240FPS Stereo Vision from Monocular Mixed Spikes

Yeliduosi Xiaokaiti[1,2] Yakun Chang[3,4] Yang Bai[1,2] Zhaojun Huang[1,2] Peiqi Duan[1,2] Boxin Shi[1,2,5*]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] Institute of Information Science, Beijing Jiaotong University

[4] Visual Intelligence +X International Cooperation Joint Laboratory of the MoE [5] PKU-AI[2] Robotics Joint Lab of Embodied AI

{yongqiye, by_baiyang, huangzhaojun}@stu.pku.edu.cn

ykchang@bjtu.edu.cn, {duanqi0001, shiboxin}@pku.edu.cn

Abstract

Stereo vision is fundamental for enabling machines to perceive and interact with the world. While monocular stereo methods offer hardware compactness, they struggle with generalization due to reliance on data-driven priors. Binocular and multi-view systems improve accuracy but incur higher hardware complexity and data inefficiency. In this paper, we introduce a monocular solution for high-frame-rate stereo vision via temporal optical modulation. The modulation directs light from two views onto a single sensor in a mixed manner, while periodically attenuating one view at 60 Hz. To capture the temporal variations introduced by this modulation, we employ a high-speed spike camera that records the mixed scene as temporally dense spikes. The high temporal resolution of these spikes enables the construction of a linear system for efficient binocular video decoupling. Consequently, we introduce a two-stage decoding methodology for achieving high-quality stereo vision: An efficient least-squares-based baseline reconstruction followed by a deep learning refinement module. Experimental results demonstrate that our approach achieves 240FPS binocular video reconstruction with superior accuracy compared to monocular systems, while maintaining the hardware compactness and data efficiency. Code is available at https://github.com/yongqiye00/MonoSpikeStereo.

1. Introduction

Stereo vision serves as a fundamental technology for enabling machine interaction with the world. In many high-level computer vision tasks, such as detection [16] and tracking [22], stereo vision provides more fine-grained scene understanding capabilities. Existing stereo vision systems can be categorized into monocular [1, 25, 31], binocular [30, 36], and multi-view [21, 33] configurations. Monocular systems [1, 25, 31] offer advantages in hardware compactness. However, their generalization to unseen scenarios remains limited as they rely heavily on data-driven priors. In contrast, binocular and multi-view

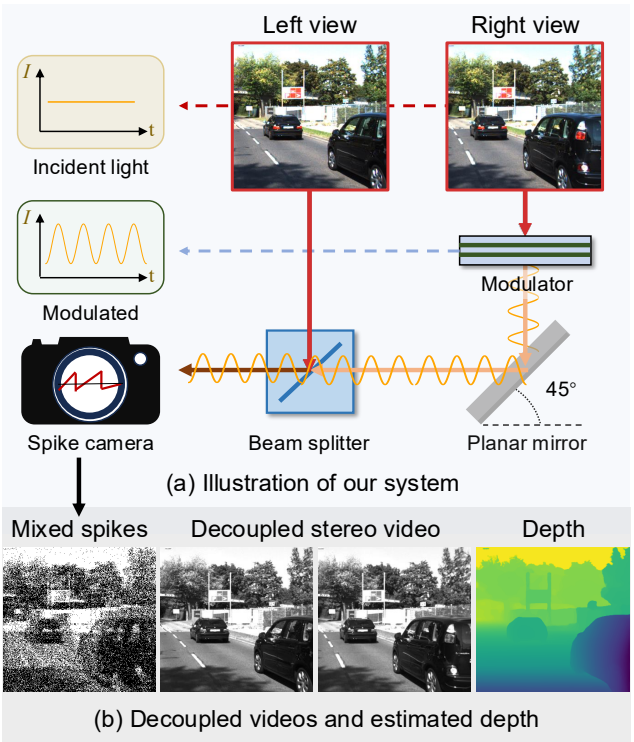

Figure 1. We propose a monocular solution capable of producing 240FPS stereo video. (a) Our optical setup mixes light from two views onto a single spike camera via a beam splitter. While one view remains unmodulated, the other view passes through an LCD modulator whose transmittance varies periodically. (b) We decouple binocular video from mixed spikes and feed the reconstructed stereo pair into an off-the-shelf stereo matcher (e.g., DEFOM-Stereo [13]) to estimate depth.

*Corresponding author.

systems [21, 30] typically employ two or more cameras to enhance accuracy. However, this performance advantage comes at the cost of increased hardware complexity. Moreover, in many high-speed scenarios such as competitive sports, autonomous driving, and aerospace applications [9, 11, 27], the demand for high frame rates in these systems adversely affects data efficiency.

To simultaneously achieve accuracy, hardware compactness, and data efficiency, one line of research utilizes two or more planar mirrors to project multiple views onto separate sub-regions of a single sensor [5, 10, 18]. In these *spatial optical modulations*, the captured sub-images are then geometrically corrected through epipolar rectification to obtain binocular video. However, epipolar rectification relies on planar scene assumptions, which may lead to geometric distortion and degraded accuracy. Therefore, another line of research projects incident light from two views onto the same sensor in a mixed manner [20]. Although this optical modulation avoids geometric distortions, decoupling binocular videos from the mixed signal still relies on a global consideration of limited disparity. For mixed signals from two views, if we introduce *temporal optical modulation* to one view, the variations will provide strong cues for effectively decoupling binocular videos. However, implementing temporal modulation requires the camera to operate at a sufficiently high speed. Given the limited frame rate of conventional digital cameras, a high-speed and data-efficient camera is highly desirable.

Recently developed spike cameras [3, 4, 12] offer readout frequencies of up to 40,000 Hz and output 1-bit data. Such characteristics make them ideally suited for our application requirements. Therefore, as shown in Fig. 1(a), we construct a temporally modulated stereo system using a monocular spike camera. In our system, incident light from one view is reflected by the mirror and optically mixed with the other view through the beam splitter. To facilitate binocular video decoupling, we maintain one view without modulation while periodically attenuating the other view at 60 Hz using an LCD modulator. Compared to existing stereo systems employing two spike cameras [14, 29], our monocular setup offers more compact hardware and efficient data.

Benefiting from the temporal optical modulation applied to one view, binocular video decoupling from mixed spikes becomes more tractable. To enable an efficient and fast solution, we initially consider that motions across consecutive frames are negligible for the high-speed spike camera. This consideration allows us to construct a linear system where the decoupling problem can be efficiently solved through least-squares optimization. For each modulation period, we decouple 4 pairs of binocular frames, thereby achieving a baseline binocular video at 240FPS[1]. These baseline results can be further refined through a learning-based module to eliminate residual artifacts induced by motions. We refer to this learning-based refinement module as SMS-Net, as our goal is **S**tereo vision from **M**ixed **S**pikes. SMS-Net is a fully end-to-end deep learning framework that refines the results through multi-scale feature fusion and temporal integration. Furthermore, we introduce a cross-view attention mechanism [6] to effectively exploit complementary information between the two views at multiple feature scales, enabling accurate and high-quality reconstruction. To validate the quality of the decoupled binocular videos, we apply them to depth estimation (Fig. 1(b)). Experimental results show that our method outperforms monocular approaches in accuracy, while maintaining the data efficiency and hardware compactness. The main contributions of this work are summarized as follows:

- a monocular solution for 240FPS binocular video reconstruction through mixed-spike decoupling;
- an SMS-Net that enables robust binocular video decoupling, which is validated on depth estimation; and
- a corresponding hardware platform that enables real-world data for testing, along with a data simulator for training.

2. Related work

Monocular stereo with mirrors. Monocular stereo vision systems commonly rely on optical modulation to capture light from multiple viewpoints [5, 10, 18]. Over the past decades, numerous such systems have been proposed. Early approaches, such as the motor-driven mirror assembly by Teoh and Zhang [24] and the rotating glass plate by Nishimoto and Shirai [19], exploit mechanical movement to acquire stereo image pairs. However, these pioneering methods are primarily suitable for static scenes due to their inherent mechanical constraints. Another line of research utilizes two or more planar mirrors to project multiple viewpoints onto separate sub-regions of a single sensor. Nene and Nayar [18] propose four stereo systems that use a single camera, establishing the practical viability of this approach. Their work is formalized by Gluckman and Nayar [10], who detail the epipolar geometry and calibration for catadioptric stereo with planar mirrors. A common requirement for these mirror-based systems is geometric correction. In contrast, Pachidis and Lygouras [20] introduce a method that projects light from two viewpoints onto the sensor in a mixed manner, thereby avoiding geometric distortions at the cost of requiring subsequent separation. Compared with existing monocular stereo vision with mirrors, the LCD modulator in our approach introduces more temporal information for binocular video reconstruction.

Stereo vision from spikes. Spike cameras have been widely used in many computer vision tasks [3, 34, 35] including stereo vision. Zhang *et al*. [32] demonstrate the fea-

[1]We drive the LCD modulator at 60 Hz in our current prototype; higher frame rates are possible with faster modulators.

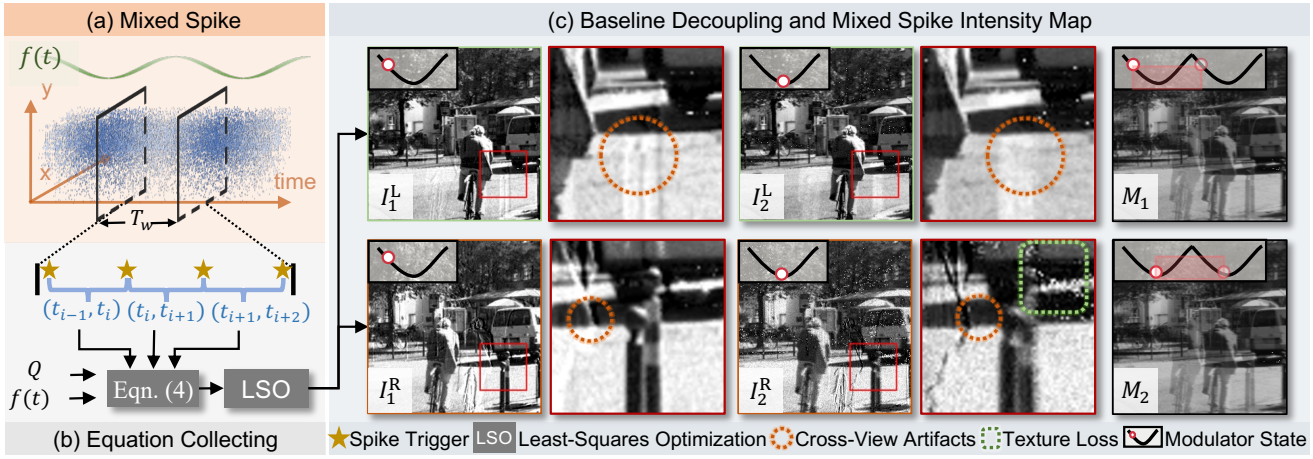

Figure 2. (a) Mixed spikes are sliced by a sliding window of length T_w. (b) We collect all spikes triggered within T_w to construct a linear system (Eqn. (4)), which is then solved using the least-squares optimization (LSO). (c) Baseline decoupling results reveal brightness inconsistency, residual cross-view artifacts, and texture loss in the right view. The mixed spike intensity maps M_t $(t = 1, 2)$ reconstructed by accumulating all spikes during a modulation period can serve as brightness-consistency guidance.

sibility of estimating depth directly from continuous spikes captured with a monocular camera. However, in regions with low texture or repeated patterns, monocular cues such as shading or perspective are insufficient to disambiguate depth. For spike-based stereo vision, constructing a binocular system is a straightforward solution [14, 29]. Li *et al*. [14] propose a dual-task depth estimation framework with joint training and uncertainty-guided fusion to combine monocular and stereo cues. Wang *et al*. [29] develop a hybrid stereo camera platform and introduce a corresponding dataset, along with a transformer-based network for dense depth prediction. Gao *et al*. [8] employ raw dual-view spike streams and iteratively refine depth estimation using a recurrent spiking neural network. However, these approaches rely on multiple cameras, which inevitably increase hardware complexity and system cost. In contrast, our monocular spike camera integrated with a temporal modulator achieves competitive performance while significantly reducing system integration complexity.

3. Method

Overview. We present a two-stage framework for achieving stereo vision using a monocular spike camera equipped with temporal optical modulation. This modulation introduces temporal inconsistency in one view within the captured mixed spikes, which serves as strong guidance for decoupling binocular videos. In the first stage (Sec. 3.1), we find that the spikes are temporally dense, which permits the consideration that motion within several adjacent frames is negligible. This consideration allows us to formulate a linear system and perform fast baseline decoupling by using least-squares optimization. In Sec. 3.2, to remove residual artifacts in the baseline decoupling results and compensate for texture loss caused by optical modulation, we design a learning-based refinement module, *i.e*., SMS-Net, to effectively remove these artifacts and restore fine details. Finally, in Sec. 3.3, we describe details of our dataset.

3.1. Baseline decoupling

Imaging model of mixed spikes. The pixels in a spike sensor operate asynchronously. Each pixel continuously integrates photo-generated electrons and triggers a spike of 1 when the accumulated electrons reach a threshold Q, after which the electrons are cleared. Due to the discrete sampling strategy, a discrepancy (ΔQ) may exist between the true number of accumulated electrons and the computed value. Therefore, the accumulated electrons are

$$Q(x,y) + \Delta Q(x,y) = \int_{t_{i-1}}^{t_i} \alpha(x,y) I(t,x,y)\, dt, \qquad (1)$$

where t_i denotes the timestamp of the i-th detected spike, α is the photoelectric conversion coefficient, I is the light intensity, and (x, y) is spatial coordinates. When we add temporal optical modulation, the incident light intensity is a mixture of two views. Omitting spatial coordinates (x, y) for brevity, the imaging model becomes:

$$Q + \Delta Q = \int_{t_{i-1}}^{t_i} \alpha \left[I^{\mathrm{L}}(t) + I^{\mathrm{R}}(t) f(t) \right] dt, \qquad (2)$$

where $I^{\mathrm{L}}(t)$ and $I^{\mathrm{R}}(t)$ denote the light from two views, and $f(t)$ (see Fig. 2(a)) is the temporal modulation .

Decoupling with least-squares optimization. When observing the accumulated electrons (*i.e*., spikes) within a temporal window T_w, from Eqn. (2) we note that both $I^L(t)$ and $I^{\mathrm{R}}(t)$ are unknown quantities. Under the assumption that motion is negligible between adjacent time windows for temporally dense spikes, we can construct a linear

system. The temporal variation of $f(t)$ guarantees that the observations across different intervals are linearly independent, rendering the system well-posed for decoupling. The linear system can be expressed as

$$Q + \Delta Q_i = \int_{t_{i-1}}^{t_i} \mathcal{G}(t)\, dt, \quad i = 1, \ldots, n, \tag{3}$$

where $\mathcal{G}(t) = \alpha[I^{\mathrm{L}}(t) + I^{\mathrm{R}}(t)f(t)]$. Since motions are negligible, light intensities across adjacent time windows can be approximated as constants, *i.e.*, $I^{\mathrm{L}}(t) \approx I^{\mathrm{L}}$, $I^{\mathrm{R}}(t) \approx I^{\mathrm{R}}$. Let $\eta_i = \Delta Q_i + \epsilon_i$ denote a composite error term, where ϵ_i arises from residual intensity inconsistency. Substituting these into the preceding formulation leads to:

$$Q + \eta_i = \alpha \left(I^{\mathrm{L}} \Delta t_i + I^{\mathrm{R}} F_i \right), \quad i = 1, \ldots, n, \tag{4}$$

where $\Delta t_i = t_i - t_{i-1}$ and $F_i = \int_{t_{i-1}}^{t_i} f(t)\, dt$. Thus, for each pixel in each window, we construct a linear system where $Q, f(t), \alpha$ and $\{t_i | i \in [0, n]\}$ are known, and the unknown quantities are I^{L} and I^{R}. The unknowns I^{L} and I^{R} are estimated via least-squares optimization:

$$\min_{I^{\mathrm{L}}, I^{\mathrm{R}}} \sum_{i=1}^{n} \left(Q - \alpha \left[I^{\mathrm{L}} \Delta t_i + I^{\mathrm{R}} F_i \right] \right)^2. \tag{5}$$

To ensure numerical stability, we augment the objective with an ℓ_2 regularization term $\lambda\big((I^L)^2 + (I^{\mathrm{R}})^2\big)$, with $\lambda = 1 \times 10^{-3}$ fixed across all experiments. The above process is shown in Fig. 2(b).

Strengths and potential refinements. The baseline decoupling method offers low computational complexity and achieves fast execution speed. When tested on a device equipped with an Intel® Core™ i9-13900KF, the average time for constructing the linear system is 0.093s, and the time for solving the least-squares optimization is only 0.0151s. This demonstrates that although our baseline method may not yet achieve real-time performance at hundreds of frames per second, it is fully capable of supporting real-time video decoupling in daily scenarios. However, as shown in Fig. 2(c), under the assumption that motions are negligible, the baseline solution may suffer from: (1) inconsistent brightness distributions across consecutive frames; (2) cross-view residual artifacts remaining in the two-view results; (3) and texture loss for the modulated view. To address these issues, we further utilize a learning-based method to refine the decoupling results and obtain high-quality binocular videos.

3.2. Refinement with SMS-Net

Since deep learning has shown strong capabilities in computer vision, we adopt a learning-based refinement network to improve the baseline and recover high-quality stereo videos. Since our network refines binocular videos from Mixed Spikes, we name the proposed architecture SMS-Net. As shown in Fig. 3, the SMS-Net consists of three modules: (1) an Adaptive Brightness Consistency (ABC) module that utilizes the intensity map reconstructed from mixed spikes over one modulation period T_f to remove temporal inconsistency; (2) a Collaborative Binocular Augment (CBA) module that employs cross-attention on two views to remove residual artifacts; (3) and a Recurrent Stereo Fusion (RSF) module that fuses the temporal information of historical frames to compensate for texture loss.

ABC module. Although brightness inconsistency is introduced by the optical modulation, its periodic nature ensures that these artifacts cancel out when the reconstruction window spans an integer multiple of the modulation period T_f. Therefore, as shown in the last column of Fig. 2(c), we construct a brightness consistency guidance by integrating spikes over one full modulation cycle:

$$M_t = \frac{Q}{T_f} \sum_{k: t_k \in (t - T_f, t]} 1. \tag{6}$$

The ABC module first extracts multi-scale features from the baseline decoupled video frames and M_t, denoted as $\boldsymbol{F}_t^{\mathrm{L}}$,$\boldsymbol{F}_t^{\mathrm{R}}$, and $\boldsymbol{F}_t^{M}$, respectively. Specifically, for each scale, we align the features of the left and right views independently using $\boldsymbol{F}_t^{M}$ as a shared reference. To align the left-view features $\boldsymbol{F}_t^{\mathrm{L}}$, we first compute the global channel-wise mean and standard deviation from both $\boldsymbol{F}_t^{\mathrm{L}}$ and $\boldsymbol{F}_t^{M}$. These concatenated statistics are fed into a lightweight prediction network of two 1×1 convolutional layers to generate a scaling factor γ_t^{L} and β_t^{L}. The alignment is then formulated as $\bar{\boldsymbol{F}}_t^{\mathrm{L}} = \gamma_t^{\mathrm{L}} \cdot \boldsymbol{F}_t^{\mathrm{L}} + \beta_t^{\mathrm{L}}$. The same procedure is applied to the right-view features $\boldsymbol{F}_t^{\mathrm{R}}$ to obtain $\bar{\boldsymbol{F}}_t^{\mathrm{R}}$.

CBA module. As shown in the regions marked by red circles in Fig. 2(c), residual artifacts manifest differently across views. In I_1^{L} the residual edges of the guardrail appear to the right of the pedestrian, whereas in I_1^{R} the residual edges appear to his left. This spatial inconsistency of residual artifacts provides crucial guidance for the refinement. Building on this observation, we design the CBA module that leverages a cross-attention mechanism to enable information interaction between the two views, thereby removing residual artifacts. At each feature scale, we adopt the Stereo Cross-Attention Module (SCAM) [6], which computes attention along corresponding epipolar lines to obtain complementary information between two views. Integrating SCAM into the CBA allows the network to leverage cross-view consistency and suppress residual artifacts. The output features are denoted as $\widetilde{\boldsymbol{F}}_t^{\mathrm{L}}, \widetilde{\boldsymbol{F}}_t^{\mathrm{R}}$.

RSF module. The RSF module (details in Fig. 4) is designed to compensate for texture degradation caused by low-transmittance phases of the LCD modulator through temporal fusion. Since we use a recurrent architecture, RSF takes $\widetilde{\boldsymbol{F}}_t^{\mathrm{L}}, \widetilde{\boldsymbol{F}}_t^{\mathrm{R}}$, and the hidden states from previous time

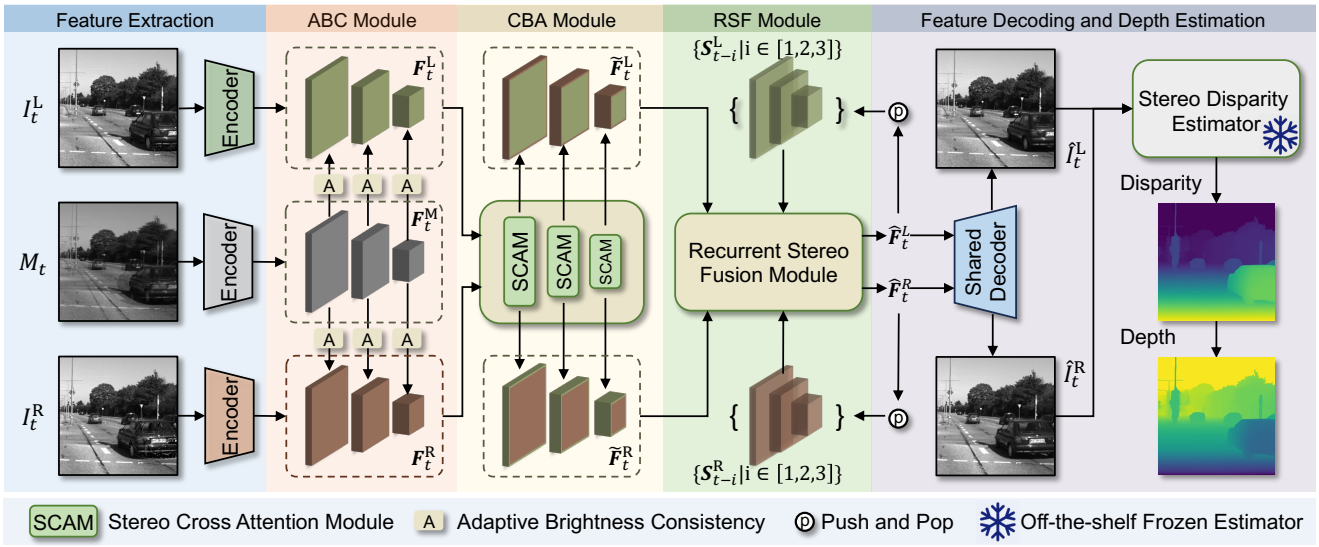

Figure 3. Overview of SMS-Net. At time t, the two-view images $\{I_t^{\mathrm{L}}, I_t^{\mathrm{R}}\}$ and brightness consistency guidance M_t are encoded into multi-scale features. We design an Adaptive Brightness Consistency (ABC) module to remove temporal inconsistency, a Collaborative Binocular Augment (CBA) module to suppress cross-view residual artifacts, and a Recurrent Stereo Fusion (RSF) module that aggregates temporal and cross-view cues and maintains a state queue. The outputs follow two paths: Push and Pop update the queue, a shared decoder produces refined stereo pair $\hat{I}_t^{\mathrm{L}}$ and $\hat{I}_t^{\mathrm{R}}$. We evaluate the reconstructed stereo video via downstream depth estimation using off-the-shelf stereo matchers (*e.g.*, DEFOM-Stereo [13] or CREstereo [15]).

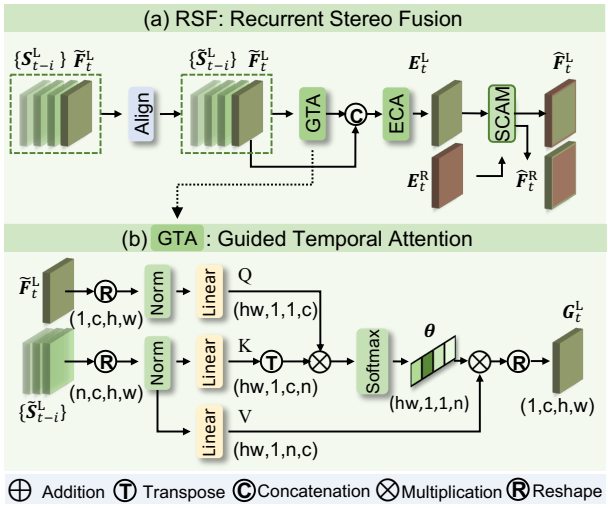

Figure 4. Architectures of (a) the Recurrent Stereo Fusion (RSF) module and (b) the Guided Temporal Attention (GTA) block.

steps $\{\boldsymbol{S}_{t-i}^{\mathrm{L}}, \boldsymbol{S}_{t-i}^{\mathrm{R}}\}$ as inputs. To achieve spatial alignment between current and historical states, the RSF module first learns spatial offsets Δ and modulation masks m. These learned parameters are then fed to deformable convolution to align historical features with the current timestep:

$$\tilde{\boldsymbol{S}}_{t-i}^{\mathrm{L}} = \mathrm{DeformConv}\left(\boldsymbol{S}_{t-i}^{\mathrm{L}}, \Delta, m\right). \tag{7}$$

Subsequently, we fuse the aligned historical states $\{\tilde{\boldsymbol{S}}_{\tau}^{\mathrm{L}}\}_{\tau=t-\ell}^{t-1}$ with the current feature $\tilde{\boldsymbol{F}}_t^{\mathrm{L}}$ via Guided Temporal Attention (GTA). Specifically, for each spatial location p, after layer normalization $\mathcal{N}(\cdot)$ and 1×1 projections W_q, W_k, W_v, we can obtain the query $\mathbf{Q}_p$, key $\mathbf{K}_p$ and the value $\mathbf{V}_p$. Next, we obtain the attention weights:

$$\boldsymbol{\theta}_p = \mathrm{softmax}\left(\frac{\mathbf{Q}_p \mathbf{K}_p^{\top}}{\sqrt{dim}}\right), \quad \hat{\mathbf{y}}_p = \sum_{i=1}^{\ell} \theta p, i\, \mathbf{V}_{p,i}, \tag{8}$$

where dim is the dimension of $\mathbf{Q}_p$, and the output of GTA is $\boldsymbol{G}_t^{\mathrm{L}}(p) = W_o\, \hat{\mathbf{y}}_p \in \mathbb{R}^C$, where W_o is a learnable output projection matrix. We then concatenate $\tilde{\boldsymbol{F}}_t^{\mathrm{L}}$ and $\boldsymbol{G}_t^{\mathrm{L}}$ along the channel dimension, and apply an Efficient Channel Attention (ECA) block [26] to obtain the enhanced feature $\boldsymbol{E}_t^{\mathrm{L}}$. We apply the identical procedure to the right branch to obtain $\boldsymbol{E}_t^{\mathrm{R}}$. The SCAM module fuses $\boldsymbol{E}_t^{\mathrm{L}}$ and $\boldsymbol{E}_t^{\mathrm{R}}$ into refined features $\hat{\boldsymbol{F}}_t^{\mathrm{L}}$ and $\hat{\boldsymbol{F}}_t^{\mathrm{R}}$. These are then passed to a shared decoder to reconstruct the refined stereo pair $\hat{I}_t^{\mathrm{L}}$ and $\hat{I}_t^{\mathrm{R}}$.

Loss and training. The network is trained end-to-end using ℓ_1 loss: $\mathcal{L} = \|\hat{I}_t^{\mathrm{L}} - I_{\mathrm{gt}}^{\mathrm{L}}\|_1 + \|\hat{I}_t^{\mathrm{R}} - I_{\mathrm{gt}}^{\mathrm{R}}\|_1$. We train for 300 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4} and the learning rate is scheduled via cosine annealing. The batch size is set to 12, and training is conducted on a single NVIDIA RTX 4090.

3.3. Dataset

Given the specialized design of our experimental setup, no publicly available dataset exists for training the proposed SMS-Net. To address this limitation, we create a synthetic dataset as illustrated in Fig. 5(a). Furthermore, to validate the practical efficacy of our approach, we developed a dedicated camera system and collected corresponding real-world test data, shown in Fig. 5(b). The details of our data

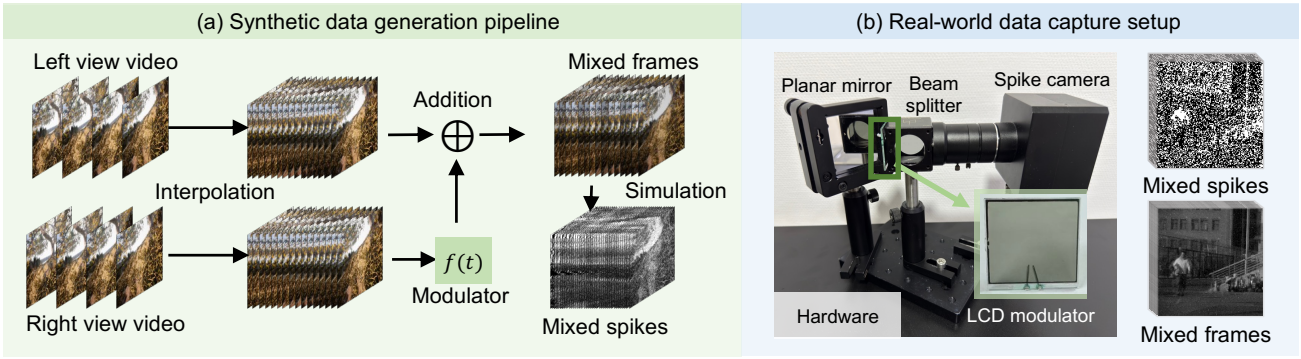

Figure 5. (a) We interpolate the binocular video, modulate the right-view video, and then mix it with the left-view video; the mixed video is subsequently simulated to generate mixed spikes. (b) The dedicated monocular camera system. A planar mirror and beam splitter optically mix two views, enabling the acquisition of mixed spikes and mixed video frames.

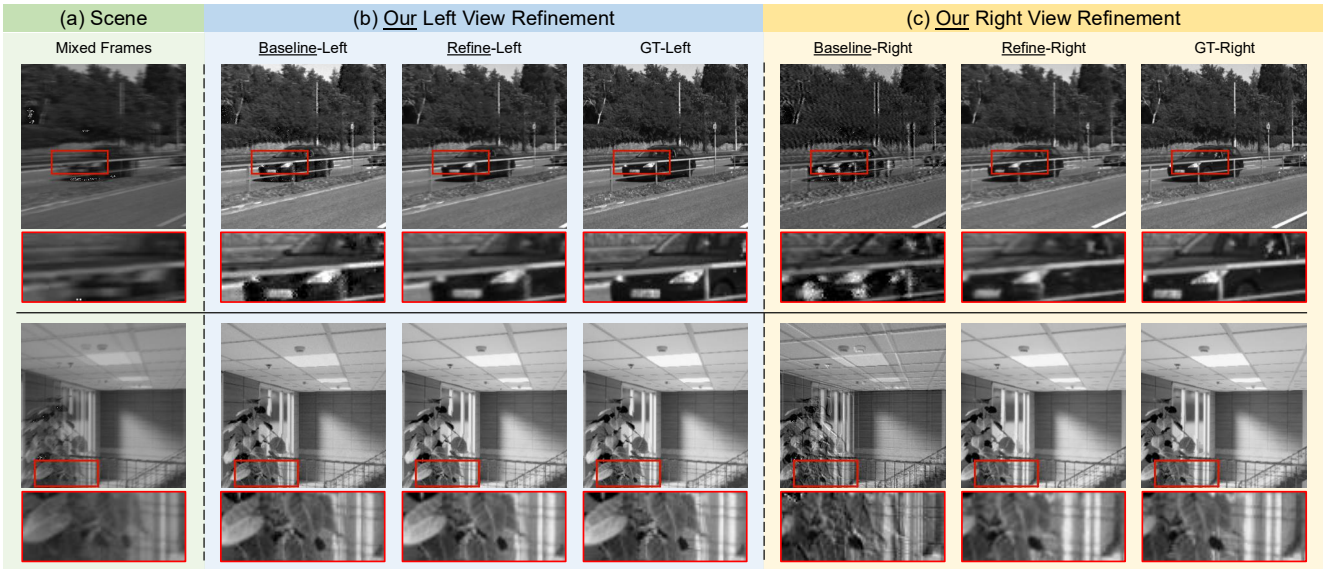

Figure 6. Decoupling binocular videos from synthetic test data. (a) The input scenes with mixed frames. (b) and (c) show the ground truth, along with the baseline and refined results. See the supplementary video for additional results.

collection methodology are described below.

Synthetic data. We generate synthetic training and testing data using TartanAir dataset [28], a large-scale photo-realistic simulation benchmark for robotic navigation, providing long-horizon binocular video sequences with precise ground-truth depth. We also generate a subset of synthetic data for testing using KITTI [17]. We first apply optical flow-based frame interpolation [23] to both views of the stereo image pair. The interpolated right view is then modulated by function $f(t)$, producing the mixed image sequence $I_{\text{mix}}(t) = I^{\text{L}}_{gt}(t) + I^{\text{R}}_{\text{gt}}(t) f(t)$. This mixed sequence is subsequently processed through spike camera simulation to generate mixed spike streams.

Real-world data. As shown in Fig. 5(b), our hardware setup employs a 45° planar mirror to redirect one view toward a beam splitter, which optically blends it with the direct view. Due to the depth displacement introduced by the planar mirror configuration, a specific calibration technique is required to establish accurate geometric correspondence. Further implementation details are provided in the supplementary material. An LCD modulator placed between the mirror and the beam splitter introduces periodic intensity variations to the reflected path. This system enables the capture of real-world test data across diverse scenes, ensuring broad scenario coverage for comprehensive evaluation.

4. Experiments

In this section, we first demonstrate the effectiveness of our method for binocular video decoupling on synthetic data. We then show that the decoupled videos serve as high-quality inputs for depth estimation across both synthetic and real-world datasets, highlighting the superior characteristics of our reconstructed binocular videos. For depth estimation, we compare against monocular methods by adapting their input to single-view frames while maintaining our two-view input configuration. We explicitly acknowl-

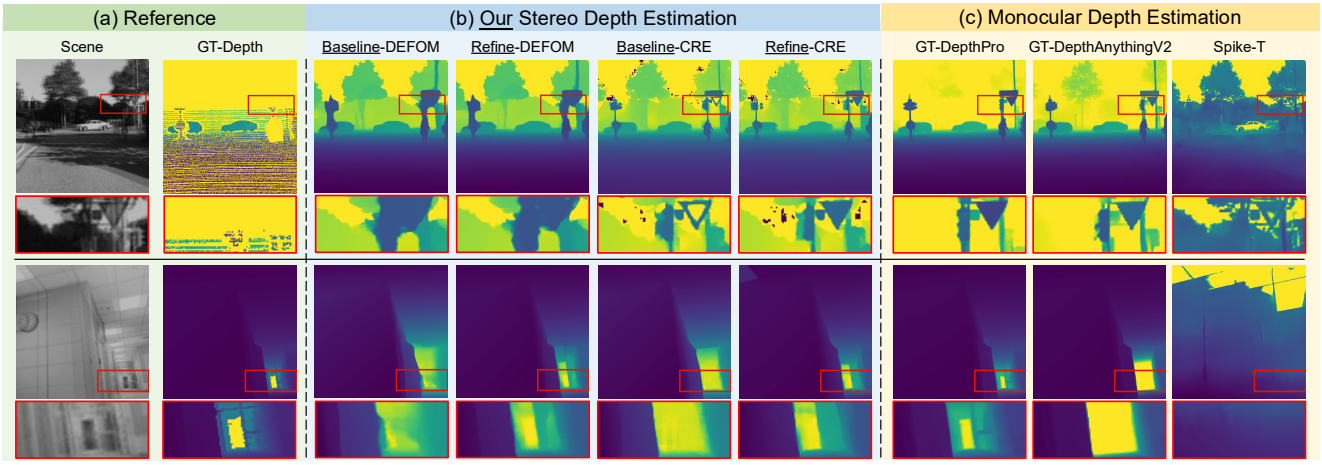

Figure 7. Depth estimation on synthetic data. We compare stereo depth maps from baseline-decoupled inputs and our refined stereo videos, and contrast them with monocular depth methods. Additional video results are provided in the supplementary material.

Table 1. Depth estimation performance on TartanAir [28] and KITTI [17] (AbsRel↓, RMSE↓, δ_1↑). ↑: higher is better; ↓: lower is better. GT-* rows indicate an upper bound when the corresponding method is fed with ground-truth grayscale inputs; these rows are excluded when determining the best and second-best results. Best is in **bold**, second-best is underlined.

Methods	TartanAir [28]			KITTI [17]		
	AbsRel↓	RMSE↓	δ_1 ↑	AbsRel↓	RMSE↓	δ_1 ↑
Base-DEFOM [13]	0.1695	2.5187	0.8795	0.4477	4.8843	0.9427
Refine-DEFOM [13]	**0.0940**	**1.7852**	**0.9202**	**0.4288**	**4.1719**	**0.9630**
GT-DEFOM [13]	0.0498	1.5839	0.9438	0.4219	4.0963	0.9655
Base-CRE [15]	0.1631	2.2064	0.8737	0.4746	6.4855	0.9262
Refine-CRE [15]	0.1126	2.3595	0.9153	0.4523	5.6236	0.9486
GT-CRE [15]	0.0609	1.6524	0.9335	0.4441	5.5583	0.9507
Spike-T [32]	0.9675	8.2092	0.2731	0.8000	13.5509	0.3233
STIR [7]-DepthAnythingV2 [31]	0.4839	14.1810	0.5483	0.4412	6.6630	0.9116
GT-DepthAnythingV2 [31]	0.4830	14.1307	0.5406	0.4304	6.4323	0.9191
STIR [7]-DepthPro [2]	0.1085	2.2625	0.8573	0.4409	5.4297	0.9302
GT-DepthPro [2]	0.0731	1.6028	0.9164	0.3399	4.8738	0.9345

edge the inherent asymmetry in this comparison. However, in the absence of an open-source video decoupling method, the improved depth estimation results obtained from our binocular videos confirm that our monocular solution achieves enhanced accuracy without increasing data requirements. For monocular depth estimation benchmarks, we selected frame-based DepthAnythingV2 [31], DepthPro [2] and spike-based Spike-T [32].

4.1. Evaluation on binocular video

We first evaluate binocular video decoupling on synthetic data. As shown in Tab. 2, our full model significantly outperforms the baseline decoupling method across all image quality metrics. This demonstrates the powerful restoration capability of our network. As shown in Fig. 6, we present the mixed frame reconstructed directly from spike streams as input. Our method successfully recovers high-quality binocular videos for both views. In the first row of Fig. 6, our approach accurately reconstructs the moving vehicle with preserved details. The second row demonstrates faithful recovery of architectural textures, while the third row shows clear reconstruction of vegetation contours.

4.2. Evaluation on depth estimation.

We now evaluate the utility of our decoupled videos for the downstream task of depth estimation.

Results on synthetic data. Tab. 1 presents the quantitative results on TartanAir and KITTI. For stereo-based approaches, we provide the baseline decoupling results, our refined results, and the ground truth grayscale stereo pairs (GT-*). For monocular methods, we evaluate them on inputs reconstructed from spike streams (e.g.,STIR [7]-DepthPro) and on ground truth grayscale images. For DepthAnythingV2 and DepthPro, we aligned their outputs with the ground truth depth for scale correction, since they take single-channel inputs. After excluding the ideal GT-* results, which serve as an upper bound, our Refine-DEFOM method demonstrates best performance. Our baseline is highly competitive on its own, and the refinement network provides a significant boost, bringing our results close to the theoretical upper bound. As shown in Fig. 7, our baseline decoupling paired with a stereo estimator already produces high-quality visual results. The refinement further enhances this performance, yielding sharper and more accurate depth maps. While DepthAnythingV2 and DepthPro also deliver strong and visually competitive outputs, their accuracy degrades for distant scenes.

Results on real-world data. Qualitative results are shown in Fig. 8. We feed the baseline decoupling and refined results into the DEFOM and CREStereo models, respectively. We also apply monocular methods to the spike streams. The results show that both the DepthPro and Spike-T methods are largely ineffective. While DepthAnythingV2 recovers the relative depth relationships well, its output lacks distinct depth stratification, resulting in a visually flat appearance. Among stereo estimators, DEFOM generally outperforms CREStereo. For DEFOM, the difference between baseline

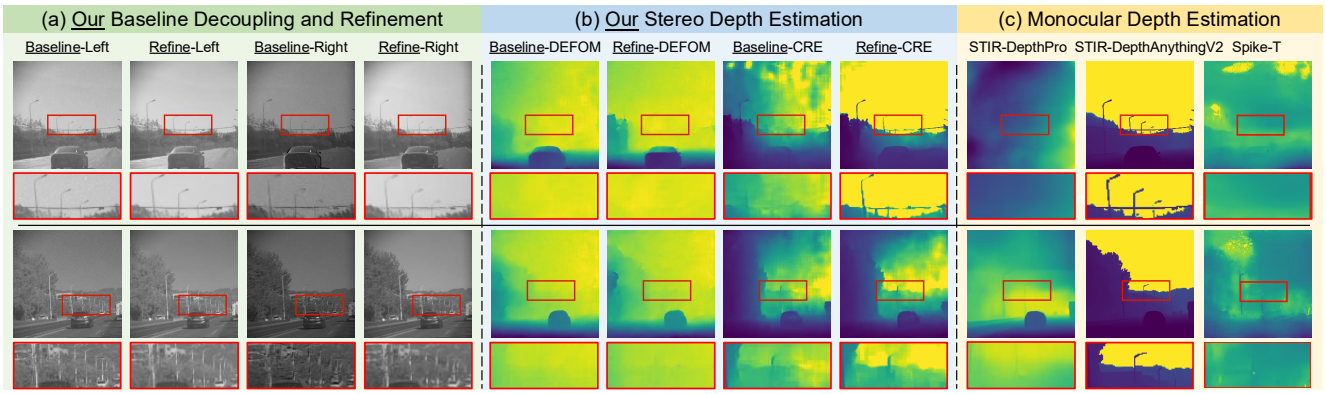

Figure 8. We show decoupled stereo results on real-world data and compare downstream depth estimation performance. Additional video results are provided in the supplementary material.

Table 2. Ablation study of the three modules (ABC, CBA, and RSF). ✓: denotes enabled. We report PSNR / SSIM / LPIPS, (Left/Right) and stereo errors (Bad 3.0/ EPE) on TartanAir [28] and KITTI [17]. Best is in **bold**, second-best is underlined.

	Proposed Modules			TartanAir [28]					KITTI [17]				
	ABC	CBA	RSF	PSNR↑	SSIM↑	LPIPS↓	Bad 3.0↓	EPE↓	PSNR↑	SSIM↑	LPIPS↓	Bad 3.0↓	EPE↓
Baseline				26.53/16.54	0.857/0.775	0.162/0.266	5.372	13.71	25.48/17.78	0.850/0.764	0.107/0.212	5.580	0.964
+ABC	✓			32.13/28.22	0.953/0.922	0.038/0.093	5.327	11.44	30.77/25.94	0.935/0.895	0.046/0.085	3.290	0.819
+CBA		✓		28.90/19.97	0.958/0.899	0.041/0.111	5.266	11.40	29.10/22.25	0.932/0.881	0.052/0.091	3.600	0.819
+RSF			✓	32.00/32.29	0.975/0.964	0.025/0.042	4.789	10.25	32.77/31.53	0.959/0.947	0.034/0.045	2.090	0.668
+ABC+CBA	✓	✓		33.66/29.73	0.964/0.926	0.038/0.084	5.232	11.13	31.26/27.33	0.937/0.897	0.049/0.078	3.630	0.808
+ABC+RSF	✓		✓	36.20/35.45	0.977/0.968	**0.023**/0.038	**4.691**	10.07	34.36/**32.88**	0.960/0.950	**0.030**/0.040	2.080	0.664
+CBA+RSF		✓	✓	31.55/32.15	0.976/0.966	0.025/0.039	4.971	**9.98**	32.72/31.60	0.960/0.949	0.032/0.041	2.010	0.658
Ours	✓	✓	✓	**36.97/35.73**	**0.978/0.968**	0.024/**0.038**	4.791	10.08	**35.00**/32.75	**0.963/0.953**	0.032/**0.039**	**1.890**	**0.644**

and refined inputs is subtle. For CREStereo, the refined output yields clearly better depth. Fig. 8(a) shows that SMS-Net restores brightness consistency and recovers fine textures in the right view.

4.3. Ablation study

Since our network consists of three functionally distinct components, we conduct systematic ablation studies to evaluate the individual and combined contributions of each module. The evaluation is performed from two perspectives: image restoration quality and disparity estimation accuracy, with experiments carried out on TartanAir and KITTI datasets. As shown in Tab. 2, compared to the baseline solution, incorporating any of the three modules, *i.e.*, ABC, CBA, or RSF brings significant performance improvements. RSF contributes most substantially to the framework. Moreover, any two-module combination outperforms using a single module alone. Integrating ABC, CBA, and RSF addresses key decoupling challenges and achieves top performance. Overall, these extensive ablation experiments robustly validate the necessity and the synergistic benefits of our architectural design choices.

5. Conclusion

We presented a monocular solution for achieving 240FPS stereo vision from mixed spikes. By introducing temporal optical modulation into one view and capturing the resulting mixed spikes, our camera system effectively encodes stereo cues in the spatial and temporal domain. A two-stage approach comprising a baseline decoupling strategy and a learning-based refinement network (SMS-Net) enables efficient and accurate reconstruction of binocular videos from mixed spikes. Experiments on both synthetic and real-world datasets demonstrate that our approach achieves high-quality stereo reconstruction while maintaining hardware compactness and data efficiency.

Limitations: In our current implementation, our LCD modulator is constrained to approximately 60 Hz due to inherent physical and operational principles. Faster modulators would enable decoupling at higher frame rates.

Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. 62136001), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and Beijing Natural Science Foundation (Grant No. L233024). Yakun Chang was supported by National Natural Science Foundation of China (Grant No. 62301009). Peiqi Duan was supported by China National Postdoctoral Program for Innovative Talents (Grant No. BX20230010) and China Postdoctoral Science Foundation (Grant No. 2023M740076).

References

[1] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 7

[3] Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, and Boxin Shi. 1000 FPS HDR video with a Spike-RGB hybrid camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[4] Yakun Chang, Yeliduosi Xiaokaiti, Yujia Liu, Bin Fan, Zhaojun Huang, Tiejun Huang, and Boxin Shi. Towards HDR and HFR video from rolling-mixed-bit spikings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[5] Yuxuan Chen, Ben Wang, Yujun Zhong, Qiongwei Li, and Yi Jin. A mirror-based compact monocular depth estimation system for environment sensing. In *International Conference on Electronic Measurement & Instruments*, 2021. 2

[6] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 2, 4

[7] Bin Fan, Jiaoyang Yin, Yuchao Dai, Chao Xu, Tiejun Huang, and Boxin Shi. Spatio-temporal interactive learning for efficient image reconstruction of spiking cameras. In *Advances in Neural Information Processing Systems*, 2024. 7

[8] Zhuoheng Gao, Yihao Li, Jiyao Zhang, Rui Zhao, Tong Wu, Hao Tang, Zhaofei Yu, Hao Dong, Guozhang Chen, and Tiejun Huang. SpikeStereoNet: A brain-inspired framework for stereo depth estimation from spike streams. In *International Conference on Learning Representations*, 2026. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012. 2

[10] Joshua Gluckman and Shree K Nayar. Catadioptric stereo using planar mirrors. *International Journal of Computer Vision*, 2001. 2

[11] Steven B Goldberg, Mark W Maimone, and Larry Matthies. Stereo vision and rover navigation software for planetary exploration. In *IEEE Aerospace Conference*, 2002. 2

[12] Tiejun Huang, Yajing Zheng, Zhaofei Yu, Rui Chen, Yuan Li, Ruiqin Xiong, Lei Ma, Junwei Zhao, Siwei Dong, Lin Zhu, et al. 1000× faster camera and machine vision with ordinary devices. *Engineering*, 2022. 2

[13] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. DEFOM-Stereo: Depth foundation model based stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1, 5, 7

[14] Jianing Li, Jiaming Liu, Xiaobao Wei, Jiyuan Zhang, Ming Lu, Lei Ma, Li Du, Tiejun Huang, and Shanghang Zhang. Uncertainty guided depth fusion for spike camera. *arXiv preprint arXiv:2208.12653*, 2022. 2, 3

[15] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5, 7

[16] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo R-CNN based 3D object detection for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[17] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 6, 7, 8

[18] Sameer A Nene and Shree K Nayar. Stereo with mirrors. In *International Conference on Computer Vision*, 1998. 2

[19] Y. Nishimoto and Y. Shirai. A feature-based stereo model using small disparities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1987. 2

[20] Theodore P Pachidis and John N Lygouras. Pseudostereo-vision system: A monocular stereo-vision system as a sensor for real-time robot applications. *IEEE Transactions on Instrumentation and Measurement*, 2007. 2

[21] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[22] Stephan Schraml, Ahmed Nabil Belbachir, Nenad Milosevic, and Peter Schön. Dynamic stereo vision system for real-time tracking. In *IEEE International Symposium on Circuits and Systems*, 2010. 1

[23] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. 6

[24] William Teoh and X. Zhang. An inexpensive stereoscopic vision system for robots. In *IEEE International Conference on Robotics and Automation*, 1984. 2

[25] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1

[26] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5

[27] Shaobo Wang, Yuan Xu, Yanghao Zheng, Mingcheng Zhu, Haodong Yao, and Zhiyong Xiao. Tracking a golf ball with high-speed stereo vision system. *IEEE Transactions on Instrumentation and Measurement*, 2018. 2

[28] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. 6, 7, 8

[29] Yixuan Wang, Jianing Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Learning stereo depth estimation with bio-inspired spike cameras. In *IEEE International Conference on Multimedia and Expo*, 2022. 2, 3

[30] Lu Yang, Baoqing Wang, Ronghui Zhang, Haibo Zhou, and Rongben Wang. Analysis on location accuracy for the binocular stereo vision system. *IEEE Photonics Journal*, 2017. 1, 2

[31] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024. 1, 7

[32] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, 2022. 2, 7

[33] Zhe Zhang, Rui Peng, Yuxi Hu, and Ronggang Wang. GeoMVSNet: Learning multi-view stereo with geometry perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[34] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

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

[36] Zhengming Zhou and Qiulei Dong. Two-in-one depth: Bridging the gap between monocular and binocular self-supervised depth estimation. In *International Conference on Computer Vision*, 2023. 1

240FPS Stereo Vision from Monocular Mixed Spikes

Yeliduosi Xiaokaiti[1,2] Yakun Chang[3,4] Yang Bai[1,2] Zhaojun Huang[1,2] Peiqi Duan[1,2] Boxin Shi[1,2,5*]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University

[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University

[3] Institute of Information Science, Beijing Jiaotong University

[4] Visual Intelligence +X International Cooperation Joint Laboratory of the MoE [5] PKU-AI[2] Robotics Joint Lab of Embodied AI

{yongqiye, by_baiyang, huangzhaojun}@stu.pku.edu.cn

ykchang@bjtu.edu.cn, {duanqi0001, shiboxin}@pku.edu.cn

In this supplementary material, we provide additional implementation details, extended experimental results, and a deeper discussion on system limitations. Specifically, Sec. 6 details the hardware calibration (Sec. 6.1) and the measurement of the LCD transmittance curve (Sec. 6.2). Sec. 7 presents further qualitative results on real indoor scenes (Sec. 7.1), compares our approach with single stereo systems (Sec. 7.2), and describes the accompanying video demonstrations (Sec. 7.3). Finally, Sec. 8 provides an extended discussion on the limitations of our current hardware prototype.

6. Implementation details

6.1. Hardware calibration details

As discussed in Section 3.3, the planar mirror configuration induces a depth disparity between the two virtual viewpoints. As illustrated in Fig. 9, our system can be modeled using two virtual cameras. For the left viewpoint, the optical path reflects off the beam splitter directly into the physical camera. Consequently, the image captured by the physical camera is spatially equivalent to the one captured by a left virtual camera positioned at a distance l_1 from the beam splitter, subject to a horizontal mirroring due to the single reflection. In contrast, the optical path for the right viewpoint reflects off the planar mirror and passes through the beam splitter. This creates a right virtual camera located at a distance of $l_1 + l_2$ from the beam splitter. As a result, there is a relative depth offset of l_2 between the right and left virtual cameras. Since both optical paths undergo a single reflection, the raw captured frames are mirror images of the scene. Therefore, a horizontal flip operation is applied to the mixed frames to align them with the coordinate systems of standard virtual cameras.

Calibration of this depth offset is essential. A significant advantage of the system is the identical intrinsic parameters of the physical camera and both virtual cameras. Moreover, the system is maintained in a fixed configuration. By capturing multiple images of a checkerboard pattern, the left and right view images can be decoupled using the baseline decoupling method described in Sec. 3.1. Subsequently, standard stereo calibration and rectification procedures are applied to these separated and horizontally flipped image pairs.

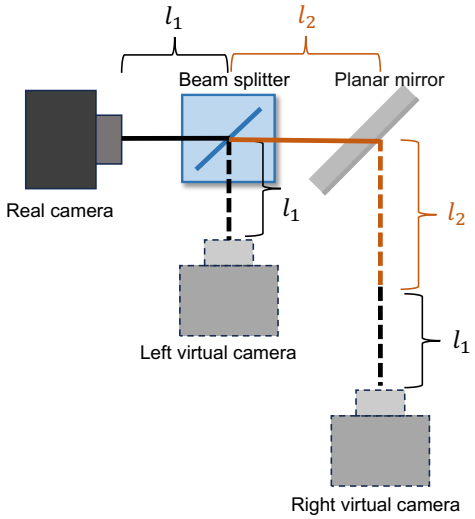

Figure 9. Schematic of the virtual camera setup for calibration. The left virtual camera is located at a distance l_1 from the beam splitter, while the right virtual camera is at a distance of $l_1 + l_2$, introducing a depth offset of l_2. Note that due to the single reflection in both optical paths, the viewpoints correspond to mirrored virtual cameras; the captured images are horizontally flipped during pre-processing to restore the natural scene orientation.

6.2. LCD modulator

Our LCD modulator features two primary states: transparent at 0 V and black when voltage is applied. The switching

*Corresponding author.

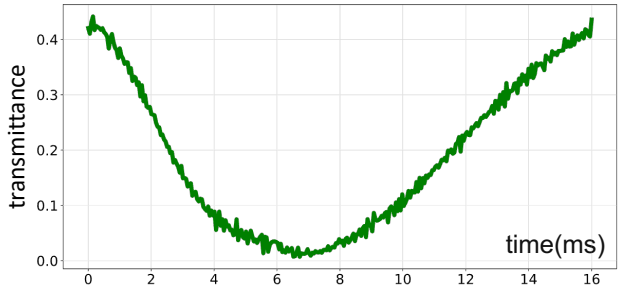

Figure 10. Reconstructed time-transmittance curve of an LCD modulator measured using a 20000 Hz spike camera. The curve shows the switching characteristics between transparent (0 V) and black states, with a total measurement duration of 16.7 ms.

time from transparent to black is a minimum of 5 ms, while the reverse transition requires at least 10 ms. This defines a minimum possible cycle time of 15 ms, or a theoretical maximum frequency of 66.7 Hz. For enhanced periodic stability, the modulator is driven at a refresh rate of 60 Hz.

We utilize a 20000 Hz spike camera to capture the detailed temporal variations in the LCD's transmittance, with the core goal of reconstructing its continuous transmittance function $f(t)$. The experiment involves placing an LCD in front of a fixed camera viewing a static scene. Upon activating the LCD, the camera records a video. In this video, the intensity of pixel p at time t is denoted by $I(t, p)$, and the trigger model follows:

$$Q(p) + \Delta Q_{p,i} \;=\; \int_{t_{i-1}}^{t_i} \alpha(p)\, I(t,p)\, f(t)\, dt, \tag{9}$$

where the notation is consistent with the main text, using p to denote a pixel instead of (x, y). Since the captured scene is static, $I(t, p)$ is time-invariant. Therefore, it can be extracted from the integral and denoted as $I(p)$:

$$Q(p) + \Delta Q_{p,i} \;=\; \alpha(p)\, I(p) \int_{t_{i-1}}^{t_i} f(t)\, dt. \tag{10}$$

To solve this integral equation, we discretize $f(t)$ over $[0, t_{\max}]$ into J segments of duration $\Delta\tau$, and approximate it as piecewise constant:

$$f(t) \approx F_j, \quad \text{for } t \in [\tau_{j-1}, \tau_j],\ \tau_j = j{\times}\Delta\tau,\ j = 1, \ldots, J. \tag{11}$$

Define the overlap length between the j-th time segment and the i-th spike interval as

$$M_{i,j} \;\triangleq\; \text{length}\Big([\tau_{j-1}, \tau_j] \cap [t_{i-1}, t_i]\Big). \tag{12}$$

Then

$$Q(p) + \Delta Q_{p,i} \;\approx\; \alpha(p)\, I(p) \sum_{j=1}^{J} M_{i,j}\, F_j. \tag{13}$$

For a given pixel p, if N_p valid spikes are recorded over $[0, t_{\max}]$, we construct $M_p \in \mathbb{R}^{N_p \times J}$ by stacking $M_{i,j}$ row-wise. Let $\boldsymbol{F} = [F_1, \ldots, F_J]^\top$ and $\mathbf{1}_{N_p} = [1, \ldots, 1]^\top$. Define the per-pixel scale

$$A_p \;\triangleq\; \frac{\alpha(p)\, I(p)}{Q(p)} \;>\; 0. \tag{14}$$

Dividing both sides by $Q(p)$ and absorbing noise into the residual yields the vector form

$$\mathbf{1}_{N_p} \;\approx\; A_p\, M_p\, \boldsymbol{F}. \tag{15}$$

Aggregating all m pixels, we solve

$$\min_{\boldsymbol{F} \geq 0,\ A_p > 0} \sum_{p=1}^{m} \big\| \mathbf{1}_{N_p} - A_p\, M_p\, \boldsymbol{F} \big\|_2^2 \quad \text{s.t. } \|\boldsymbol{F}\|_2 = 1. \tag{16}$$

We solve this non-convex problem using an Alternating Least Squares (ALS) algorithm, which iteratively updates the variables $\{\boldsymbol{F}, A_p\}$ through the following steps:

(i) Fixing $\boldsymbol{F}$ and updating A_p:
We first update the scale A_p for each pixel while holding $\boldsymbol{F}$ fixed. The optimal solution has a closed-form:

$$A_p \;=\; \frac{\mathbf{1}_{N_p}^\top \boldsymbol{s}_p}{\boldsymbol{s}_p^\top \boldsymbol{s}_p + \varepsilon}, \tag{17}$$

where $\boldsymbol{s}_p = M_p \boldsymbol{F}$ and $\varepsilon > 0$ is a small constant for numerical stability.

(ii) Fixing $\{A_p\}$ and updating $\boldsymbol{F}$:
Next, we update $\boldsymbol{F}$ while holding all $\{A_p\}$ fixed. We first stack all observation equations into a single global system. Let $N = \sum_{p=1}^{m} N_p$ be the total number of observations. We form a stacked matrix $M \in \mathbb{R}^{N \times J}$ and a target vector $\boldsymbol{b} \in \mathbb{R}^N$, where its k-th element is $b_k = \frac{1}{A_{g_k}}$ (with g_k being the pixel index for observation k). The update for $\boldsymbol{F}$ is then found by solving the following Non-Negative Least Squares (NNLS) problem:

$$\min_{\boldsymbol{F} \geq 0} \; \| M\boldsymbol{F} - \boldsymbol{b} \|_2^2.$$

(iii) Normalization:
To enforce the constraint $\|\boldsymbol{F}\|_2 = 1$, we normalize the vector after each update:

$$\boldsymbol{F} \leftarrow \boldsymbol{F} / \|\boldsymbol{F}\|_2. \tag{18}$$

These three steps are repeated iteratively until the solution converges. The final vector $\boldsymbol{F}$ as shown in Fig. 10 represents the reconstructed discrete LCD transmittance response, while the set of scaling factors $\{A_p\}$ indicates the relative gain of each pixel.

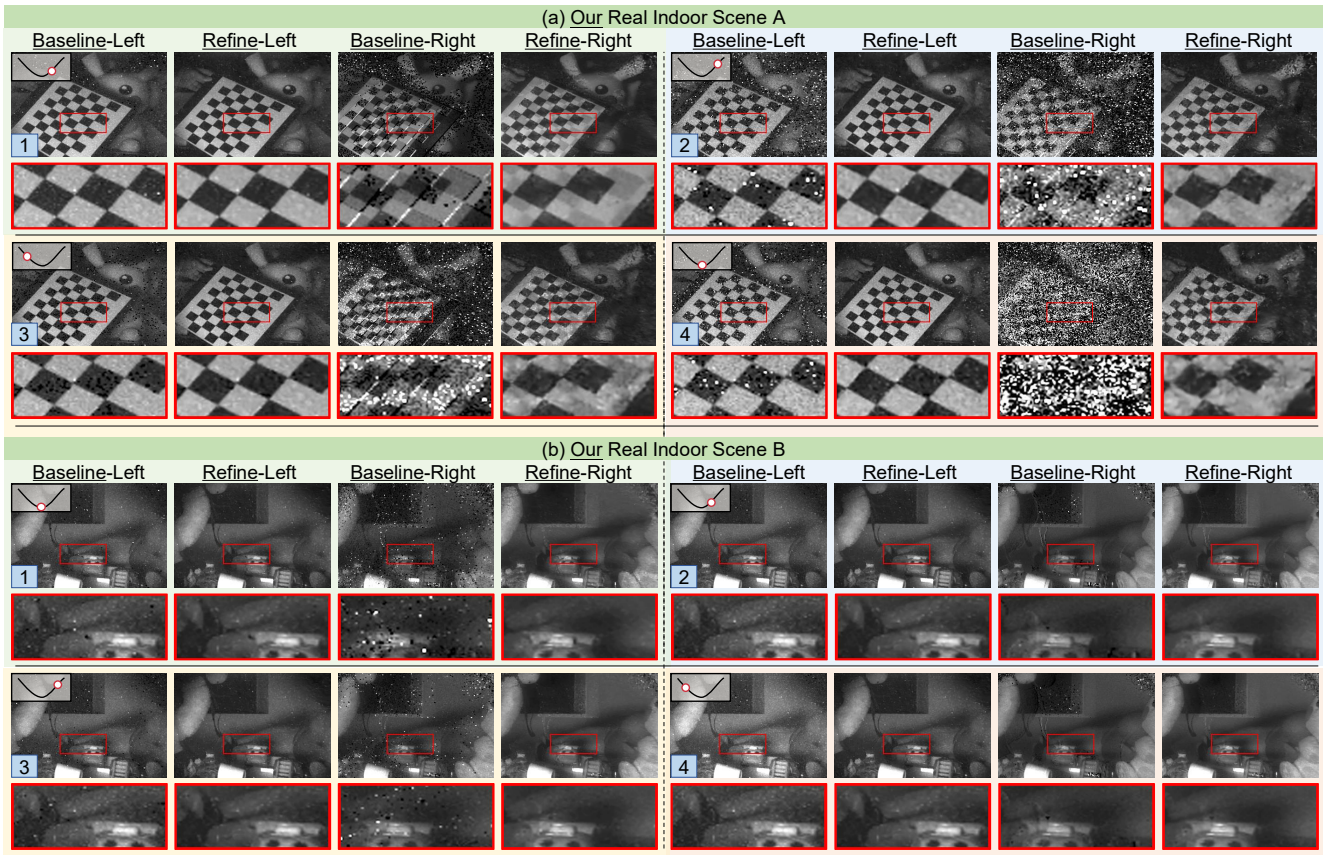

Figure 11. Qualitative results on real indoor scenes. We compare baseline decoupling and SMS-Net across left and right views for two scenes at four time instances (1-4). Red boxes denote zoomed-in patches highlighting noise reduction and detail preservation. The top-left overlay indicates the current LCD transmittance state.

7. Additional Experimental Results

7.1. Experiments on real indoor scenes

To validate real-world effectiveness, we captured indoor data. We evaluate performance based on both the visual quality of the decoupled images and the estimated depth.

Qualitative image evaluation. Fig. 11 presents the qualitative comparison between the baseline decoupling method and our proposed refinement approach. As observed in the zoomed-in patches, the baseline method suffers from noise and cross-view artifacts. Our SMS-Net effectively suppresses these artifacts while preserving high-frequency texture details. Notably, in Fig. 11(a) the right view produced by the baseline method is severely corrupted by noise; by contrast, our result consistently recovers fine details.

Depth estimation. Fig. 13 shows a comparison of depth estimation for two scenes. We adopt the same comparison method as used in Sec. 4.2. For stereo matching, we utilize the DEFOM [4] and CREStereo [5] algorithms. We also include results from monocular depth estimation models: DepthPro [1], DepthAnythingV2 [6], and Spike-T [7]. We present the depth results for four timestamps.

7.2. Comparison with single stereo systems

We compare our method with the coded-aperture method [3] and the dual-pixel method [2] on close-range scenes. For a fair comparison, we provide these methods with the grayscale ground-truth views as input. Evaluation metrics are computed only for the close-range region (1–5 m). As shown in Tab. 3 and Fig. 12, both quantitative and qualitative results show that our method achieves superior performance.

Table 3. Quantitative comparison on TartanAir dataset.

Method	AbsRel ↓	RMSE ↓	δ_1 ↑
Ikoma21 [3]	0.1167	0.4864	0.8488
He25 [2]	0.1392	0.5749	0.8188
Ours	**0.0299**	**0.2284**	**0.9795**

7.3. Video demonstrations

We provide a video in the supplementary material archive to demonstrate the performance of our method in dynamic scenes. The video showcases the following:

Dynamic illustration of our system: The video begins with a dynamic illustration of our hardware setup, showing how the LCD modulator, planar mirror, and beam splitter

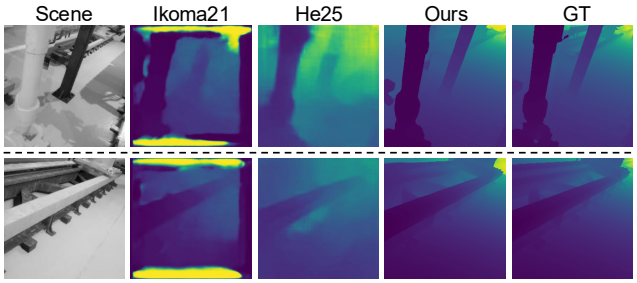

Figure 12. Qualitative comparison on the TartanAir dataset.

merge the videos from two views into one, resulting in a time-modulated, view-mixed spike stream.

Decoupling and depth estimation videos: The video corresponds to the qualitative results presented in Fig. 6, Fig. 7 and Fig. 8 of the main paper. The video also includes the corresponding results for the indoor scenes shown in Fig. 11 and Fig. 13. It shows side-by-side comparisons of our decoupled binocular videos and the resulting depth maps against baseline methods. It is worth noting that since the two depth estimation methods we employed process stereo frames independently and are not specifically designed for stereo video, some temporal flickering may be observed in the depth sequences.

8. Extended Discussion on Limitations

As briefly discussed in the conclusion of the main paper, our current hardware prototype employs an LCD modulator operating at approximately 60 Hz. This introduces a practical trade-off between the high temporal resolution of the spike camera and the "zero motion" assumption required by our decoupling strategy.

Specifically, to capture sufficient intensity variations in the modulated view, the effective time window for decoupling is bottlenecked by the 60 Hz refresh rate of the LCD, rather than the inherent speed of the scene motion. Consequently, when objects move significantly within this relatively wide time window, the zero-motion assumption is violated. This violation manifests as localized artifacts in the reconstructed output.

In future iterations, upgrading to a higher-speed optical modulator would allow for a substantially narrower time window. This would satisfy the zero-motion assumption even under extreme high-speed scenarios, fully unlocking the 240 FPS potential of our monocular spike camera system without introducing motion-related artifacts.

References

[1] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 3

[2] Fengchen He, Dayang Zhao, Hao Xu, Tingwei Quan, and Shaoqun Zeng. Simulating dual-pixel images from ray tracing for depth estimation. In *International Conference on Computer Vision*, 2025. 3

[3] Hayato Ikoma, Cindy M. Nguyen, Christopher A. Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *International Conference on Computational Photography*, 2021. 3

[4] Hualie Jiang, Zhiqiang Lou, Laiyan Ding, Rui Xu, Minglang Tan, Wenjie Jiang, and Rui Huang. DEFOM-Stereo: Depth foundation model based stereo matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[5] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[6] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024. 3

[7] Jiyuan Zhang, Lulu Tang, Zhaofei Yu, Jiwen Lu, and Tiejun Huang. Spike transformer: Monocular depth estimation for spiking camera. In *European Conference on Computer Vision*, 2022. 3

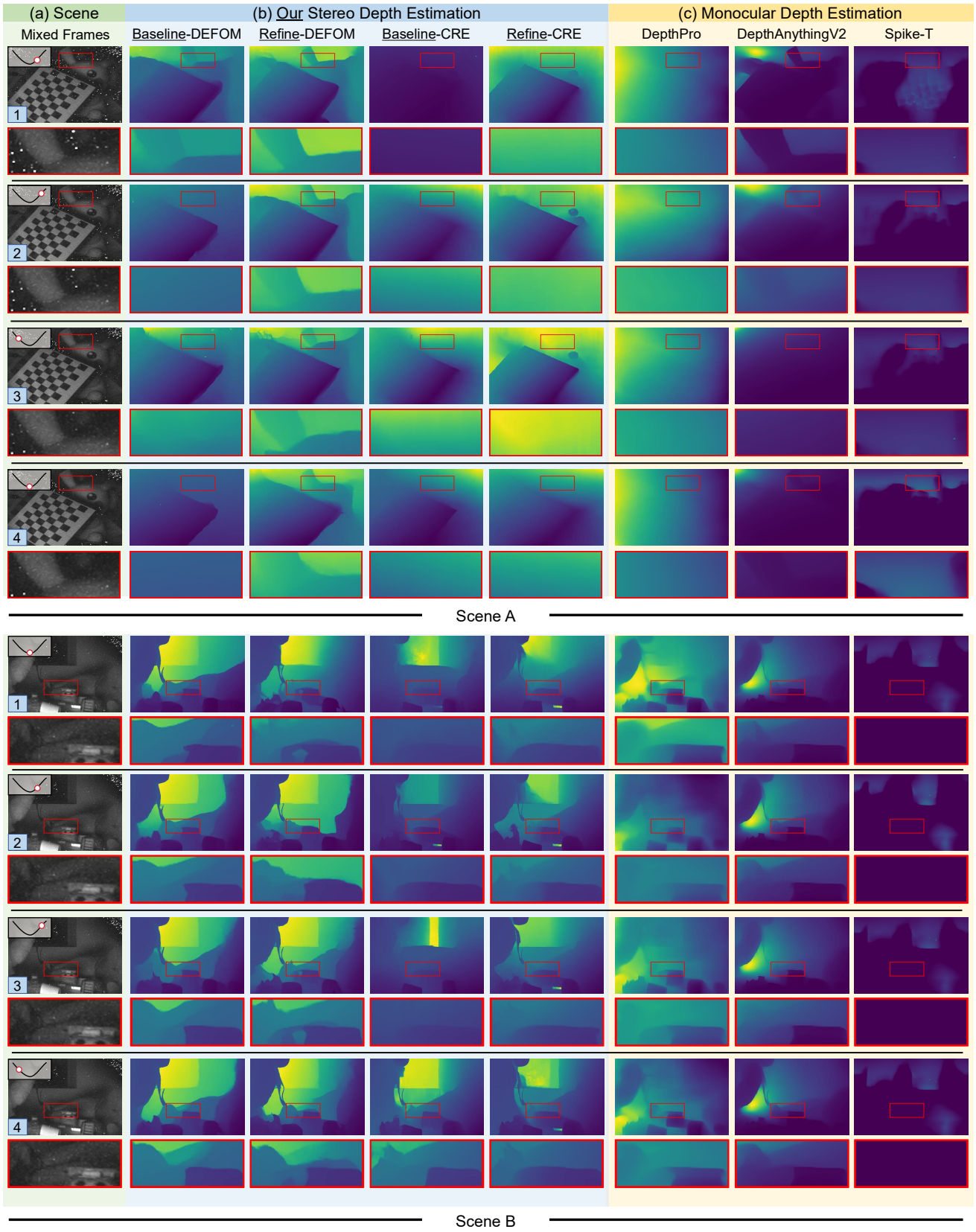

Figure 13. Qualitative evaluation of downstream depth estimation on Scene A and Scene B. (a) Mixed input frames captured at four LCD modulation states (labeled 1–4); the overlay plot shows the LCD transmittance. (b) Depth maps produced by stereo matching algorithms comparing inputs decoupled by the Baseline Decoupling and by our SMS-Net. (c) Monocular depth estimates for reference. The bottom row shows cropped regions (red boxes) highlighting that our Refined decoupling yields cleaner depth discontinuities and substantially fewer artifacts than the Baseline, while monocular methods lack fine geometric detail.