Supplementary Materials for “Unfolding Framework with Prior of Convolution-Transformer Mixture and Uncertainty Estimation for Video Snapshot Compressive Imaging”

Anonymous ICCV submission

Paper ID 5951

Abstract

In this supplementary material, we provide the details of network structure, comparison of different algorithms' adaptability and full original results of both simulation and real experiments (in the .Zip file).

1. Details of Network Design

Assume x to be the input feature of each phase. x is first fed into 3D convolutional layers for feature extraction (downsampling). Before the BDA and DSA module, we both conduct the operation of **LayerNorm**. Unlike MaxViT[2] applying a standard MLP network after calculating the attention, we feed the output feature in to the FF module. Finally the Feature up sample block restores the same resolution as input. To upsample the feature map, ConvTranspose3D in conducted. There are four 3D conv layers in the feature extraction part. The corresponding kernel sizes are $5\times5\times5$, $3\times3\times3$, $1\times1\times1$, and $3\times3\times3$. There are one 3D convTranspose layer and three 3D conv layers in the decoder part for upsampling. The corresponding kernel sizes are $1\times3\times3$, $3\times3\times3$, $1\times1\times1$, and $3\times3\times3$. We employ 3D-CNN layers To extract the feature map of uncertainty. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper.

2. Ablation Study and Experimental Results

Ablation Study: In the paper, to test the efficiency of each module, we directly remove each part of the module separately. However, we should not ignore the effect brought by the reduction of parameter count. As shown in Tab. S1, the parameter count of FF is the largest among the three parts of CMT. Hence it brings the most obvious drop.

BDA	DSA	FF	Parameter	Flops
×	✓	✓	55M	4389G
✓	×	✓	55M	4389G
✓	✓	×	19M	2050G

Table S1. Computational complexity of different models.

Experimental Results: As shown in Tab. S2, all previous learning-based SOTA algorithms can not be adapted to different masks directly. Though MetaSCI claims that it can adapt to different masks, it still needs additional training. The adaptability of the algorithm greatly promotes the application of the algorithm in practical scenarios.

Algorithms	Trained mask	New mask 1	New mask 2
Ours	**36.52 0.985**	**36.51 0.985**	**36.53 0.985**
DUN-3DUnet	35.32 0.968	31.58 0.934	31.75 0.935
BIRNAT	33.31 0.951	23.13 0.730	23.10 0.730
RevSCI	33.92 0.956	18.99 0.537	18.93 0.535

Table S2. Quantitative comparison with different masks.

The measurements and all frames of the real data **Water Balloon** and **Domino** are shown in Fig. S1 and Fig. S2. We reconstruct multiple frames from the single measurement and achieve cleaner background, sharper edges, and more detail information.

3. Model for Color Video SCI

Compared to the grayscale model, the color model shares almost the same architecture. However, since the captured measurement of color video SCI system is a Bayer mosaicked measurement, we do not fuse the RF and uncertainty map into the model to avoid the unknown adverse factors. Different from previous methods such as [1, 3] which divide the measurement into four individual parts corresponding to the Bayer-filter, our proposed model directly takes the entire Bayer measurement as input and outputs the desired color videos at each stage. In addition, the output of the first stage is sent to the Bayer filter again before being fed into the next projection.

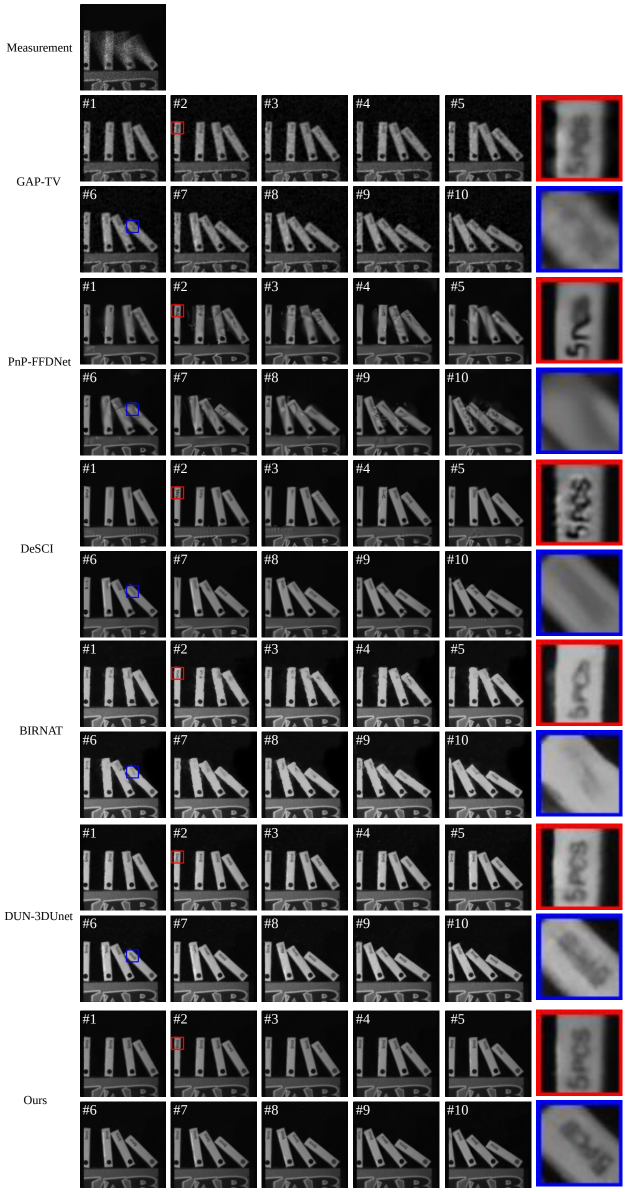

Figure S1. All reconstruction frames of **Domino**. Zoom in for better view.

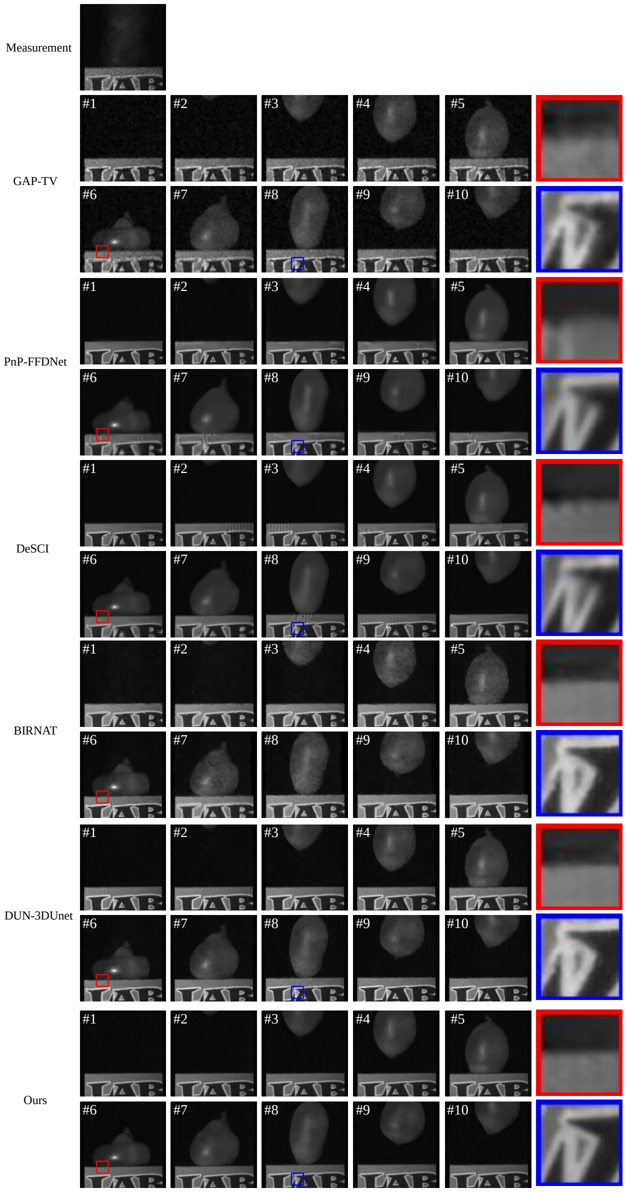

Figure S2. All reconstruction frames of **Water Balloon**. Zoom in for better view.

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*, pages 16246–16255, 2021. 1

[2] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 1

[3] Xin Yuan, Patrick Llull, Xuejun Liao, Jianbo Yang, David J. Brady, Guillermo Sapiro, and Lawrence Carin. Low-cost compressive sensing for color video and depth. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3318–3325, 2014. 1