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.

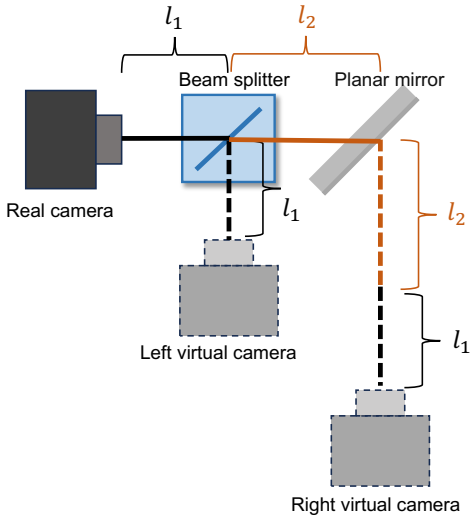

Figure 9. Schematic of the virtual camera setup for calibration. The left virtual camera is located at a distance l_1 from the beam splitter, while the right virtual camera is at a distance of $l_1 + l_2$, introducing a depth offset of l_2. Note that due to the single reflection in both optical paths, 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.

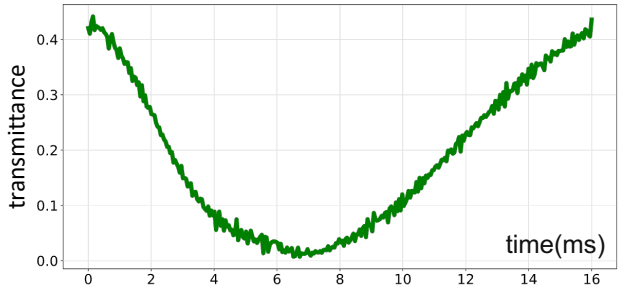

Figure 10. Reconstructed time-transmittance curve of an LCD modulator measured using a 20000 Hz spike camera. The curve shows the switching characteristics between transparent (0 V) and black states, with a total measurement duration of 16.7 ms.

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.

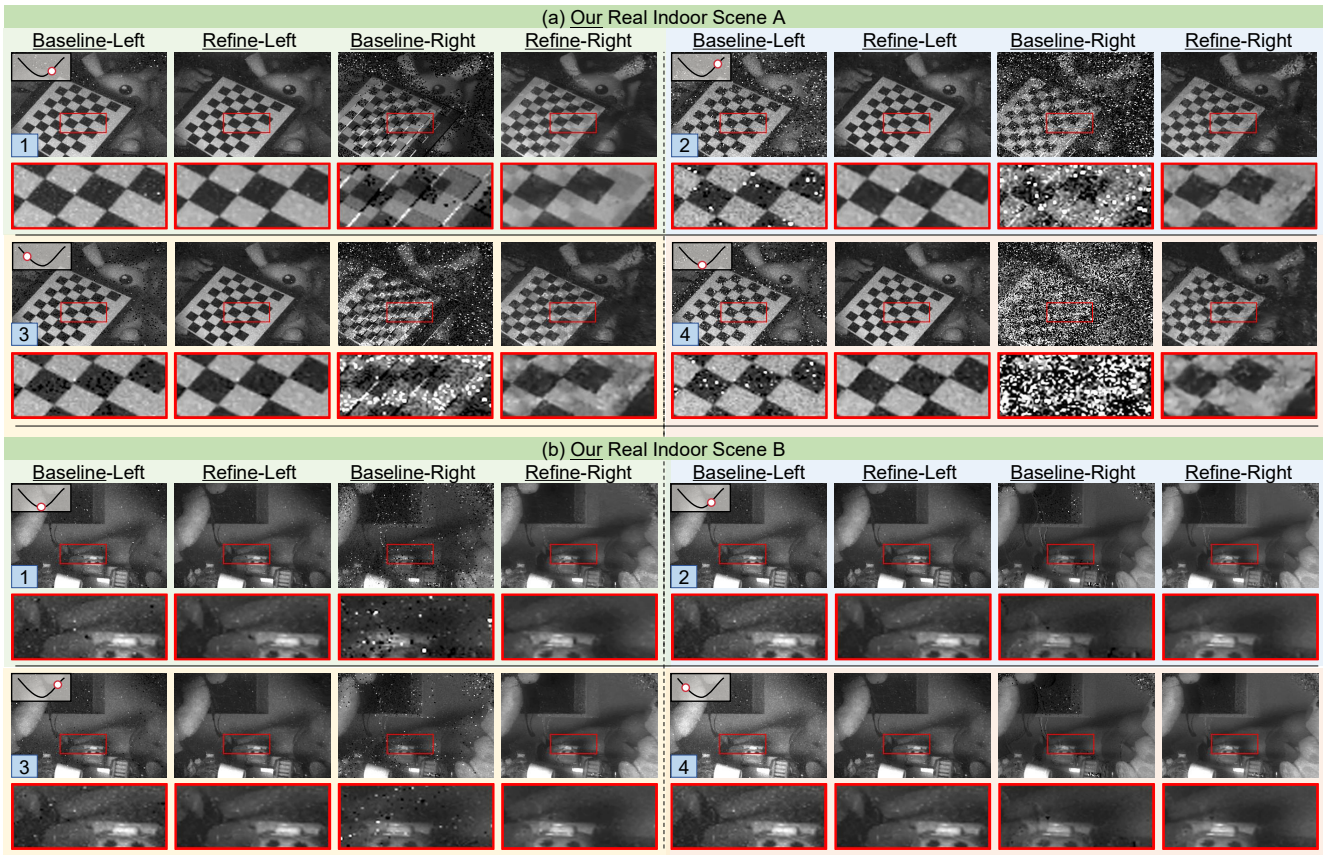

Figure 11. Qualitative results on real indoor scenes. We compare baseline decoupling and SMS-Net across left and right views for two scenes at four time instances (1-4). Red boxes denote zoomed-in patches highlighting noise reduction and detail preservation. The top-left overlay indicates the current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

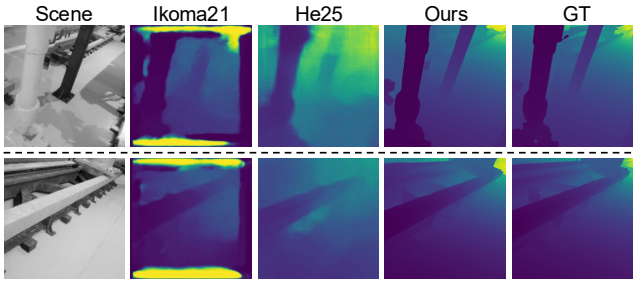

Figure 12. Qualitative comparison on the TartanAir dataset.

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

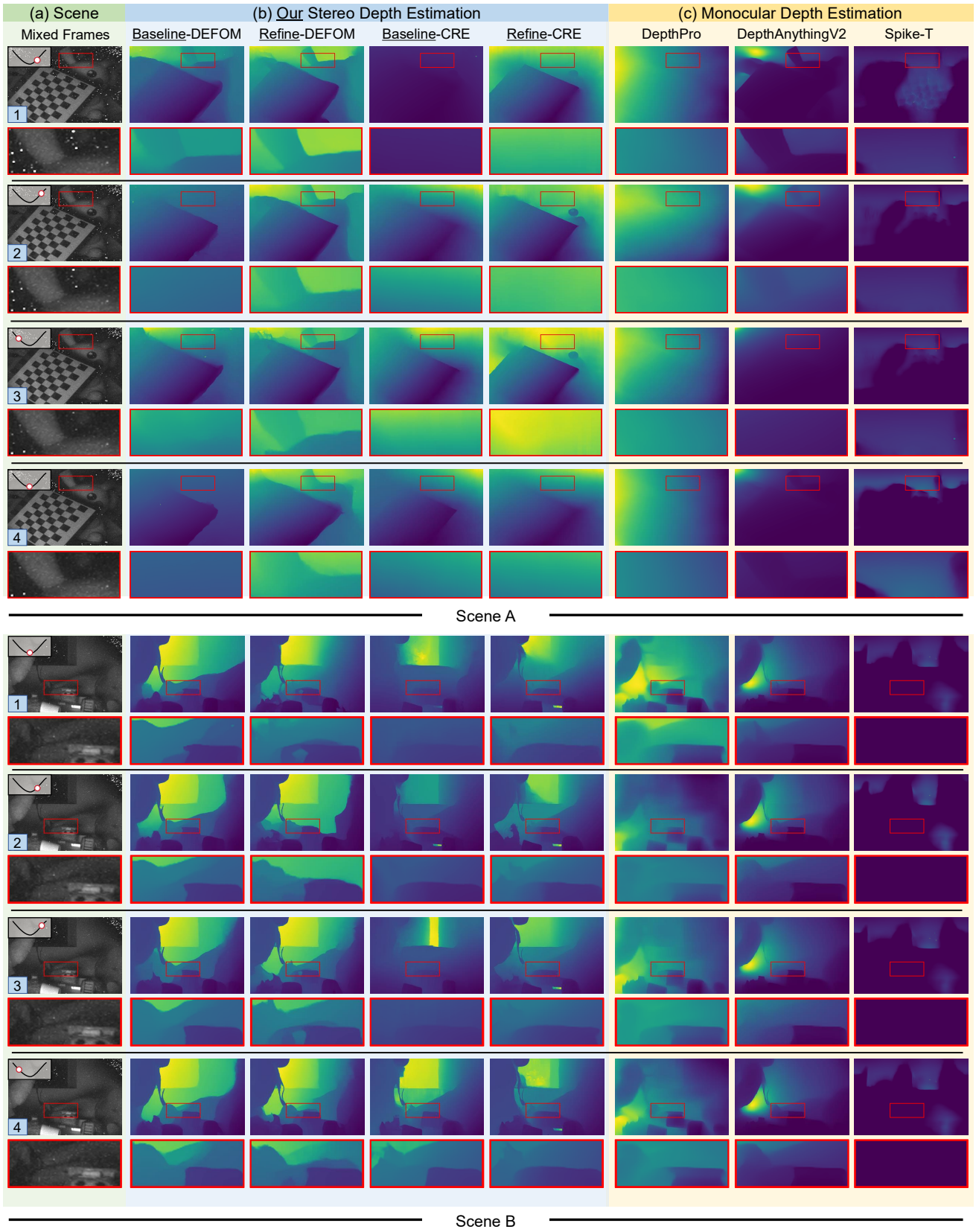

Figure 13. Qualitative evaluation of downstream depth estimation on Scene A and Scene B. (a) Mixed input frames captured at four LCD modulation states (labeled 1–4); the overlay plot shows the LCD transmittance. (b) Depth maps produced by stereo matching algorithms comparing inputs decoupled by the Baseline 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.