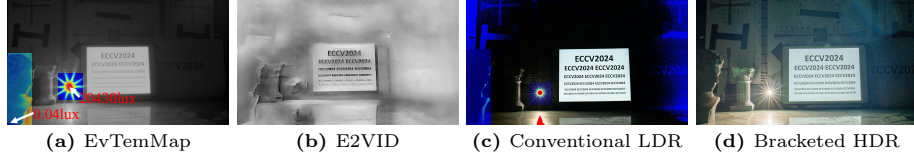


Fig. 5: Comparison in high dynamic range scenes. The brightest and darkest areas of (a) and (c) are rendered in pseudo-color. Overexposed areas and underexposed areas in (c) are marked in red and blue, respectively. EvTemMap presents exceptional HDR photography with minimal noise. Conversely, E2VID exhibits a loss of many texture features, while Conventional LDR suffers from inadequate information capture in both underexposed and overexposed regions.

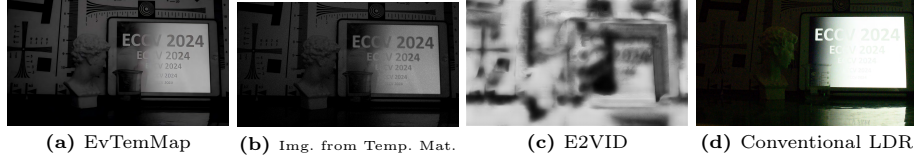


Fig. 6: Comparison in $<0.01\text{lux}$ environment. (a): EvTemMap shows clear textures and natural grayscale variations. (b): The raw intensity image from Temporal Matrix contains a lot of random noise. (c): E2VID loses a lot of texture. (d): Conventional LDR contains a lot of noise.

restore textures with high contrast. Compared with the single-exposure result in Fig. 5c, our method captures more details in both shadow and highlight grayscale variations than the conventional LDR image. The outcome from the conventional camera using bracketed HDR is depicted in Fig. 5d for reference.

4.5 Low-light Photography

We challenge the limits of EvTemMap in extremely dark environments (less than 0.01lux). In Fig. 6a, the texts on the screen and the resolution board line pairs in the dark background are clearly visible. Comparing Fig. 6a to Fig. 6b highlights the denoising capability of our image restoration model in low-light conditions. As shown in Fig. 6c, E2VID presents a worse result in extremely dark environments. It loses much of the texture detail in the scene, making it challenging to locate the target. Compared to Fig. 6d, we highlight the short “exposure time” in extremely dark conditions: the conventional LDR image requires an exposure time of up to 1.3 s to capture images in the low-light environment, whereas our AT-DVS only needs 78 ms for the capture of Temporal Matrix, which **reduces 94%** of the conventional exposure time. Note that low light condition also brings more noise to conventional LDR results.

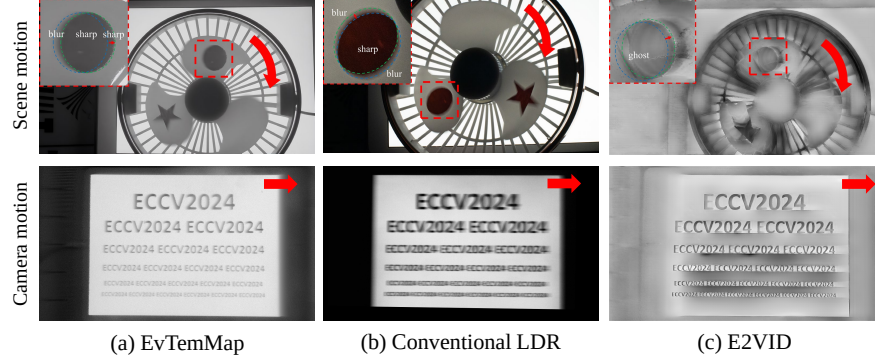


Fig. 7: Comparison in motion scenes.(a) Motion blur in EvTemMap occurs only on **one** side. (b) Motion blur in conventional image occurs on both sides. (c) Motion blur in E2VID exhibits “ghosting” effects.

4.6 Motion Photography

Because the time for capturing Temporal Matrix is not ignorable, EvTemMap also experiences motion blur in motion photography, but it presents different characteristics than motion blur in conventional and E2VID images.

Figure 7 shows results from EvTemMap, a traditional camera and E2VID. The top row shows an example of scene motion, where the fan rotates at a fixed speed. Comparing EvTemMap images with conventional LDR images, the motion blur in the EvTemMap image occurs on a single side of the circle during the motion from blue to green, whereas the motion blur in conventional image occurs on both sides of the circle. In another example of camera motion, AT-DVS and a conventional SLR camera are combined into a coaxial system with a beam splitter to capture a fixed screen while the system moves. With both cameras set to an exposure time of 40 ms, the bottom row demonstrates that EvTemMap produces less motion blur compared to the conventional LDR image. In Fig. 7c, while E2VID also reconstructs grayscale for motion scenes, it suffers from “ghosting” effects (*cf.* Sec. 2.1) and fails to handle static areas of the scene effectively (*e.g.*, the static background in the top row).

EvTemMap’s unique motion blur characteristics arise from its temporal mapping principle. Unlike conventional cameras that use integral imaging to capture the entire image during a fixed integration period, EvTemMap varies its grayscale acquisition speed based on the brightness of different areas in the image. Brighter areas are captured faster, resulting in less motion blur. Theoretically, EvTemMap’s maximum motion blur occurs in the darkest areas. Although these areas experience the same amount of motion as in conventional images for the same exposure time, the motion blur is only half as prominent because it occurs on only one side, which demonstrates the advantage of EvTemMap over conventional SLR in motion photography.

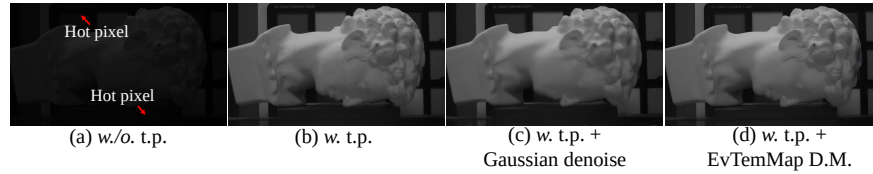


Fig. 8: Visual results of ablation experiment. t.p.: temporal pre-processing, D.M.: Degradation Model. (a): The presence of the hot pixels results in an overall darker image. (b): Hot pixel removed. (c): Conventional Gaussian noise reduction degrades image details. (d): With our proposed degradation model, the image still maintains faithful details.

4.7 Ablation Study

We analyze the effectiveness of the method mentioned in Sec. 3.3 through an ablation experiment, shown in Fig. 8. Before temporal pre-processing, there are several hot pixels in the original image, and these extremely bright white dots pull down the overall brightness of the picture during the grayscale normalization process, which also degrades the grayscale resolution. Based on the pre-calibrated hot pixel table, we perform temporal pre-processing to remove this fixed pattern noise. In Fig. 8b, the overall brightness of the picture is acceptable but presents a lot of random noise. With traditional Gaussian noise reduction method [24], (c) exhibits less noise but it also makes some details, such as text in the background, blurry. In (d), by applying a degradation model for both temporal domain and spatial domain, we not only remove a large amount of random noise but also retain most of the details of the texture in the image.

5 Conclusion

In this work, we have successfully achieved high-quality event-to-image conversion in static scenes using a stationary event camera. We design a unique event acquisition device, the Adjustable Transmittance Dynamic Vision Sensor (AT-DVS), to capture positive events generated during the gradual increase of transmittance, forming a Temporal Matrix. Subsequently, we map the Temporal Matrix to an adaptive dynamic range grayscale image using the proposed EvTemMap. We analyze the degradation factors for reconstruction and designed a temporal and spatial degradation model for training. Experimental results on collected TemMat dataset demonstrate the excellent photographic performance of EvTemMap in high dynamic range and low-light scenarios, along with enhanced performance in computer vision tasks.

Acknowledgements

This work was supported in part by the National Key R&D Program of China, under Grant No.2022YFF0705500, in part by the National Natural Science Foundation of China under Grant No.12174341.

References

1. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 126–135 (2017)
2. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing* **54**(11), 4311–4322 (2006)
3. Bao, Y., Sun, L., Ma, Y., Gu, D., Wang, K.: Improving fast auto-focus with event polarity. *Opt. Express* **31**(15), 24025–24044 (Jul 2023). <https://doi.org/10.1364/OE.489717>, <https://opg.optica.org/oe/abstract.cfm?URI=oe-31-15-24025>
4. Barua, S., Miyatani, Y., Veeraraghavan, A.: 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)
5. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240×180 130 dB 3 μ s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits* (2014)
6. Cadena, P.R.G., Qian, Y., Wang, C., Yang, M.: Spade-e2vid: Spatially-adaptive denormalization for event-based video reconstruction. *IEEE Transactions on Image Processing* **30**, 2488–2500 (2021)
7. Chaudhari, P., Schirmacher, F., Maier, A., Riess, C., Köhler, T.: Merging-isp: Multi-exposure high dynamic range image signal processing. In: DAGM German Conference on Pattern Recognition. pp. 328–342. Springer (2021)
8. Chen, S.Y., Chuang, Y.Y.: Deep exposure fusion with deghosting via homography estimation and attention learning. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1464–1468. IEEE (2020)
9. Cook, M., Gugelmann, L., Jug, F., Krautz, C., Steger, A.: Interacting maps for fast visual interpretation. In: The 2011 International Joint Conference on Neural Networks. pp. 770–776. IEEE (2011)
10. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2758–2766 (2015)
11. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)* **36**(6), 1–15 (2017)
12. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. *ACM Trans. Graph.* **36**(6), 177–1 (2017)
13. Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A.J., Conradt, J., Daniilidis, K., Scaramuzza, D.: Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022)

14. Gehrig, D., Scaramuzza, D.: Are high-resolution event cameras really needed? arXiv preprint arXiv:2203.14672 (2022)
15. Hu, Y., Liu, S.C., Delbruck, T.: v2e: From video frames to realistic dvs events. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1312–1321 (2021)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
17. Jiang, Z., Zhang, Y., Zou, D., Ren, J., Lv, J., Liu, Y.: Learning event-based motion deblurring. In: Proc. CVPR. pp. 3320–3329 (2020)
18. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
19. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* **36**(4), 144–1 (2017)
20. Kaur, H., Koundal, D., Kadyan, V.: Image fusion techniques: a survey. *Archives of computational methods in Engineering* **28**, 4425–4447 (2021)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
22. KS, G.R., Biswas, A., Patel, M.S., Prasad, B.P.: Deep multi-stage learning for hdr with large object motions. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 4714–4718. IEEE (2019)
23. Lee, S.H., Chung, H., Cho, N.I.: Exposure-structure blending network for high dynamic range imaging of dynamic scenes. *IEEE Access* **8**, 117428–117438 (2020)
24. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
25. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 136–144 (2017)
26. Lin, S., Zhang, J., Pan, J., Jiang, Z., Zou, D., Wang, Y., Chen, J., Ren, J.: Learning event-driven video deblurring and interpolation. In: ECCV (2020)
27. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17627–17638 (2023)
28. Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J., Liu, S.: Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 463–470 (2021)
29. Ma, K., Duanmu, Z., Wu, Q., Wang, Z., Yong, H., Li, H., Zhang, L.: Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing* **26**(2), 1004–1016 (2016)
30. Messikommer, N., Georgoulis, S., Gehrig, D., Tulyakov, S., Erbach, J., Bochicchio, A., Li, Y., Scaramuzza, D.: Multi-bracket high dynamic range imaging with event cameras. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 547–557 (2022)
31. Munda, G., Reinbacher, C., Pock, T.: Real-time intensity-image reconstruction for event cameras using manifold regularisation. *International Journal of Computer Vision* **126**, 1381–1393 (2018)

32. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing* **30**, 3885–3896 (2021)
33. Paredes-Vallés, F., de Croon, G.C.: Back to event basics: Self-supervised learning of image reconstruction for event cameras via photometric constancy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3446–3455 (2021)
34. Posch, C., Matolin, D., Wohlgenannt, R.: A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds. *IEEE Journal of Solid-State Circuits* **46**(1), 259–275 (2010)
35. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: Events-to-video: Bringing modern computer vision to event cameras. In: *CVPR* (2019)
36. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence* **43**(6), 1964–1980 (2019)
37. Scheerlinck, C., Barnes, N., Mahony, R.: Continuous-time intensity estimation using event cameras. In: *Asian Conference on Computer Vision*. pp. 308–324. Springer (2018)
38. Scheerlinck, C., Rebecq, H., Gehrig, D., Barnes, N., Mahony, R., Scaramuzza, D.: Fast image reconstruction with an event camera. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 156–163 (2020)
39. Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., Mahony, R.: Reducing the sim-to-real gap for event cameras. In: *ECCV* (2020)
40. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: *European Conference on Computer Vision*. pp. 412–428. Springer (2022)
41. Sun, L., Sakaridis, C., Liang, J., Sun, P., Cao, J., Zhang, K., Jiang, Q., Wang, K., Van Gool, L.: Event-based frame interpolation with ad-hoc deblurring. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 18043–18052 (2023)
42. Timofte, R., Agustsson, E., Van Gool, L., Yang, M.H., Zhang, L.: Ntire 2017 challenge on single image super-resolution: Methods and results. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. pp. 114–125 (2017)
43. Tulyakov, S., Gehrig, D., Georgoulis, S., Erbach, J., Gehrig, M., Li, Y., Scaramuzza, D.: Time lens: Event-based video frame interpolation. In: *CVPR* (2021)
44. Wang, L., Yoon, K.J.: Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE transactions on pattern analysis and machine intelligence* **44**(12), 8874–8895 (2021)
45. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 1905–1914 (2021)
46. Wang, Z., Ng, Y., van Goor, P., Mahony, R.E.: Event camera calibration of per-pixel biased contrast threshold. *ArXiv abs/2012.09378* (2020), <https://api.semanticscholar.org/CorpusID:229298059>
47. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 117–132 (2018)

- 48. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019)
- 49. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 41–50. IEEE (2019)
- 50. Zhang, K., Liang, J., Van Gool, L., Timofte, R.: Designing a practical degradation model for deep blind image super-resolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4791–4800 (2021)
- 51. Zhang, S., Zhang, Y., Jiang, Z., Zou, D., Ren, J., Zhou, B.: Learning to see in the dark with events. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 666–682. Springer (2020)
- 52. Zhang, X., Yu, L.: Unifying motion deblurring and frame interpolation with events. In: Proc. CVPR. pp. 17765–17774 (2022)