Towards HDR and HFR Video from Rolling-Mixed-Bit Spikings

Yakun Chang[3,4#†] Yeliduosi Xiaokaiti[1,2#] Yujia Liu[1,2] Bin Fan[5]
Zhaojun Huang[1,2] Tiejun Huang[1,2] Boxin Shi[1,2*]
[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] Institute of Information Science, School of Computer Science, Beijing Jiaotong University
[4] Beijing Key Laboratory of Advanced Information Science and Network Technology
[5] National Key Lab of General AI, School of Intelligence Science and Technology, Peking University
ykchang@bjtu.edu.cn, {yujia_liu, binfan, tjhuang, shiboxin}@pku.edu.cn

Abstract

The spiking cameras offer the benefits of high dynamic range (HDR), high temporal resolution, and low data redundancy. However, reconstructing HDR videos in high-speed conditions using single-bit spikings presents challenges due to the limited bit depth. Increasing the bit depth of the spikings is advantageous for boosting HDR performance, but the readout efficiency will be decreased, which is unfavorable for achieving a high frame rate (HFR) video. To address these challenges, we propose a readout mechanism to obtain rolling-mixed-bit (RMB) spikings, which involves interleaving multi-bit spikings within the single-bit spikings in a rolling manner, thereby combining the characteristics of high bit depth and efficient readout. Furthermore, we introduce RMB-Net for reconstructing HDR and HFR videos. RMB-Net comprises a cross-bit attention block for fusing mixed-bit spikings and a cross-time attention block for achieving temporal fusion. Extensive experiments conducted on synthetic and real-synthetic data demonstrate the superiority of our method. For instance, pure 3-bit spikings result in 3 times of data volume, whereas our method achieves comparable performance with less than 2% increase in data volume.

1. Introduction

Real-world scenes possess a significantly wider dynamic range that exceeds the capability of conventional sensors. Typical high dynamic range (HDR) video reconstruction methods [3, 23, 24, 53] with conventional sensors encode exposure times to capture images with alternating exposures. And by fusing the low dynamic range (LDR) images taken under different exposures, the pitfalls of underexposure and overexposure are alleviated. This kind of approach has a dilemma between the frame rate and exposure time [20], *i.e.*, long exposure restricts the improvement of frame rate [2], which makes it challenging to capture high frame rate (HFR) videos with conventional sensors in high-speed scenes.

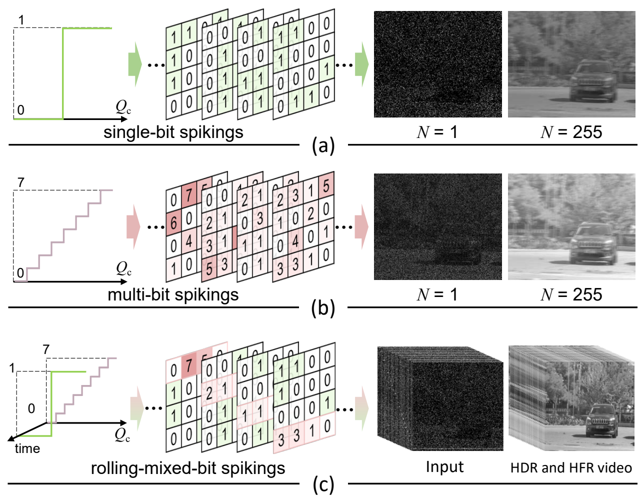

Figure 1. The HDR performance of a spiking camera is closely tied to the bit depth of the spikings. (a) From left to right: Single-bit quantization (Q_c is the accumulated photon electrons), diagram of single-bit spiking planes, and reconstructed image by accumulating N spiking planes. Increasing N to 255 significantly boosts HDR performance. (b) HDR can be boosted by reading out multi-bit spikings. However, multi-bit spikings decrease the readout efficiency, which is not conducive to obtaining HFR videos. (c) The proposed RMB spikings with time-varying quantization. We further reconstruct HDR and HFR videos from RMB spikings.

Recent advancements in the field of HDR and HFR photography have benefited from the integration of *neuromorphic sensors* such as event cameras [6, 29, 30] and spiking cameras [2, 21]. These sensors offer appealing characteris-

[#] Equal contribution. [*] Corresponding author.
[†] Majority of this work was done at Peking University.
Project page: https://github.com/yongqiye00/RMB-Net

tics such as high dynamic range (>90dB), high temporal resolution (μs), and low redundancy of single-bit data. In contrast to difference-based event cameras that solely detect changes in brightness [30], spiking cameras are more adept at reconstructing fine texture details as they continuously accumulate photon electric signals [60]. However, single-bit spikings are less compatible with the human visual system, necessitating reconstruction algorithms to convert them into video frames [57, 58, 61].

As illustrated in Fig. 1 (a), when electrons generated by accumulated photons in a pixel reach a predetermined threshold, a binary spiking of 1 is read out and the pixel is reset to 0. The video frame reconstruction can be easily achieved by accumulating a number N of spiking planes. In Fig. 1 (a), a smaller N leads to lower bit depth of the reconstructed image ($N = 1$ for 1 bit), while a larger N results in a higher bit depth ($N = 255$ for 8 bit). In static and low-speed conditions, one could set a large N to reconstruct an HDR image [19]. This image can then be used to compensate for missing details in LDR images captured by conventional sensors. However, these methods [19, 55] suffer from limitations in high-speed conditions, as longer accumulation time introduces more severe motion blur. An alternative approach [2] is to build a hybrid system consisting of a spiking camera and an alternating exposure camera, which enables both HDR and HFR video reconstruction. Unfortunately, such hybrid systems require cumbersome synchronization and optical alignment, and the space occupied by the beam splitter poses challenges in building a compact device.

As a result, it would be advantageous if we could simultaneously achieve HDR and HFR video reconstruction with a standalone spiking camera. To achieve higher dynamic range without introducing motion blur, as illustrated in Fig. 1 (b), it is theoretically feasible to increase the bit depth of spiking signals. This goal can be achieved along two paths: modifying the analog-to-digital converter (ADC) to read out multi-bit spikings, *e.g.*, the cyclic ADC technology [26], or accumulating more spikings within a limited time interval. Nevertheless, higher bit depth inevitably brings increased transmission pressure, which may reduce the readout efficiency of spikings. Hence, the reconstruction of video frames from spikings faces a trade-off between HDR and HFR.

To balance the trade-off between HDR and HFR, a reasonable solution is to alternately read out mixed-bit data stream that comprises both multi-bit and single-bit spikings. From the perspective of hardware manufactures and practical deployment, it is desirable that the design principles of the readout mechanism for mixed-bit spikings incorporate advantages such as simplified implementation, low bandwidth requirements, and fast sampling rates. Compared to the full-frame readout mechanism, the rolling readout mechanism can naturally take the above advantages [1, 11, 50, 52] and thus has been widely employed in HDR [17, 31] and HFR [8–10, 12] video reconstruction. Taking inspiration from these approaches, as shown in Fig. 1 (c), we propose to design the rolling-mixed-bit (RMB) spikings to take the best of both worlds: high bit depth and efficient readout. Further, we propose an HDR and HFR video reconstruction framework, termed *RMB-Net*, that leverages a *cross-bit attention* and a *cross-time attention* block, to effectively reconstruct an HDR and HFR video from RMB spikings.

Through experiments conducted on both synthetic and real-synthetic data, our RMB-Net demonstrates comparable performance to using pure multi-bit spikings while maintaining the advantage of fewer data redundancy. For instance, compared to pure single-bit data, in an RMB spiking plane with a row count of 500 and containing 4 rows of 3-bit spikings, the increase in data volume is less than 2%, whereas pure 3-bit spikings result in three times of data volume. Our main contributions can be summarized as:

- discovering a promising solution to boost the HDR performance through a comprehensive analysis of the relationship between dynamic range and bit depth of spikings;
- designing a novel RMB spiking mechanism that effectively balances both bit depth and readout efficiency;
- proposing an effective RMB-Net to reconstruct HDR and HFR videos from the RMB spikings.

2. Related Work

HDR with conventional sensors. Conventional sensors often fail to capture HDR ambient light with a single exposure. The methods for HDR image reconstruction from these single LDR images [4, 7, 27, 33] cannot restore the missing details in under-exposure and over-exposure regions. One solution for HDR reconstruction is to fuse a set of LDR images with different exposures [5, 41]. This approach often leads to ghosting artifacts in dynamic scenes. To address this issue and enhance the sharpness of HDR images, techniques such as image alignment [35, 46] and deep learning [22, 54] are employed. Lee and Song [28] utilize motion information from high frame rate sequences to improve HDR image synthesis and minimize ghosting artifacts. Merging sequences of alternating-exposure frames is feasible to reconstruct HDR videos with frame rates ranging from 20 to 60FPS [16, 23–25, 37, 38]. Chen *et al*. [3] propose a coarse-to-fine network that performs alignment and fusion sequentially in both image and feature space.

HDR/HFR with unconventional sensors. Numerous unconventional sensors have been investigated for the purpose of capturing HDR videos, such as scanline exposure [18], per-pixel exposure [44], or multiple sensors [40, 49]. Many unconventional sensors, including event cameras [30], spiking cameras [21], single photon avalanche diodes (SPAD) [45], and quanta image sensors (QIS) [14], have emerged with the capability to capture HDR signals even in high-speed conditions. The QIS and spiking cameras have similar imaging models and primarily output binary sequences. But they differ in several characteristics, *e.g.*, the QIS16TS camera [34] features very small pixels ($1.1\mu m \times 1.1\mu m$) and a relatively low frame rate (62FPS with a frame size of 1024×1024), whereas spiking cameras have larger pixels ($17\mu m \times 17\mu m$) and higher frame rate (20,000FPS with a frame size of 1000×1000) [13, 21]. HDR images are reconstructed by [15, 36] in the context of QIS cameras. Han *et al*. [19] and Yang *et al*. [55] leverage the intensity map reconstructed from events or spiking signals to compensate for LDR images. Liu *et al*. [32] present the single-photon camera guided HDR imaging. Messikommer *et al*. [42] and Shaw *et al*. [47] explore the motion information within events to promote image alignment of alternating exposure images. Chang *et al*. [2] build a hybrid spike-RGB camera system to recover 1000FPS HDR video. However, the hybrid camera requires synchronization and optical alignment, and the space taken up by the beam splitter presents challenges in constructing a compact device.

3. Dynamic Range of Spikings

In this paper, we investigate the reconstruction of HDR videos from mixed-bit spikings, allowing for the simultaneous achievement of HDR and HFR. Toward this goal, we begin by providing a concise description of the emission model for both single-bit and multi-bit spikings in Sec. 3.1. Subsequently, in Sec. 3.2, we analyze the correlation between HDR performance and bit depth of the spikings.

3.1. Spiking emission model

Single-bit spikings. For each pixel in the spiking camera, the electrons generated by photons are continuously accumulated as long as the electrons does not reach the threshold Q_{th}. Simultaneously, the readout circuit samples the pixel value at a fixed frequency and the value of 0 is read out at each readout point. Once the accumulated electrons reach Q_{th}, a signal of 1 is read out and the electrons of the pixel is reset to 0. We denote the accumulated electrons at a read out point t as $Q_{\mathrm{c}}(t)$, the single-bit spiking $S(t)$ at t is

$$S(t) = \begin{cases} 1, & Q_{\mathrm{c}}(t) \geq Q_{\mathrm{th}}, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $Q_{\mathrm{c}}(t)$ consists of three components: the accumulated electrons $Q_{\mathrm{a}}(t)$ in the previous accumulation interval, the photo-generated electrons $Q_{\mathrm{p}}(t)$, and the dark electrons $Q_{\mathrm{d}}(t)$. That is, $Q_{\mathrm{c}}(t) = Q_{\mathrm{p}}(t) + Q_{\mathrm{a}}(t) + Q_{\mathrm{d}}(t)$. Q_{th} is also affected by the deviation of the capacitor capacitance C^S, the voltage deviation V^S when resetting the voltage, and the voltage deviation $V^{T0}(t)$ caused by temperature[1].

How to increase bit depth. Single-bit (also called 1-bit) spikings are limited in representing textures, hence, increasing the bit depth is necessary to encode richer texture information. There are two viable solutions to increase bit depth and obtain finer textures. The first solution employs higher-level quantization of the detected electrons, while the second solution focuses on accumulating more spikings in the temporal domain. For the first solution, we define the bit depth of a spiking as B, and the quantization level as L, where L is equal to $2^B - 1$. The formulation for generating multi-bit spiking, denoted as $S^L(t)$ is given by

$$S^L(t) = \begin{cases} L, & Q_{\mathrm{c}}(t) \geq Q_{\mathrm{th}}, \\ H, & Q_{\mathrm{c}}(t) \in [\frac{HQ_{\mathrm{th}}}{L}, \frac{(H+1)Q_{\mathrm{th}}}{L}), \end{cases} \tag{2}$$

where H is an integer within the range of $[0, L-1]$. The accumulator is reset to 0 when $0 < S^L(t) \leq L$. For the second solution, we denote τ as the time interval between two adjacent spikings, and accumulate a sequence of N consecutive readout spikings:

$$S_N(t) = \sum_{n=1}^{N} S(t + n\tau). \tag{3}$$

Then, by combining Eqns. (2) and (3), the bit depth can be jointly increased as $S_N^L(t) = \sum_{n=1}^{N} S^L(t + n\tau)$.

3.2. HDR with multi-bit spikings

SNR vs. dynamic range of spikings. The dynamic range of camera sensors is typically defined as the range of the exposure where the signal-to-noise ratio (SNR) surpasses a threshold of 1. Let $S^L(t)$ refer to a spiking with an arbitrary bit depth at a pixel, $\mu = \mathbb{E}(S^L(t))$ is the expectation, and $\sigma^2 = \mathrm{Var}(S^L(t))$ is the variance of the spiking signal. Similar with the definition in [15], the SNR of spiking camera can be defined as $\mathrm{SNR}_\lambda(S^L(t)) = \frac{\lambda}{\sigma} \frac{d\mu}{d\lambda}$, where λ denotes the exposure, *i.e.*, the average number of photons. The expectation of $S_N^L(t)$ is linearly related to $\mathbb{E}(S^L(t))$, *i.e.*, $\mathbb{E}(S_N^L(t)) = N \cdot \mathbb{E}(S^L(t))$. The variance of $S_N^L(t)$ is $N \cdot \mathrm{Var}(S^L(t))$ when the spiking signals between intervals are independent of each other. For spiking cameras, information inheritance[2] occurs between intervals due to continuous electron accumulation. This leads to a reduction in information loss and variance. Thus, we have $\mathrm{Var}(S_N^L(t)) \leq N \cdot \mathrm{Var}(S^L(t))$. And consequently, $\mathrm{SNR}_\lambda(S_N^L(t)) \geq \sqrt{N} \cdot \mathrm{SNR}_\lambda(S^L(t))$. Since the expectation and variance of spiking signals are the same for each accumulation interval t, we omit t in the subsequent formulations. $\mathbb{E}(S^L)$ and $\mathrm{Var}(S^L)$ can be formulated as

$$\mathbb{E}(S^L) = \sum_{H=0}^{L} H \, \mathbb{P}(S^L = H), \tag{4}$$

$$\mathrm{Var}(S^L) = \sum_{H=0}^{L} H^2 \, \mathbb{P}(S^L = H) - \mathbb{E}^2(S^L). \tag{5}$$

Here, $\mathbb{P}(S^L = H)$ denotes the probability of the read out value being H, and it is formulated as follows:

$$\mathbb{P}(S^L = H) = \sum_{Q'=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \frac{P_{Q'} P_k \sum_{n=1}^{+\infty} P_n^H}{\sum_{h=1}^{L} \sum_{n=1}^{+\infty} n P_n^h}, \tag{6}$$

[1]Details of the Q_{c} and Q_{th} are available in the supplementary material.

[2]Explanations are available in the supplementary material.

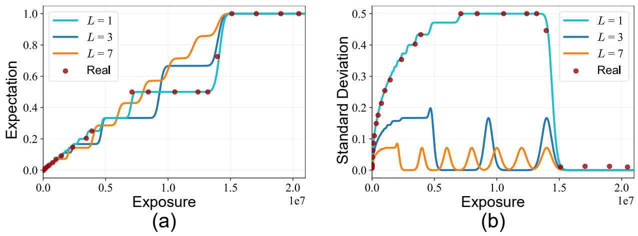

Figure 2. We show the curves of expectation and standard deviation of normalized S_N^L, where N is set to 1. (a) The expectation exhibits a stepwise increase when the bit depth is low. The real-captured spikings are marked by red points. In each flat region of the step, single-bit spikings cannot differentiate between different exposures, leading to the loss of some texture information. As the quantization level L increases, the range of flat regions in each step decreases, allowing for better preservation of textures. (b) Higher-level quantization leads to a smaller standard deviation.

where $P_{Q'} = \mathbb{P}(Q_{\text{th}} = Q')$ represents the probability of the threshold electrons Q_{th} being Q', $P_k = \mathbb{P}(Q_{\text{d}} = k)$ denotes the probability of dark electrons being k [3]. P_n^H represents the probability that, when initiating the accumulation with $Q_a = 0$, no spikings are emitted within the time interval $[0, (n-1)\tau]$, and the quantization level within the range $((n-1)\tau, n\tau]$ is H. P_n^H can be expressed as follows:

$$P_n^H = \sum_{u \in \mathcal{N}_u} \mathbb{P}(Q_{\text{p}} = u | n_{\text{p}} \alpha \lambda) \sum_{v \in \mathcal{N}_v} \mathbb{P}(Q_{\text{p}} = v | \alpha \lambda), \quad (7)$$

where $n_{\text{p}} = n - 1, \mathcal{N}_u = [-n_{\text{p}}k, \lfloor Q'/L \rfloor - n_{\text{p}}k) \cap \mathbb{N}$, and $\mathcal{N}_v = [\lceil Q'H/L \rceil - u - nk, \lceil Q'(H+a)/L \rceil - u - nk) \cap \mathbb{N}$, $a = 1$ when $H < L$ and $a \to +\infty$ when $H = L$. Note that the photons arrived at a pixel can be modeled as a Poisson process. For any $z \in \mathbb{N}$, the probability for photo-generated electrons Q_{p} being z in n intervals is:

$$\mathbb{P}(Q_{\text{p}} = z | n\alpha\lambda) = (n\alpha\lambda)^z \exp(-n\alpha\lambda)/(z!), \quad (8)$$

where α denotes the photoelectric conversion rate.

Simulation and validation. By jointly solving Eqns. (4)-(8), the mapping from λ to $\mathbb{E}(S^L)$ and $\text{Var}(S^L)$ can be obtained correspondingly. Denote the normalized spiking which equals to S^L/L as $\boldsymbol{S^L}$. The curves of $\mathbb{E}(\boldsymbol{S^L})$ and $\sqrt{\text{Var}(\boldsymbol{S^L})}$ with respect to exposure are shown in Fig. 2. To validate our analysis, we conducted measurements on the response of real-captured spikings[4] to exposure and fit the theoretical curves to the actual responses. In Fig. 2 (a), when $L = 1$, $\mathbb{E}(\boldsymbol{S^L})$ demonstrates a stepwise increasing curve as λ increases. As L gradually increases, the stepwise pattern becomes more refined until the relationship between $\mathbb{E}(\boldsymbol{S^L})$ and λ approaches a smoothly linear response. In Fig. 2 (b), $\sqrt{\text{Var}(\boldsymbol{S^L})}$ shows a trend of initially increasing and then decreasing with the increase of λ. As L increases, $\sqrt{\text{Var}(\boldsymbol{S^L})}$ tends to decrease. $\mathbb{E}(S^L)$ and $\sqrt{\text{Var}(S^L)}$ collectively determine the SNR_λ, as shown in Fig. 3, thereby influencing the dynamic range of the spiking camera.

[3]Explanations of Eqn. (6) are available in the supplementary material.
[4]Details of this experiment are available in the supplementary material.

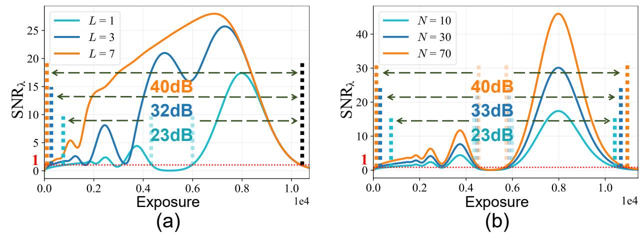

Figure 3. The dynamic range of the spikings ($S_N^L(t)$) is jointly determined by L and N. We illustrate the mapping curve between the exposure λ and signal-to-noise ratio SNR_λ. (a) We set $N = 10$ and L is varied as 1, 3, and 7. (b) We set $L = 1$ and N is varying as 10, 30, and 70. Both increasing L and N boost the extension of the dynamic range. Here, the unit for exposure is the number of photons per square micron per second (photon$/(\mu m^2 \cdot s)$).

Table 1. Dynamic range for varying bit depth and number of intervals N. Here, the unit for dynamic range is decibels (dB).

L \ N	1	10	30	70	1000	10000	20000
1 (1-bit)	9.07	22.70	32.73	40.33	63.84	84.25	90.38
3 (2-bit)	13.91	32.32	42.35	49.95	73.48	93.80	99.91
7 (3-bit)	21.79	39.79	49.83	64.09	80.83	101.70	107.71

Theoretical bound of dynamic range. In Fig. 3, we illustrate the dynamic range of the spiking signal ($S_N^L(t)$) with respect to L and N. In Fig. 3 (a), where N is set to a fixed value of 10, the dynamic range is boosted from 23dB to 40dB when L is increased from 1 to 7. In Fig. 3 (b), where L is fixed at 1, increasing N from 10 to 70 also boosts the dynamic range. To provide a more detailed overview of the relationship between dynamic range and $\{L, N\}$, we list the detailed numbers in Table 1. When N takes a very large value, *i.e.*, 20,000, the theoretical bound of single-bit spikings is around 90dB, while the theoretical bound of 3-bit spikings is around 107dB.

4. RMB-Net for HDR and HFR Video

In high-speed scenes, given the constraint of limited bandwidth, increasing L from 1 to 7 results in 3 times of data volume, which poses a disadvantage for HFR video reconstruction. Therefore, further consideration is required for the theoretical bound of HDR. To balance the bit depth and readout efficiency, we design a rolling-mixed-bit (RMB) readout mechanism in Sec. 4.1. This approach yields a significant reduction of data volume compared to pure multi-bit spikings, while still retaining the capability to reconstruct HDR and HFR videos. In Sec. 4.2, we propose an effective RMB-Net to reconstruct HDR and HFR videos from RMB spikings. The RMB-Net utilizes a cross-bit attention block to merge the single-bit signals and multi-bit spikings. Meanwhile, as there are high-speed motions, naive accumulation with a large N yields motion blur. And it becomes increasingly challenging with larger values of N. Intuitively, the accumulation of N spikings is better determined by a weighting scheme that takes into account the temporal information. Thus, RMB-Net employs a cross-time attention block that

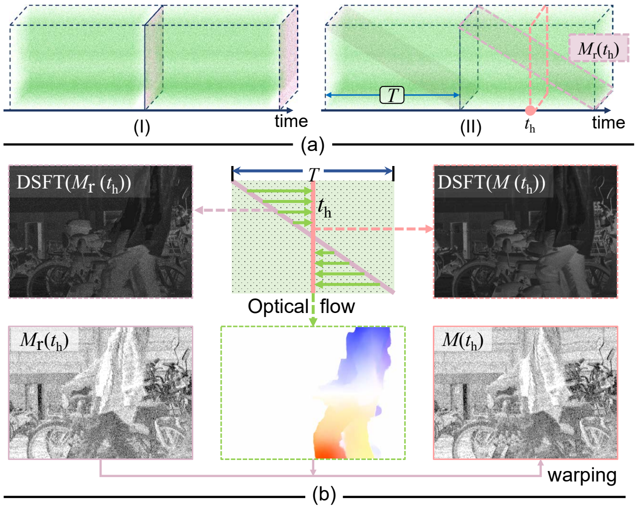

Figure 4. (a) Two conceptual designs for mixed-bit spikings are proposed: (I) Reading out full-frame multi-bit spiking planes intermittently; (II) sequentially mixing multi-bit spikings row by row within single-bit spikings. T is the scanning cycle of multi-bit spikings, t_{h} is the timestamp that the scan precisely undergoes a half cycle, $M_{\mathrm{r}}(t_{\mathrm{h}})$ is the corresponding spiking plane by accumulating the multi-bit spikings read out within T. (b) We rectify the shape distortion of $M_{\mathrm{r}}(t_{\mathrm{h}})$ to obtain $M(t_{\mathrm{h}})$. The warping operation is performed with the optical flow estimated from single-bit spikings.

learns weight masks cross a sequence of spiking frames to facilitate the merging process.

4.1. Preprocessing for RMB spikings

We now describe two conceptual designs for the mixed-bit spikings. The first design shown on the left side of Fig. 4 (a) reads out full-frame multi-bit spiking planes intermittently. These full-frame multi-bit spikings instantaneously impose high transmission pressure. The RMB mechanism shown in the right side of Fig. 4 (a) employs a time-varying readout mechanism to evenly distribute the transmission pressure generated from multi-bit spikings. In this work, the size of full frame is 500×500, RMB mechanism reads out 4 rows of three-bit spikings at each readout point. Thus, compared to pure single-bit spikings, the increase in data volume is $((496 + 3 \times 4)/500 - 1) \times 100\% = 1.6\%$.

Upsampling to dense multi-bit spikings. As illustrated in Fig. 4 (b), during scanning multi-bit spikings over time, motions often cause shape distortion in $M_{\mathrm{r}}(t_{\mathrm{h}})$, similar to the jelly effects of a rolling shutter sensor. Rectifying the shape distortion is solvable since the full-time motion information for each pixel is captured by single-bit spikings. We denote the target multi-bit spiking plane at t_h as $M(t_{\mathrm{h}})$. As shown in Fig. 4 (b), we firstly estimate the differential of spike firing time (DSFT) [59] corresponding to $M_{\mathrm{r}}(t_{\mathrm{h}})$ and $M(t_{\mathrm{h}})$. Then, we estimate the optical flow between DSFT$(M_{\mathrm{r}}(t_{\mathrm{h}}))$ and DSFT$(M(t_{\mathrm{h}}))$ using Spike2Flow [59]. This is followed by a warping operation to obtain $M(t_{\mathrm{h}})$. Similarly, we can estimate the bidirectional optical flows between each pair of $M(t_{\mathrm{h}})$ and $M(t_{\mathrm{h}} + T)$, enabling us to upsample dense multi-bit spikings with the linear interpolation.

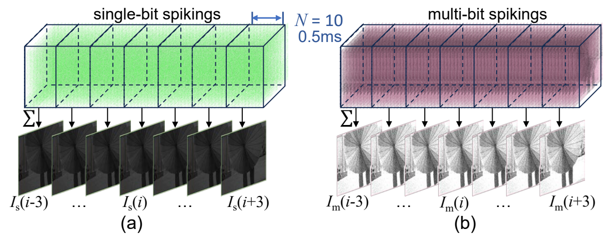

Figure 5. (a) and (b) show the preliminary reconstruction from single-bit and upsampled multi-bit spikings, respectively. $\sum$ denotes the temporal accumulation in this figure.

Preliminary reconstruction. To reconstruct videos with a frame rate of 2,000FPS from 20,000Hz spikings, as illustrated in Fig. 5, we initially split the spiking data with a fixed interval of 0.5ms ($N = 10$) in the time domain. Then, we accumulate the spikings in each interval to preliminarily reconstruct spiking frames. The spiking frames accumulated from single-bit and upsampled multi-bit spikings are denoted as $\mathbb{I}_{\mathrm{s}} = \{I_{\mathrm{s}}(i) | i = 0, 1, ...K\}$ and $\mathbb{I}_{\mathrm{m}} = \{I_{\mathrm{m}}(i) | i = 0, 1, ...K\}$, where i is the index of the spiking frames, and K is the total number of frames in a sequence.

Input of RMB-Net. For the current step i, since the accumulation interval N for preliminary reconstruction is 10, the theoretical dynamic ranges of $I_{\mathrm{s}}(i)$ and $I_{\mathrm{m}}(i)$ are limited to 22.7dB and 39.79dB, respectively (see Table 1). Larger N has been proven to be effective for boosting HDR performance. Hence, for the current reference image $I_{\mathrm{s}}(i)$, we select a bunch of frames $\mathbf{I}_{\mathrm{s}}(i) = \{I_{\mathrm{s}}(j) | j \in \mathcal{N}\}$ as the input of RMB-Net, where $\mathcal{N} = [i - w, i + w]$ and w is the number of subsequent or previous frames. Hence, the accumulation interval N for each step is $(2w + 1) \times 10$. To balance the trade-off between HDR performance and motions in high-speed scenes, we set the value of w to 3, which is equivalent to $N = 70$ and is well suited for video reconstruction at 2,000FPS. Simultaneously, in correspondence with $\mathbf{I}_{\mathrm{s}}(i)$, we also select a bunch of upsampled multi-bit spiking frames $\mathbf{I}_{\mathrm{m}}(i) = \{I_{\mathrm{m}}(j) | j \in \mathcal{N}\}$ to further boost the HDR reconstruction process. In this condition, assuming that multi-bit spikings can be ideally upsampled from the rolling multi-bit spikings, the theoretical limit of HDR video is 64.09dB.

4.2. Architecture of RMB-Net

As shown in Fig. 6, RMB-Net firstly tackles the issue of spatial misalignment with the optical flows estimated from single-bit spikings. Next, two encoders are utilized to extract multi-scale features from the two-stream input $\mathbf{I}_{\mathrm{s}}(i)$ and $\mathbf{I}_{\mathrm{m}}(i)$ in parallel, and the corresponding features are denoted as $\mathbf{F}_{\mathrm{s}}(i) = \{F_{\mathrm{s}}(j) | j \in \mathcal{N}\}$ and $\mathbf{F}_{\mathrm{m}}(i) = \{F_{\mathrm{m}}(j) | j \in \mathcal{N}\}$. To reconstruct a single frame at step i by merging $\mathbf{F}_{\mathrm{s}}(i)$ and $\mathbf{F}_{\mathrm{m}}(i)$, the design of a two-stage fusion process is reasonable: We firstly accomplish the fusion of single-bit and multi-bit features, and then deal with the fusion of the tempo-

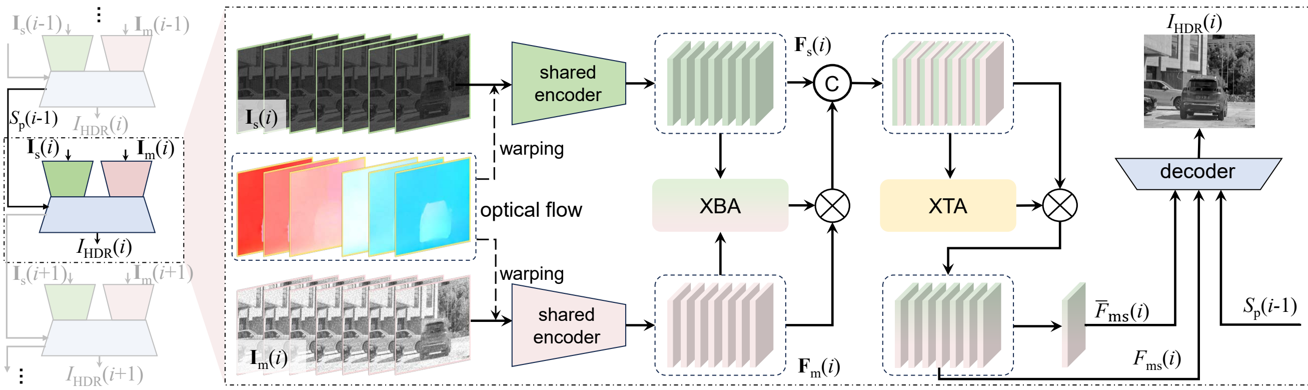

Figure 6. RMB-Net reconstructs HDR and HFR videos in a step-wise manner. The input at step i consists of two bunches of single-bit frames $\mathbf{I}_s(i)$ and multi-bit frames $\mathbf{I}_m(i)$. We align the images to the reference timestamp i using Spike2Flow [59]. $\mathbf{F}_s(i)$, $\mathbf{F}_m(i)$, and $\bar{F}_{sm}(i)$ are internal features. The cross-bit attention (XBA, details in Fig. 7) and cross-time attention (XTA) are designed to facilitate the merging process. $S_p(i-1)$ is the previous states of step $i-1$. $I_{HDR}(i)$ is the output image.

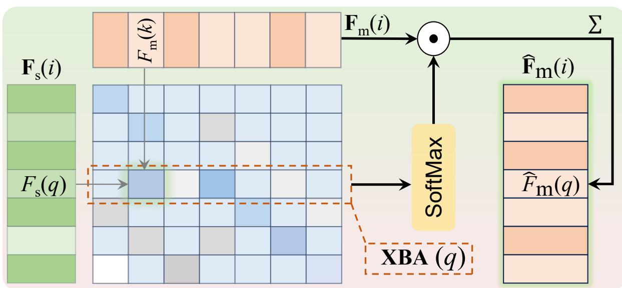

Figure 7. The cross-bit attention learns weight masks from each pair of $F_s(q)$ and $F_m(k)$. The weight masks are normalized with SoftMax and then applied to $\mathbf{F}_{ms}(i)$.

ral sequence within the bunch $\mathcal{N}$. In the first stage, RMB-Net achieves the fusion by merging multi-bit and single-bit features through a cross-bit attention (XBA) block. The output of XBA is denoted as $\hat{\mathbf{F}}_m(i)$. Subsequently, RMB-Net concatenates $\hat{\mathbf{F}}_m(i)$ and $\mathbf{F}_s(i)$ to obtain the mixed-bit feature collection. In the second stage, RMB-Net achieves fusion by leveraging temporal fusion with the cross-time attention block (XTA). The output of it is denoted as $\mathbf{F}_{ms}(i)$. Then, an average operation is applied to $\mathbf{F}_{ms}(i)$ in temporal domain to obtain $\bar{F}_{ms}(i)$. Finally, we use a decoder to reversely map $\bar{F}_{ms}(i)$ to the output frame $I_{HDR}(i)$, and we add residual links from the reference feature $F_{ms}(i)$ to the decoder. In order to output flicker-free video frames, we add three ConvLSTM layers [48] to feed previous states $S_p(i-1)$ forward in the temporal domain.

Fusing with cross-bit attention. Since motions are continuous during a bunch of frames, there are spatial correlations between $F_s(q)$ and $F_m(k)$ for a given query index q and any key index $k \in \mathcal{N}$. To measure these correlations, the cross-bit attention block that learns weighted masks between all the pairs of $F_s(q)$ and $F_m(k)$ is designed as illustrated in Fig. 7. We denote the collection of weighted masks for index q as $\mathbf{XBA}(q)$:

$$\mathbf{XBA}(q) = \{\mathcal{F}(\mathcal{C}(F_s(q), F_m(k)))|q \in \mathcal{N}\}, \qquad (9)$$

where $\mathbf{XBA}(q)$ comprises $2w$+1 attention masks, $\mathcal{F}(\cdot)$ denotes the projection function composed of multiple convolutional layers, $\mathcal{C}(\cdot)$ signifies the concatenation operation. Subsequently, we perform normalization with SoftMax:

$$\mathbf{XBA}(q) \leftarrow \mathrm{SoftMax}(\mathbf{XBA}(q)). \qquad (10)$$

The weighted masks are then applied to $\mathbf{F}_m(i)$ in order to obtain the refined multi-bit $\hat{F}_m(q)$:

$$\hat{F}_m(q) = \sum_{k \in \mathcal{N}} \mathbf{XBA}(q, k) \odot F_m(k), \qquad (11)$$

where $\odot$ is element-wise multiplication. Next, by concatenating each $\hat{F}_m(q)$ to $F_s(q)$ and reducing the channel of features with 1×1 convolutional layers, we obtain the mixed-bit feature collection: $\mathbf{F}_{ms}(i) = \{F_{ms}(q)|q \in \mathcal{N}\}$.

Fusing with cross-time attention. The cross-time attention block is proposed to deal with the temporal fusion of the feature collection $\mathbf{F}_{ms}(i)$. Similar to cross-bit attention, cross-time attention is performed to measure the inter-correlations between all pairs of $F_{ms}(q)$ and $F_{ms}(k)$. The attention mask for each $F_{ms}(q)$ is:

$$\mathbf{XTA}(q) = \{\mathcal{F}(\mathcal{C}(F_{ms}(q), F_{ms}(k))) + B(i)|q \in \mathcal{N}\}, \qquad (12)$$

where $B(i)$ is the bias that equals the temporal average of $\mathbf{F}_{ms}(i)$, $\mathcal{F}(\cdot)$ and $\mathcal{C}(\cdot)$ are the same as Eqn. (9). Also, we adopt SoftMax to normalize the weight masks like Eqn. (10). $\mathbf{XTA}(q)$ is then applied to $\mathbf{F}_{ms}$ in the same manner as Eqn. (11), and the output feature is denoted as $\hat{F}_{ms}(q)$. Finally, we merge all the $\hat{F}_{ms}(q)$ for the current step i as: $\bar{F}_{ms}(i) = \frac{1}{2w+1}\sum_{q\in\mathcal{N}} \hat{F}_{ms}(q)$.

4.3. Implementation details

Data preparation. RMB-Net incorporates Spike2Flow [59] for optical flow estimation. Since Spike2Flow has not been trained for HDR scenes, we finetune it with our synthetic dataset. The dataset utilized to train RMB-Net includes three components: RMB spikings, ground truth optical flows, and ground truth HDR video frames. Following Chang *et al.* [2] and Zhao *et al.* [59], we synthesize HDR and HFR videos

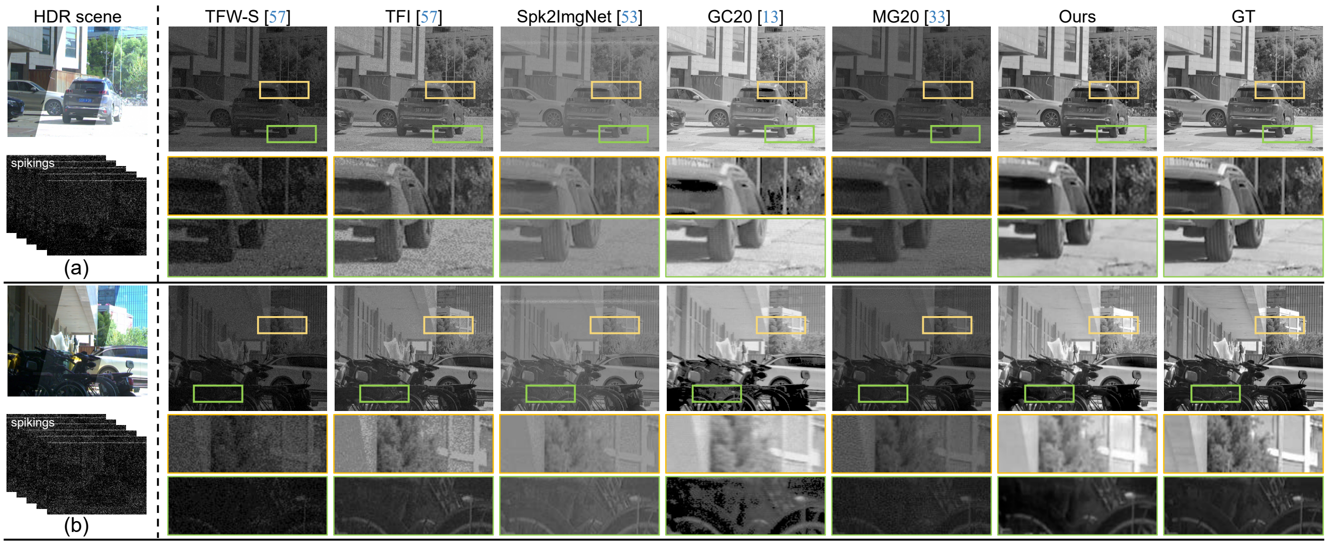

Figure 8. Visual equality comparison of synthetic data between the proposed method and compared methods. The HDR scene is captured by alternating exposures. Please zoom-in electronic versions for better details, and watch the videos in the project page.

Table 2. Quantitative results and ablation studies on our synthetic data. ↑ (↓) indicates larger (smaller) values are better.

	Comparison with state-of-the-art methods							Ablation studies					
Method	TFW-S [61]	TFW-L [61]	TFI [61]	Spk2ImgNet [57]	GC20 [15]	MG20 [36]	Ours	$w = 0$	Pure-S	Pure-M	Full-M	w/o XBA	w/o XTA
PSNR↑	9.04	13.33	17.67	15.91	19.22	16.16	**24.06**	19.95	19.37	<u>26.62</u>	*24.50*	22.29	21.81
SSIM↑	0.408	0.778	0.697	0.717	0.800	0.748	**0.895**	0.807	0.867	<u>0.904</u>	*0.901*	0.881	0.858
HDR-VDP3↑	6.383	7.564	7.048	7.664	7.249	7.623	**8.103**	7.531	7.502	<u>8.182</u>	*8.120*	7.971	7.957
HDR-VQM↓	0.939	0.621	0.849	0.753	0.618	0.741	**0.096**	0.259	0.315	<u>0.084</u>	*0.093*	0.112	0.106

with the alternating-exposure videos in [2]. Then, we synthesize the RMB spikings with the mechanism described in Sec. 3. T is 12.5ms in this work. We also collect 10 groups of real data. As the spiking camera at our disposal has not yet undergone hardware upgrades to enable the RMB readout, we simulate RMB spikings through spatial and temporal aggregation, *i.e.*, a 3-bit spiking is obtained by aggregating the spikings in a $2 \times 2 \times 2$ (height, width, and time) binning. Similarly, the single-bit spiking is obtained by retaining only one pixel in the binning. Since the RMB spikings are generated from real data, there is no ground truth for them.

Loss and training. It has been confirmed that training the network on the tone-mapped images is more effective than training directly in the HDR domain [3, 19, 22]. We compress the range of the ground truth G by applying the μ-law function: $\mathcal{T}(G) = \log(1 + \mu G)/\log(1 + \mu)$. To train our merging module, we employ the l_2 loss, Structure similarity (SSIM) loss [51], and Learned Perceptual Image Patch Similarity (LPIPS) loss [56]. The total loss at step i is:

$$\mathcal{L}(i) = \mathcal{L}_{l_2}(i) + \beta_1 \mathcal{L}_{\text{SSIM}}(i) + \beta_2 \mathcal{L}_{\text{LPIPS}}(i), \quad (13)$$

where $\beta_1 = \beta_2 = 1$. We train the merging module at each step i. We set the batch size to 2 and use Adam optimizer for training. The merging module is trained with 50 epochs.

5. Experiments

5.1. Quantitative evaluation on synthetic data

We compare our method with existing spiking-based video reconstruction methods, *i.e.*, TFW [61], TFI [61], and Spk2ImgNet [57]. Note that they reconstruct videos from single-bit spikings. While we acknowledge that it is not entirely fair to compare our method with these approaches because they are designed only for single-bit spikings, they serve as the best baselines for showcasing the potential of introducing multi-bit spikings. TFW-S indicates the TFW with a small temporal window (10), while TFW-L indicates the TFW with a long temporal window (70). Given the similar data stream of quanta image sensors (QIS) and spiking cameras, HDR methods developed for QIS can be adapted to spiking cameras. Hence, we choose GC20 [15] and MG20[5] [36] for comparison. All the inputs of the compared methods are RMB spikings. We further conduct ablation studies to demonstrate the effectiveness of each module in our framework. For "$w = 0$", we only input a single I_s and a single I_m to RMB-Net, instead of frame bunches; for "Pure-S", we feed pure single-bit spikings to RMB-Net; for "Pure-M", we feed pure 3-bit spikings to RMB-Net; for "Full-M", we feed the first type of data in Fig. 4 (a) to the model; for "w/o XBA", we replace cross-bit attention with a simple concatenation operation; for "w/o XTA", we remove the cross-time attention with a simple temporal average operation.

Fig. 8 shows the reconstruction results on synthetic data of the proposed method and compared methods. TFW and TFI tend to reconstruct noisy video frames. Spk2ImgNet [57] reconstructs low-contrast video frames. GC20 [15] performs well in preserving textures in bright regions, but fails to preserve textures in dark regions. MG20 [36] successfully

[5]We use "first letter of first names of first two authors + year" as the abbreviation for the comparison methods in this section.

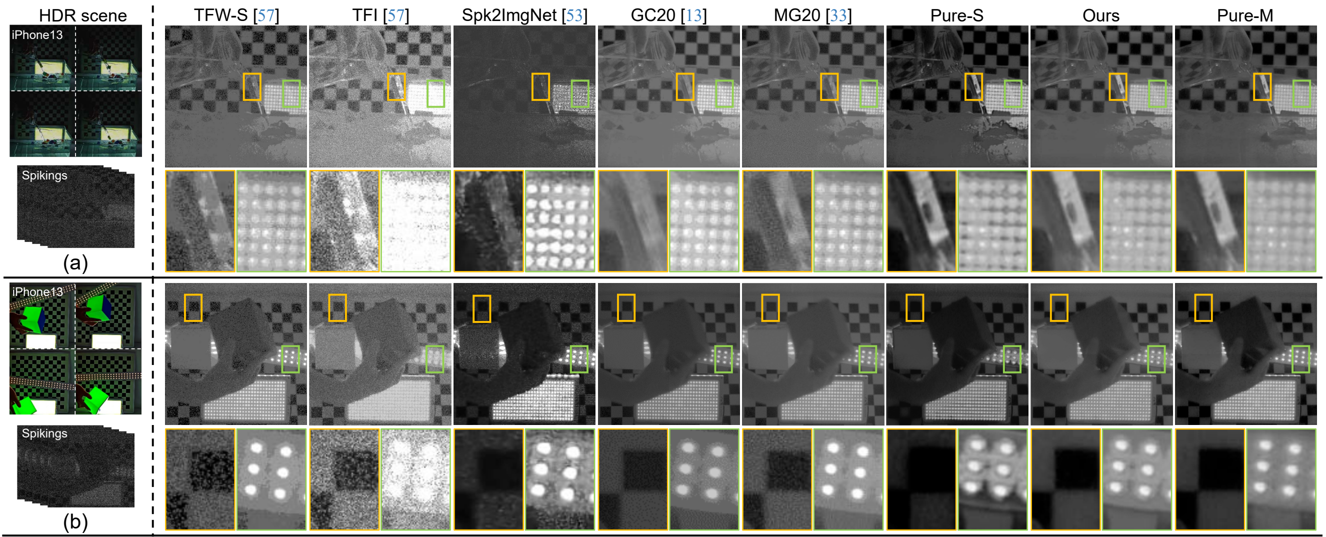

Figure 9. Visual equality comparison of real-synthetic data. The four frames captured by iPhone13 are used to illustrate HDR scenes. Please zoom-in electronic versions for better details, and watch the corresponding videos on the project page.

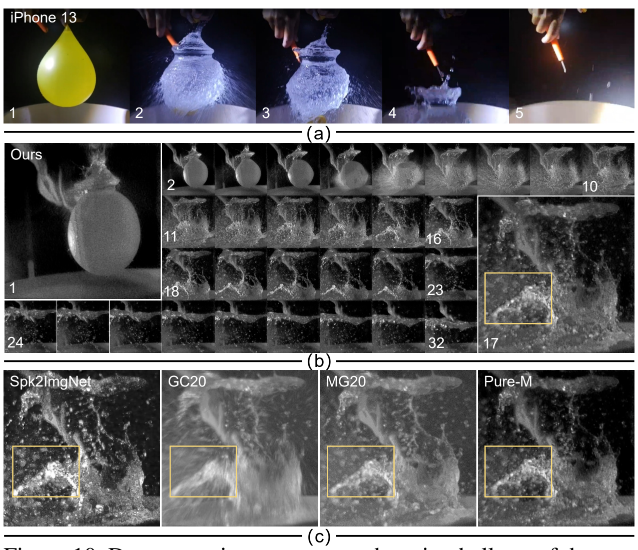

Figure 10. Demonstration to capture a bursting balloon of the proposed method and compared methods. (a) The five frames captured by iPhone13 are used to illustrate the HDR scene. (b) We sample 32 frames from 260 frames for illustration, and select the 17-th image for comparison. (c) Results of Spk2ImgNet [57], GC20 [15], MG20 [36], and Pure-M.

recovers sharp textures, but the frames are still contaminated by noise. The proposed method is capable of reconstructing rich texture details for both dark and bright regions. We evaluate the reconstructed HDR videos in terms of PSNR, HDR-VDP-3 [39], HDR-VQM [43], and SSIM [51] in Table 2, showing that our method consistently achieves state-of-the-art performance.

5.2. Qualitative evaluation on real-synthetic data

In order to demonstrate the effectiveness of the proposed framework on real-world scenes, we perform experimental comparisons on real-synthetic data. In Fig. 9 (a), we position an LED light array (about 8,000LUX) in front of the spiking camera and pour out a cup of water, causing small objects to flow out along with the water. In Fig. 9 (b), we rapidly wave an LED light strip and a cuboid. Through observing the regions marked by bounding boxes, we can see that the RMB approach outperforms state-of-the-art methods in reconstructing finer details and achieves comparable performance with Pure-M. As shown in Fig. 10, in a dimly-lit environment (about 200LUX), a balloon is bursting and the water is splashing around. Both Spk2ImgNet [57] and Pure-S yield unsatisfactory results with a loss of detail in the water splash reflections. Our approach consistently achieves comparable results with Pure-M in capturing the details.

6. Conclusion

In this paper, through a comprehensive analysis of the dynamic range of spiking cameras, we identified that increasing the bit depth of spikings can boost HDR performance. In high-speed conditions, compared to the pure multi-bit readout mechanism that results in multifold data volume, RMB spikings only increase the data volume by less than 2% in our setting, yet still enable comparable reconstruction of HDR frames. Accordingly, we developed an effective RMB-Net to achieve HDR and HFR video reconstruction. Under the joint action of the proposed cross-bit and cross-time attention blocks, our RMB-Net demonstrates excellent performance on both synthetic and real data simulations.

Acknowledgments. This work was supported by National Science and Technology Major Project (Grant No. 2021ZD0109803), and National Natural Science Foundation of China (Grant No. 62301009, 62088102, and 62136001). Bin Fan was also supported by National Postdoctoral Program for Innovative Talents of China (Grant No. BX20230013). The authors thank the anonymous reviewers and the area chairs for their helpful comments.

References

[1] Nick Antipa, Patrick Oare, Emrah Bostan, Ren Ng, and Laura Waller. Video from stills: Lensless imaging with rolling shutter. In *Proc. of International Conference on Computational Photography*, pages 1–8, 2019. 2

[2] Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 FPS HDR video with a Spike-RGB hybrid camera. In *Proc. of Computer Vision and Pattern Recognition*, pages 22180–22190, 2023. 1, 2, 3, 6, 7

[3] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *Proc. of International Conference on Computer Vision*, pages 2502–2511, 2021. 1, 2, 7

[4] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. HDRUNet: Single image HDR reconstruction with denoising and dequantization. In *Proc. of Computer Vision and Pattern Recognition*, pages 354–363, 2021. 2

[5] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. of ACM SIGGRAPH*, pages 1–10, 2008. 2

[6] Tobi Delbruck. Frame-free dynamic digital vision. In *Proc. of International Symposium on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, pages 21–26. Citeseer, 2008. 1

[7] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics*, 36(6):1–15, 2017. 2

[8] Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: Bring rolling shutter images to high framerate global shutter video. In *Proc. of International Conference on Computer Vision*, pages 4228–4237, 2021. 2

[9] Bin Fan, Yuchao Dai, and Hongdong Li. Rolling shutter inversion: Bring rolling shutter images to high framerate global shutter video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6214–6230, 2022.

[10] Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *Proc. of Computer Vision and Pattern Recognition*, pages 17572–17582, 2022. 2

[11] Bin Fan, Yuchao Dai, and Mingyi He. Rolling shutter camera: Modeling, optimization and learning. *Machine Intelligence Research*, 20(6):783–798, 2023. 2

[12] Bin Fan, Yuchao Dai, and Hongdong Li. Learning bilateral cost volume for rolling shutter temporal super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2024. 2

[13] Kexiang Feng, Chuanmin Jia, Siwei Ma, and Wen Gao. Spike-Codec: An end-to-end learned compression framework for spiking camera. *arXiv preprint arXiv:2306.14108*, 2023. 3

[14] Eric R Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8):1260, 2016. 2

[15] Abhiram Gnanasambandam and Stanley H Chan. HDR imaging with quanta image sensors: Theoretical limits and optimal reconstruction. *IEEE Transactions on Computational Imaging*, 6:1571–1585, 2020. 3, 7, 8

[16] Yulia Gryaditskaya, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel. Motion aware exposure bracketing for HDR video. In *Proc. of Computer Graphics Forum*, pages 119–130, 2015. 2

[17] Sheetal B Gupta, AN Rajagopalan, and Gunasekaran Seetharaman. HDR recovery under rolling shutter distortions. In *Proc. of International Conference on Computer Vision Workshops*, pages 8–15, 2015. 2

[18] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualISO HDR reconstruction. *EURASIP Journal on Image and Video Processing*, 2015(1):1–13, 2015. 2

[19] 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 *Proc. of Computer Vision and Pattern Recognition*, pages 1730–1739, 2020. 2, 3, 7

[20] Ankur Handa, Richard A Newcombe, Adrien Angeli, and Andrew J Davison. Real-time camera tracking: When is high frame-rate best? In *Proc. of European Conference on Computer Vision*, pages 222–235. Springer, 2012. 1

[21] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 25:110–119, 2023. 1, 2, 3

[22] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):144–1, 2017. 2, 7

[23] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR video from sequences with alternating exposures. In *Proc. of Computer Graphics Forum*, pages 193–205, 2019. 1, 2

[24] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Transactions on Graphics*, 32(6):202–1, 2013. 1

[25] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics*, 22(3):319–325, 2003. 2

[26] Kazuya Kitamura, Toshihisa Watabe, Takehide Sawamoto, Tomohiko Kosugi, Tomoyuki Akahori, Tetsuya Iida, Keigo Isobe, Takashi Watanabe, Hiroshi Shimamoto, Hiroshi Ohtake, Satoshi Aoyama, Shoji Kawahito, and Norifumi Egami. A 33-megapixel 120-frames-per-second 2.5-watt CMOS image sensor with column-parallel two-stage cyclic analog-to-digital converters. *IEEE Transactions on Electron Devices*, 59(12):3426–3433, 2012. 2

[27] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-image HDR reconstruction by multi-exposure generation. In *Proc. of Winter Conference on Applications of Computer Vision*, pages 4063–4072, 2023. 2

[28] Byungju Lee and Byung Cheol Song. Multi-image high dynamic range algorithm using a hybrid camera. *Signal Processing: Image Communication*, 30:37–56, 2015. 2

[29] Juan Antonio Leñero-Bardallo, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco. A 3.6 μs latency asynchronous frame-free event-driven dynamic-vision-sensor.

IEEE Journal of Solid-State Circuits, 46(6):1443–1455, 2011. 1

[30] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576, 2008. 1, 2

[31] Guixu Lin, Jin Han, Mingdeng Cao, Zhihang Zhong, and Yinqiang Zheng. Event-guided frame interpolation and dynamic range expansion of single rolling shutter image. In *Proc. of the ACM International Conference on Multimedia*, pages 3078–3088, 2023. 2

[32] Yuhao Liu, Felipe Gutierrez-Barragan, Atul Ingle, Mohit Gupta, and Andreas Velten. Single-photon camera guided extreme dynamic range imaging. In *Proc. of Winter Conference on Applications of Computer Vision*, pages 1575–1585, 2022. 3

[33] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proc. of Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 2

[34] Jiaju Ma, Dexue Zhang, Dakota Robledo, Leo Anzagira, and Saleh Masoodian. Ultra-high-resolution quanta image sensor with reliable photon-number-resolving and high dynamic range capabilities. *Scientific Reports*, 12(1):13869, 2022. 3

[35] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: A structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017. 2

[36] Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics*, 39(4):79–1, 2020. 3, 7, 8

[37] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Proc. of Applications of Digital Image Processing*, pages 307–314. SPIE, 2010. 2

[38] Stephen Mangiat and Jerry Gibson. Spatially adaptive filtering for registration artifact removal in HDR video. In *Proc. of International Conference on Image Processing*, pages 1317–1320, 2011. 2

[39] Rafal K Mantiuk, Dounia Hammou, and Param Hanji. HDR-VDP-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*, 2023. 8

[40] Morgan McGuire, Wojciech Matusik, Hanspeter Pfister, Billy Chen, John F Hughes, and Shree K Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 27(2):32–42, 2007. 2

[41] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Proc. of Pacific Conference on Computer Graphics and Applications*, pages 382–390, 2007. 2

[42] Nico Messikommer, Stamatios Georgoulis, Daniel Gehrig, Stepan Tulyakov, Julius Erbach, Alfredo Bochicchio, Yuanyou Li, and Davide Scaramuzza. Multi-Bracket high dynamic range imaging with event cameras. In *Proc. of Computer Vision and Pattern Recognition*, pages 547–557, 2022. 3

[43] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 35:46–60, 2015. 8

[44] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proc. of Computer Vision and Pattern Recognition*, pages 472–479, 2000. 2

[45] Cristiano Niclass, Alexis Rochas, P-A Besse, and Edoardo Charbon. Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9):1847–1854, 2005. 2

[46] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Transactions on Graphics*, 31(6):203–1, 2012. 2

[47] Richard Shaw, Sibi Catley-Chandar, Ales Leonardis, and Eduardo Perez-Pellitero. HDR reconstruction from bracketed exposures and events. *arXiv preprint arXiv:2203.14825*, 2022. 3

[48] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Proc. of Advances in Neural Information Processing Systems*, 28, 2015. 6

[49] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile HDR video production system. *ACM Transactions on Graphics*, 30(4):1–10, 2011. 2

[50] Esteban Vera, Felipe Guzmán, and Nelson Díaz. Shuffled rolling shutter for snapshot temporal imaging. *Optics Express*, 30(2):887–901, 2022. 2

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7, 8

[52] Gil Weinberg and Ori Katz. 100,000 frames-per-second compressive imaging with a conventional rolling-shutter camera by random point-spread-function engineering. *Optics Express*, 28(21):30616–30625, 2020. 2

[53] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proc. of Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 1

[54] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. 2

[55] Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, and Boxin Shi. Learning event guided high dynamic range video reconstruction. In *Proc. of Computer Vision and Pattern Recognition*, pages 13924–13934, 2023. 2, 3

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[57] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic

scene from continuous spike stream. In *Proc. of Computer Vision and Pattern Recognition*, pages 11996–12005, 2021. 2, 7, 8

[58] Jing Zhao, Ruiqin Xiong, Jiyu Xie, Boxin Shi, Zhaofei Yu, Wen Gao, and Tiejun Huang. Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. *IEEE Transactions on Computational Imaging*, 8:12–27, 2021. 2

[59] Rui Zhao, Ruiqin Xiong, Jing Zhao, Zhaofei Yu, Xiaopeng Fan, and Tiejun Huang. Learning optical flow from continuous spike streams. *Proc. of Advances in Neural Information Processing Systems*, 35:7905–7920, 2022. 5, 6

[60] Lin Zhu, Siwei Dong, Tiejun Huang, and Yonghong Tian. A retina-inspired sampling method for visual texture reconstruction. In *Proc. of International Conference on Multimedia and Expo*, pages 1432–1437, 2019. 2

[61] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. Retina-like visual image reconstruction via spiking neural model. In *Proc. of Computer Vision and Pattern Recognition*, pages 1438–1446, 2020. 2, 7