Supplementary Material
1000 FPS HDR Video with a Spike-RGB Hybrid Camera

Yakun Chang[1,2] Chu Zhou[3] Yuchen Hong[1,2] Liwen Hu[2] Chao Xu[3] Tiejun Huang[1,2] Boxin Shi[1,2*]
[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] National Key Laboratory of General AI, School of Intelligence Science and Technology, Peking University
{yakunchang, zhou_chu, huliwen, tjhuang, shiboxin}@pku.edu.cn
yuchenhong.cn@gmail.com, xuchao@cis.pku.edu

In the supplementary material, we provide details of optical flow estimation (Sec. 3.2),our real-synthetic data (Sec. 3.5), and show additional comparisons with the state-of-the-art method [2] (Sec. 4). We further provide a supplementary video to show the motivation of our methodology and results for both synthetic and real-world scenes.

6. Details of Optical Flow Estimation

The source code of SC-Flow [3] outputs optical flow vectors every 1 ms. However, $\mathbf{F}_{j\rightarrow i}$ in Eqn. 4 requires flow vectors estimated from longer time interval. To obtain more accurate optical flow vectors, we first initialize the flow vectors by accumulating the optical flows from j to i, then we conduct refinement by secondly feeding the initial flows to SC-Flow [3].

7. Details of Real-Synthetic Data

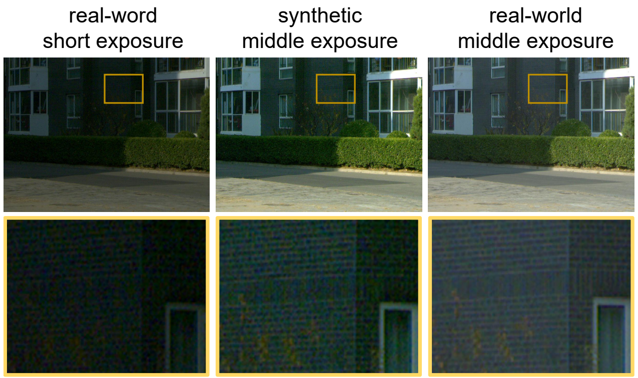

Figure 11. Visualization for noise levels of a short-exposure image (1 ms), a synthetic middle-exposure image by merging 4 short-exposure images, and a real-world middle-exposure image.

Since the dataset provided by Chen *et al.* [2] only contains low frame rate (LFR) alternating-exposure RGB sequence, we cannot synthesize high-speed spike trains from such an LFR dataset. To collect alternating-exposure images and spike trains in high-speed conditions, it may be feasible to synthesize them from HFR videos captured with a short exposure. Theoretically, a middle-exposure image can be synthesized by merging several short-exposure images if there is no camera noise. However, as shown in Fig. 11, we find that the short-exposure image (left) captured with 1 ms exposure contains strong noise, which cannot be effectively suppressed by merging a burst of short-exposure images. In the middle column of Fig. 11, the synthetic middle-exposure image contains strong noise as well, whereas the real-world middle-exposure image (right) captured with 4 ms contains less noise. The reason is that images captured with short exposures are more severely contaminated by camera noise, and the camera noise (the mean is not zero) is also accumulated when we merge short-exposure images. Since it is infeasible to synthesize longer-exposed images by merging a sequence of short-exposure images, we design a method to synthesize blurry longer-exposed images.

In this work, we collect the real-synthetic dataset from alternating-exposure RGB sequences captured in slow-motion conditions. Our pipeline for the synthesis of middle-exposure images is shown in Fig. 12. Firstly, we set the alternating exposures to 1 ms, 4 ms, and 12 ms, which are consistent with our real-world data. Then we capture RGB sequences in slow-motion conditions with a frame rate of 80 FPS (the largest frame rate in this exposure setting). To synthesize the ground truths, we treat each 3 adjacent alternating-exposure frames as a group and synthesize a well-exposed image (ground truth) using exposure fusion [5]. Since the total time T of each frame in the original RGB sequence is 12.5 ms, and we fuse 3 adjacent

[*] Corresponding author.

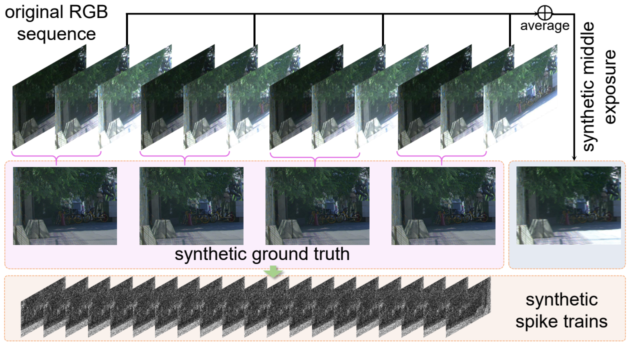

Figure 12. We capture alternating-exposure RGB sequence in slow-motion conditions, and compress the time with a ratio of 37.5 $ms \rightarrow 1\ ms$. We synthesize a blurry middle-exposure image by averaging 4 neighbored middle-exposure images.

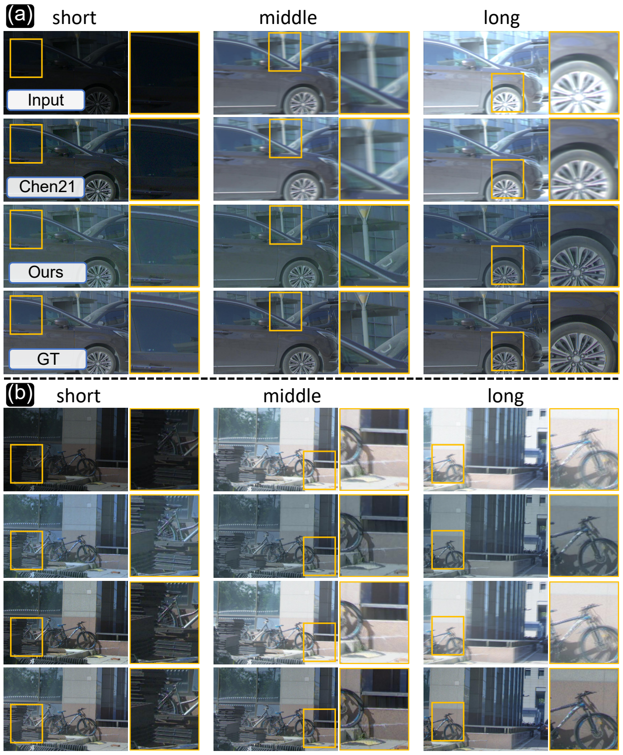

Figure 13. This is the following of Fig. 7 in the main paper. Visual quality comparison of real-synthetic data between the proposed method and the state-of-the-art HDR video reconstruction method: Chen 21 [2].

frames as a ground truth image (1 ms), the compression ratio of time is 37.5 $ms \rightarrow 1\ ms$. We select the synthetic data according to the temporal relationship of real-world data. For example, since the start time of the first three RGB frames in real-world data are at the first, the 17th, and the 34th ms, the first three synthetic alternating-exposure RGB frames are generated from group 1, group 17 to 20, and group 34 to 45, respectively. A short-exposure image is obtained by directly selecting the short-exposure image in group 1. To synthesize a blurry middle-exposure image, we average four middle-exposure images selected from group 17 to 20. Similarly, a blurry long-exposure image is synthesized by averaging the 12 long-exposure images of group 34 to 45. Finally, we synthesize spike trains by converting ground truth RGB frames to intensity maps and generate spike trains with the integrate-and-fire methodology [6]. Thanks to this method, the domain gap between the synthetic alternating-exposure images and real-world ones captured in high-speed condition is small.

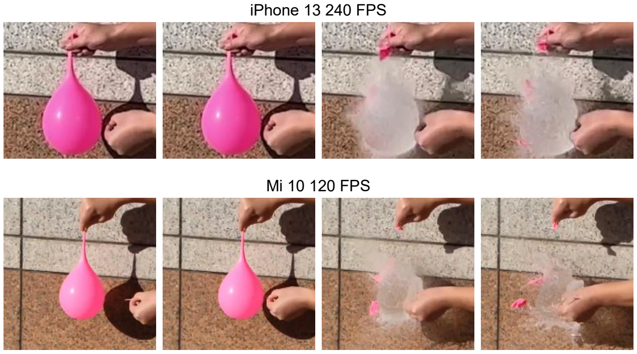

Figure 14. This is the following of Fig. 10 in the main paper. Please view this figure together with Fig. 10. In this figure, we compare our results on the balloon bursting with the slow-motion capability of iPhone 13 and Mi 10. We show 4 adjacent frames captured by the smartphones.

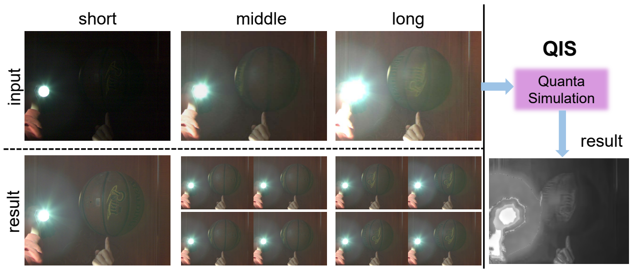

Figure 15. Testing in a scene with strong light. And the simulation of an HDR method that designed for quanta image sensors [1].

8. Additional qualitative results

In this section, we present additional two sets of visual comparisons on real-synthetic data. As shown in Fig. 13, our method recovers well-exposed color frames with less motion blur. Figure 14 is the following of Fig. 10 in the main paper, which presents the visual comparison of the balloon bursting with the slow-motion capability of two commercial cameras. We can see that the two cameras also fail to capture continuous motions of the balloon bursting. In Fig. 15, we present a set of results (3 RGB images and 9 output images) to validate HFR&HDR performance in conditions with strong light source. We can see a basketball is spinning rapidly at the fingertip beside a hand-held strong light. Our method successfully captures the texture details of the basketball without motion blur. Since quanta image sensors (QIS) [4], *e.g.*, the SPAD camera and Gigajot QIS series share similar imaging model with the spiking camera, we conduct comparison with QIS. And for the reason that we do not have a QIS camera on hand, the comparison is performed through a simulation with the source code provided by Abhiram and Chan [1].

References

[1] Gnanasambandam Abhiram and Chan Stanley H. HDR imaging with quanta image sensors: Theoretical limits and optimal reconstruction. *IEEE Transactions on Computational Imaging*, 6:1571–1585, 2020. 2, 3

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

[3] Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, and Tiejun Huang. Optical flow estimation for spiking camera. In *Proc. of Computer Vision and Pattern Recognition*, pages 17844–17853, 2022. 1

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

[5] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Pacific Conference on Computer Graphics and Applications*, pages 382–390, 2007. 1

[6] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2ImgNet: Learning to reconstruct dynamic scene from continuous spike stream. In *Proc. of Computer Vision and Pattern Recognition*, pages 11996–12005, 2021. 2