Supplementary Material for "EfficientSCI: Densely Connected Network with Space-time Factorization for Large-scale Video Snapshot Compressive Imaging"

Lishun Wang [1,2,*], Miao Cao [3,4,*] , and Xin Yuan [3,†]
[1] Chengdu Institute of Computer Application Chinese Academy of Sciences,
[2] University of Chinese Academy of Sciences, [3] Westlake University, [4] Zhejiang University

1. Results on Mid-scale Color Datasets

We compare our method with SOTA model-based methods (GAP-TV [6], DeSCI [3], PnP-FFDNet [7] and PnP-FastDVDnet [8]) and deep learning-based method (BIRNAT-color [2]) on six benchmark mid-scale color datasets (`Beauty`, `Bosphorus`, `Jockey`, `Runner`, `ShakeNDry` and `Traffic` with a size of $512 \times 512 \times 3 \times 8$). Among them, PnP-FFDNet and PnP-FastDVDnet have grayscale and color versions, which are used to indicate that they use a grayscale denoiser and a color denoiser, respectively. Table 2 shows the quantitative comparison results, it can be observed that our proposed EfficientSCI-B ('Base' version) can achieve the highest reconstruction quality and good real-time performance. In particular, the PSNR value of our method surpasses the existing best method BIRNAT-color by 2.02 dB on average. In addition, our proposed EfficientSCI-S ('Small' version) achieves high reconstruction quality with the best real-time performance. Fig. 1 shows the visual reconstruction results of some simulation data. By zooming in some local areas, we can observe that our method can recover sharper edges and more detailed information compared to previous state-of-the-art (SOTA) methods (with artifacts or over-smoothing).

Table 1. Comparison of memory consumption (MB) between our proposed method and other Transformer architectures during runtime on large-scale color datasets.

Method	Resolution	Memory (MB)
Swin	$1080 \times 1920 \times 3 \times 8$	13281
VSwin^{d2}	$1080 \times 1920 \times 3 \times 8$	19589
VSwin^{d4}	$1080 \times 1920 \times 3 \times 8$	> 24000
TimeSformer^{1}	$512 \times 512 \times 3 \times 8$	> 24000
TimeSformer^{2}	$1080 \times 1920 \times 3 \times 8$	> 24000
Ours	$1080 \times 1920 \times 3 \times 8$	8995

*Equal Contribution, † Corresponding Author

2. Comparison with Other Transformer Network Architectures

To further validate the memory effectiveness of our proposed CFormer block on large-scale color datasets, we compare with some current SOTA Transformer networks, including Swin Transformer (Swin) [4], Video Swin Transformer (VSwin) [5] and TimeSformer [1]. We have verified that our proposed method can achieve higher reconstruction quality in previous experiments, so here we only compare the memory consumption of different networks during runtime.

Table 1 shows the memory consumption of different network blocks during runtime, where VSwin^{d2} indicates that the local window depth is 2, VSwin^{d4} indicates that the local window depth is 4, and the local window space size for Swin and VSwin is 7×7. Compared to other Transformer network blocks, our proposed CFormer block has a lower memory consumption. On large-scale color data with a size of $1080 \times 1920 \times 3 \times 8$, the CFormer block only needs 8995 MB memory consumption, which is 35% less than Swin Transformer and 85% less than Video Swin Transformer. For TimeSformer and VSwin^{d4}, they cannot be applied to large-scale video reconstruction tasks due to the memory constraint.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1

[2] Ziheng Cheng, Bo Chen, Ruiying Lu, Zhengjue Wang, Hao Zhang, Ziyi Meng, and Xin Yuan. Recurrent neural networks for snapshot compressive imaging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[3] Yang Liu, Xin Yuan, Jinli Suo, David J Brady, and Qionghai Dai. Rank minimization for snapshot compressive imaging. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2990–3006, 2018. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Pro-*

Table 2. The average PSNR in dB (left entry), SSIM (right entry) and running time per measurement of different algorithms on 6 benchmark Mid-scale color datasets. Best results are in bold and the second-best results are underlined.

Method	Beauty	Bosphorus	Jockey	Runner	ShakeNDry	Traffic	Average	Running time(s)
GAP-TV	33.08, 0.964	29.70, 0.914	29.48, 0.887	29.10, 0.878	29.59, 0.893	19.84, 0.645	28.47, 0.864	10.80 (CPU)
DeSCI	34.66, 0.971	32.88, 0.952	34.14, 0.938	36.16, 0.949	30.94, 0.905	24.62, 0.839	32.23, 0.926	92640 (CPU)
PnP-FFDNet-gray	33.21, 0.963	28.43, 0.905	32.30, 0.918	30.83, 0.888	27.87, 0.861	21.03, 0.711	28.93, 0.874	13.20 (GPU)
PnP-FFDNet-color	34.15, 0.967	33.06, 0.957	34.80, 0.943	35.32, 0.940	32.37, 0.940	24.55, 0.837	32.38, 0.931	97.80 (GPU)
PnP-FastDVDnet-gray	33.01, 0.963	30.95, 0.934	33.51, 0.928	32.82, 0.900	29.92, 0.892	22.81, 0.776	30.50, 0.899	19.80 (GPU)
PnP-FastDVDnet-color	35.27,0.972	37.24, 0.971	35.63,0.950	38.22, 0.965	33.71, 0.949	27.49, 0.915	34.60, 0.953	52.2 (GPU)
BIRNAT-color	36.08, 0.975	38.30, 0.982	36.51, 0.956	39.65, 0.973	34.26, 0.951	28.03, 0.915	35.47, 0.959	0.98 (GPU)
EfficientSCI-S	37.39, 0.978	40.52, 0.987	38.09, 0.967	42.24, 0.984	35.03, 0.951	29.71, 0.938	37.16, 0.968	0.61 (GPU)
EfficientSCI-B	**37.51**, **0.979**	**40.89**, **0.988**	**38.49**, **0.969**	**42.73**, **0.985**	**35.19**, **0.953**	**30.13**, **0.943**	**37.49**, **0.970**	1.31 (GPU)

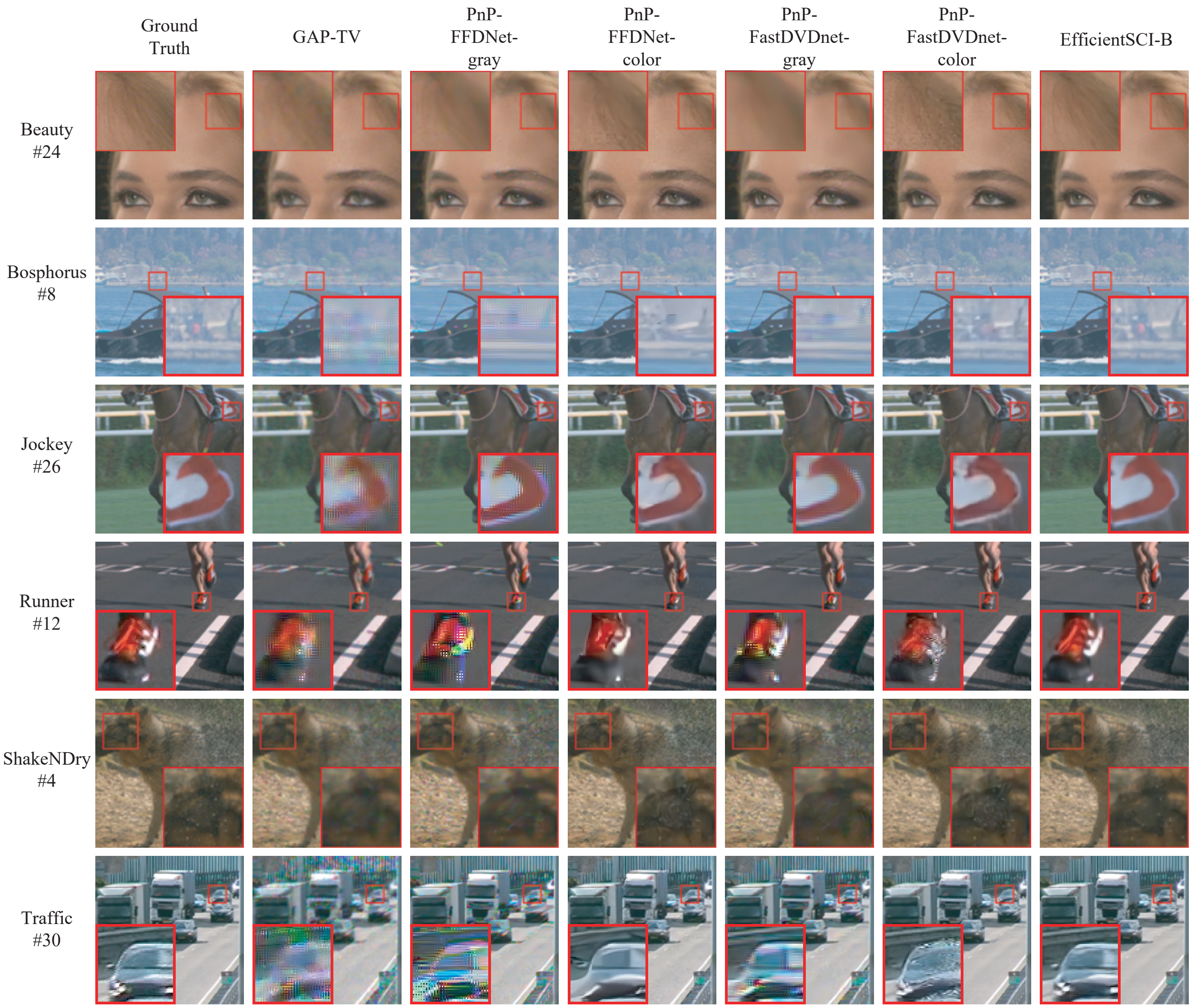

Figure 1. Selected reconstruction frames of simulated color data. Zoom in for better view.

ceedings of the IEEE/CVF International Conference on Computer Vision, pages 10012–10022, 2021. 1

[5] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 1

[6] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543. IEEE, 2016. 1

[7] Xin Yuan, Yang Liu, Jinli Suo, and Qionghai Dai. Plug-and-play algorithms for large-scale snapshot compressive imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1447–1457, 2020. 1

[8] Xin Yuan, Yang Liu, Jinli Suo, Fredo Durand, and Qionghai Dai. Plug-and-play algorithms for video snapshot compressive imaging. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–1, 2021. 1