Real-data-driven 2000 FPS Color Video from Mosaicked Chromatic Spikes

Siqi Yang[1,#], Zhaojun Huang[2,3,#], Yakun Chang[5,6], Bin Fan[4], Zhaofei Yu[1], and Boxin Shi[2,3,1,*]

[1] Institute for Artificial Intelligence, Peking University
[2] State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[3] Nat'l Eng. Research Ctr. of Visual Technology, School of Computer Science, Peking University
[4] Nat'l Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[5] Institute of Information Science, Beijing Jiaotong University
[6] Visual Intelligence +X International Cooperation Joint Laboratory of the Ministry of Education
{yousiki, huangzhaojun, binfan, yuzf12, shiboxin}@pku.edu.cn, ykchang@bjtu.edu.cn

Abstract. The spike camera continuously records scene radiance with high-speed, high dynamic range, and low data redundancy properties, as a promising replacement for frame-based high-speed cameras. Previous methods for reconstructing color videos from monochromatic spikes are constrained in capturing full-temporal color information due to their reliance on compensating colors from low-speed RGB frames. Applying a Bayer-pattern color filter array to the spike sensor yields mosaicked chromatic spikes, which complicates noise distribution in high-speed conditions. By validating that the noise of short-term frames follows a zero-mean distribution, we leverage this hypothesis to develop a self-supervised denoising module trained exclusively on real-world data. Although noise is reduced in short-term frames, the long-term accumulation of incident photons is still necessary to construct HDR frames. Therefore, we introduce a progressive warping module to generate pseudo long-term exposure frames. This approach effectively mitigates motion blur artifacts in high-speed conditions. Integrating these modules forms a real-data-driven reconstruction method for mosaicked chromatic spikes. Extensive experiments conducted on both synthetic and real-world data demonstrate that our approach is effective in reconstructing 2000FPS color HDR videos with significantly reduced noise and motion blur compared to existing methods.

Keywords: Spike camera· High-speed video · High dynamic range video · Color demosaicking · Self-supervised learning

1 Introduction

As a new class of image sensors inspired by the fovea of retina, the spike camera [3, 14, 46] offers attractive characteristics such as high temporal resolution (20000Hz), high dynamic range (HDR, > 90dB), and low data redundancy [14]. In comparison with conventional digital cameras that capture discrete frames,

# Equal contributions. * Corresponding author.

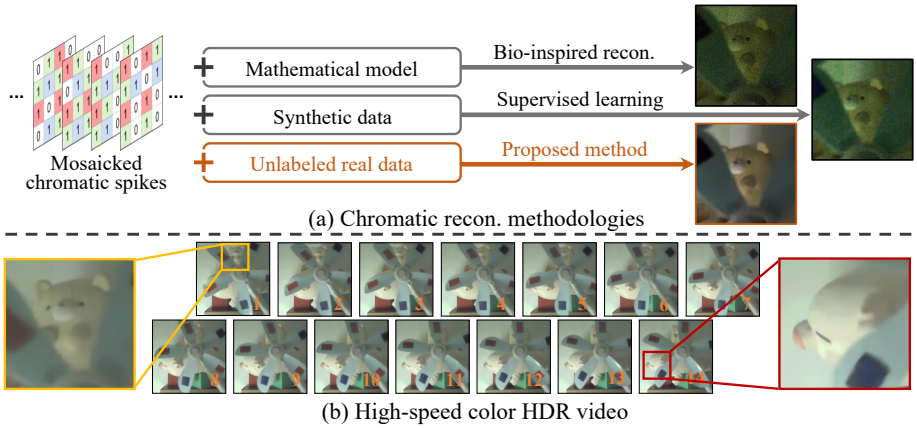


Fig. 1: High-speed color HDR video reconstruction from mosaicked chromatic spikes. (a) Comparison of different chromatic spike reconstruction (recon.) methodologies, including the adaptation of bio-inspired [45] and supervised learning [39] monochromatic methods to chromatic spikes and our proposed method. (b) A high-quality color HDR video with a frame rate of 2000FPS reconstructed by our method. The corresponding video is available in the supplementary material.

spike cameras continuously accumulate the photons and trigger a single-bit spike of 1 when the detected photons reach a predefined threshold [14]. High-speed readout of spikes provides the spike camera with the potential for photon counting, and the high-dynamic-range ambient light can be estimated through continuous accumulation of spikes in the temporal domain. Therefore, spike cameras exhibit significant advantages in handling low-level vision tasks in high-speed conditions, such as video frame interpolation [33], HDR imaging [11] and high-frame-rate HDR video reconstruction [3].

With the higher demand for visual quality in images and videos, there is a strong desire that images or videos reconstructed from spike cameras can reproduce the real world to the greatest extent, with realistic color appearance and extended dynamic range at high frame rate. However, existing approaches [39, 40, 43, 44, 46] mainly focus on image or video reconstruction with monochromatic spikes, and cannot yet recover a high-speed color HDR video. To introduce color, one alternative practice [11] is to build a hybrid spike-RGB camera system, which exploits the colors captured by the conventional RGB camera to colorize the image reconstructed from spikes. Unfortunately, the color information provided by conventional RGB cameras often suffers from underexposure and overexposure, which inevitably lead to color deviations from the real world. Another strategy [3] to alleviate this problem is to replace the conventional RGB camera with an alternate exposure RGB camera to obtain HDR color, thereby enabling high-speed HDR color video reconstruction. However, the above methods have limitations in high-speed scenes, as the RGB camera has a low frame rate (*e.g.*, 60FPS), which results in the loss of colors during the

intervals between frames. In addition, the corresponding hybrid camera system relies on complex temporal synchronization and optical alignment, and the beam splitter with large space footprints would be detrimental to the construction of compact devices.

Reconstructing high-speed color HDR videos *based only on a single spike camera* not only facilitates the acquisition of full-temporal color information but also offers the advantage of lightweight equipment. To this end, a tractable solution is to assemble a Bayer-pattern color filter array (CFA) in the spike camera, thereby producing mosaicked chromatic spikes. Consequently, a pressing challenge that arises is how to reconstruct high-speed color HDR videos from this new data representation. Nevertheless, directly adapting existing monochromatic spike reconstruction methods [39,40,43,44] to mosaicked chromatic spikes suffers from many issues, as shown in Fig. 1 (a). On the one hand, methods based on biology-inspired models [40,43,44] cannot take advantage of deep neural networks and are susceptible to varying degrees of noise contamination. On the other hand, supervised learning methods (*e.g.*, [39]) trained using synthetic datasets have limited generalization ability when transferring to real-world data, because of the imperfection of spike noise modeling and simulation. These issues hinder the development of high-speed color HDR video reconstruction methods based on chromatic spike cameras and greatly limit the effective deployment of existing methods in real scenes.

In this paper, we propose a *real-data-driven* mosaicked chromatic spike reconstruction method to recover high-speed color HDR videos, trained solely on readily accessible unlabelled spike streams. The mathematical analysis behind this requires no specific spike noise modeling or calibration, relying only on insights into the relationship between short-term and long-term frames accumulated from chromatic spikes. We design a deep learning framework that allows training directly on real-world data, naturally avoiding the influence of imperfect spike simulation and generalization gap. Specifically, we first propose a self-supervised denoising module to suppress the noise in short-term accumulation of chromatic spikes, which is trained on real-world spike streams without any annotation. Next, considering the limited dynamic range of short-term accumulated frames, we further present a progressive warping module to simulate a pseudo-long exposure, which is effective in reconstructing HDR video frames without motion blur. Finally, we integrate these two modules to achieve higher-quality color HDR video reconstruction.

In summary, our approach is able to achieve *2000FPS color HDR video* reconstruction from mosaicked chromatic spikes with a single device in a real-data-driven manner, as illustrated in Fig. 1 (b), because of the technical contributions from these two complementary aspects:

- We design a self-supervised denoising module to eliminate the noise from short-term accumulation (with high-speed motion cues) without relying on chromatic spike noise modeling and simulation.

– We propose a progressive warping module with pseudo long-term exposure (with expanded dynamic range) to reconstruct HDR frames without introducing motion blur in dynamic scenes.

2 Related Work

Video reconstruction with monochromatic spikes. To better satisfy human visual perception and facilitate direct processing by machines, reconstructing the corresponding video frame from spikes has been widely studied. Prevalent image reconstruction methods typically employ the temporal statistical properties of spike cameras [43–46]. Zhu *et al*. [45] explored the spike generation principle and proposed estimating the firing frequency or firing interval of each pixel to reconstruct dynamic scenes, which are known as texture from playback (TFP) and texture from inter-spike-intervals (TFI), respectively. Recently, another line of work has enhanced firing frequency estimates by mimicking human physiological mechanisms, such as retina-like visual imaging [46] and short-term plasticity [43, 44]. Although these methods bring better interpretability, they require careful window selection to balance noise and motion blur, and the reconstruction effect is not yet satisfactory. The powerful representation capabilities of deep learning also inject vitality into spike-based image reconstruction. Spk2ImgNet [39] is the first spike-to-image deep network architecture based on deformable convolution, achieving impressive results. Further, several self-supervised methods [5,6] have also been developed to reduce the dependence on synthetic datasets. However, these methods are tailored for monochromatic spikes and thus cannot be directly applied to mosaicked chromatic spikes.

Demosaicking and denoising with raw images. The purpose of raw image demosaicking is to recover a full-color image from a sub-sampled mosaicked image that contains potential noise. To deal with the ill-posedness of image demosaicking, various image priors have been introduced to regularize the recovery process, such as heuristics [2, 24], sparsity [23, 36], total variation [7, 13], integral gradient [28], self-similarity [22], and residual interpolation [15]. In addition, researchers designed deep neural networks [16, 29, 30] to automatically learn the prior knowledge required for image demosaicking. Moreover, considering the coupling between demosaicking and denoising, some deep learning-based joint demosaicking and denoising network models [9, 18, 34, 38] have also been developed to efficiently remove mosaicks from noisy raw images.

HDR video with non-conventional sensors. In recent years, a series of non-conventional sensors, such as event cameras [17], spike cameras [14], and quanta image sensors (QIS) [8], have been invented, empowering the ability to capture HDR information in high-speed motion scenes. QIS-based HDR video reconstruction methods [10, 19, 20] were developed to accomplish perceptual HDR illumination. Additionally, the intensity image reconstructed by event cameras [35] or spike cameras [11] can serve as a proxy for enhancing details in low dynamic range (LDR) images. More recently, a hybrid camera system consisting of a spike camera and an alternating-exposure RGB camera is constructed in [3] for 1000×

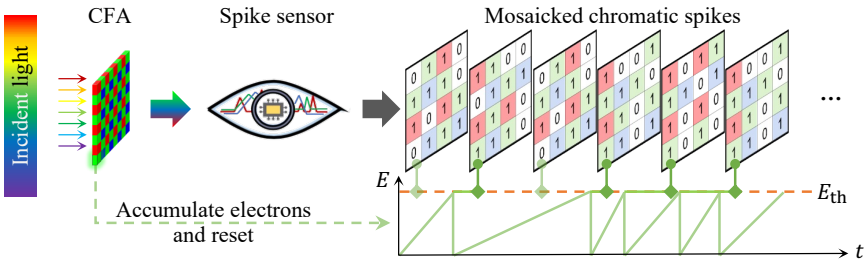


Fig. 2: Chromatic spike accumulation and firing mechanism. We put a Bayer pattern CFA in the spike camera to filter incident light before the photons arrive at the sensor. Each individual pixel corresponds to one of red, green, and blue channels, accumulates the filtered photons, and triggers a spike signal, then resets when it surpasses the threshold E_{th}. Reading out the chromatic spikes forms the chromatic spike planes.

frame rate HDR video reconstruction. Nevertheless, these methods also fall short when dealing with mosaicked chromatic spikes.

3 Preliminaries

We initially present a concise review of the operational principles of spike cameras, encompassing both monochromatic spike cameras and those equipped with Color Filter Array (CFA). Subsequently, we discuss the baseline reconstruction techniques for monochromatic spikes.

Monochromatic spikes. The spike camera persistently accumulates electrons produced by incoming photons. When the electrons accumulated at a pixel surpass the predefined threshold E_{th}, it triggers a spike signal, immediately resetting the accumulator to initiate a new cycle of accumulation:

$$E(\mathbf{x}, t) = \int \alpha L(\mathbf{x}, t) dt \mod E_{\text{th}}, \tag{1}$$

where α denotes the conversion ratio and $L(\mathbf{x}, t)$ denotes the incoming light intensity of pixel $\mathbf{x}$ at time t. Consequently, the spike signal $S(\mathbf{x}, t)$ can be defined as:

$$S(\mathbf{x}, t) = \begin{cases} 1, & \text{if } E(\mathbf{x}, t) \geq E_{\text{th}}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

As all the pixels of a spike sensor are arranged in an array, reading out spikes from the pixel array at each time stamp t yields a spike plane of dimensions $h \times w$, where h and w are the height and width of the spike data, respectively. Given the readout frequency of 20000Hz, the spike camera [14] allows for capturing spike planes at a frame rate of 20000FPS.

Chromatic spikes. To incorporate color information while capturing HDR scenes with high-speed motion at the same temporal resolution, we can put a

CFA in front of the sensor of the spike camera. In a similar way as conventional cameras, the CFA applied to spike sensors presents a Bayer pattern. Each pixel of the spike camera sensor with CFA corresponds to only one channel among red, green, and blue, while the firing mechanism of spikes remains unchanged. As illustrated in Fig. 2, under the influence of the CFA, a spike signal is generated when the cumulative electrons $E_c(\mathbf{x}, t)$ produced by the incident photons of the corresponding spectral band meets the threshold E_{th}. The chromatic spike signal $S_c(\mathbf{x}, t)$ can be recorded as:

$$S_c(\mathbf{x}, t) = \begin{cases} 1, & \text{if } E_c(\mathbf{x}, t) \geq E_{\text{th}}, \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $c \in \{R, G, B\}$. Consequently, similar to the firing process of monochromatic spikes, we can obtain a mosaicked chromatic spike plane at each time stamp t.

Baseline for chromatic frame reconstruction. Preliminary chromatic frame reconstruction can be achieved similarly to [45] by accumulating the spike signals within a defined temporal window:

$$I_c(\mathbf{x}, t) = \frac{M_c}{W} \cdot \sum_{\tau=t-W/2}^{t+W/2} S_c(\mathbf{x}, \tau), \tag{4}$$

where W is the size of the accumulation temporal window, M_c represents the maximum dynamic range that can be perceived in channel c. Directly applying Eq. (4) to chromatic spikes as a baseline reconstruction algorithm can only obtain mosaicked chromatic frames, which is similar to the RAW image of conventional frame-based sensors. Demosaicking the chromatic frame is challenging since there is a trade-off between noise and superposition of colors. A short-term temporal window W results in severe noise, whereas a long-term W leads to the spatial superposition of varying color information. Furthermore, both noise distribution of chromatic spikes and object motions are more complex, making the direct application of off-the-shelf demosaicking methods insufficient for recovering high-quality color HDR videos.

4 Real-data-driven chromatic spikes to video

Overview. As shown in Fig. 3, we propose a progressive mechanism for high-speed color HDR video reconstruction. The pipeline starts with the accumulation of mosaicked chromatic spikes to reconstruct short-term frames. While these short-term frames are free of motion blur, they suffer from significant noise contamination. Consequently, as shown in Fig. 3 (a), we propose the chromatic spike denoising module (Sec. 4.1) to suppress the noise. The key insight of the denoising module is that a sufficiently long-term frame (*e.g.*, $W = 10000$, corresponding to 0.5s) of chromatic spikes can produce promising reconstructions in static (or low-speed) scenes, effectively representing the expectations of short-term frames. Given the limited incident photons in short-term frames, which

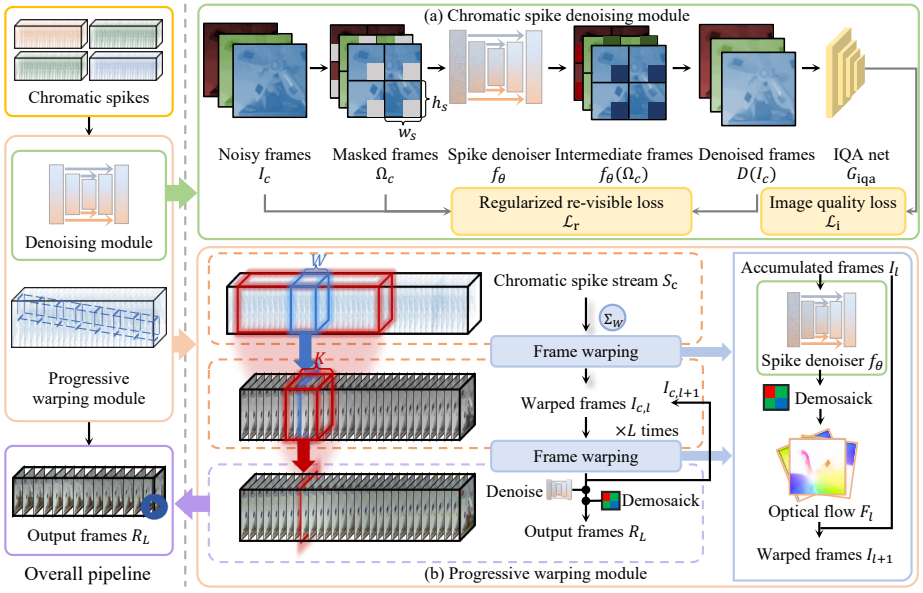


Fig. 3: Illustration of the overall pipeline. For each sequence of mosaicked chromatic spikes, we split them into four channels according to the RGGB pattern. The four-channel spikes are fed to our progressive warping module, which aligns multiple short-term frames to achieve "pseudo-long exposure". (a) Our chromatic spike denoising module utilizes a self-supervised training strategy, which is pretrained on real-world data. (b) Our progressive warping module consists of L steps of frame warping operations, iteratively achieving high-quality HDR frames. At each step of the progressive warping module, we leverage the pretrained chromatic spikes denoising module to facilitate optical flow estimation. R_L is the final output color HDR video with 2000FPS.

lack the HDR characteristic, extending the accumulation time is a feasible strategy for HDR reconstruction. As shown in Fig. 3 (b), we design the progressive warping module in Sec. 4.2 to establish a pseudo long exposure. At each step of the warping module, chromatic spikes frames are warped according to the estimated optical flow, providing pseudo long-term frames with less noise and higher dynamic range. To leverage a wide range of existing conventional RGB optical flow estimation approaches, we integrate the denoising module into each iteration step of the warping module. After L times of iteration, the progressive warping module produces an HDR color video with a frame rate of 2000FPS.

4.1 Denoising of chromatic spikes

Noise distribution of chromatic spikes noise. The noise in monochromatic spikes is primarily categorized into two types: diffuse noise and inherent noise. For the mosaicked chromatic spikes, although the incorporation of a CFA slightly alters the noise distribution for pixels under different filters, pixels corresponding

to the same color exhibit similar noise distribution patterns as monochromatic spikes, except for its expectation and variance. Given the chromatic spikes S_c, for each $c \in \{R, G, B\}$, we use Q_d to denote the number of diffuse noise photons arriving at a pixel in a time period of t, and γ to denote the likelihood of Q_d diffuse noise photons. The probability of diffuse noise can be modeled as the Poisson distribution [41]:

$$\mathcal{P}(Q_d|t,\gamma) = \frac{(\gamma t)^{Q_d}}{Q_d!}\exp(-\gamma W). \tag{5}$$

Hence, the expected number of diffuse noise spikes in temporal window W is

$$\mathbb{E}[N_d] = (\alpha\gamma W)/E_{\text{th}}. \tag{6}$$

For inherent noise, which triggers spikes in the absence of light, we use T_i to indicate the interval between the noise spikes. Then, the interval between such noise spikes can be represented by a Gaussian model [41]:

$$\mathcal{N}(T_i|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\cdot\exp(-\frac{1}{2}\cdot(\frac{T_i-\mu}{\sigma})^2). \tag{7}$$

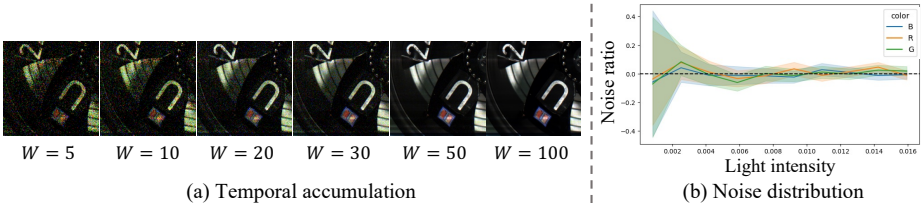
We approximate the expectation of the number of inherent noise spikes in the temporal window as

$$\mathbb{E}[N_i] = \int_1^{+\infty}\frac{W}{T_i}\cdot\frac{1}{\sigma\sqrt{2\pi}}\cdot\exp(-\frac{1}{2}\cdot(\frac{T_i-\mu}{\sigma})^2)dT_i = C_iW, \tag{8}$$

where C_i denotes the integral after factoring out W from the integral expression. Combining Eq. (6) and Eq. (8), the expectation of the fraction of noise spikes can be formulated as:

$$\mathbb{E}[N]/W = (\mathbb{E}[N_d] + \mathbb{E}[N_i])/W = \alpha\gamma/E_{\text{th}} + C_i. \tag{9}$$

Instead of the actual incoming light intensity, we estimate the expectation of short-term accumulations $\mathbb{E}[I_c(\mathbf{x},t)] = L(\mathbf{x},t) + \mathbb{E}[N]/W$. Thus, the noise between short-term frames and their expectations naturally follows a zero-mean distribution. We demonstrate that the expected fraction of noise spikes is relatively small, especially with sufficient light illumination, and doesn't disturb the visual semantic. As shown in Fig. 4 (a), the noise ratio $(I_c(\mathbf{x},t)/\mathbb{E}[I_c(\mathbf{x},t)]-1)$ in the accumulated images $(I_c(\mathbf{x},t))$gradually decreases as the accumulation window size W increases. Note that the larger window accumulation is equivalent to the average of multiple smaller windows. The convergence demonstrates the existence of expectation and the result faithfully reflects the original light intensity with negligible disturbance. We further validate the zero-mean distribution hypothesis by collecting a series of static scene spike streams with varying light illuminations, and visualize the distribution of noise ratio in Fig. 4 (b). The result shows that the zero-mean assumption holds with sufficient illumination, which supports our self-supervised learning mechanism.



(a) Temporal accumulation (b) Noise distribution

Fig. 4: Validation of noise on real-world chromatic spikes. As shown in (a), as the accumulation window size W increases, the noise ratio in the accumulated images gradually decreases, making the images more closely resemble the static scene. In (b), we present the statistical noise fraction distribution across three different color channels and different illumination levels.

Self-supervised chromatic spike denoising module. Taking the inspiration of Batson *et al.* [1] and Wang *et al.* [31], which do not require paired noise-contaminated images or clean targets, we design a chromatic spike denoising module trained on only accessible real-world data with self-supervision learning strategy. Specifically, our denoising module operates based on the zero-mean characteristic of short-term frame noise and relies on the blind spot architecture to prevent the network from learning the identity mapping. Hence, the denoising module transforms the noise-contaminated images to their expectations to achieve minimal losses, producing desired clean images. Therefore, as shown in Fig. 3 (a), we begin by extracting the four channels of the chromatic spikes according to the RGGB pattern of the CFA, and the size of each channel is $\frac{h}{2} \times \frac{w}{2}$. For each channel c, we preliminarily reconstruct a chromatic frame I_c by accumulating the chromatic spikes in a temporal window W. Then, I_c is further partitioned into $\frac{h}{2h_s} \times \frac{w}{2w_s}$ cells, where h_s and w_s denote height and width of each cell. We sequentially mask one pixel for each cell following the order from top to bottom and left to right. Consequently, we obtain a masked volume with a size of $C \times (\frac{h}{2} \times \frac{w}{2})$ corresponding to I_c, where C is the channel of the masked volume that equals $h_s \times w_s$. Each channel of the mask volume acts as the blind spot operation to avoid identity mapping. We then employ a U-Net-inspired fully-convolutional neural network $f_\theta(\cdot)$ to denoise each masked frame in this volume and merge the denoised frames to the clean image:

$$D(I_c) = \Phi(f_\theta(\Omega_{c,0}), f_\theta(\Omega_{c,1}), \ldots, f_\theta(\Omega_{c,C})), \tag{10}$$

where $\Omega_{c,i}$ denotes the i-th masked frame in volume Ω_c, $\Phi(\cdot)$ represents a function for mask-aware averaging:

$$\Phi(f_\theta(\Omega_{c,0}), \ldots) = \sum_i M_i f_\theta(\Omega_{c,i}) / \sum_i M_i, \tag{11}$$

and M_i denotes the mask for $\Omega_{c,i} (\Omega_{c,i} = (1 - M_i) I_c)$. At last, the denoised four-channel images are recombined to form the final mosaicked chromatic frame. It is remarkable that the zero-mean hypothesis holds for arbitrary temporal window

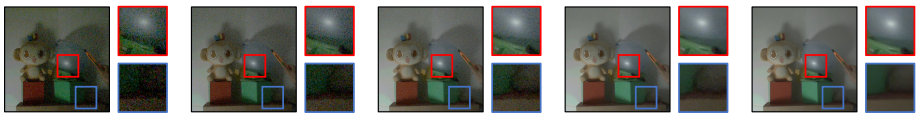


Fig. 5: Progressive warping results. We visualize the intermediate frames from our progressive warping module. As the iterative step l increases, the noise in the reconstructed frames is gradually suppressed and the dynamic range is extended.

size W, indicating that our neural network can be trained and applied on varying chosen W.

Loss function. The chromatic spike denoising module is trained using real-world data with realistic noise. Similar to Wang *et al.* [31], we adopt the regularized re-visible loss:

$$\mathcal{L}_{\mathrm{r}} = \mathcal{L}_{\mathrm{rev}} + \eta\mathcal{L}_{\mathrm{reg}} = \|D(I_c) + \lambda f_\theta(I_c) - (\lambda+1)I_c\|_2^2 + \eta\|D(I_c) - I_c\|_2^2 \quad (12)$$

where η is a constant hyper-parameter, and λ is adjusted in the training progress. We further enhance the denoising module by incorporating an image quality assignment loss, which encourages the predicted score to be close to 1: $\mathcal{L}_{\mathrm{i}} = \|1 - G_{\mathrm{iqa}}(D(I_c))\|_2^2$, where G_{iqa} denotes a pre-trained image quality assignment model [4]. The total training objective combines these losses: $\mathcal{L} = \mathcal{L}_{\mathrm{r}} + \lambda_{\mathrm{i}}\mathcal{L}_{\mathrm{i}}$. Note that the training objectives require no clean frames or human annotations, but only readily accessible real-world mosaicked chromatic spike streams.

4.2 Color and HDR via progressive warping

Although our chromatic spike denoising module can suppress the noise in frames reconstructed from short-term temporal windows, the dynamic range of each short-term frame is still restricted, as the number of photons arriving at the pixels within an extremely short period (*e.g.*, $W = 10$) is limited. Moreover, though our denoising module can fit arbitrary temporal window W theoretically, longer-term input frames lead to better output quality. To reconstruct images with less noise and higher dynamic range, it is necessary to select longer windows to preserve sufficient photons. However, directly warping the frames over the longer temporal window (*e.g.*, $W = 100$) is challenging since high-speed motions may span a wide range of pixels. Thus, as illustrated in Fig. 3 (b), we develop the progressive warping module, which only warps a small number of K chromatic frames in each step. By progressively applying the warping operation to the output of the previous step, at the L-th step, each output frame is equivalent to the weighted accumulation from $[(K-1) \times L + 1] \times W$ aligned chromatic spike planes. Setting a large value of L for the progressive warping module is equivalent to creating a pseudo-long exposure. The pseudo-long exposure leverages optical flow estimation at each step to obtain HDR frames without motion blur.

As shown in Fig. 3 (b), the progressive warping module involves the application of the pretrained chromatic spike denoising module from Sec. 4.1 and demosaicking for better optical flow estimation. In the initial 0-th step, we split T chromatic spike planes using a short-term temporal window W and accumulate the chromatic spikes in each window to form $\frac{T}{W}$ mosaicked frames with Eq. (4). We denote these initial mosaicked frames as $I_{c,0}(\mathbf{x}, t)$, where $c \in \{R, G, B\}$, which is contaminated by noise given a small temporal window. Afterward, for more accurate optical flow estimation, we feed $I_{c,0}(\mathbf{x}, t)$ to the self-supervised denoising module and the demosaicking module that employs linear interpolation to generate intermediate reconstruction frames:

$$R_0(t) = \Psi[D(I_{R,0}(t)), D(I_{G,0}(t)), D(I_{B,0}(t))], \tag{13}$$

where Ψ denotes linear interpolation demosaicking and $\mathbf{x}$ is omitted. We then estimate the optical flow F_0 with off-the-shelf methods from R_0 [12]. Note that in this initial step, $I_{c,0}(\mathbf{x}, t)$ is not warped and is mainly used to estimate a better initial optical flow F_0. Subsequently, F_0 is leveraged to warp $K \times W$ spike planes towards the reference frame in the 1-th step, described as:

$$I_{c,1}(\mathbf{x}, t) = \frac{M_c}{KW} \cdot \sum_{t'=t-KW/2}^{t+KW/2} S_c(\mathbf{x} + F_0(\mathbf{x}, t, t'), t'). \tag{14}$$

Note that we utilize chromatic spike planes in the first two steps and mosaicked frames are warped in the later steps. Specifically, for the l-th iteration step, we continuously pass the accumulated frame $I_{c,l-1}$ to the denoising module and the demosaicking module to facilitate the estimation of optical flow F_{l-1} from the previously generated frames R_{l-1}, and warp $I_{c,l-1}$ to the reference time. This operation can further promote noise suppression and dynamic range extension. In particular, we average K accumulated frames $I_{c,l-1}$ obtained in the $(l-1)$-th step, yielding:

$$I_{c,l}(\mathbf{x}, t) = \frac{1}{K} \cdot \sum_{i=-K/2}^{K/2} I_{c,l-1}(\mathbf{x} + F_{l-1}(\mathbf{x}, t, t'), t'), \tag{15}$$

where $t' = t + iW$. The entire procedure is iteratively repeated for L steps. In Fig. 5, we show a set of examples to demonstrate the effectiveness of the progressive warping module. It is evident that as the iteration step l increases, the details in both highlighted dark regions become clearer.

Finally, we obtain color HDR video frames R_L with a frame rate of $\frac{20000}{W}$FPS. Note that W is adjustable, which leads to higher frame rates for smaller values and lower frame rates for larger values. In our most experiments, we find that $W = 10$ offers a good balance between temporal resolution and reconstruction quality, which ensures 2000FPS high-speed color HDR video reconstruction.

(a) Synthetic chromatic spike streams and reconstructions (b) Real chromatic spike streams and reconstructions

Fig. 6: Illustration of our real-world collected chromatic spike streams, synthetic data derived from conventional cameras, and corresponding reconstructions from our proposed method.

Table 1: Quantitative evaluation of the image reconstruction quality. The scores are averaged across our synthetic dataset. ↑ (↓) means higher (lower) is better.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	HDR-VDP3 ↑	HDR-VQM ↓
TFI [46]	19.215	0.246	0.716	4.645	0.7897
TFP(W=10) [46]	23.253	0.331	0.609	5.778	0.7799
TFP(W=100) [46]	28.791	0.720	0.278	7.545	0.4994
TFSTP [43]	20.945	0.253	0.670	4.587	0.7843
MS23 [21]	28.410	0.635	0.359	7.180	0.5765
Ours	**30.780**	**0.883**	**0.208**	**7.991**	**0.3023**

5 Experiment

5.1 Data collection

Synthetic Data. In line with the methodologies of monochromatic spike simulators (*e.g.*, Zhao *et al.* [41], SpikeCV [42]), we develop a chromatic spike simulator, following the same Bayer pattern as the real-world chromatic spike camera. We leverage the frames in the GoPro dataset [26] and their corresponding synthetic spike streams to train a supervised learning model and compare it with our self-supervised learning method, which addresses the superiority of our strategy. We interpolate the HDR videos collected by Chang *et al.* [3] as the evaluation data for quantitative evaluation in our experiments. Some examples of these synthetic scenes are shown in Fig. 6 (a). For the detailed design of our chromatic spike simulator, please refer to our supplementary material.

Real-world data. To accomplish real-data-driven self-supervised learning and validate the effectiveness of our proposed method, we capture a collection of real-world chromatic spikes using a spike camera equipped with the Bayer pattern. Our dataset comprises over 100 sets of mosaicked chromatic spikes captured from 30 diverse scenes, characterized by varying degrees of high-speed motion and high dynamic range, as depicted in Fig. 6 (b). The real-world data inherently contain the noise characteristics of a real spike camera prototype, posing significant challenges for denoising and demosaicking algorithms. Our real-world dataset is available at https://huggingface.co/datasets/YOUSIKI/chromatic-spikes.

5.2 Quantitative evaluation on synthetic data

We compare our method with existing spike-based video reconstruction methods, *i.e.*, TFI [46], TFP [46], and TFSTP [43]. We acknowledge that it may not be entirely fair to compare our method with these approaches, as they are designed for monochromatic spikes. However, they can serve as baselines to demonstrate the effectiveness of our denoising and progressive warping modules. Since the data format of the spike camera and the single photon avalanche diodes (SPAD) are both single-bit signals, the color video reconstruction method can be adapted to chromatic spikes, and MS23 [21] is chosen for comparison[7]. The quantitative evaluation is performed on synthetic data. We select PSNR, SSIM [32], LPIPS [37], HDR-VDP3 [25], and HDR-VQM [27] as the metrics. The quantitative evaluation is shown in Tab. 1. We can see that the proposed method demonstrated superior performance over the competing techniques across all metrics.

5.3 Qualitative evaluation on real data

To underscore the advantages of our method, we conduct a qualitative evaluation on real and synthetic chromatic spike streams. Fig. 7 presents a quadruple comparison. Our approach not only reconstructs motion blur free frames and recovers color information but also achieves the best balance between noise suppression and detail preservation across diverse scenes. In Fig. 7 (a), we create a colorful and dynamic environment by placing a high-speed electric fan between our chromatic spike camera and a set of toys. The reconstruction from TFP ($W = 100$) exhibits noticeable motion blur in the blue bounding box. Conversely, TFI, TFP ($W = 10$), TFSTP and MS23 [21] suffer from significant noise. On the contrary, our method yields a clean outcome. Fig. 7 (b) showcases flowing water in glass and a rotating toy duck. As indicated by red bounding boxes, TFI, TFP ($W = 10$), TFSTP, and MS23 [21] contain different levels of noise. And TFP ($W = 100$) suffers from heavy motion blur in blue bounding box. Meanwhile, our proposed method produce clean reconstruction without blur. Fig. 7 (c, d) further demonstrate the denoising advantage of our method on synthetic data.

6 Conclusion

In this paper, we propose a novel real-data-driven approach for high-speed color HDR video reconstruction from mosaicked chromatic spikes. Through the analysis on the noise of chromatic spikes, we find the noise in short-term frames can be regarded as a zero-mean distribution. According to this characteristic, we design a self-supervised denoising module that can be trained on real-world chromatic spikes. We further propose a progressive warping module that enables the reconstruction of HDR frames from a pseudo long exposure. Experiments on

[7] As the source code for MS23 [21] is not publicly accessible, we conduct re-implementation and modification for the evaluation.

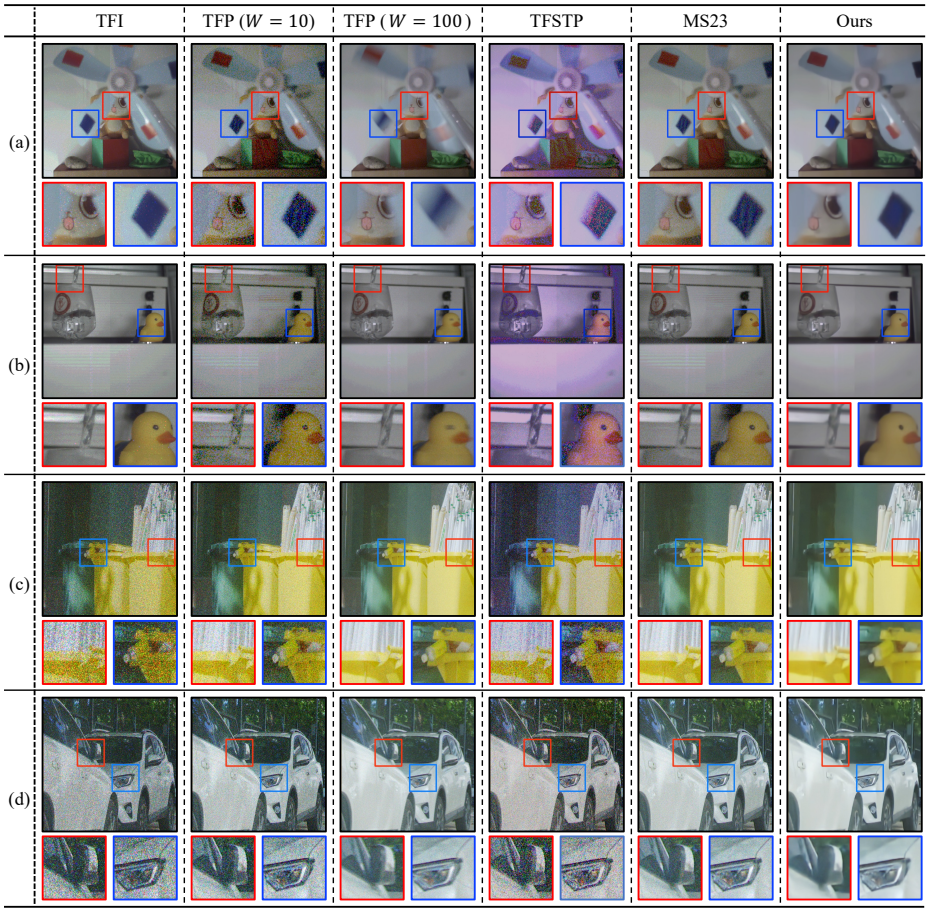


Fig. 7: Visual quality comparison of real (a, b) and synthetic (c, d) data between the proposed method and compared methods. Please zoom-in electronic versions for better details, and watch the videos in the supplementary material.

real-world data demonstrate the superiority of our approach for reconstructing 2000FPS color HDR videos.

Limitations and future work. The Bayer-pattern CFA introduces chromatic information to the spikes but also reduces the intensity of incident light. As a result, compared to monochromatic spikes, the ability to preserve textures in low-light regions is relatively compromised in high-speed conditions. In future research, exploring different types of CFAs to enhance low-light performance would be worth investigating. And the excessive focus on denoising can result in over-smooth downsides, which may lead to the loss of fine details. Future work could involve the development of a more sophisticated denoising module to balance noise suppression and detail preservation.

Acknowledgements

This work was supported by National Science and Technology Major Project (Grant No. 2021ZD0109803) and National Natural Science Foundation of China (Grant No. 62088102, 62136001, 62301009). Bin Fan was supported by National Postdoctoral Program for Innovative Talents of China (Grant No.BX20230013).

References

1. Batson, J., Royer, L.: Noise2Self: Blind denoising by self-supervision. In: ICML. pp. 524–533 (2019)
2. Buades, A., Coll, B., Morel, J.M., Sbert, C.: Self-similarity driven color demosaicking. IEEE TIP **18**(6), 1192–1202 (2009)
3. Chang, Y., Zhou, C., Hong, Y., Hu, L., Xu, C., Huang, T., Shi, B.: 1000 FPS HDR video with a spike-RGB hybrid camera. In: CVPR. pp. 22180–22190 (2023)
4. Chen, C., Mo, J., Hou, J., Wu, H., Liao, L., Sun, W., Yan, Q., Lin, W.: TOPIQ: A top-down approach from semantics to distortions for image quality assessment. IEEE Transactions on Image Processing (2024)
5. Chen, S., Duan, C., Yu, Z., Xiong, R., Huang, T.: Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In: IJCAI. p. 2859–2866 (2022)
6. Chen, S., Yu, Z., Huang, T.: Self-supervised joint dynamic scene reconstruction and optical flow estimation for spiking camera. In: AAAI. vol. 37, pp. 350–358 (2023)
7. Condat, L., Mosaddegh, S.: Joint demosaicking and denoising by total variation minimization. In: ICIP. pp. 2781–2784 (2012)
8. Fossum, E.R., Ma, J., Masoodian, S., Anzagira, L., Zizza, R.: The quanta image sensor: Every photon counts. Sensors **16**(8), 1260 (2016)
9. Gharbi, M., Chaurasia, G., Paris, S., Durand, F.: Deep joint demosaicking and denoising. ACM TOG **35**(6), 1–12 (2016)
10. Gnanasambandam, A., Chan, S.H.: HDR imaging with quanta image sensors: Theoretical limits and optimal reconstruction. IEEE TCI **6**, 1571–1585 (2020)
11. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: CVPR. pp. 1730–1739 (2020)
12. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM TOG **35**(6), 1–12 (2016)
13. Heide, F., Rouf, M., Hullin, M.B., Labitzke, B., Heidrich, W., Kolb, A.: High-quality computational imaging through simple lenses. ACM TOG **32**(5), 1–14 (2013)
14. Huang, T., Zheng, Y., Yu, Z., Chen, R., Li, Y., Xiong, R., Ma, L., Zhao, J., Dong, S., Zhu, L., Li, J., Jia, S., Fu, Y., Shi, B., Wu, S., Tian, Y.: 1000× faster camera and machine vision with ordinary devices. Engineering **25**, 110–119 (2022)
15. Kiku, D., Monno, Y., Tanaka, M., Okutomi, M.: Beyond color difference: Residual interpolation for color image demosaicking. IEEE TIP **25**(3), 1288–1300 (2016)
16. Kokkinos, F., Lefkimmiatis, S.: Deep image demosaicking using a cascade of convolutional residual denoising networks. In: ECCV. pp. 303–319 (2018)
17. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128×128 120 db 15μs latency asynchronous temporal contrast vision sensor. IEEE JSSC **43**(2), 566–576 (2008)

18. Liu, L., Jia, X., Liu, J., Tian, Q.: Joint demosaicing and denoising with self guidance. In: CVPR. pp. 2240–2249 (2020)
19. Liu, Y., Gutierrez-Barragan, F., Ingle, A., Gupta, M., Velten, A.: Single-photon camera guided extreme dynamic range imaging. In: WACV. pp. 1575–1585 (2022)
20. Ma, S., Gupta, S., Ulku, A.C., Bruschini, C., Charbon, E., Gupta, M.: Quanta burst photography. ACM TOG **39**(4), 79–1 (2020)
21. Ma, S., Sundar, V., Mos, P., Bruschini, C., Charbon, E., Gupta, M.: Seeing photons in color. ACM TOG **42**(4), 1–16 (2023)
22. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: ICCV. pp. 2272–2279 (2009)
23. Mairal, J., Elad, M., Sapiro, G.: Sparse representation for color image restoration. IEEE TIP **17**(1), 53–69 (2007)
24. Malvar, H.S., He, L.w., Cutler, R.: High-quality linear interpolation for demosaicing of bayer-patterned color images. In: ICASSP. vol. 3, pp. iii–485 (2004)
25. Mantiuk, R.K., Hammou, D., Hanji, P.: 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)
26. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR. pp. 3883–3891 (2017)
27. Narwaria, M., Da Silva, M.P., Le Callet, P.: HDR-VQM: An objective quality measure for high dynamic range video. Signal Processing: Image Communication **35**, 46–60 (2015)
28. Pekkucuksen, I., Altunbasak, Y.: Gradient based threshold free color filter array interpolation. In: ICIP. pp. 137–140 (2010)
29. Tan, D.S., Chen, W.Y., Hua, K.L.: Deepdemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. IEEE TIP **27**(5), 2408–2419 (2018)
30. Tan, R., Zhang, K., Zuo, W., Zhang, L.: Color image demosaicking via deep residual learning. In: ICME. pp. 793–798 (2017)
31. Wang, Z., Liu, J., Li, G., Han, H.: Blind2unblind: Self-supervised image denoising with visible blind spots. In: CVPR. pp. 2027–2036 (2022)
32. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
33. Xia, L., Zhao, J., Xiong, R., Huang, T.: SVFI: Spiking-based video frame interpolation for high-speed motion. In: AAAI. vol. 37, pp. 2910–2918 (2023)
34. Xing, W., Egiazarian, K.: End-to-end learning for joint image demosaicing, denoising and super-resolution. In: CVPR. pp. 3507–3516 (2021)
35. Yang, Y., Han, J., Liang, J., Sato, I., Shi, B.: Learning event guided high dynamic range video reconstruction. In: CVPR. pp. 13924–13934 (2023)
36. Yu, G., Sapiro, G., Mallat, S.: Solving inverse problems with piecewise linear estimators: From gaussian mixture models to structured sparsity. IEEE TIP **21**(5), 2481–2499 (2011)
37. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
38. Zhang, Z., Zheng, H., Hong, R., Xu, M., Yan, S., Wang, M.: Deep color consistent network for low-light image enhancement. In: CVPR. pp. 1899–1908 (2022)
39. Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T.: Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In: CVPR. pp. 11996–12005 (2021)

40. Zhao, J., Xiong, R., Xie, J., Shi, B., Yu, Z., Gao, W., Huang, T.: Reconstructing clear image for high-speed motion scene with a retina-inspired spike camera. IEEE TCI **8**, 12–27 (2021)
41. Zhao, J., Zhang, S., Ma, L., Yu, Z., Huang, T.: SpikingSIM: A bio-inspired spiking simulator. In: IEEE ISCAS. pp. 3003–3007 (2022)
42. Zheng, Y., Zhang, J., Zhao, R., Ding, J., Chen, S., Xiong, R., Yu, Z., Huang, T.: SpikeCV: Open a continuous computer vision era. arXiv preprint arXiv:2303.11684 (2023)
43. Zheng, Y., Zheng, L., Yu, Z., Huang, T., Wang, S.: Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. IEEE TPAMI **45**(7), 8127–8142 (2023)
44. Zheng, Y., Zheng, L., Yu, Z., Shi, B., Tian, Y., Huang, T.: High-speed image reconstruction through short-term plasticity for spiking cameras. In: CVPR. pp. 6358–6367 (2021)
45. Zhu, L., Dong, S., Huang, T., Tian, Y.: A retina-inspired sampling method for visual texture reconstruction. In: ICME. pp. 1432–1437 (2019)
46. Zhu, L., Dong, S., Li, J., Huang, T., Tian, Y.: Retina-like visual image reconstruction via spiking neural model. In: CVPR. pp. 1438–1446 (2020)

Real-data-driven 2000 FPS Color Video from Mosaicked Chromatic Spikes (Supplemental Material)

Siqi Yang[1,#], Zhaojun Huang[2,3,#], Yakun Chang[5,6], Bin Fan[4], Zhaofei Yu[1], and Boxin Shi[2,3,1,*]

[1] Institute for Artificial Intelligence, Peking University
[2] State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[3] Nat'l Eng. Research Ctr. of Visual Technology, School of Computer Science, Peking University
[4] Nat'l Key Lab of General AI, School of Intelligence Science and Technology, Peking University
[5] Institute of Information Science, Beijing Jiaotong University
[6] Visual Intelligence +X International Cooperation Joint Laboratory of the Ministry of Education
{yousiki, huangzhaojun, binfan, yuzf12, shiboxin}@pku.edu.cn, ykchang@bjtu.edu.cn

In the supplementary material, we provide chromatic spike camera details, method implementation details, analysis of the hyperparameters (K, L, and W in Sec. 4), and additional results. We further provide a supplementary video to show the high-speed color videos reconstructed from mosaicked chromatic spike streams.

7 Mosaicked chromatic spike camera

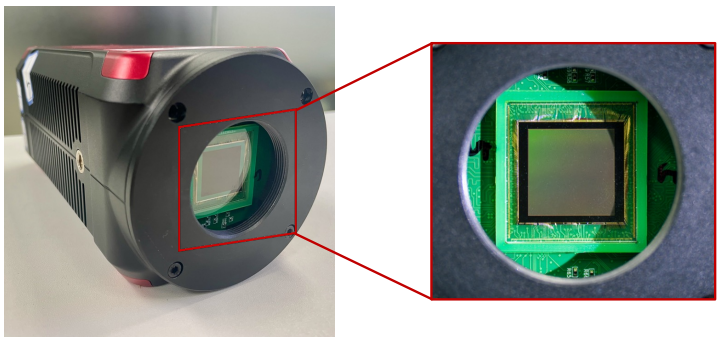


Fig. 8: Mosaicked chromatic spike camera.

We describe more specifics of the mosaicked chromatic spike camera. As we mentioned in Sec. 3, a color filter array (CFA) is applied to the sensor to capture mosaicked chromatic spike streams, adhering to the widely used Bayer pattern (RGGB). The chromatic spike frames are transmitted to the main computer via optical fiber, and are then stored to solid-state drives.

# Equal contributions. * Corresponding author.

8 Implementation details

In this section, we provide implementation details about our method.

Chromatic spikes denoising. During the training stage of our chromatic spike denoiser, we employ a sampling strategy where a subset of the masked volume is randomly selected for each iteration. The loss function remains nearly the same, except for the replacement of mask-aware averaging with summation. In terms of experimental setup, we configured the parameters as follows. We set $h_s = w_s = 4$ for global-aware masking, $\eta = 1$ and $\lambda_i = 0.05$ for loss weighting, and λ gradually increasing from 2 to 20 aligned with the training progress. Furthermore, to accommodate a wide range of signal-to-noise ratios, the accumulation temporal window size W is randomly drawn from the range 5 to 200, enhancing the tolerance of the denoising module to varied noise levels. To address the scarcity of real-world chromatic spike streams, we augment our training data with common means, *e.g.*, randomly flipping and cropping accumulated frames. As previously described in Sec. 4.1, our approach is based on the zero-mean assumption of the noise distribution of chromatic spikes. For the chromatic spike denosing module, we modify the U-Net [4] to restore clean frames from the noise contaminated spike frames. The architecture of our network comprises 5 blocks to extract multi-scale features, and there are 5 blocks in the decoder, which reversely map the multi-scale features to an output video frame. To preserve the texture information in low-level features, we add skip connections between the encoder and decoder. Each block in the encoder and the decoder consist of 2 convolutional layers, and the output of each convolutional layer is activated by LeakyReLU [5]. Thus, the denoising module consists of a total of 25 convolutional layers, including head and tail processing layers.

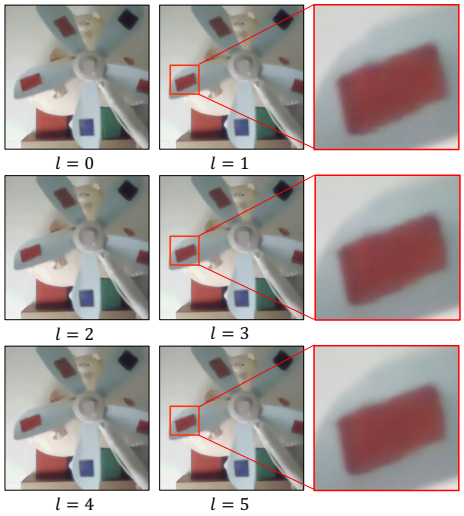


Fig. 9: Additional results of progressive warping. We visualize the intermediate frames corresponding to increasing l (from left to right, from up to down) and enlarge some regions in red bounding boxes for detailed observation.

Progressive warping. In the progressive warping module, we capitalize on the existing method's capacity to align multiple adjacent frames [2]. This approach offers enhanced robustness against potential noise in frames accumulated over short durations from chromatic spike streams, in comparison to other optical flow estimation techniques. The initially accumulated frames, recovered from a

small window size (*e.g.*, $W = 10$), may exhibit some degree of noise even after denoising, which is detrimental to the alignment of small patches. Consequently, we employ larger patch sizes for these initial frames, progressively decreasing the patch size as the frames' reliability improves. In our experiments, setting $L = 3$, $W = 10$, and $K = 3$ is proved sufficient for most of the scenes. To achieve robust progressive warping, we estimate the optical flow from multi-scale maps. At the first step, we obtain multi-scale maps by downsampling the video frames with the scales of $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$. Thus, including the original resolution frame, each group of multi-scale maps contains 4 frames. The process entails progressively searching for the most suitable match from the pyramid maps to references, starting from the lowest resolution and ascending to the highest. The culmination of this process is the identification of the optimal optical flow required to align K frames with the reference frame, specifically the $(K + 1)/2$-th frame. This procedure is elaborated in Sec. 4.2 and represents a singular warping process. The entire progressive warping pipeline is composed of L such warping steps. It is noteworthy that the initial step in this sequence operates on spike planes (S_c), as opposed to intermediate frames ($I_{c,l}$).

Spike simulator. Our chromatic spike simulator mainly follows the noise modeling design of existing works (*e.g.*, [6,7]), including dark-current estimation and perturbed stimulation threshold, and also introduces the simulation of Bayer-pattern CFA.

Inference time. Our chromatic spike denoising module requires approximately 10 hours for training. The proposed method currently functions as an *off-line solver* for 2000 FPS video reconstruction, with an inference speed of 16.7 FPS, in the condition of $L = 3, W = 10$. It is worth noting that the primary time consumption is optical flow estimation ($\sim 65\%$), which is independent from our main pipeline and can be independently optimized. We compare the inference speed of our method with other methods in Tab. 2. All metrics are benchmarked with an RTX3090 GPU, except SJDD [1], which utilizes an A6000 GPU due to its higher memory requirements.

Table 2: Comparison of inference speed.

Method	Ours	TFP	TFI	TFSTP	MS23	SJDD
FPS	16.7	1k	22	13.9	2.5	0.32

9 Analysis of hyperparameters

As discussed in Sec. 8, we empirically found that the set of hyperparameters $L = 3, W = 10, K = 3$ is sufficient for our testing scenes. The three hyperparameters jointly determine the pseudo-long exposure, that is, the exposure time is

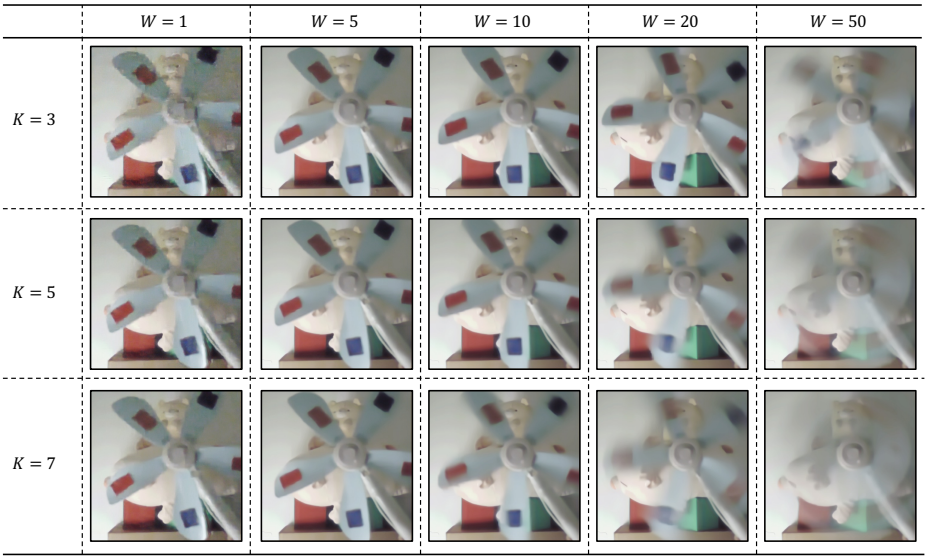


Fig. 10: Analysis of hyperparameters (K and W). We illustrate the reconstruction results from different combinations of K and W.

Fig. 7 (c) Fig. 7 (d) Fig. 12 (e) Fig. 12 (f) Fig. 12 (g) Fig. 12 (h)

Fig. 11: The ground truth of synthetic scenes.

equivalent to $(K-1) \times L \times W + 1$. We further conduct an ablation study to analyze the impact of these hyperparameters on the reconstruction quality, as shown in Fig. 9 and Fig. 10. We adjust the hyperparameters and evaluate the reconstruction results qualitatively. With W and K increasing, our proposed method obtains better performance at static regions, while leading to more potential blur at motion regions. Specifically, small W (*e.g.*, $W \leq 5$) makes optical flow estimation almost unpractical, given that the detected photons are extremely limited. Large W (*e.g.*, $W > 20$) introduces motion blur before flow estimation and warping, leading to irretrievable blurry artifacts. We can empirically conclude that W between 10 and 20 fits most of the cases, both static scenes and dynamic objects. While the increase of W doesn't change the computation time very much (because the dimensions of I remain unchanged), K linearly affects the computation costs. While greater K accumulates more spike planes and suppresses noise better, we observe that K between 3 and 5 is sufficient for most of the cases. As shown in Fig. 5 and Fig. 9, with the increase of L, our

proposed method can refine the reconstruction results with adjacent frames. In our experiment, we find that when $L = 3$, the quality of the reconstructed color image can converge to a stable value. Note that the hyperparameters are not totally fixed and can be customized based on user preference.

10 Additional results

Furthermore, additional results on both real and synthetic data are illustrated in Fig. 12, ground truth images for all synthetic scenes are shown in Fig. 11, and a comparison video is also uploaded with the supplementary material. Please refer to the video for a more comprehensive comparison.

11 Compared to supervised learning on synthetic data

One significant challenge in current spike camera research lies in the substantial disparity between synthetic data and real data, which makes models trained on synthetic data perform poorly in real-world applications. Consequently, we utilize real spike streams for self-supervised training to avoid the domain gap issue in this paper. To substantiate this, we employ an identical neural network to conduct supervised training on the GoPro dataset [3], leveraging the video-to-spike simulators to generate synthetic spike streams. Subsequently, we compare the performance of the supervised-learning (SL) model trained on synthetic data with that of our self-supervised learning (SSL) model trained on real data, as shown in Fig. 13. Our proposed method performs significantly better in evaluation.

12 Compared to concurrent work

As SJDD [1] was published after our submission, it should be treated as concurrent work, and we did not include a comparison in our main figures. We compare our proposed method with SJDD as shown in Fig. 14, which demonstrates the superiority of our method in terms of noise supervision.

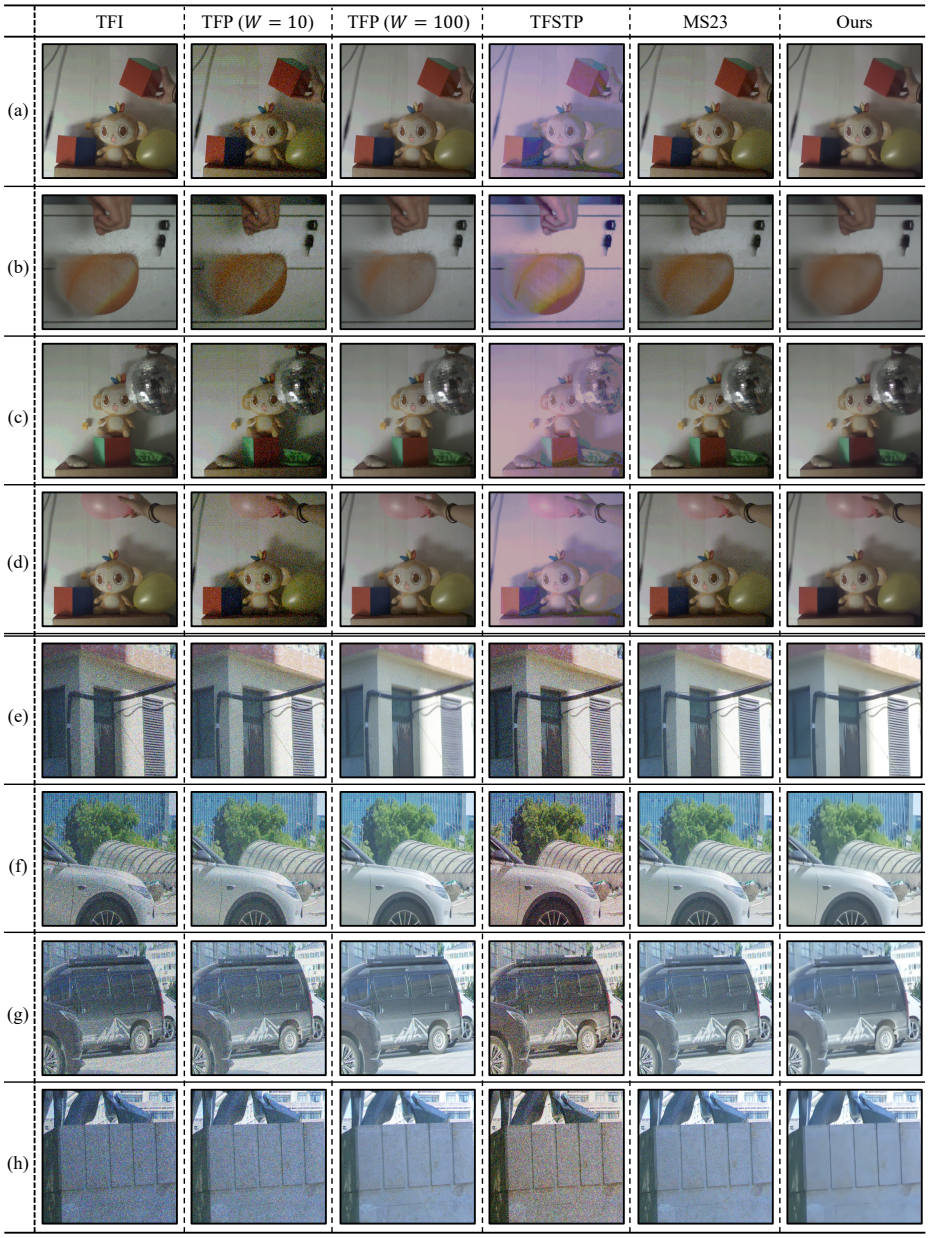


Fig. 12: Additional reconstruction results for visual equality comparison of real (a-d) and synthetic (e-h) data between the proposed method and compared methods.

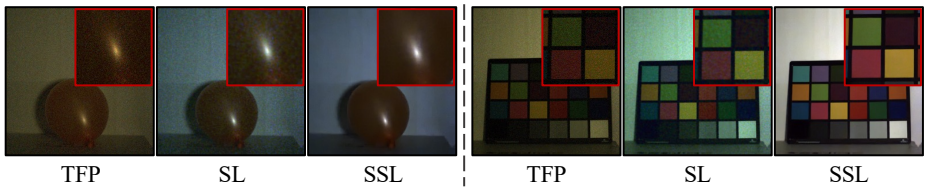


Fig. 13: Comparing the performance of supervised-learning denoiser (SL) and our self-supervised learning denoiser (SSL). The noise in short-term temporal window accumulation (TFP) is better removed by our real-data-driven self-supervised learning denoising module.

Fig. 14: Comparing our method with SJDD [1].

References

1. Dong, Y., Xiong, R., Zhao, J., Zhang, J., Fan, X., Zhu, S., Huang, T.: Joint demosaicing and denoising for spike camera. AAAI (2024)
2. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM TOG **35**(6), 1–12 (2016)
3. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR. pp. 3883–3891 (2017)
4. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241 (2015)
5. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
6. Zheng, Y., Zhang, J., Zhao, R., Ding, J., Chen, S., Xiong, R., Yu, Z., Huang, T.: SpikeCV: Open a continuous computer vision era. arXiv preprint arXiv:2303.11684 (2023)
7. Zheng, Y., Zheng, L., Yu, Z., Huang, T., Wang, S.: Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. IEEE TPAMI **45**(7), 8127–8142 (2023)