Learn to See the Unseen in Low-light Spike Streams

Liwen Hu[1], Yang Li[2], Mianzhi Liu[3], Yijia Guo[1], Shenghao Xie[1], Wenqiang Zu[4], Ziluo Ding[1], Tiejun Huang[1], and Lei Ma[3⋆]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
{huliwen, ziluoding, tjhuang}@pku.edu.cn, 2301112015@stu.pku.edu.cn
[2] School of Software and Microelectronics, Peking University
ly0376@stu.pku.edu.cn
[3] College of Future Technology, Peking University
liumianzhi@stu.pku.edu.cn, lei.ma@pku.edu.cn
[4] University of Chinese Academy of Sciences
zuwenqiang2022@ia.ac.cn

Abstract. Spike camera, a type of neuromorphic sensor with high-temporal resolution, shows great promise for high-speed visual tasks. Unlike traditional cameras, spike camera continuously accumulates photons and fires asynchronous spike streams. Due to unique data modality, spike streams require reconstruction methods to become perceptible to the human eye. However, under low-light high-speed conditions, spike streams become highly sparse and noisy, making faithful pixel-wise recovery intrinsically difficult. In this work, we propose **Diff-SPK**, a diffusion-based framework for perceptual reconstruction from low-light spike streams. Diff-SPK leverages generative priors to produce visually plausible reconstructions while remaining constrained by spike-derived structural conditions. Specifically, it first employs an **E**nhanced **T**exture **f**rom Inter-spike **I**nterval (ETFI) to aggregate sparse structural information from low-light spike streams. Then, the encoded ETFI by a suitable encoder serves as the input of ControlNet for high-speed scenes generation. To improve the quality of results, we introduce an ETFI-based feature fusion module during the generation process.

Keywords: Neuromorphic vision · Reconstruction · Spike camera

1 Introduction

With the development of vehicles, drones, and robotics, visual perception in high-speed scenes have gained significant attention. Since traditional cameras often suffer from motion blur in high-speed scenes, many researchers have started turning to neuromorphic sensors [1, 4, 11, 17, 26]. In this context, spike camera, a type of neuromorphic sensor, shows remarkable potential in visual tasks, e.g.,

⋆ Corresponding author.

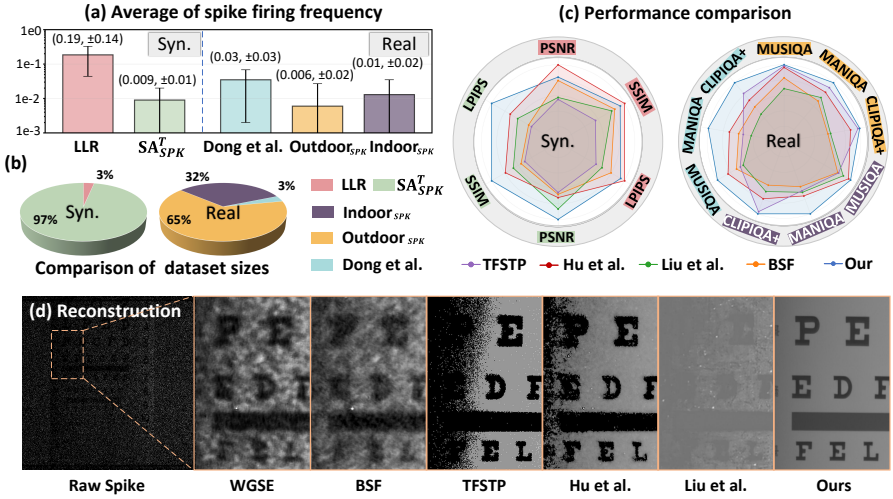

Fig. 1: Analysis of low-light high-speed spike stream datasets and performance comparison of methods. SA^T_{SPK} ($Outdoor_{SPK}$ & $Indoor_{SPK}$) is proposed synthetic (real) dataset. (a) Average Spike Firing Frequency (ASFF): The darker the scene, the lower the ASFF. Our datasets have darker illumination conditions compared to LLR [14] and Dong et al. [8]. (b) Dataset size comparison: Ours are $\sim$ **30** $\times$ larger than others. (c) Performance comparison: The background color of metrics corresponds to different datasets in (b). TFSTP [48], Hu et al. [14]&Liu et al. [29], and BSF [47] represent the state-of-the-art for traditional, deep learning-based low-light, and deep learning-based normal-light reconstruction. (d) Reconstruction: Our method clearly reconstructs letters in the low-light dynamic real scene.

optical flow estimation [16, 40, 46] and depth estimation [25, 28, 42]. Different from traditional cameras, spike camera accumulates photons at an extremely high frame rate for each pixel, and a spike is fired once the accumulated value exceeds the fixed threshold. To perceive high-speed scenes from binary spike data, also known as spike stream, numerous reconstruction methods [3, 14, 41, 44, 47, 50] are proposed and achieve significant progress. However, as illumination intensity diminishes, vision sensors inevitably encounter a fundamental performance bottleneck [9, 10, 23, 24]. The insufficient photon flux coupled with inherent sensor noise results in critically low signal-to-noise ratios (SNR) in captured data. Under such severe degradation, fine-grained appearance details may no longer be uniquely determined by the input spike stream, making strict pixel-wise recovery inherently ill-posed. Existing reconstruction methods based on regression frameworks, e.g., CNNs [22], fail to preserve recognizable visual structures. Recently, diffusion models [13, 31, 43] present a novel paradigm. By leveraging the strong generative priors, diffusion models demonstrate remarkable capability in recovering a perceptually plausible image from severely degraded observations. A natural question arises: Can we use powerful generative models to improve perceptual quality

while preserving consistency with sparse information of low-light high-speed spike streams?

We propose the diffusion-based reconstruction methods, Diff-SPK, to address the limitations of existing methods in low-light high-speed scenes. Diff-SPK is mainly divided into two stages. First, low-light spike stream passes through an **E**nhanced **T**exture **f**rom Inter-spike **I**nterval module (ETFI). ETFI can estimate the illumination of scenes based on spike intervals and enhance excessively dark regions into a suitable range. Next, Diff-SPK utilizes Stable Diffusion (SD) [31] as a generative prior and finetunes a ControlNet [43], where results from ETFI serve as a condition. Results from ETFI need to be encoded into the latent space before being fed to ControlNet. Diffusion models [27,35,38] in image vision tasks often use VAE encoder for this operation. However, VAE encoder is not suitable for data modality of ETFI. To reduce distortion, we introduce an encoding module to extract and compress the features of input condition. Furthermore, in order to improve the guidance of input condition, i.e., ETFI, on the denoising process, a fusion module has been proposed. It can inject the information of condition into the noise latent variable during each denoising timestep. To verify the effectiveness of Diff-SPK, we propose low-light and high-speed spike streams datasets (synthetic & real). As shown in Fig. 1, they have a larger scale and more extreme lighting conditions compared to the existing datasets. Our method performs well on multiple datasets and can be generalized to real low-light high-speed scenes. The contributions in this paper can be summarized as follows:

- We propose Diff-SPK, a diffusion-based framework for perceptual reconstruction from low-light high-speed spike streams. By leveraging generative priors under spike-derived structural constraints, the proposed method produces visually plausible reconstruction results.
- We introduce an **E**nhanced **T**exture **f**rom Inter-spike **I**nterval (ETFI) representation, together with a tailored conditional encoding module and an ETFI-based feature fusion module, to improve structural conditioning and strengthen guidance during the denoising process.
- We propose first bona fide benchmark for low-light high-speed spike streams reconstruction. It has a larger scale ($\sim$ **30**$\times$) and more challenging illumination conditions than the existing dataset. Experiments show Diff-SPK improves perceptual quality and robustness to severe low-light degradation.

2 Related Work

2.1 Reconstruction for Spike Camera

Traditional method Early spike-based reconstruction methods, TFI and TFP [49], build on the unique statistical characteristics of spike streams. After that, TFSTP [48] introduces neurobiological-inspired short-term plasticity to imporve reconstruction of motion region. Subsequent research by Dong et al. [8] expand the application scope to challenging low-light, high-speed environments.

Deep learning-based method To improve the performance of traditional methods, Zhao et al. [44] introduced Spk2ImgNet (S2I), the first deep learning-based reconstruction framework. SSML [3] pioneers a self-supervised architecture integrating blind-spot networks, demonstrating similar performance to S2I. Concurrently, Zhang et al. [41] develops the WGSE framework that employs hierarchical wavelet transforms for systematic noise suppression in reconstructed results. Zhao et al. [47] further improves reconstructed results by using the framework of transformer. Dong et al. [5–7] develop an image reconstruction network for color spike cameras. Hu et al. [14] uses RNN [39] to better handle low-light spike streams. However, the performance of the former can severely deteriorate under extreme low-light conditions. Liu et al. [29] try to explore diffusion models for reconstruction. However, this method struggles to generalize across high-speed and diverse low-light conditions because its input is based on the average number of spikes in a large temporal window, introducing severe motion blur. In this work, we present the tailored diffusion-based reconstruction method, Diff-SPK, unlocking the potential of spike cameras in low-light high-speed scenarios.

2.2 Diffusion for Low-level Vision

The emergence of diffusion models [13, 31, 43] has provided new insights for low-level vision tasks, such as image enhancement [18, 19, 37], dehazing [30, 36], and super-resolution [2, 12, 27, 35, 38]. Recent studies leveraging powerful generative models have demonstrated remarkable performance in human visual perception. Many approaches, e.g., SUPIR [38] and DiffBIR [27], typically employ conventional networks to obtain preliminary high-quality images, which are then refined through conditional diffusion models to improve details. They find VAE encoder particularly effective for comprehensive feature extraction and compression when high-quality images are encoded as condition. However, these established paradigms in above methods are inapplicable to spike streams due to data modality disparities.

3 Preliminary

3.1 Spike Camera Model

Spike camera [17] is a kind of neuromorphic sensor with high-temporal resolution. In spike camera, each pixel initially transforms the incoming light signal into a corresponding current signal. This current signal serves as the foundational input for the sensor. As shown in Fig. 2, spike camera can continuously accumulate current. When the accumulation at a pixel location $\boldsymbol{x} = (x, y)$ attains a predefined threshold ϕ, a spike is fired, and the accumulation is reset. This behavior is mathematically represented as,

$$A(\boldsymbol{x}, t) = \int_0^t I_{in}(\boldsymbol{x}, \tau) d\tau \bmod \phi, \tag{1}$$

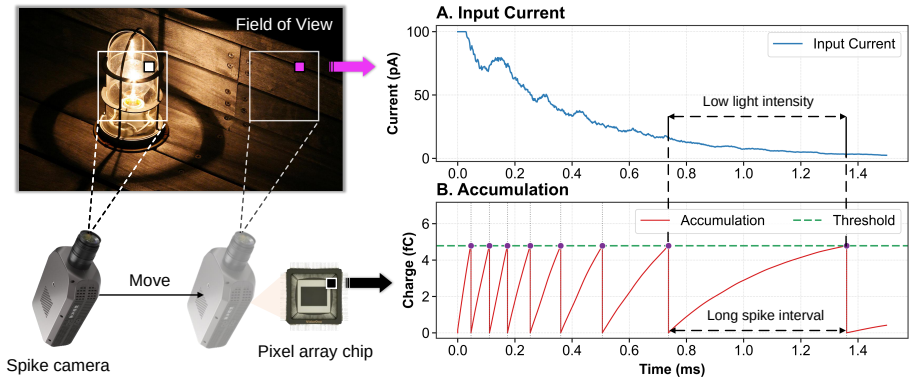

Fig. 2: Illustration of spike camera model. (A) Input current. (B) Accumulation. Low-light intensity leads to long intervals between adjacent spike, i.e., sparse information.

where $A(\boldsymbol{x}, t)$ denotes the accumulated current at time t, and $I_{\text{in}}(\boldsymbol{x}, \tau)$ represents the input current at time τ, which is directly proportional to the incident light intensity. Additionally, due to hardware constraints, the spikes are sampled at discrete intervals nT, where $n \in \mathbb{N}$ and T is on the order of microseconds. Consequently, the output of the spike camera manifests as a binary spatio-temporal sequence $\{S_i \mid i = 0, 1, 2, \dots\}$, with each $S_i \in \{0, 1\}^{W \times H}$ corresponding to the i-th spike frame. Here, H and W denote the resolution's height and width, respectively.

3.2 Diffusion Model

Diffusion models [13] primarily consist of a forward process and a reverse process. By introducing text or images during above processes, the diffusion model can produce high-quality images that adhere to the specified conditions. Our method is based on conditional diffusion models, i.e., **L**atent **D**iffusion **M**odel (LDM) [31] with **C**ontrolNet [43] (LDMC). LDMC is performed in the latent space. Specifically, the image x_0 is first encoded into the latent space as $z_0 = \mathcal{E}(x_0)$. Furthermore, in the forward process, the noisy latent variable at timestep t, denoted as z_t, can be expressed as,

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \tag{2}$$

where $\alpha_t = 1 - \beta_t$ is the noise schedule, $\beta_t \in [0, 1)$, $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise. In the reverse process, the noisy latent z_t can be denoised to z_{t-1} as,

$$z_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(z_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(z_t, t, c, c_{text}) \right) + \sigma_t \epsilon, \tag{3}$$

where $\epsilon_\theta(\cdot)$ is the learned noise predictor, c is the condition input to ControlNet and c_{text} the text prompt for LDM. After obtaining the clean latent z_0, the

generated image can be decoded from z_0, i.e., $y = \mathcal{D}(z_0)$. The optimization of the noise predictor is defined as follows,

$$\mathcal{L} = \mathbb{E}_{z_t,t,c,c_{text},\epsilon}\left[\|\epsilon_\theta(z_t, t, c, c_{text}) - \epsilon\|_2^2\right], \tag{4}$$

where t is randomly sampled timestep, (z_t, c, c_{text}) is from training data (z_t is generated according to the forward process), and $\|\cdot\|_2$ denotes the ℓ_2-norm.

4 Method

Our goal is to perform perceptual reconstruction from low-light high-speed spike streams by combining spike-derived structural conditions with the generative prior of diffusion models. As shown in Fig. 3, we propose a tailored method, Diff-SPK, which uses the Latent Diffusion Model with ControlNet (LDMC) as the architecture. Next, we first briefly introduce the problem statement, and then elaborate on the motivation and details of the main modules including **E**nhanced **T**exture **f**rom Inter-spike **I**nterval (ETFI), condition encoding and fusion module.

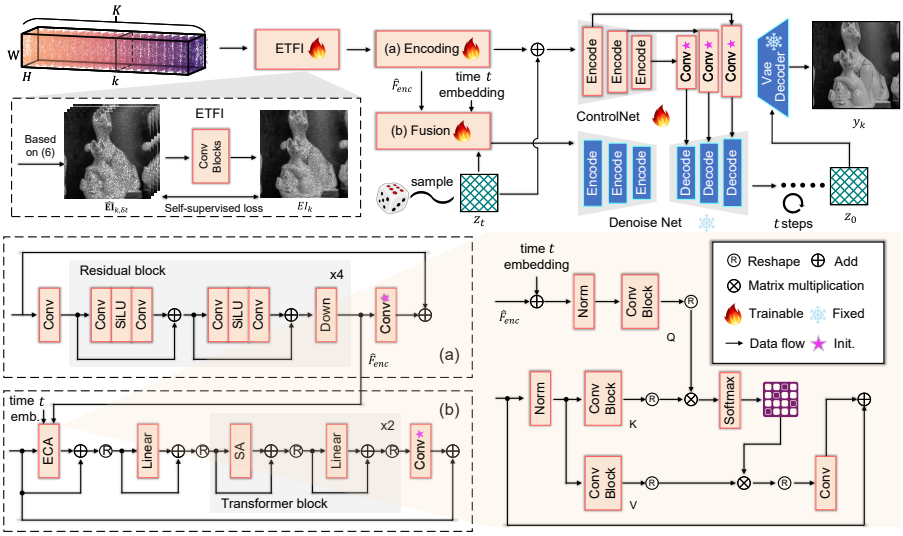

Fig. 3: Framework of Diff-SPK. Diff-SPK first uses ETFI to compute and enhance light intensity information from a spike stream. The condition is through encoding (a) before being fed into ControlNet. To improve the guidance on denoising process, we introduce a fusion module (b).

4.1 Problem Statement

We denote $\mathbf{S}_{k,\delta t} = \{S_{k+i} \in \{0,1\}^{W\times H} \mid i \in \mathbb{Z}, -\delta t \le i \le \delta t\}$ as a spike stream within a temporal window $2\delta t$. The reconstruction of spike streams can be expressed as,

$$y_k = \mathcal{M}(\mathbf{S}_{k,\delta t}), \tag{5}$$

where y_k is the light intensity information at the k-th frame and $\mathcal{M}$ is reconstruction model. In low-light high-speed scenes, the input spike stream is typically sparse and noisy. Therefore, the objective is not to assume exact recovery of all missing details, but to infer a visually plausible image under the structural constraints provided by the observed spike stream.

4.2 ETFI

Motivation Diff-SPK builds a two-stage reconstruction pipeline. We first derive a coarse but stable spike stream representation and then use the representation as a condition to generate the final reconstruction. To obtain light intensity information, a straightforward approach is to employ a reconstruction network. However, in extremely low-light settings, the outputs of existing methods often degrade irregularly across illumination levels, which makes them less suitable as conditioning signals for generative models. In contrast, we seek a representation whose degradation is predictable and structurally consistent, even if it remains noisy or visually imperfect. We propose Enhanced Texture from Inter-spike Interval (ETFI) that computes and enhances light intensity information in low-light spike streams by fully exploiting temporal features. ETFI does not aim to serve as the final reconstruction. Instead, it provides a spike-derived structural condition for subsequent diffusion-based refinement.

Details The proposed ETFI explicitly encodes and enhances light intensity information from low-light spike streams. Specifically, ETFI first uses global inter-spike intervals (ISI) to compute light intensity, following prior methods [14, 49]. The operation can fully extract the temporal information of spike streams. However, under low-light illumination, some ISI are so long that the encoded light intensity becomes imperceptible to human observers when visualized as images. To address this, ETFI enhances the light intensity using the maximum ISI. Assuming the output images are represented with 8-bit intensity values, i.e., 256 gray levels, the enhanced light intensity can be written as,

$$\hat{EI}_k = \frac{max(ISI_k)}{ISI_k}, \tag{6}$$

where $\hat{EI}_k$ is enhanced light intensity information at the k-th spike frame, ISI_k is a matrix composed of ISI in all pixels at the k-th spike frame. The advantage of this operation lies in ensuring that the minimum light intensity information is not be lost in the image. Note that the above operation may cause some pixels to be overexposed (exceeding 256), and we will further adjust the overall brightness (see appendix). Besides, to incorporate additional temporal evidence and reduce quantization noise, we stack the local $\hat{EI}_k$ within a temporal window as $\hat{\mathbf{EI}}_{k,\delta t} \in \mathbb{R}^{T\times H\times W}$ where $T = 2\delta t + 1$. A convolution block is then used to compress this temporal stack into a single-channel aggregated representation:

$$\tilde{EI}_k = ConvB(\hat{\mathbf{EI}}_{k,\delta t}), \quad \tilde{EI}_k \in \mathbb{R}^{1\times H\times W}. \tag{7}$$

The convolution block is trained with the self-supervised objective, i.e.,

$$\mathcal{L}_{ETFI} = \mathbb{E}_{\hat{\mathbf{EI}}_{k,\delta t}} \|\hat{\mathbf{EI}}_{k,\delta t} - \mathcal{B}(\tilde{EI}_k)\|_2^2, \tag{8}$$

where $\mathcal{B}(\cdot)$ denotes channel-wise broadcasting along the temporal dimension.

4.3 Condition Encoding

Motivation Diff-SPK is based on a popular architecture, the Latent Diffusion Model with ControlNet, LDMC. For LDMC, previous work in the image vision task [27, 35, 38] has demonstrated that the Variational Autoencoder (VAE) is well-suited to encode input images and can effectively improve generation results. However, through observation and analysis, we find that the VAE encoder are not suitable for processing results from ETFI, i.e., a large amount of spatial information is lost after encoding (see Fig. 4(b)). Hence, it is necessary to introduce an ETFI encoding module.

Details To better preserve the spatial structures in ETFI, as shown in Fig. 3(a), we propose a lightweight condition encoding module. It uses residual structures to effectively capture key information from ETFI, i.e.,

$$\mathcal{F}_{enc} = Conv_{init}(\underbrace{Res(Conv(E\tilde{I}_k))}_{\hat{\mathcal{F}}_{enc}}) + E\tilde{I}_k \tag{9}$$

where $Res(\cdot)$ denotes residual blocks, $Conv_{init}$ represents a convolutional layer initialized with zero weights and $\hat{\mathcal{F}}_{enc}$ represents the feature fed into fusion module. Compared with directly using the VAE encoder, our encoder can better preserve structurally informative cues.

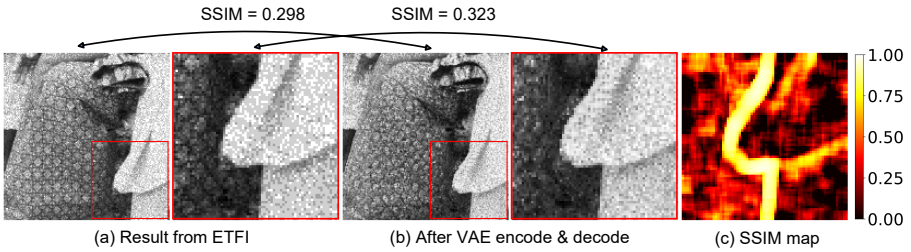

Fig. 4: The influence of VAE encoder on ETFI. The SSIM map visualizes the per-pixel (local-window) structural similarity between (a) and (b). SSIM ranges from 0 to 1 and the larger the value, the more similar they are.

4.4 Fusion Module

Motivation In the standard ControlNet design, condition features are primarily injected into the denoising U-Net through decoder-side residual connections [43]. While effective for many image-based conditions, this design can be insufficient for ETFI, since the condition guidance is introduced relatively late and may not adequately constrain the full denoising trajectory. As a result, the model may rely excessively on generative priors, especially under severe low-light degradation where the input information is highly sparse.

Details To strengthen structural guidance throughout generation, as shown in Fig. 3(b), we propose a fusion module to directly inject the information of condition, i.e., ETFI, into the noise latent variable z_t at different timesteps. The module first employs a **E**TFI-guided **c**ross-**a**ttention (ECA) and a linear layer to fuse ETFI, timestep t and z_t where queries are from z_t and t, and key and value are from the feature of ETFI, $\hat{\mathcal{F}}_{enc}$. The process can be written as,

$$\mathcal{F}_{ECA} = Linear(\underbrace{ECA(t_{emb}, \hat{\mathcal{F}}_{enc}, z_t)}_{\hat{\mathcal{F}}_{ECA}} + z_t)) + \hat{\mathcal{F}}_{ECA} \tag{10}$$

where t_{emb} is the encoding of time t. Then, the feature $\mathcal{F}_{ECA}$ are through transformer blocks [32] to further dig the information. Finally, fused features can be fed to U-Net for denoising. The above process can be written as,

$$\mathcal{F}_{fuse} = Conv_{init}(Trans(\mathcal{F}_{ECA})) + z_t \tag{11}$$

where $Trans(\cdot)$ denotes transformer blocks. Furthermore, the optimization of the noise predictor in (4) can be rewritten as,

$$\mathcal{L} = \mathbb{E}\left[\|\epsilon_\theta(\mathcal{F}_{fuse}, t, \mathcal{F}_{enc}, c_{text}) - \epsilon\|_2^2\right], \tag{12}$$

5 Dataset

Datasets are important for both evaluation (testing set) and optimization (training set) of models. As shown in Table 1, we propose the first bona fide benchmark. Next, we will introduce these datasets respectively.

Name	Data Kind	Class	Frames / Class	Resolution	Motion	Illum.	GT
Training - SA_{SPK}	Synthetic	111, 860	$\approx$ 256	(512, 512)	Random movement	✗	✔
Testing - SA^T_{SPK}	Synthetic	300	$\approx$ 2,000	(512, 512)	Random movement	✗	✔
Testing - $Outdoor_{SPK}$	Real	90	$\approx$ 20,000	(1000, 1000)	Driving	✗	✗
Testing - $Indoor_{SPK}$	Real	50	$\approx$ 20,000	(1000, 1000)	Lens shake	✔	✗

Table 1: Details of high-speed low-light spike stream datasets. Illum. (GT) denotes illumination information (scenes images).

5.1 Training Dataset

We randomly sample 111,860 images (cropped to 512×512) from the SAM dataset [21] as ground truth in the training set. Further, these high-quality images are used to synthesize spike streams. The implementation comprises the following steps: (a) Degrade image quality through interpolation. (b) Simulate spike streams from the processed images. Notably, our simulation consider both the primary noise patterns of spike camera [15,45] and hot pixel noise. (c) Convert each spike stream into ETFI to serve as input for Diff-SPK. See the appendix for more simulation details.

5.2 Testing Dataset

The captured scenarios by spike camera typically involve high-speed motion, making it infeasible to obtain clear images as ground truth (GT) for real-world datasets. Therefore, we have introduced both synthetic and real-world data as testing datasets. The synthetic data is used to evaluate the fidelity of reconstruction methods, while the real-world data is employed to evaluate the generalization of methods.

Synthetic data Beyond the training set scenarios, we randomly sample 300 scenes from the SAM dataset [21]. To simulate low-light conditions, each image was darkened by multiplying a random decay factor. For high-speed motion, we first randomly generated a motion trajectory, and then the cropping box moves along this path. We denote the dataset as SA^{T}_{SPK}. Compared to existing high-speed low-light dataset, LLR [14], our dataset features a larger scale and more extreme illumination conditions as shown in Fig. 1.

Real data We use the spike camera with 1000×1000 resolution to establish a low-light high-speed benchmark, covering both outdoor and indoor scenes. For indoor scenes (Indoor_{SPK}), by adjusting the light source power in Fig. 5(b), we control spike camera to record 5×10 spike streams under different illumination intensities, where 5 (10) denotes illumination levels (scene categories). Each spike stream is record for 1 second, corresponding to 20,000 frames. For outdoor scenes (Outdoor_{SPK}), due to the difficulty in quantitatively controlling illumination conditions, we only collect qualitative low-light data. Specifically, spike camera is placed on a high-speed driving car to capture street scenes at dusk. The outdoor dataset contains 90 spike streams, each with a 1-second recording duration. As shown in Fig. 1, the size of our datasets is ∼ 30 × that of the existing low-light high-speed real dataset [8]. Moreover, our illumination conditions are much lower than [8].

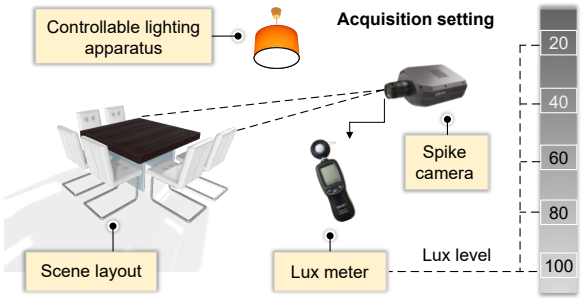

Fig. 5: Instructions for obtaining illumination in Indoor_{SPK}. We collect five different low-light conditions (20/40/60/80/100 Lux) for each dynamic scene.

6 Experiment

6.1 Implementation detail

Diff-SPK is based on the popular architecture, the Latent Diffusion Model with ControlNet (LDMC). We use SD1.5 [31] as our Latent Diffusion Model and ControlNet is initialized with control_v11f1e_sd15_tile.ckpt. We first pre-train the ETFI module for 6,000 iterations. Then, we freeze the ETFI module and

finetune ControlNet, encoding module and fusion module for 30,000 iterations (batch size=96). The learning rate is set to 10^{-5} and the optimizer is AdamW with default parameters. The training process is conducted on 4 NVIDIA A800-80Gb GPUs. For inference of Diff-SPK, we use DDPM scheduler [13] with 50 sampling steps. We use classifier-free guidance (CFG) and CFG scale is set as 2. The positive prompt is set to empty and negative prompts are set as "low quality", "unsharp", "noisy", "weird textures", "artifacts", etc.

6.2 Reconstruction Results

We compare the reconstruction methods including traditional methods (TFI [49], TFP [49], TFSTP [48]) and state-of-the-art deep learning-based methods (SSML [3], S2I [44], WGSE [41], Hu et al. [14], BSF [47] and Liu et al. [29]). These methods are evaluated on low-light high-speed datasets including real-world and synthetic data. For real-world datasets, clear ground-truth images are infeasible to obtain due to the high-speed capture setting. Therefore, we report no-reference image quality assessment (NRIQA) metrics, i.e., MUSIQA [20], MANIQA [34] and CLIP-IQA$^+$ [33], which are popular metrics in the image restoration task. For synthetic datasets, we use the reference image quality assessment metrics (PSNR, SSIM and LIPIS) and NRIQA metrics (MUSIQA, MANIQA and CLIP-IQA$^+$). All methods are tested using the official source code and model weights. Besides, retraining on our training set does not improve baseline methods. Related details are in the appendix.

Dataset	Metric	Traditional Method			Deep learning-based Method						
		TFI	TFP	TFSTP	SSML	S2I	WGSE	BSF	Hu et al.	Liu et al.	Ours
LLR [14]	PSNR (↑)	31.409	32.443	24.882	38.432	36.435	36.094	35.877	**45.075**	26.018	37.898
	SSIM (↑)	0.723	0.793	0.555	0.899	0.835	0.843	0.841	**0.987**	0.796	0.923
	LPIPS (↓)	0.360	0.359	0.441	0.185	0.231	0.225	0.222	**0.022**	0.355	0.079
	MUSIQA(↑)	41.445	35.322	43.667	41.560	46.171	44.603	49.614	49.927	32.472	**52.780**
	MANIQA(↑)	0.328	0.210	0.343	0.287	0.362	0.324	0.367	0.371	0.257	**0.391**
	CLIP-IQA$^+$ (↑)	0.415	0.304	0.475	0.399	0.443	0.417	0.460	0.470	0.308	**0.513**
SA^T_{SPK}	PSNR(↑)	21.909	19.315	15.315	16.194	15.043	17.571	15.898	16.876	20.346	**23.499**
	SSIM(↑)	0.403	0.261	0.342	0.182	0.393	0.378	0.381	0.542	0.460	**0.683**
	LPIPS (↓)	0.651	0.785	0.688	0.710	0.660	0.642	0.655	0.439	0.527	**0.223**
	MUSIQA(↑)	39.090	39.049	36.175	41.314	38.551	42.005	31.453	36.551	28.394	**56.565**
	MANIQA(↑)	0.298	0.337	0.194	0.241	0.229	0.266	0.266	0.281	0.250	**0.359**
	CLIP-IQA$^+$ (↑)	0.301	0.341	0.368	0.261	0.246	0.376	0.284	0.328	0.235	**0.492**
$Outdoor_{SPK}$	MUSIQA(↑)	42.755	29.314	48.690	41.396	36.554	33.192	41.056	47.777	34.146	**49.494**
	MANIQA(↑)	0.377	0.230	0.374	0.329	0.269	0.267	0.281	0.342	0.305	**0.408**
	CLIP-IQA$^+$ (↑)	0.402	0.347	0.469	0.338	0.319	0.274	0.278	0.422	0.295	**0.480**
$Indoor_{SPK}$	MUSIQA (↑)	33.345	24.736	39.310	38.212	37.138	33.366	38.766	46.559	42.737	**48.046**
	MANIQA (↑)	0.345	0.233	0.297	0.307	0.362	0.263	0.272	0.328	0.306	**0.436**
	CLIP-IQA$^+$ (↑)	0.390	0.344	0.484	0.344	0.324	0.292	0.301	0.395	0.345	**0.499**
Dong et al. [8]	MUSIQA (↑)	35.584	34.338	36.286	34.461	37.999	45.440	39.483	46.456	33.917	**55.268**
	MANIQA (↑)	0.199	0.199	0.256	0.234	0.281	0.317	0.284	0.342	0.206	**0.474**
	CLIP-IQA$^+$ (↑)	0.248	0.333	0.368	0.242	0.259	0.311	0.266	0.317	0.210	**0.450**

Table 2: Quantitative comparison on low-light high-speed datasets. Red / Orange / Yellow indicate best / second / third performance.

Synthetic datasets On LLR, Diff-SPK does not achieve the best scores on all reference metrics. This is expected because LLR contains relatively brighter scenes (see Fig. 1), and some samples are close to normal-light conditions [14]. In such cases, the input spike streams contain moderate information, allowing regression-based methods [14,41,44,47,50] to achieve pixel-wise fidelity. In contrast, Diff-SPK is designed to improve perceptual quality under severe degradation by leveraging generative priors. As a result, it may trade some pixel-level fidelity for better perceptual quality, which is reflected in its superior NRIQA performance as shown in Table 2. For SA^T_{SPK}, its scenes are all dim and the regression-based methods are difficult to reconstruct scenes from the low-quality sparse spike streams. In this context, the fine details produced by diffusion-based methods are vital. As shown in Fig. 6, our method has demonstrated obvious advantages. Besides, although Liu et al. [29] also use the diffusion model. Limited by its design, this method is subject to a large amount of motion blur in high-speed scenes.

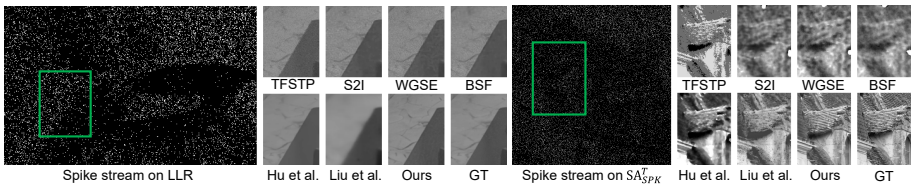

Fig. 6: Visual comparison for reconstruction on LLR [14] and SA^T_{SPK}. Due to low-light conditions, we perform brightness alignment between the reconstruction results and ours for viewing. **Better viewed when zoomed in.**

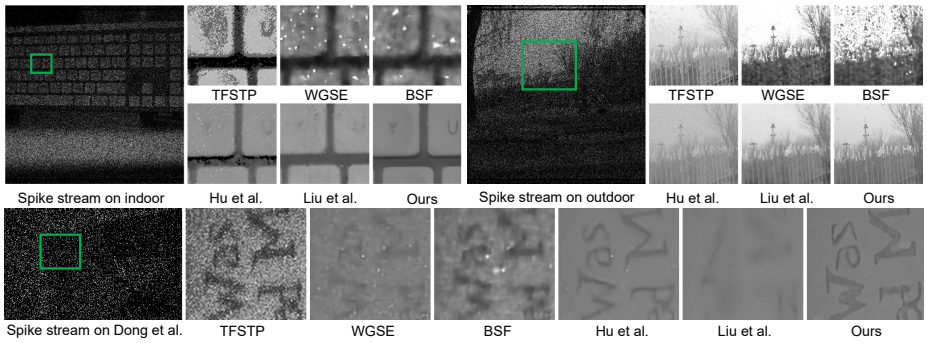

Fig. 7: Visual comparison for reconstruction on Indoor_{SPK}, Outdoor_{SPK}, and Dong et al. **Better viewed when zoomed in.**

Real datasets For Outdoor_{SPK}&Indoor_{SPK}, as shown in Table 2, we can find that Diff-SPK has achieved an overwhelming advantage. Fig. 7 shows the reconstruction results on the two real datasets. For Outdoor_{SPK}, our method clearly restore branches and flags. while the branches reconstructed by Liu et al. have obvious motion blur. For Indoor_{SPK}, our method detects the letters on the keyboard that are "invisible" by other state-of-the-art methods.

For Dong et al. [8], the dataset was captured by Spike Camera-001T-Gen2, and it differs significantly from the spike camera used in this paper, e.g., resolution

and noise level. As shown in Table 2, our method still has the best overall performance. This means that our method can be well generalized to different version of spike camera. Fig. 7 shows the visualization results. We find that Liu et al. [29] suffers from obvious motion blur. Only our method and Hu et al. [14] can restore all letters and the results of our method are clearer.

6.3 Ablation

Different illumination To verify that the advantage of Diff-SPK is not limited to a single illumination setting, we further evaluate it under multiple illumination levels in Indoor$_{SPK}$. As shown in Table 3, Diff-SPK consistently outperforms competing methods across 40, 60, and 80 Lux. These results indicate that the proposed method remains effective across a broad range of low-light conditions, rather than benefiting only from a particular illumination condition.

Method	∼ 40 Lux			∼ 60 Lux			∼ 80 Lux		
	MUS.(↑)	MAN.(↑)	CLIP.$^+$(↑)	MUS.(↑)	MAN.(↑)	CLIP.$^+$(↑)	MUS.(↑)	MAN.(↑)	CLIP.$^+$(↑)
TFI	32.251	<u>0.338</u>	0.389	34.702	<u>0.350</u>	0.396	36.245	<u>0.358</u>	0.410
TFP	23.716	0.238	0.333	23.458	0.234	0.327	25.915	0.240	0.351
TFSTP	38.862	0.287	<u>0.490</u>	36.160	0.271	<u>0.475</u>	37.494	0.292	<u>0.479</u>
SSML	36.580	0.304	0.327	37.449	0.303	0.342	35.783	0.301	0.343
S2I	36.522	0.257	0.321	39.465	0.266	0.317	30.279	0.266	0.335
WGSE	32.502	0.243	0.286	35.599	0.277	0.301	34.162	0.271	0.305
BSF	38.131	0.264	0.290	38.702	0.270	0.301	38.588	0.284	0.335
Hu et al.	<u>47.056</u>	0.317	0.447	<u>48.010</u>	0.314	0.383	<u>45.980</u>	0.326	0.382
Liu et al.	44.317	0.307	0.343	42.101	0.286	0.341	44.377	0.311	0.364
Ours	**48.223**	**0.445**	**0.506**	**50.057**	**0.429**	**0.503**	**50.506**	**0.448**	**0.520**

Table 3: Performance comparison of different methods under varying illumination conditions. The best/ second result in each column is bolded/underlined. MUS./MAN./CLIP.$^+$ denotes MUSIQA/MANIQA/CLIPIQA$^+$.

Method	Indoor$_{SPK}$			Outdoor$_{SPK}$			SA$^T_{SPK}$		
	MUS.(↑)	MAN.(↑)	CLIP.$^+$(↑)	MUS.(↑)	MAN.(↑)	CLIP.$^+$(↑)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
Baseline$_1$	46.268	0.375	0.430	**50.895**	0.316	0.411	15.605	0.429	0.474
Baseline$_2$	39.917	0.372	0.444	47.134	**0.419**	<u>0.465</u>	21.891	0.583	0.353
Baseline$_3$	<u>47.867</u>	<u>0.408</u>	<u>0.471</u>	48.955	0.381	0.451	<u>23.365</u>	<u>0.679</u>	<u>0.233</u>
Ours	**48.046**	**0.436**	**0.499**	<u>49.494</u>	<u>0.408</u>	**0.480**	**23.499**	**0.683**	**0.223**

Table 4: The influence of ETFI module (Baseline$_1$), encoder module (Baseline$_2$), and fusion module (Baseline$_3$) on the proposed datasets. Bold/underline numbers indicate the best/second results in each column. All methods are trained with the **same** configuration. MUS./MAN./CLIP.$^+$ denotes MUSIQA/MANIQA/CLIPIQA$^+$.

ETFI module As mentioned in Sec. 4.2 (motivation), we study the baseline, i.e., first applying a state-of-the-art reconstruction method [14] of low-light high-speed spike streams to obtain an initial image, followed by refinement using Diff-SPK. The results, denoted as Baseline$_1$ in Table 4, demonstrate that employing the reconstructed results as the condition leads to a performance degradation. This is because the method shows highly irregular degradation

patterns under different low-light conditions, which adversely affects Diff-SPK's inference capability.

Encoder module Baseline$_2$ replaces our tailored condition encoder with a VAE encoder that is fine-tuned for ETFI conditioning. As shown in Table 4, this variant consistently underperforms our full model. We attribute this to the intrinsic modality mismatch: ETFI is a spike-derived structural representation with strong quantization artifacts and non-natural-image statistics. VAE tends to distort structurally information leading to degraded perceptual quality.

Fusion module We investigate the impact of the fusion module on Diff-SPK. As shown in Table 4 (Baseline$_3$), Diff-SPK without the fusion module inferior performance compared to our full method. As discussed in Sec. 4.4 (motivation), this performance gap stems from the fact that ControlNet only injects conditional information, i.e., ETFI, in the decoding stages of the U-Net, thereby weakening the guidance of ETFI throughout the denoising process.

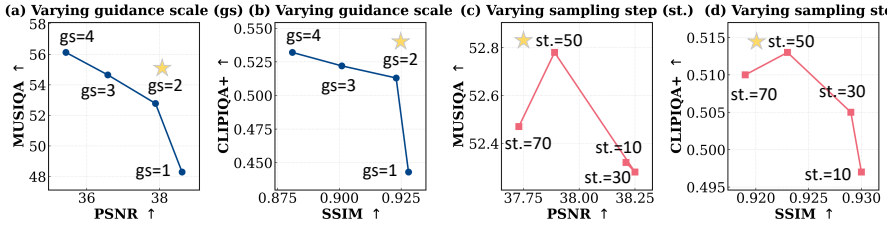

Fig. 8: Pareto curve of sampling steps and guidance scale on LLR [14] dataset. Stars denote our setting.

Hyperparameter trade-off analysis Fig. 8 illustrates the fidelity-perceptual trade-off induced by diffusion sampling hyperparameters. When increasing the guidance scale (gs), we observe a clear improvement in perceptual metrics (MUSIQA/CLIPIQA$^+$), while reference fidelity (PSNR/SSIM) tends to degrade. Similarly, varying the sampling step (st) changes the balance between enforcing the conditional signal and maintaining perceptual naturalness. We choose the starred setting as a practical operating point that provides a favorable trade-off between fidelity and perceptual quality.

7 Conclusion

We present **Diff-SPK**, a diffusion-based framework specifically designed for the perceptual reconstruction from low-light spike streams. Combined with generative priors under spike-derived structural constraints, our method effectively improves perceptual quality and visual interpretability under different low-light conditions. Besides, we establish the first bona fide benchmark for low-light high-speed spike stream reconstruction which has a larger scale than the existing real dataset ($\sim$ **30**$\times$). Experiments show that Diff-SPK produces more structurally consistent reconstructions than existing methods in challenging low-light high-speed scenes.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant No. 62088102), and the Beijing Natural Science Foundation (Grant Nos. F251020 and JQ24023).

References

1. Brandli, C., Berner, R., Yang, M., Liu, S.C., Delbruck, T.: A 240 × 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. IEEE Journal of Solid-State Circuits (JSSC) **49**(10), 2333–2341 (2014) 1
2. Chen, B., Li, G., Wu, R., Zhang, X., Chen, J., Zhang, J., Zhang, L.: Adversarial diffusion compression for real-world image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 28208–28220 (June 2025) 4
3. chen, S., Duan, C., Yu, Z., Xiong, R., Huang, T.: Self-supervised mutual learning for dynamic scene reconstruction of spiking camera. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI. pp. 2859–2866 (2022) 2, 4, 11
4. Dong, S., Zhu, L., Tian, Y., Huang, T.: An efficient coding method for spike camera using inter-spike intervals. In: Data Compression Conference (DCC). pp. 437–437 (2017) 1
5. Dong, Y., Xiong, R., Fan, X., Yu, Z., Tian, Y., Huang, T.: Self-supervised learning for color spike camera reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6231–6240 (June 2025) 4
6. Dong, Y., Xiong, R., Zhao, J., Zhang, J., Fan, X., Zhu, S., Huang, T.: Joint demosaicing and denoising for spike camera. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1582–1590 (2024) 4
7. Dong, Y., Xiong, R., Zhao, J., Zhang, J., Fan, X., Zhu, S., Huang, T.: Learning a deep demosaicing network for spike camera with color filter array. IEEE Transactions on Image Processing **33**, 3634–3647 (2024) 4
8. Dong, Y., Zhao, J., Xiong, R., Huang, T.: High-speed scene reconstruction from low-light spike streams. In: 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP). pp. 1–5. IEEE (2022) 2, 3, 10, 11, 12
9. Graca, R., McReynolds, B., Delbruck, T.: Optimal biasing and physical limits of dvs event noise. arXiv preprint arXiv:2304.04019 (2023) 2
10. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1780–1789 (2020) 2
11. Guo, M., Huang, J., Chen, S.: Live demonstration: A 768× 640 pixels 200meps dynamic vision sensor. IEEE International Symposium on Circuits and Systems (ISCAS) pp. 1–1 (2017) 1
12. Hadji, I., Noroozi, M., Escorcia, V., Zaganidis, A., Martinez, B., Tzimiropoulos, G.: Edge-sd-sr: Low latency and parameter efficient on-device super-resolution with stable diffusion via bidirectional conditioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12789–12798 (2025) 4
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851 (2020) 2, 4, 5, 11

14. Hu, L., Ding, Z., Liu, M., Ma, L., Huang, T.: Learning to robustly reconstruct dynamic scenes from low-light spike streams. In: European Conference on Computer Vision (ECCV). pp. 88–105 (2024) 2, 4, 7, 10, 11, 12, 13, 14
15. Hu, L., Ma, L., Guo, Y., Huang, T.: Scsim: A realistic spike cameras simulator. arXiv preprint arXiv:2405.16790 (2024) 9
16. Hu, L., Zhao, R., Ding, Z., Ma, L., Shi, B., Xiong, R., Huang, T.: Optical flow estimation for spiking camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17844–17853 (2022) 2
17. Huang, T., Zheng, Y., Yu, Z., Chen, R., Li, Y., Xiong, R., Ma, L., Zhao, J., Dong, S., Zhu, L., et al.: 1000× faster camera and machine vision with ordinary devices. Engineering (2022) 1, 4
18. Jiang, H., Luo, A., Fan, H., Han, S., Liu, S.: Low-light image enhancement with wavelet-based diffusion models. ACM Transactions on Graphics (TOG) **42**(6), 1–14 (2023) 4
19. Jiang, H., Luo, A., Liu, X., Han, S., Liu, S.: Lightendiffusion: Unsupervised low-light image enhancement with latent-retinex diffusion models. In: European Conference on Computer Vision (ECCV). pp. 161–179 (2024) 4
20. Ke, J., Wang, Q., Wang, Y., Milanfar, P., Yang, F.: Musiq: Multi-scale image quality transformer. In: IEEE International Conference on Computer Vision (ICCV). pp. 5128–5137 (2021) 11
21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023) 9, 10
22. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998) 2
23. Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.M., Gu, J., Loy, C.C.: Low-light image and video enhancement using deep learning: A survey. IEEE transactions on pattern analysis and machine intelligence **44**(12), 9396–9416 (2021) 2
24. Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(8), 4225–4238 (2021) 2
25. Li, J., Liu, J., Wei, X., Zhang, J., Lu, M., Ma, L., Du, L., Huang, T., Zhang, S.: Uncertainty guided depth fusion for spike camera. arXiv preprint arXiv:2208.12653 (2022) 2
26. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128 × 128 120 db 15 μs latency asynchronous temporal contrast vision sensor. IEEE Journal of Solid-State Circuits (JSSC) **43**(2), 566–576 (2008) 1
27. Lin, X., He, J., Chen, Z., Lyu, Z., Dai, B., Yu, F., Qiao, Y., Ouyang, W., Dong, C.: Diffbir: Toward blind image restoration with generative diffusion prior. In: European Conference on Computer Vision (ECCV). pp. 430–448 (2024) 3, 4, 8
28. Liu, J., Zhang, Q., Li, J., Lu, M., Huang, T., Zhang, S.: Unsupervised spike depth estimation via cross-modality cross-domain knowledge transfer. arXiv preprint arXiv:2208.12527 (2022) 2
29. Liu, R., Zhu, L., Xiang, X., Wang, L., Huang, H.: Noise-modeled diffusion models for low-light spike image restoration. In: IEEE International Conference on Computer Vision (ICCV). pp. 4080–4089 (2025) 2, 4, 11, 12, 13
30. Liu, Y., Ke, Z., Liu, F., Zhao, N., Lau, R.W.: Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4197–4208 (2024) 4

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) 2, 3, 4, 5, 10
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017) 9
33. Wang, J., Chan, K.C., Loy, C.C.: Exploring clip for assessing the look and feel of images. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023) 11
34. Yang, S., Wu, T., Shi, S., Lao, S., Gong, Y., Cao, M., Wang, J., Yang, Y.: Maniqa: Multi-dimension attention network for no-reference image quality assessment. arXiv preprint arXiv:2204.08958 (2022) 11
35. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European Conference on Computer Vision (ECCV). pp. 74–91 (2024) 3, 4, 8
36. Yang, Z., Yu, H., Li, B., Zhang, J., Huang, J., Zhao, F.: Unleashing the potential of the semantic latent space in diffusion models for image dehazing. In: European Conference on Computer Vision (ECCV). pp. 371–389 (2024) 4
37. Yi, X., Xu, H., Zhang, H., Tang, L., Ma, J.: Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In: IEEE International Conference on Computer Vision (ICCV). pp. 12302–12311 (October 2023) 4
38. Yu, F., Gu, J., Li, Z., Hu, J., Kong, X., Wang, X., He, J., Qiao, Y., Dong, C.: Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 25669–25680 (2024) 3, 4, 8
39. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint arXiv:1409.2329 (2015) 4
40. Zhai, M., Ni, K., Xie, J., Gao, H.: Spike-based optical flow estimation via contrastive learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 2
41. Zhang, J., Jia, S., Yu, Z., Huang, T.: Learning temporal-ordered representation for spike streams based on discrete wavelet transforms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 137–147 (2023) 2, 4, 11, 12
42. Zhang, J., Tang, L., Yu, Z., Lu, J., Huang, T.: Spike transformer: Monocular depth estimation for spiking camera. In: European Conference on Computer Vision (ECCV) (2022) 2
43. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE Conference on International Conference on Computer Vision (ICCV). pp. 3836–3847 (2023) 2, 3, 4, 5, 8
44. Zhao, J., Xiong, R., Liu, H., Zhang, J., Huang, T.: Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11996–12005 (2021) 2, 4, 11, 12
45. Zhao, J., Zhang, S., Ma, L., Yu, Z., Huang, T.: Spikingsim: A bio-inspired spiking simulator. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 3003–3007. IEEE (2022) 9
46. Zhao, R., Xiong, R., Zhao, J., Yu, Z., Fan, X., Huang, T.: Learninng optical flow from continuous spike streams. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS) (2022) 2
47. Zhao, R., Xiong, R., Zhao, J., Zhang, J., Fan, X., Yu, Z., Huang, T.: Boosting spike camera image reconstruction from a perspective of dealing with spike fluctuations.

In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24955–24965 (2024) 2, 4, 11, 12
48. Zheng, Y., Zheng, L., Yu, Z., Huang, T., Wang, S.: Capture the moment: High-speed imaging with spiking cameras through short-term plasticity. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(7), 8127–8142 (2023) 2, 3, 11
49. Zhu, L., Dong, S., Huang, T., Tian, Y.: A retina-inspired sampling method for visual texture reconstruction. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1432–1437 (2019) 3, 7, 11
50. Zhu, L., Li, J., Wang, X., Huang, T., Tian, Y.: Neuspike-net: High speed video reconstruction via bio-inspired neuromorphic cameras. In: IEEE International Conference on Computer Vision (ICCV). pp. 2400–2409 (2021) 2, 12