Deep Optics for Video Snapshot Compressive Imaging

Ping Wang[1,2] Lishun Wang[2] Xin Yuan[2,*]
[1]Zhejiang University [2]School of Engineering, Westlake University
{wangping,wanglishun,xyuan}@westlake.edu.cn

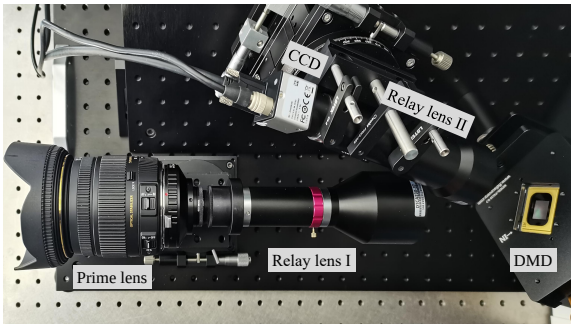


Figure 1. Our video SCI prototype.

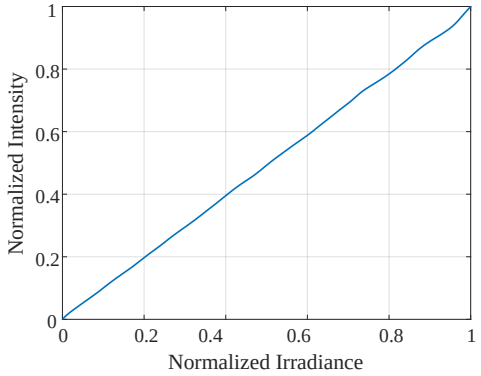


Figure 2. Camera response function in our prototype.

1. Details of Our Video SCI Prototype

1.1. Optical Setup

Our video SCI prototype is shown in Fig. 1, mainly composed of a prime lens, two relay lenses, an 8-bit charge coupled device (CCD) camera, and a digital micro-mirror device (DMD) with an array size of 768×1024. The micro-mirror size of DMD is 13.68 µm $\times$ 13.68 µm and the pixel size of CCD is 5.5 µm $\times$ 5.5 µm. Dynamic scene is first relayed through a prime lens and a relay lens onto an intermediate plan, where DMD is employed to introduce time-varying spatial modulation (masks) to the scene. The modulated scene is subsequently relayed to the CCD camera through another relay lens. Magnification of the second relay lens is chosen to provide approximately 1-to-1 element mapping between the DMD and CCD. Masks with 2-by-2 element size are implemented on the DMD. The synchronization between CCD and DMD is realized by an external trigger line and the response delay between them is empirically set to 45 us.

1.2. Camera Response Function

We measure camera response function f of the used CCD camera according to [3]. As shown in Fig. 2, there is an approximate linearity between scene irradiance and image intensity and thus it is modeled as $f(x) = x$ in our experiments.

*Corresponding author.

2. Experiment Details

2.1. Testing and Re-training with Sensor Response

Sensor response is overlooked in previous works [1, 2, 4, 10, 5, 6, 7, 8, 9] such that their excellent results on simulation could generally be reproduced in real video SCI systems. To evaluate different methods' real performance quantitatively, the simulated experiments are conducted with an 8-bit, approximately linear, and off-the-shelf camera and its sensor response $\mathcal{R}$ is modeled as

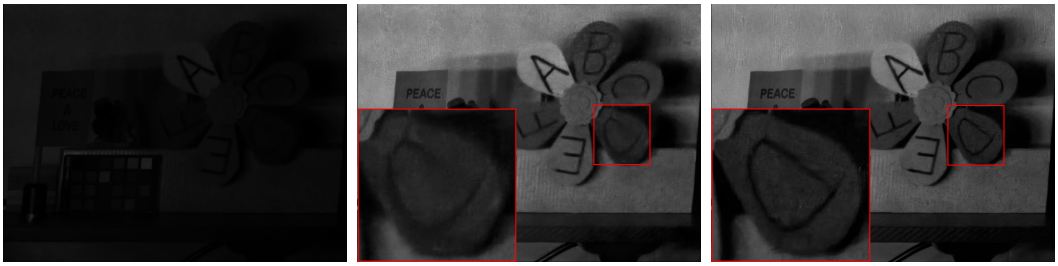
$$y = \lfloor 255 \cdot x + 0.5 \rfloor / 255, \tag{1}$$

where x and y is the normalized irradiance and the normalized image intensity, respectively.

During previous networks' testing (Tab. 2) and re-training (Tab. 3) in the main paper, global scaling which simulates automatic aperture is performed before sensor response to avoid measurement over-exposure (because the over-exposured region is irreversible). Therefore, the testing and re-training pipeline follows

$$\begin{aligned} \textbf{Encoder}: &\quad \boldsymbol{y} = \mathcal{R} \circ \mathcal{H}(\boldsymbol{x})/\gamma, \\ \textbf{Decoder}: &\quad \hat{\boldsymbol{x}} = \mathcal{D}(\boldsymbol{y}), \end{aligned} \tag{2}$$

where $\gamma = \max(\mathcal{H}(\boldsymbol{x}))$ and $\mathcal{D}$ denotes one of the well-trained or re-trained networks. As a result of considering sensor response, the captured dynamic range is severely limited to avoid measurement overexposure. The re-trained



(a) (b) (c)

Figure 3. The reconstructed results of the first `Windmill` frame. The raw result (a) is retrieved from an 8-bit raw measurement and (b) is a result of brightening (a). Compared with (a) and (b), a better result (c) can be retrieved from a magnified measurement.

networks in Tab. 3 of main paper are simultaneously faced with two tasks: i) compressed video reconstruction and ii) dynamic range reconstruction, and thus lead to worse results.

Note that PSNR is susceptible to luminance distortion as a metric of estimating the absolute errors between distortion-free image and the recovered one. To evaluate different networks' performance of retrieving structural information, we compute $\mathrm{PSNR}(\boldsymbol{x}, \hat{\boldsymbol{x}}/\max(\hat{\boldsymbol{x}}) \cdot \max(\boldsymbol{x}))$ for Tab. 2 and Tab. 3 in main paper (friendlier than absolute-error based PSNR).

2.2. Dynamic Range Augmentation

As mentioned in main paper, previous reconstruction networks for video SCI (with random binary mask) did not consider the limited dynamic range of sensor. Taking a 10-frame video SCI (almost 5 scene images need to be integrated into a single-shot measurement image) as an example, only about 51 pixel values are available for the video frames given an 8-bit (256 pixel values) camera sensor. Therefore, it is said that a serious dynamic range degradation is rooted in previous works.

When applying previous methods in a real system, the recovered video frames are too dark as shown in Fig. 3 (a). To address this problem, an intuitive solution is to brighten Fig. 3 (a) into Fig. 3 (b). Another solution is to multiply measurement by $B/2$ (B denotes compressed frames) after inputting into a reconstruction network. The second solution guarantees the testing input close to the training input in terms of value range and thus leads to a better result as shown in Fig. 3 (c). We employ the second solution to get the real and simulated results of previous methods in this work.

3. More Results

Please refer to `Car.avi`, `Windmill.avi`, and `Kobe.avi` files in the same path. The detailed information is described in Tab. 1

References

[1] Ziheng Cheng, Bo Chen, Guanliang Liu, Hao Zhang, Ruiying Lu, Zhengjue Wang, and Xin Yuan. Memory-efficient network for large-scale video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16246–16255, June 2021. 1

[2] Ziheng Cheng, Ruiying Lu, Zhengjue Wang, Hao Zhang, Bo Chen, Ziyi Meng, and Xin Yuan. BIRNAT: Bidirectional recurrent neural networks with adversarial training for video snapshot compressive imaging. In *Proceedings of the European conference on computer vision (ECCV)*, August 2020. 1

[3] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 1

[4] Jiawei Ma, Xiaoyang Liu, Zheng Shou, and Xin Yuan. Deep tensor admm-net for snapshot compressive imaging. In *IEEE/CVF Conference on Computer Vision (ICCV)*, 2019. 1

[5] M. Qiao, Z. Meng, J. Ma, and X. Yuan. Deep learning for video compressive sensing. *APL Photonics*, 5(3):030801, 2020. 1

[6] Lishun Wang, Miao Cao, Yong Zhong, and Xin Yuan. Spatial-temporal transformer for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2022. 1

[7] Zhengjue Wang, Hao Zhang, Ziheng Cheng, Bo Chen, and Xin Yuan. Metasci: Scalable and adaptive reconstruction for video compressive sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2083–2092, June 2021. 1

[8] Zhuoyuan Wu, Jian Zhang, and Chong Mou. Dense deep unfolding network with 3d-cnn prior for snapshot compressive imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4892–4901, 2021. 1

[9] Chengshuai Yang, Shiyu Zhang, and Xin Yuan. Ensemble learning priors unfolding for scalable snapshot compressive sensing. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1

[10] Shirin Jalali Ziyi Meng and Xin Yuan. Gap-net for snapshot compressive imaging. *arXiv: 2012.08364*, December 2020. 1

Table 1. Description of the attached `.avi` files.

Filename	Real / Simulated	Compressed Ratio	Total Frames	Description
`Car.avi`	Real	1/10	100	i) 10-frame measurement is captured using our learned structural mask or widely-used random binary mask by our video SCI prototype. ii) 100-frame video is reconstructed using previous SOTA STFormer or our Res2former. iii) The reconstructed videos are at 1/25x speed.
`Windmill.avi`	Real	1/10	100	
`Kobe.avi`	Simulated	1/8	32	Considering sensor response, different methods are used toConsidering sensor response, different methods are used to retrieve an 32-frame video from an 4-framemeasurement. retrieve an 32-frame video from an 4-frame measurement.