Dual Prior Unfolding for Snapshot Compressive Imaging

Supplementary Material

Summary

This document supplies further analysis and comparisons of our method for comprehensive instructions. This additional content is structured as follows: Section 1 provides more details on the module design, including asymmetric backbone and improvement for SCI Transformer. Sections 2 and 3 present the visual and quantitative results on Focused Attention and additional SCI reconstruction results, providing more evidence of the efficacy of our modules.

1. Further Analysis of Module Design

In this section, we provide a more detailed analysis of various modules and some ablation experiment results.

1.1. Asymmetric Backbone for Hierarchical Module

As shown in Fig. 1(a), when we use hierarchical modules such as Swin attention in Unet, there is an effective non-local information modeling capability but also some computational burdens. To solve this problem, the proposed asymmetric backbone utilizes the skip connection of Unet to reduce computation without destroying the properties of the hierarchy, as shown in Fig. 1(b). In the attention ablation experiment of the main paper, we implement Swin attention through the half-split operation in DAUHST[4] and our asymmetric backbone respectively. The asymmetric backbone's effectiveness is evidenced by its superior performance, reduced computation, and fewer parameters. An ablation study under the setting of baseline-2 and Swin* in the main paper, replacing the asymmetric backbone's Swin modules with basic modules, further supports this. Detailed in Table 1, the results highlight the benefits of Swin attention in capturing non-local similarity and the function of the asymmetric backbone in maintaining the hierarchy.

Table 1. Ablation Comparison of Different Modules.

Module	PSNR	SSIM	Params (M)	FLOPs (G)
All Basic	35.95	0.951	1.38	22.57
Basic+Swin	**36.27**	**0.953**	1.38	22.57

1.2. Multi-Pattern MLP (MPMLP)

In this section, we aim to illustrate the differences in neuronal interaction between ordinary MLP and MPMLP as

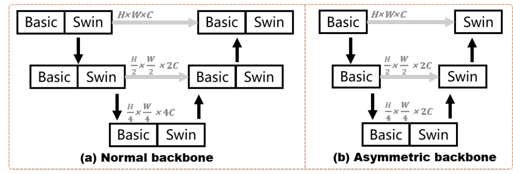

Figure 1. Maintain the hierarchy by different backbones.

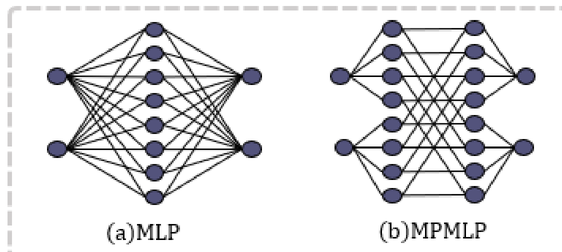

Figure 2. Neuronal Interaction in Different MLPs.

shown in Fig. 2. In vision transformers [6, 10], for input $X \in \mathbb{R}^{N\times C}$, the dimension of the middle feature in general tow-layer MLP is set to $4C$. Then we could get that the parameters of MLP are $8C^2$ while that of MPMLP is $8C + 4C^2$. Thus, the computation complexity of MLP is $8NC^2$ while that of MPMLP is $8NC + 4NC^2$. This means that MPMLP almost halves the number of parameters and computations when $C \gg 2$ is the general case. To verify this conclusion, we replace the MPMLP of DPU-5stg with the ordinary tow-layer MLP. Other experimental settings are the same as in the main paper and the results are reported in Table 2. Better performance, less computational overhead, and memory requirements demonstrate the effectiveness of MPMLP.

Table 2. Ablation Study of MLPs.

Modules	PSNR	SSIM	Params (M)	FLOPs (G)
MLP	39.23	0.971	1.85	31.49
MPMLP	**39.62**	**0.973**	**1.59**	**27.41**

1.3. Improvement for SCI Transformer

In a normal transformer, the feature dimensions of Q, K are the same as that of input $X \in R^{N\times C}$, which ensures the accuracy of similarity. In the SCI transformers, the feature dimension of X is expanded from the basic spectral data Λ bands after downsampling, that is, $C = k\Lambda, k \in \{1, 2, 4\}$.

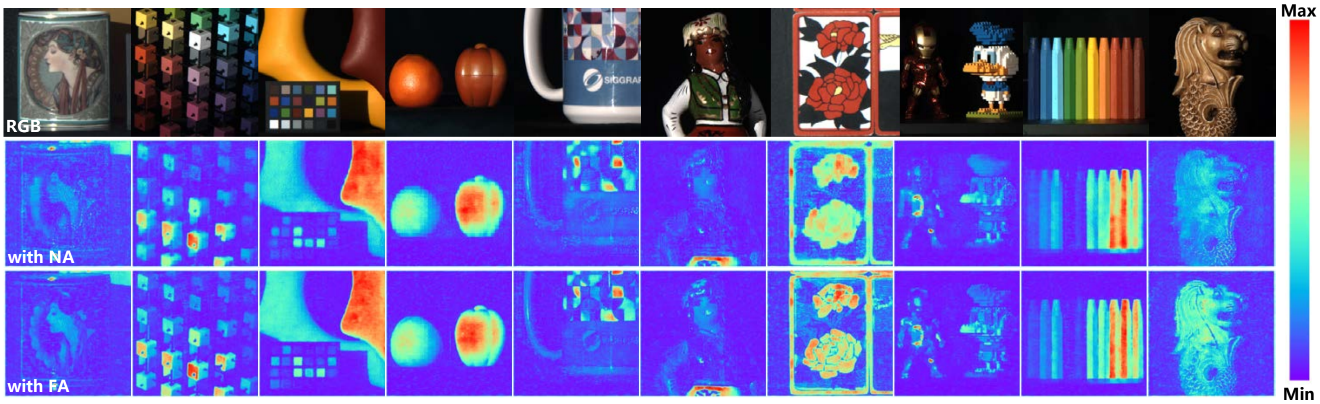

Figure 3. Constructed images with Normal Attention (NA) and Focused Attention (FA) on KAIST.

Therefore, we consider that the effect of projecting X onto the Q, K with Λ feature dimension to calculate the similarity may not be worse than that of ordinary Q, K. Then we simply verified this under the setting of baseline-2 and Swin* in the main paper. The results are shown in Tab 3 and $W_Q, W_K \in R^{C\times C}$ are improved to $W_Q, W_K \in R^{C\times\Lambda}$ in our method, which could reduce the computation cost and parameters in the SCI transformer with negligible performance degradation.

Table 3. Ablation Comparison of Different Settings.

Modules	PSNR	SSIM	Params (M)	FLOPs (G)
Normal	**36.27**	**0.953**	1.44	23.51
Improved	36.21	**0.953**	**1.38**	**22.57**

2. The Visualization of Focused Attention (FA)

To intuitively show the advantages of FA, we visualize the feature maps of the last Focued Attention Block (FAB) in the first stage of DPU-5stg. As depicted in Fig. 3, the top row shows the RGB images of the 10 scenes. The middle and bottom rows exhibit the feature maps with Normal Attention (NA) and FA, respectively. We can see that the feature maps with NA produce more blurred details and distorted deformations and focus attention on less critical backgrounds. In contrast, thanks to principal component projection enlarging the attention weight of key features and threshold filtering removing the attention of irrelevant components, the feature maps with FA restore accurate textures, complete shapes, and clear details, and focus more attention on objects that need to be reconstructed, which demonstrates the effectiveness of FA.

3. More Comparison Results

3.1. Comparison with Model-based Methods

To further strengthen the evaluation, we add the comparison with 3 model-based methods, i.e., TwIST[1], GAP-TV[13], and DeSCI[9] here, and the results are shown in Table 4. As we can see, DPU has significant performance advantages over model-based approaches, which demonstrate the superiority of the deep learning approach.

Table 4. Comparison with Model Methods.

Method	TwIST[1]	GAP-TV[13]	DeSCI[9]	DPU-5stg	DPU-9stg
PSNR	23.12	24.36	25.27	39.62	**40.52**
SSIM	0.669	0.669	0.721	0.973	**0.977**

3.2. More Visual Comparison Results

Figs. 4-11 show more visual comparison results of the state-of-the-art competing methods, including: HDNet [7], TSA-Net [11], BIRNAT [5], and unfolding methods: DGSMP [8], GAP-Net [12], DAUHST [4], and Transformer methods: MST [3], CST [2]. Constructed images with 4 out of 28 spectral bands for other simulated and real scenes, simulated ground truth, measurements, and RGB images are shown for reference. It can be intuitively observed that our DPU yields more detailed content, cleaner textures, and fewer artifacts than the other competing methods. Meanwhile, compared with other unfolding methods, our DPU requires the least single-stage parameters and computation costs, demonstrating the effectiveness and efficiency of our DPU method.

References

[1] José M. Bioucas-Dias and Mário A. T. Figueiredo. A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004, 2007. 2

[2] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool.

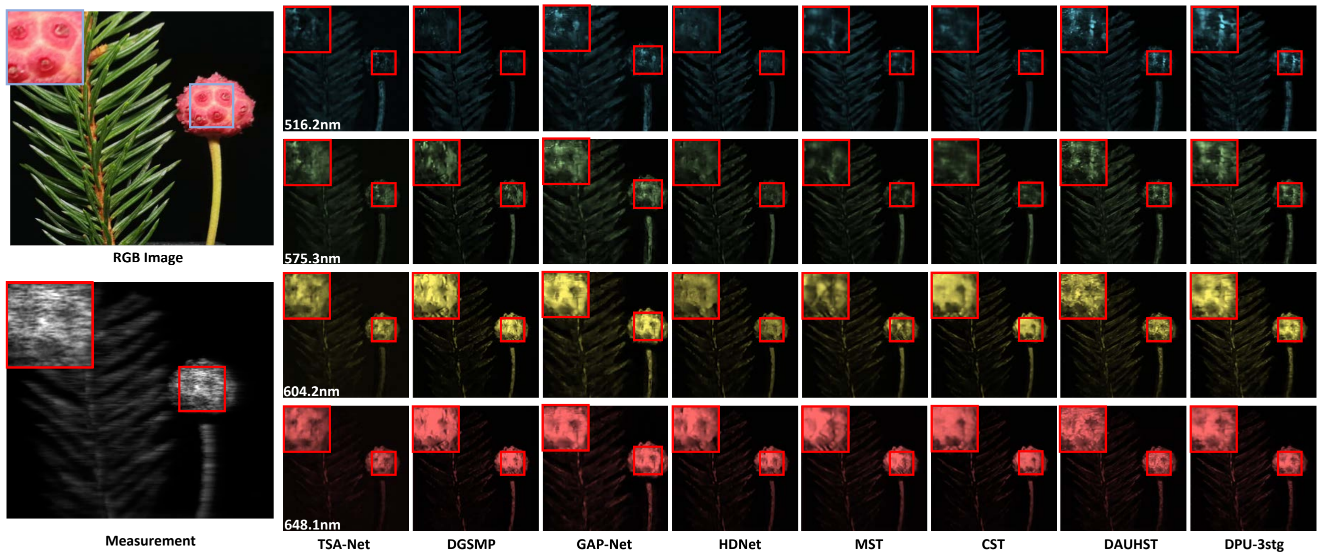

Figure 4. Constructed images of real scene 2 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

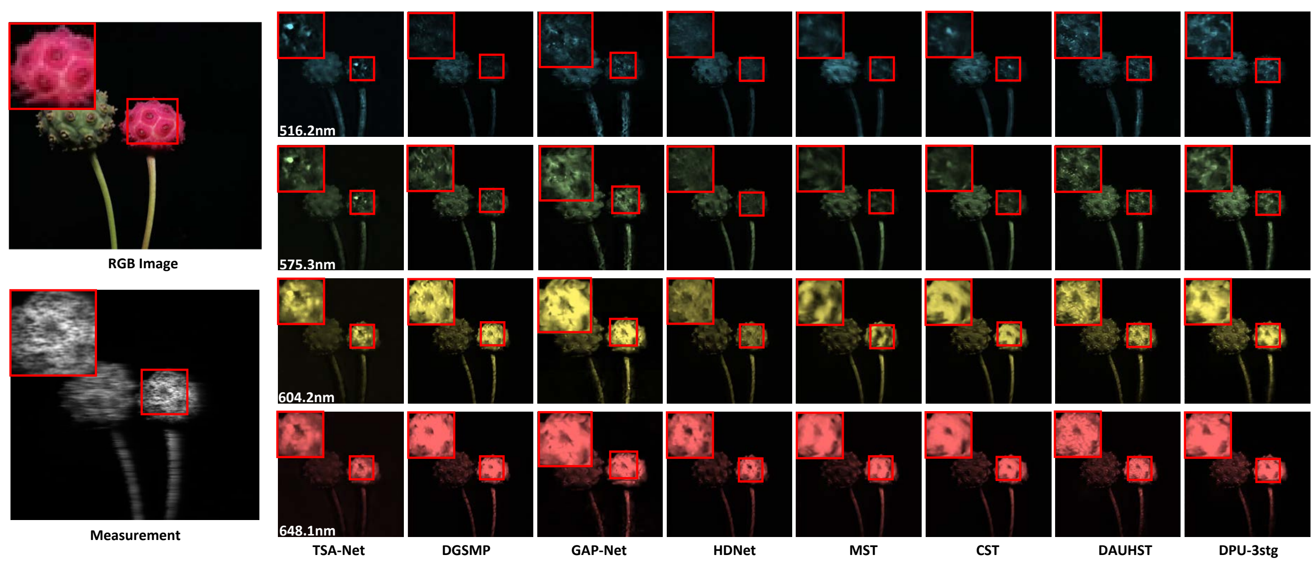

Figure 5. Constructed images of real scene 4 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

Coarse-to-fine sparse transformer for hyperspectral image reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 686–704. Springer, 2022. 2

[3] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17481–17490, 2022. 2

[4] Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Henghui Ding, Yulun Zhang, Radu Timofte, and Luc Van Gool. Degradation-aware unfolding half-shuffle transformer for spectral compressive imaging. In *Advances in Neural Information Processing Systems*, 2022. 1, 2

[5] 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*, 45(2):2264–2281, 2023. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

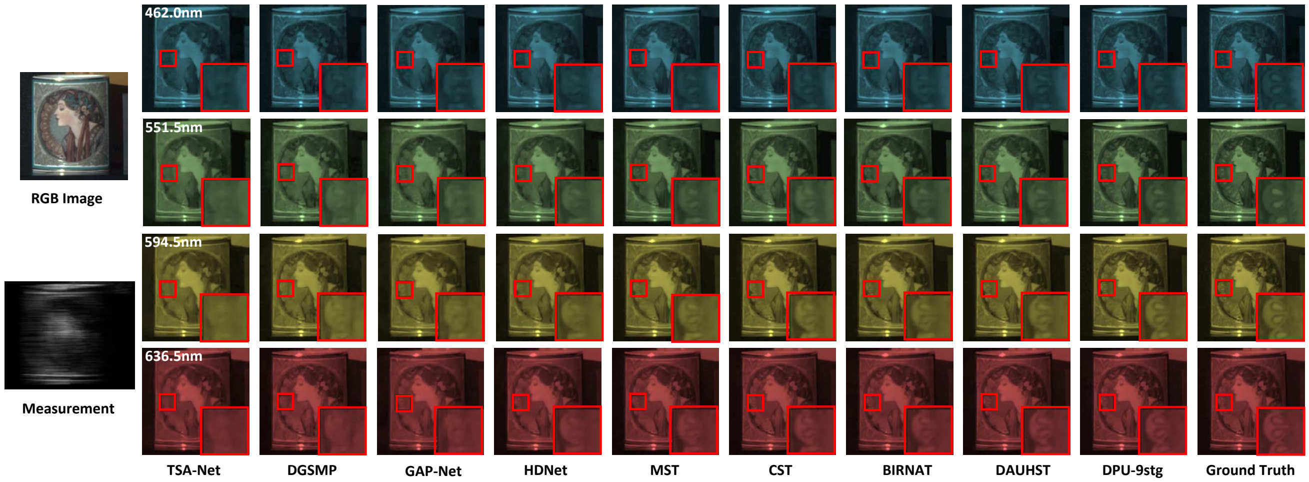

Figure 6. Constructed images of simulated scene 1 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

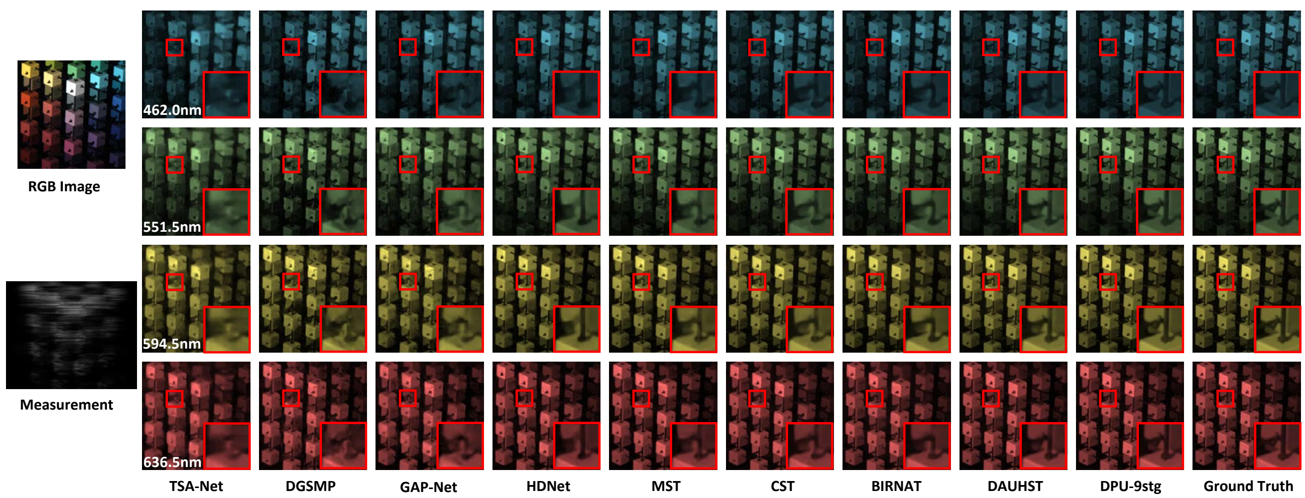

Figure 7. Constructed images of simulated scene 2 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[7] Xiaowan Hu, Yuanhao Cai, Jing Lin, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Hdnet: High-resolution dual-domain learning for spectral compressive imaging. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17521–17530, 2022. 2

[8] Tao Huang, Weisheng Dong, Xin Yuan, Jinjian Wu, and Guangming Shi. Deep gaussian scale mixture prior for spectral compressive imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16211–16220, 2021. 2

[9] 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*, 41:2990–3006, 2019. 2

[10] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[11] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-

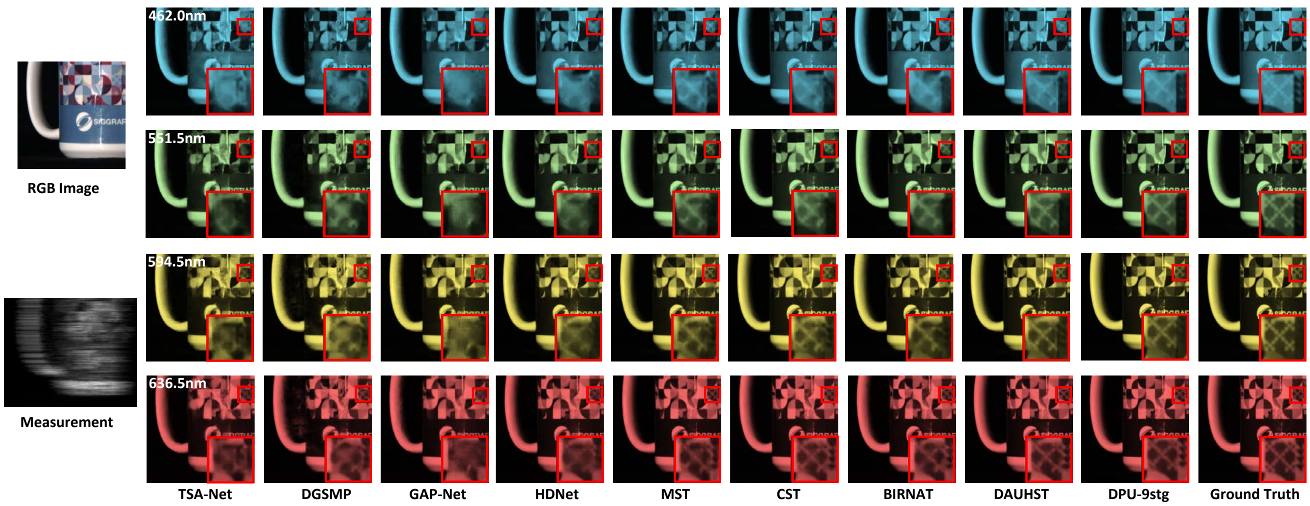

Figure 8. Constructed images of simulated scene 5 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

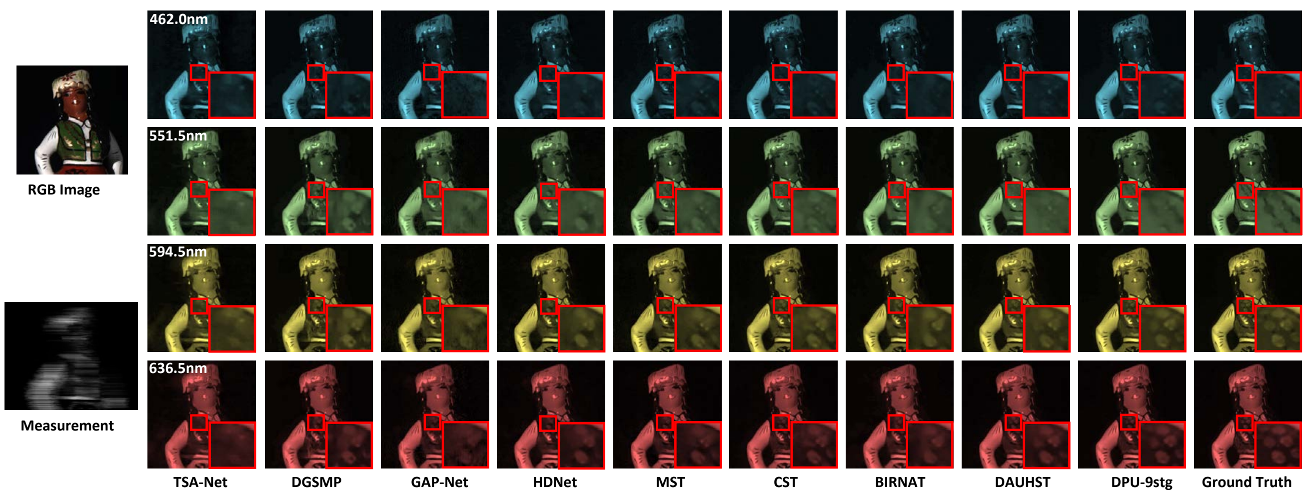

Figure 9. Constructed images of simulated scene 6 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

attention. In *ECCV*, 2020. 2

[12] Ziyi Meng, Xin Yuan, and Shirin Jalali. Deep unfolding for snapshot compressive imaging. *International Journal of Computer Vision*, 131(11):2933–2958, 2023. 2

[13] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, 2016. 2

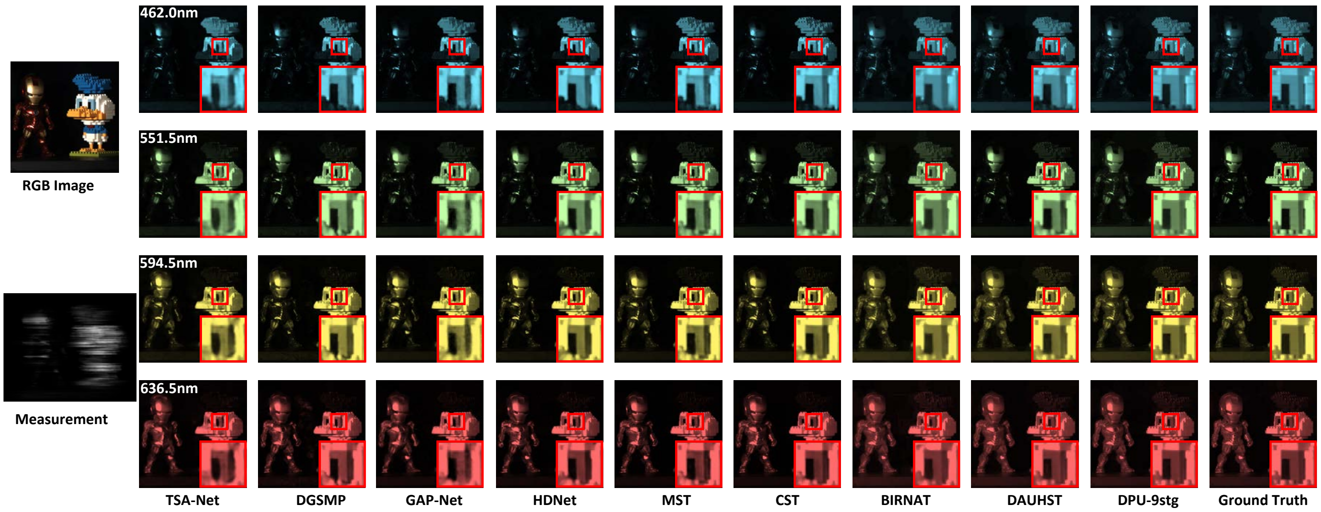

Figure 10. Constructed images of simulated scene 8 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.

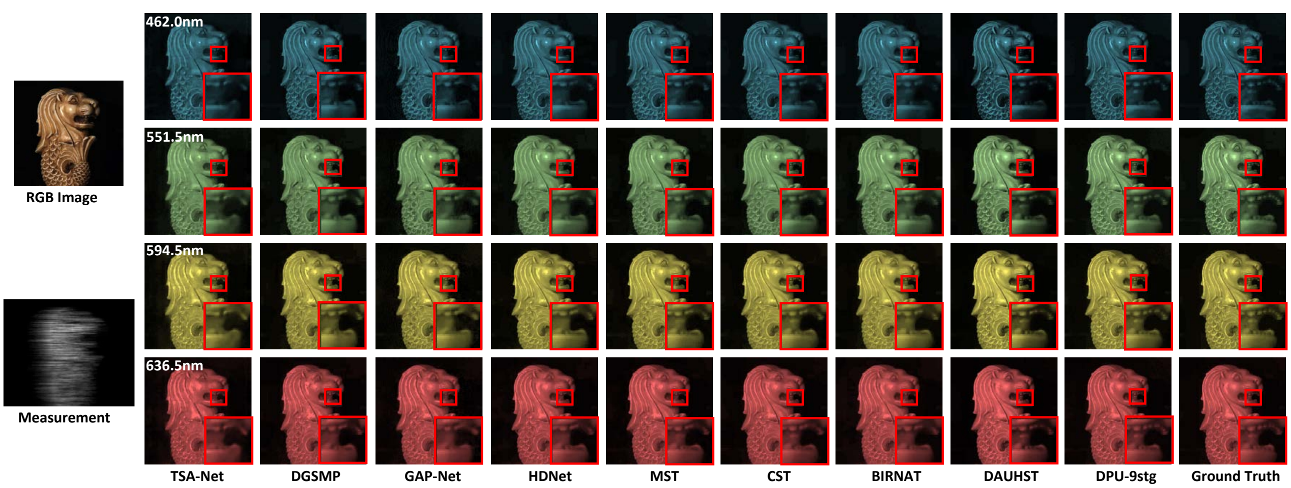

Figure 11. Constructed images of simulated scene 10 with 4 out of 28 spectral channels by the state-of-the-art methods. Zoom in for a better view.