Supplementary Material for "Self-supervised Neural Networks for Spectral Snapshot Compressive Imaging"

Ziyi Meng Zhenming Yu Kun Xu
Beijing University of Posts and Telecommunications, Beijing, China
{mengziyi,yuzhenming,xukun}@bupt.edu.cn

Xin Yuan*
Westlake University, Hangzhou, Zhejiang, China
xyuan@westlake.edu.cn

In this supplementary material, we provide the derivation details of the solution for Single Fidelity Formulation and show additional results of the ablation study. We further show more results of the synthetic and real data recovered by our methods and compare it with other algorithms.

1. Derivation of Single Fidelity Formulation

Derivation of the Solution for Single Fidelity Formulation in Eq. (8) in the main paper: ADMM solves the (8) by splitting it into the following subproblems:

- Given $\boldsymbol{\Theta}$ and $\boldsymbol{b}$, $\boldsymbol{x}$ is solved by

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e}) - \boldsymbol{b}\|_2^2 + \frac{\lambda}{\mu} R(\boldsymbol{x}). \tag{1}$$

  This is a traditional denoising problem and can be solved by the PnP algorithm given the prior $R(\boldsymbol{x})$, *i.e.*,

$$\hat{\boldsymbol{x}} = \mathscr{D}_{\sigma}(\boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e}) + \boldsymbol{b}). \tag{2}$$

  where $\mathscr{D}_{\sigma}$ denotes the denoising operator being used and σ is the estimated noise level depending on λ/μ.

- Given $\boldsymbol{x}$ and $\boldsymbol{b}$, optimizing $\boldsymbol{\Theta}$ leads to the following problem:

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e})\|_2^2 + \mu\|\boldsymbol{x} - \boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e}) - \boldsymbol{b}\|_2^2, \tag{3}$$

  which can be solved by the back-propagation optimization as in DIP, modified by a proximity regularization that forces $\boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e})$ to be close to $\boldsymbol{x} - \boldsymbol{b}$. For the U-net being used in our implementation, instead of only minimizing the first term in (3) as in the loss function, we used both terms as the loss function. This learned $\boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e})$ is thus playing the role of: i) denoising $\boldsymbol{x} - \boldsymbol{b}$, and ii) minimizing the measurement loss $\boldsymbol{y} - \boldsymbol{H}\boldsymbol{T}_{\boldsymbol{\Theta}}(\boldsymbol{e})$.

- Optimizing $\boldsymbol{b}$ is given by

$$\boldsymbol{b}^{k+1} = \boldsymbol{b}^k - (\boldsymbol{x}^k - \boldsymbol{T}_{\boldsymbol{\Theta}^k}(\boldsymbol{e})), \tag{4}$$

  where the superscript k denotes the iteration number.

Note that these three steps are performed iteratively and each of them can have their own inner loops such as the $\boldsymbol{\Theta}$ optimization.

2. Ablation Study Results

2.1. DIP vs. Deep Decoder

We visualize the results of the proposed self-supervised methods using DIP [8] and deep decoder (DD) [3] as the prior (PnP-DIP and PnP-DD), as shown in Fig. M1. It can be seen seen that PnP-DD provide a good reconstruction on some smooth regions of the images, but for the regions with many spatial details, there are significant artifacts and over-smoothness. As mentioned in the main paper, this might be caused by the lack of the network parameters of the deep decoder.

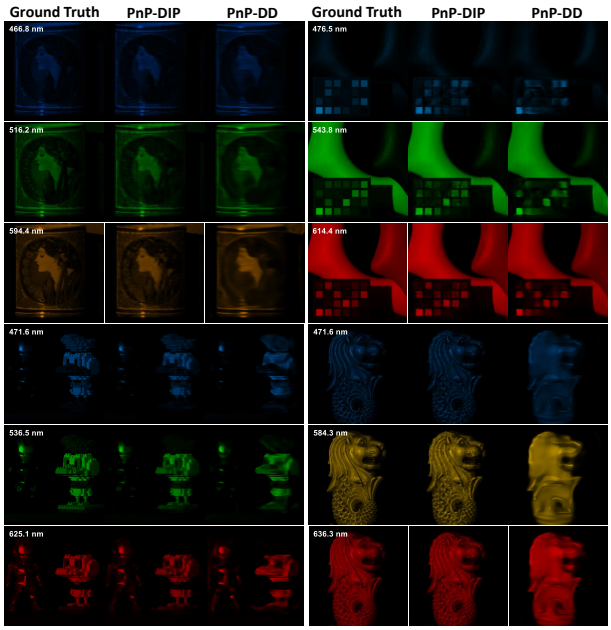

Figure M1. Reconstructed results of 4 synthetic data with 3 spectral channels by the proposed self-supervised methods using DIP and deep decoder as the prior, respectively.

*Corresponding author.

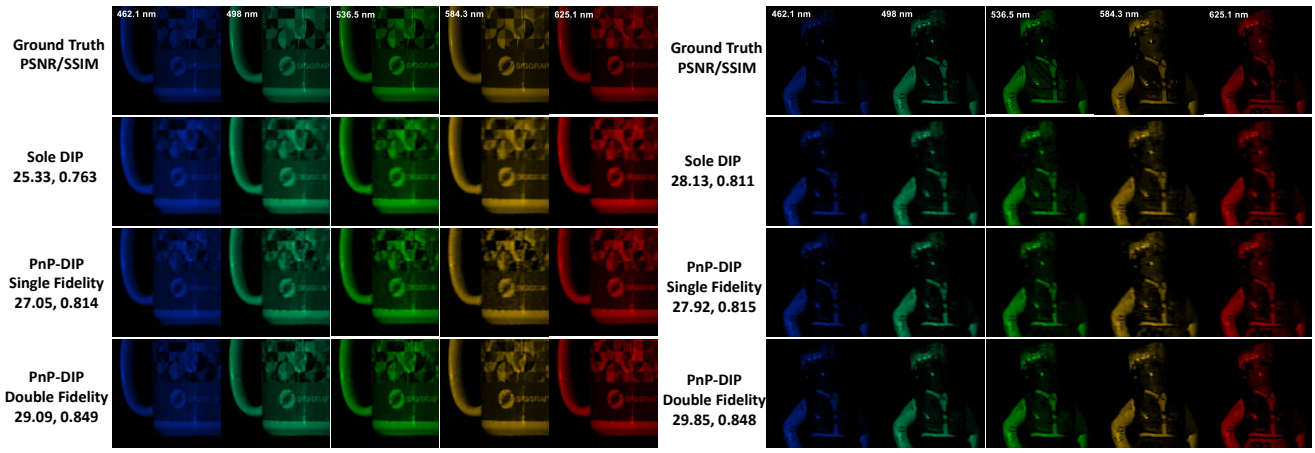

Figure M2. Reconstructed results of two synthetic data with 5 spectral channels by the sole DIP, PnP-DIP with single fidelity term and the proposed PnP-DIP with double fidelity terms.

2.2. Single Fidelity vs. Dual Fidelity in DIP

Fig. M2 compare the results of the sole DIP (the directly using DIP), PnP-DIP with single fidelity term and the proposed PnP-DIP with double fidelity terms. It can be seen that the reconstructed results of the proposed PnP-DIP with double fidelity terms have clearer detials, as well as less noise and artifacts.

2.3. Incorporating DIP with TV Prior

Without considering pre-trained HSI denoiser, the previous self-supervised results are obtained by only using DIP as the prior in our proposed PnP framework (PnP-DIP). Here we incorporate the widely used TV prior with DIP to form a joint PnP framework, namely PnP-DIP-TV. We initial the parameters by $\{\boldsymbol{u} = \boldsymbol{H}^\top \boldsymbol{y},\ \boldsymbol{v} = 0,\ \eta = 0.01\}$, and other parameters keep the same as before. We reduce the effect of TV prior gradually as the increasing of iterations by scaling η by 0.95 each ADMM iteration. Finally, the average results of PnP-DIP-TV (30.44dB, 0.852) is very close to PnP-DIP (30.48dB, 0.854). This gives us the following observations; *i*) Though TV is widely used and can achieve good results in most tasks, DIP is powerful to learn a stronger prior. Similar case will happen for other pre-trained priors such as sparsity. *ii*) Even at the first few iterations, TV will help the reconstruction, the final results will rely on DIP. Therefore, we recommend that in our spectral SCI reconstruction, PnP-DIP can be used as a new baseline without any training data. However, we do notice that in real data experiments, TV will help the reconstruction, which may be due to the measurement noise.

2.4. Choice of the Up/downsampling

U-net is an encoder-decoder scenario, and thus up/downsampling is playing a pivotal role. As the crucial components of U-net, pooling and upsampling change the scale and depth of the feature maps. Upsampling is usually implemented by unlearned forms, (such as bilinear and PixelShuffle [7]) and learned convolutional filters (transposed convolution or ConvTranspose [10]). We compare the results using different upsampling in the DIP network, shown in the upper part Table M1. It can be seen that the network using ConvTranspose achieves the highest performance. For the downsampling, we find that the average-pooling provide a better results compared with max-pooling with comparison shown in the lower part in Table M1. Therefore, ConvTranspose and average-pooling are used in our experiments.

Table M1. Average PSNR and SSIM of the DIP network using different up/downsampling.

Upsampling	Conv2DTranspose	Bilinear	PixelShuffle
PSNR/SSIM	30.48, 0.854	29.69, 0.821	27.59, 0.824
Downsampling	Average-pooling	Max-pooling	
PSNR/SSIM	30.48, 0.854	30.26, 0.848	

3. Supplementary Results

3.1. PnP-DIP vs. PnP-HSI

As shown in Table 1 in the main paper, the average result of PnP-HSI [11] ([42] in the main paper) has an about 5dB gap with the proposed PnP-DIP. The main reason causing the less-than-perfect results of PnP-HSI is the simulation setting. We used the *real captured mask* and **a larger mask-shift range (54 pixels)**. PnP-HSI is heavily dependent on the initialization results of ADMM-TV, which is not good in our simulation setting. The results of PnP-HSI usually cannot converge well (generating artifacts) when using a bad initialization. This is why we use HSI denoiser in only the last few ADMM iterations. Our simulation setting is closer to the real systems compared with [5], and our results indicate that PnP-HSI is getting degraded when the shifting

Table M2. Average PSNR and SSIM of ADMM-TV, PnP-HSI and PnP-DIP on two datasets used in [11] under two different simulation settings.

Dataset	Simulation setting	ADMM-TV	PnP-HSI	PnP-DIP
ICVL	Binary mask, 30-pixel shift	32.56, 0.899	39.43, **0.974**	**40.72**, 0.970
	Real mask, 60-pixel shift	29.01, 0.867	32.91, 0.930	**37.27**, **0.954**
KAIST	Binary mask, 30-pixel shift	37.25, 0.957	39.15, **0.974**	**41.79**, **0.974**
	Real mask, 60-pixel shift	34.25, 0.941	34.92, 0.954	**38.74**, **0.962**

pixels are larger.

For verifying the analysis, we give the results of the datasets used by [11] in Table M2 and Fig. M3. Specifically, when the mask shifting is small, PnP-DIP and PnP-HSI are providing similar results, but when the mask shifting is big, our proposed PnP-DIP outperforms PnP-HSI by 4.36dB and 3.82dB, respectively on the two datasets, respectively.

This has also been verified by the real data results (Fig. 5 in the main paper and Fig. M4 in this SM) where a large mask shifting was used.

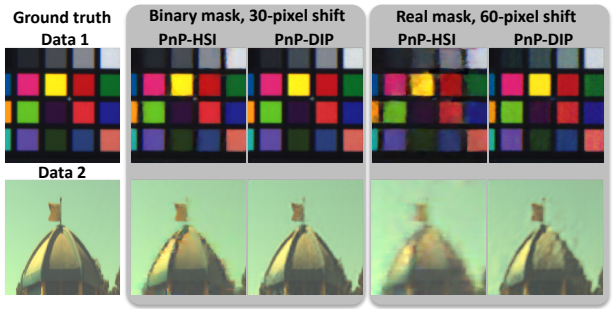

Figure M3. Result comparison of PnP-HSI and PnP-DIP on two data from [11] under two different simulation settings.

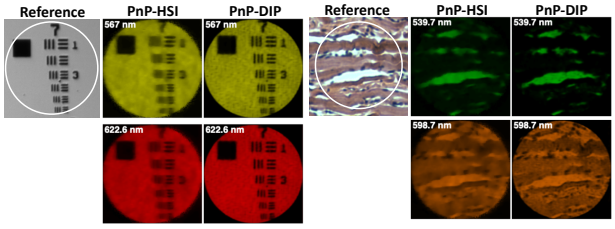

Figure M4. Result comparison of PnP-HSI and PnP-DIP on endomicroscapy data in [6].

3.2. PnP-DIP vs. Autoencoder

We compare our proposed PnP-DIP with Autoencoder [2] on the datasets used in [11]. We use binary mask in simulation, and the shift range is 30-pixel. Fig. M5 shows the sRGB results of ADMM-TV [9], Autoencoder [2] and our PnP-DIP. It can be seen that our method achieves much better results. Autoencoder suffers from the spatial blur in this single-disperser CASSI model, which is different from the dual-disperser CASSI model mainly used in [2]. We will put the results into the final paper.

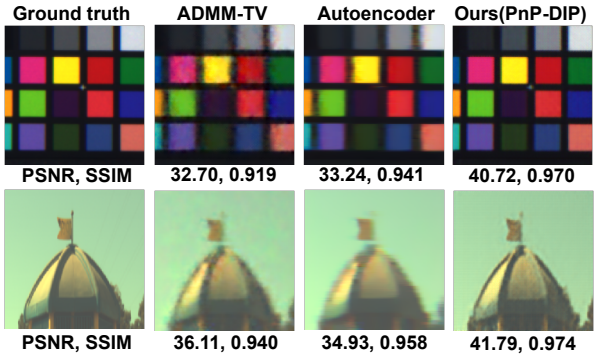

Figure M5. Comparison (sRGB) of ADMM-TV [37], Autoencoder [6] and our PnP-DIP on two datasets used in [11].

3.3. Results on the Synthetic Data

Fig. M6-M15 show the reconstructed results of the synthetic data with 10 out of 28 spectral channels. We compare the proposed self-supervised method (PnP-DIP) and the method using HSI prior (PnP-DIP-HSI) with the state of the art supervised algorithm (TSA-Net [5]) and list the corresponding PSNR and SSIM.

3.4. Results on the Real Data

CASSI Data Set 1 We show more results on the datasets captured by the recently bulit CASSI system in [5]. The 2D measurements have a spatial size of 550×604, and the recovered spectral cube contains 28 spectral channels with the size of 550×550. The specific wavelengths are {453.3, 457.6, 462.1, 466.8, 471.6, 476.5, 481.6, 486.9, 492.4, 498.0, 503.9, 509.9, 516.2, 522.7, 529.5, 536.5, 543.8, 551.4, 558.6, 567.5, 575.3, 584.3, 594.4, 604.2, 614.4, 625.1, 636.3, 648.1}nm. Fig. M16-M19 show the reconstructed results of the 4 scenes with 10 out of 28 spectral channels. We compare the proposed self-supervised method (PnP-DIP) and the method using HSI prior (PnP-DIP-HSI) with ADMM-TV and the supervised algorithm TSA-Net [5].

CASSI Data Set 2 We show more results on the datasets captured by the original CASSI system [4]. The reconstructed spectral image contains 33 spectral channels with the size of 210×256. The specific wavelengths are {454, 458, 462, 465, 468, 472, 475, 479, 483, 487, 491, 496, 500, 505, 509, 514, 520, 525, 531, 537, 543, 549, 556, 564, 571, 579, 587, 596, 605, 615, 626, 637, 650}nm. Fig. M20

show the reconstructed results of the data *Object* with 10 out of 33 spectral channels. We compare the proposed self-supervised method (PnP-DIP) and the method using HSI prior (PnP-DIP-HSI) with Twist [1], ADMM-TV and the deep PnP method (PnP-HSI) [11].

Endomicroscopy Data We show more results on the datasets captured by the compressive multispectral endomicroscopy system [6].The captured measurements are of spatial size of 660×706, which are used to reconstruct the multispectral endoscopic images with the size of $660 \times 660 \times 24$. The specific wavelengths are {454.4, 459.5, 464.9, 470.5, 476.2, 482.1, 488.4, 494.8, 501.5, 508.5, 515.8, 523.4, 531.4, 539.7, 548.4, 557.5, 567.0, 577.0, 587.6, 598.7, 610.3, 622.6, 635.6, 649.3}nm. Fig. M21-M26 show the reconstructed results of the 4 scenes with 10 out of 28 spectral channels. We compare the proposed self-supervised method (PnP-DIP) and the method using HSI prior (PnP-DIP-HSI) with TwIST and the supervised deep neural network [6].

References

[1] J.M. Bioucas-Dias and M.A.T. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004, December 2007. 4

[2] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Trans. Graph.*, 36(6), Nov. 2017. 3

[3] Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. *International Conference on Learning Representations*, 2019. 1

[4] David Kittle, Kerkil Choi, Ashwin Wagadarikar, and David J Brady. Multiframe image estimation for coded aperture snapshot spectral imagers. *Applied optics*, 49(36):6824–6833, 2010. 3

[5] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In *European Conference on Computer Vision (ECCV)*, August 2020. 2, 3

[6] Ziyi Meng, Mu Qiao, Jiawei Ma, Zhenming Yu, Kun Xu, and Xin Yuan. Snapshot multispectral endomicroscopy. *Optics Letters*, 45(14):3897–3900, 2020. 3, 4

[7] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 2

[8] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[9] Xin Yuan. Generalized alternating projection based total variation minimization for compressive sensing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2539–2543, Sept 2016. 3

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[11] Siming Zheng, Yang Liu, Ziyi Meng, Mu Qiao, Zhishen Tong, Xiaoyu Yang, Shensheng Han, and Xin Yuan. Deep plug-and-play priors for spectral snapshot compressive imaging. *Photonics Research*, 9(2):B18–B29, 2021. 2, 3, 4

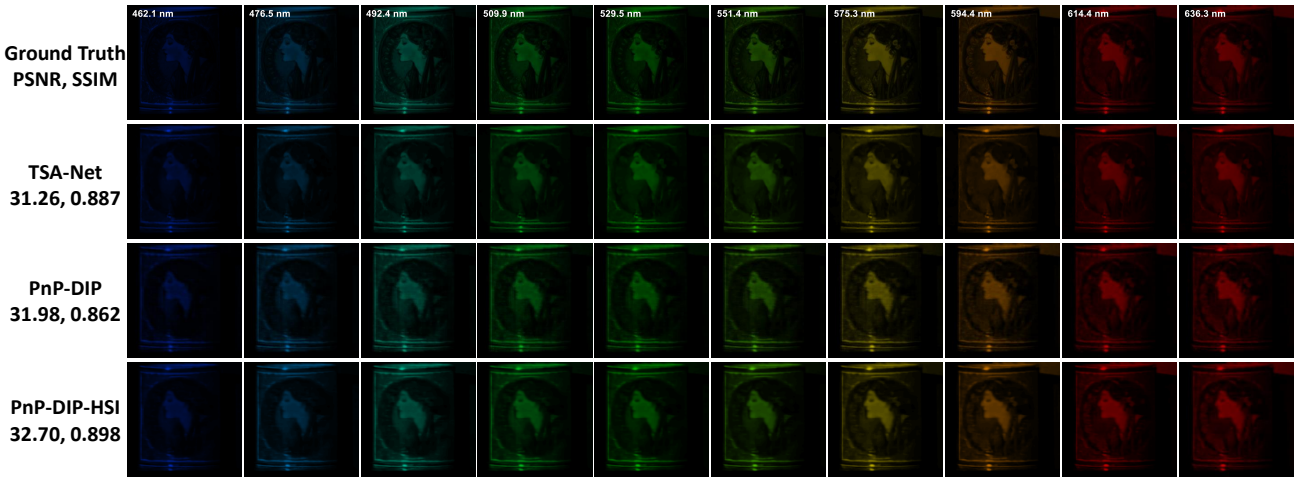

Figure M6. The results of the synthetic data *Scene 1* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

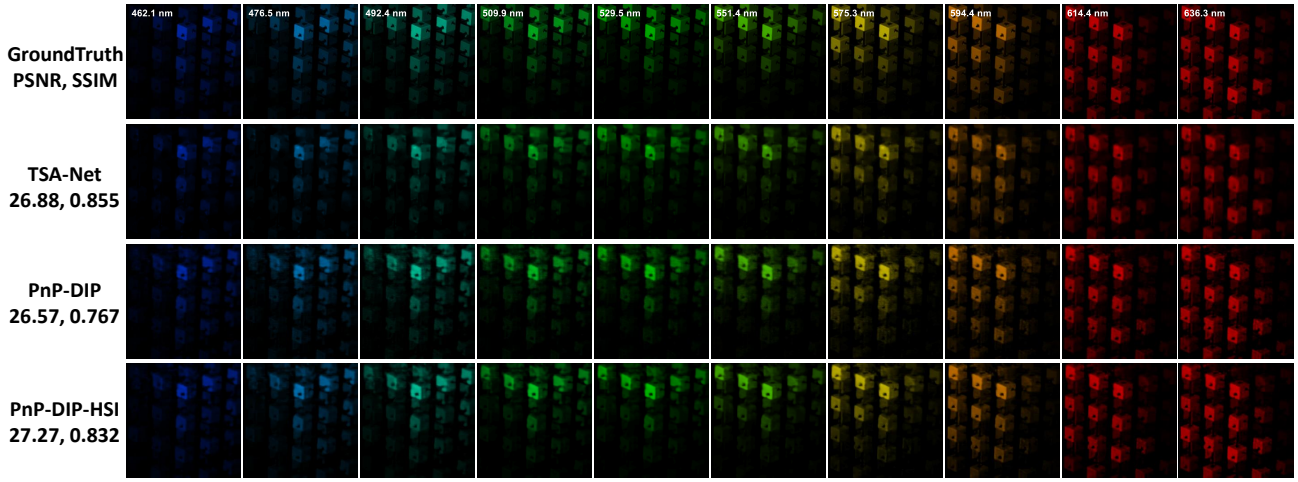

Figure M7. The results of the synthetic data *Scene 2* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

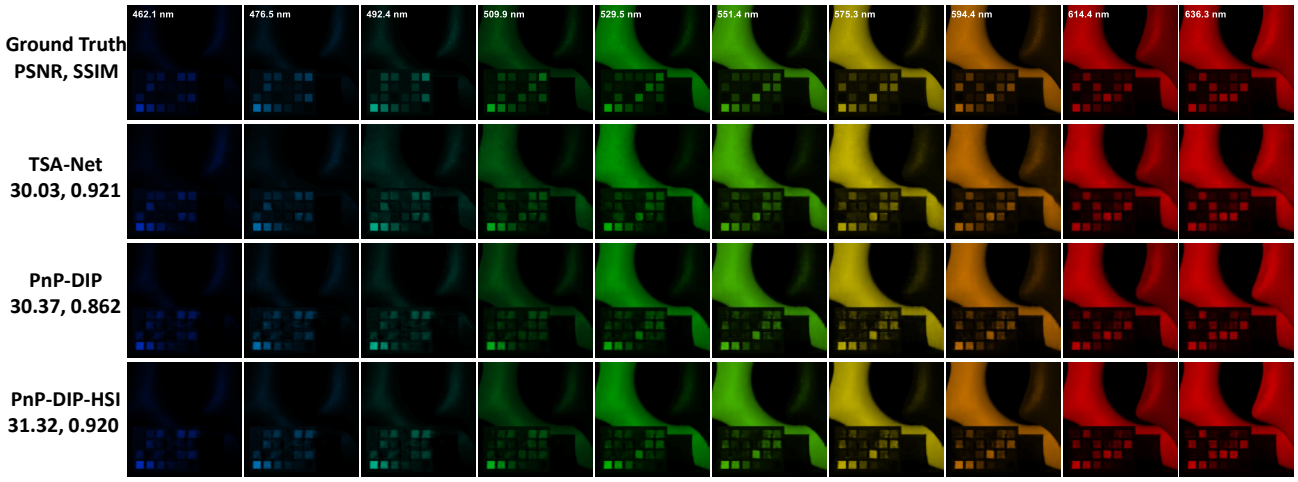

Figure M8. The results of the synthetic data *Scene 3* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

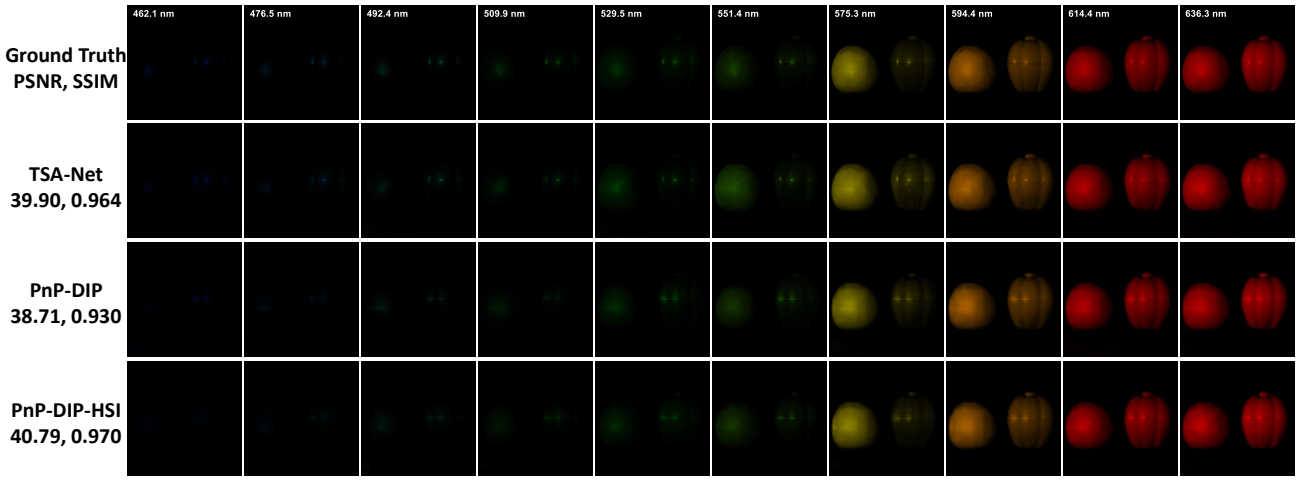

Figure M9. The results of the synthetic data *Scene 4* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

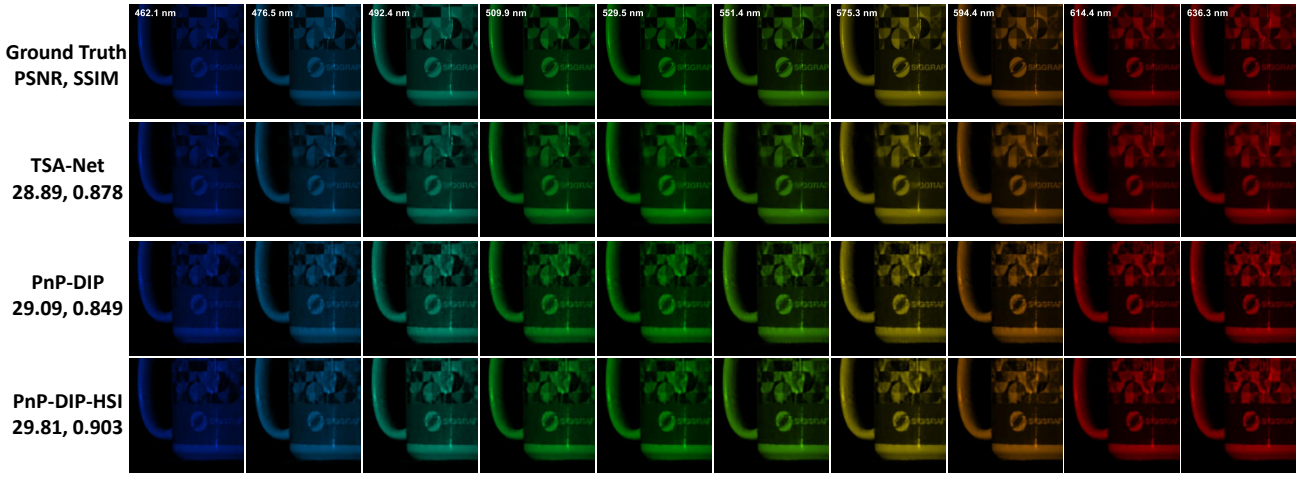

Figure M10. The results of the synthetic data *Scene 5* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

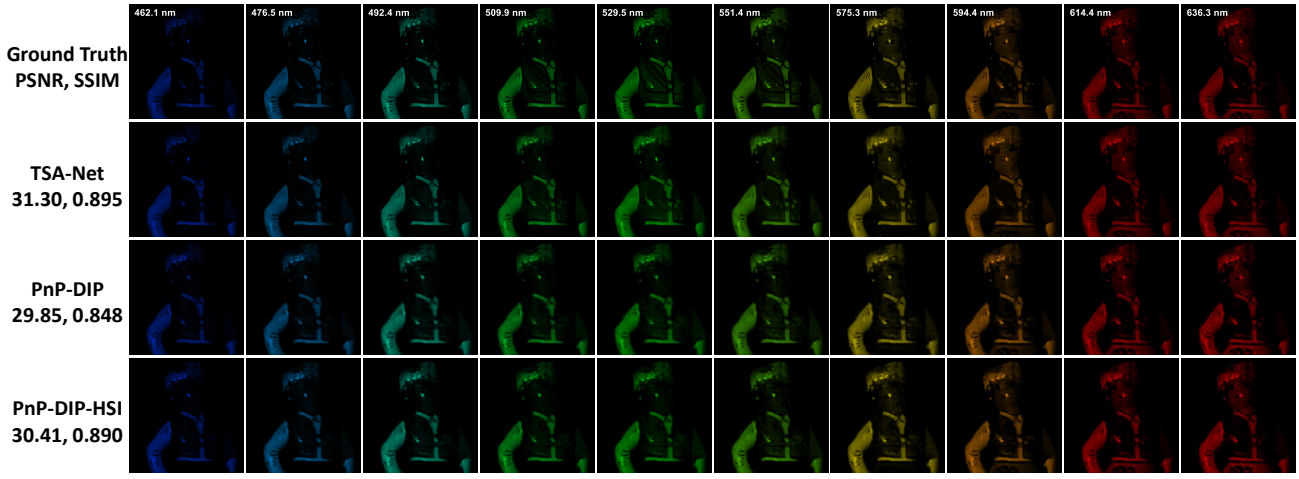

Figure M11. The results of the synthetic data *Scene 6* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

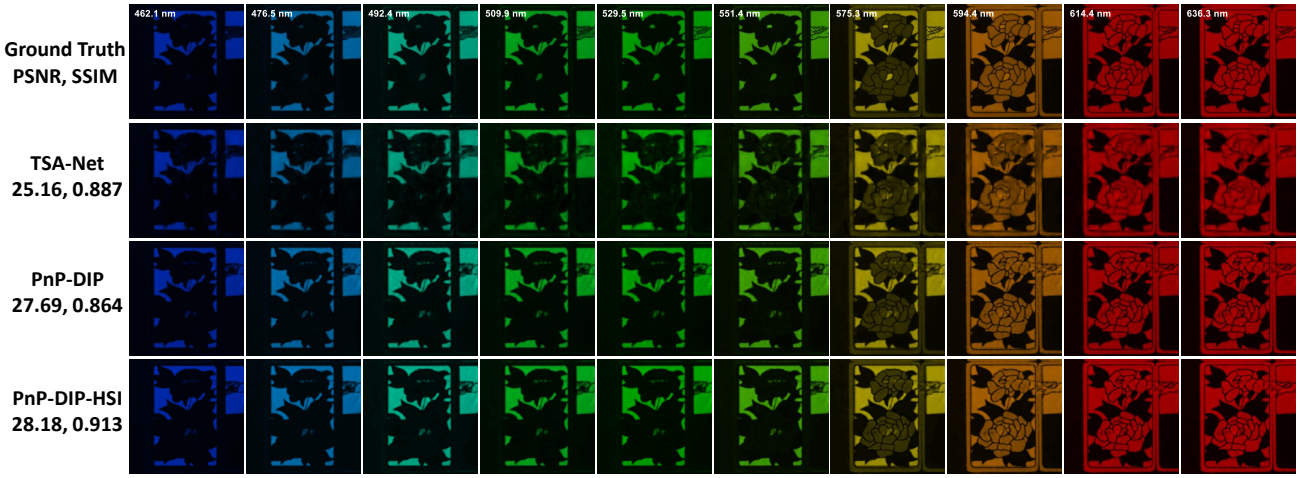

Figure M12. The results of the synthetic data *Scene 7* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

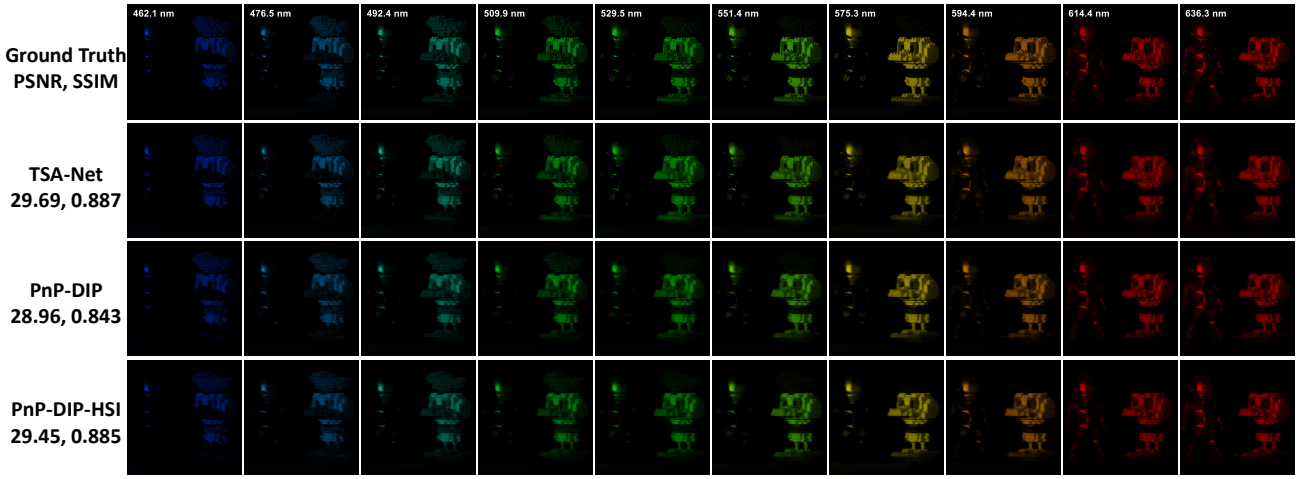

Figure M13. The results of the synthetic data *Scene 8* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

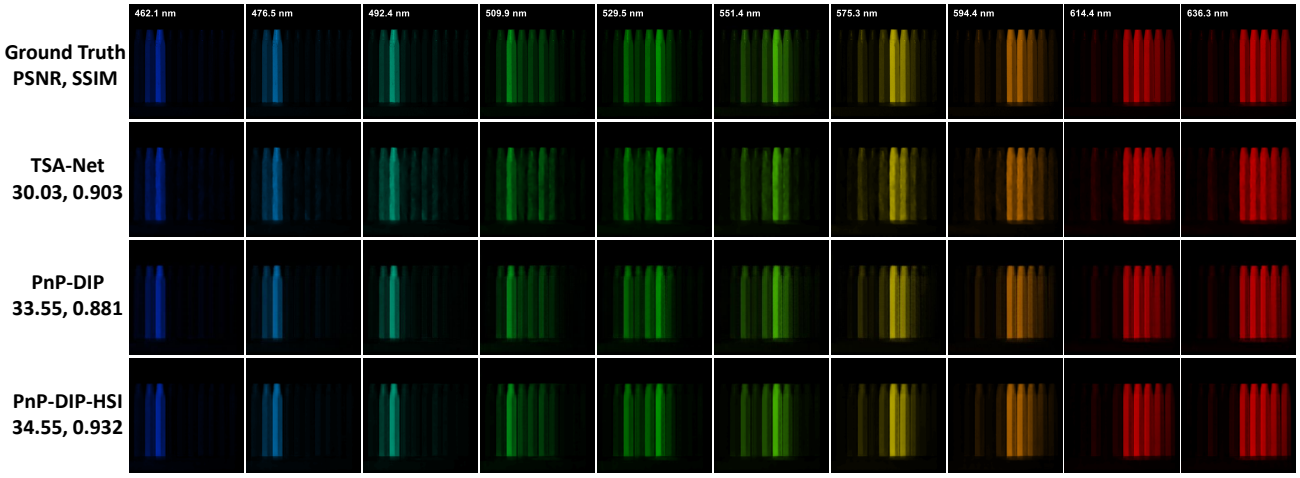

Figure M14. The results of the synthetic data *Scene 9* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

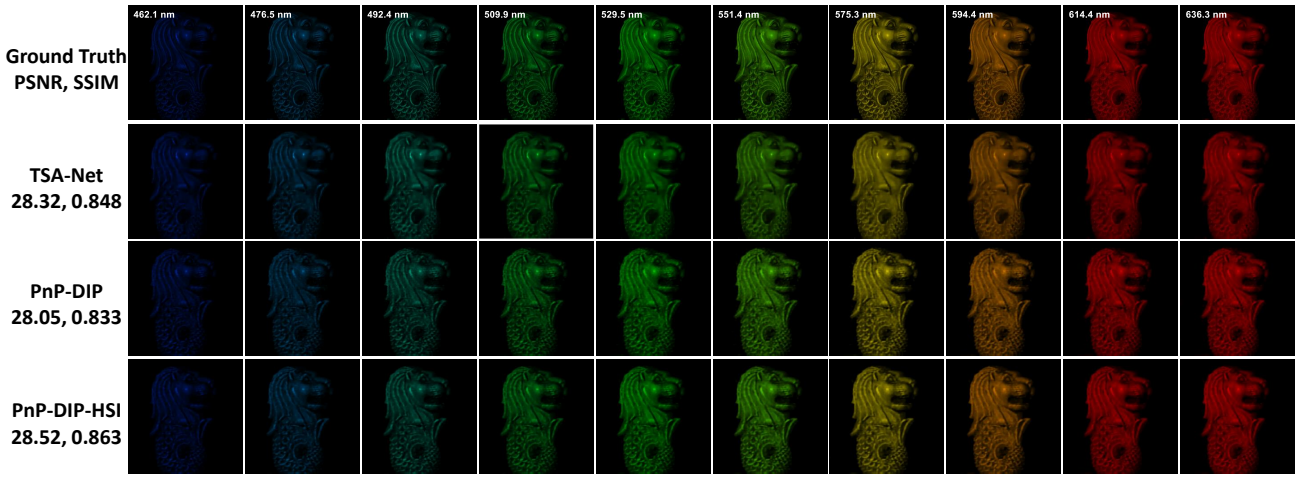

Figure M15. The results of the synthetic data *Scene 10* with 10 spectral channels reconstructed by TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

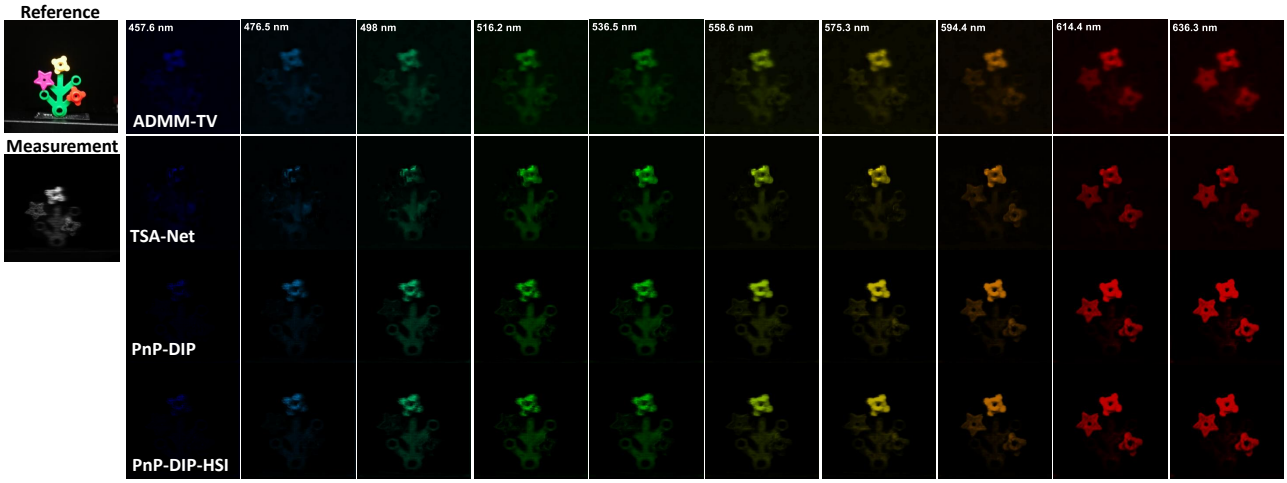

Figure M16. The results of the real data *Lego plant* with 10 spectral channels reconstructed by ADMM-TV, TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

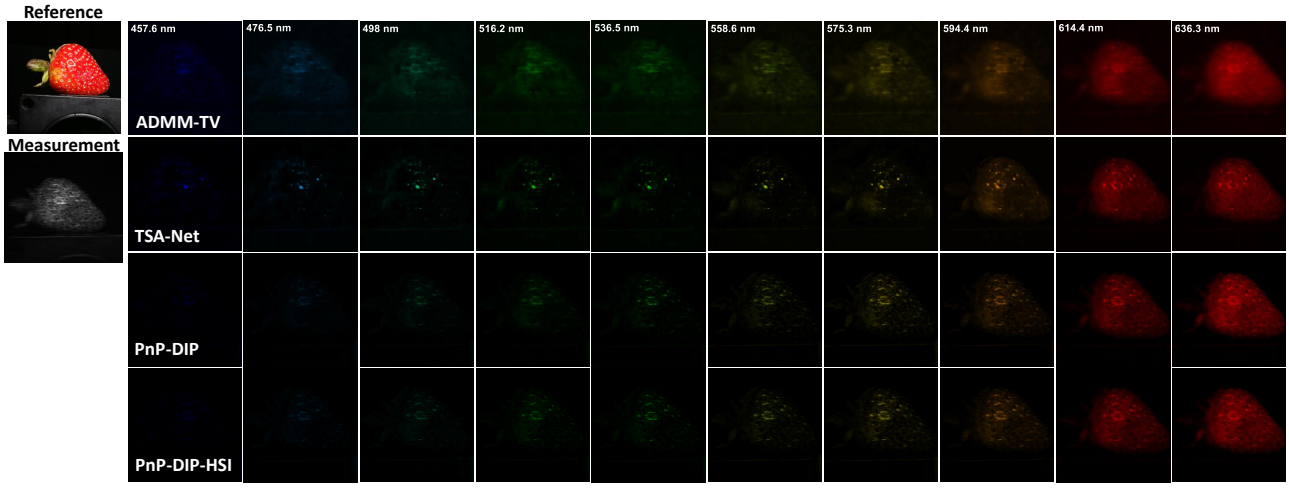

Figure M17. The results of the real data *Strawberry* with 10 spectral channels reconstructed by ADMM-TV, TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

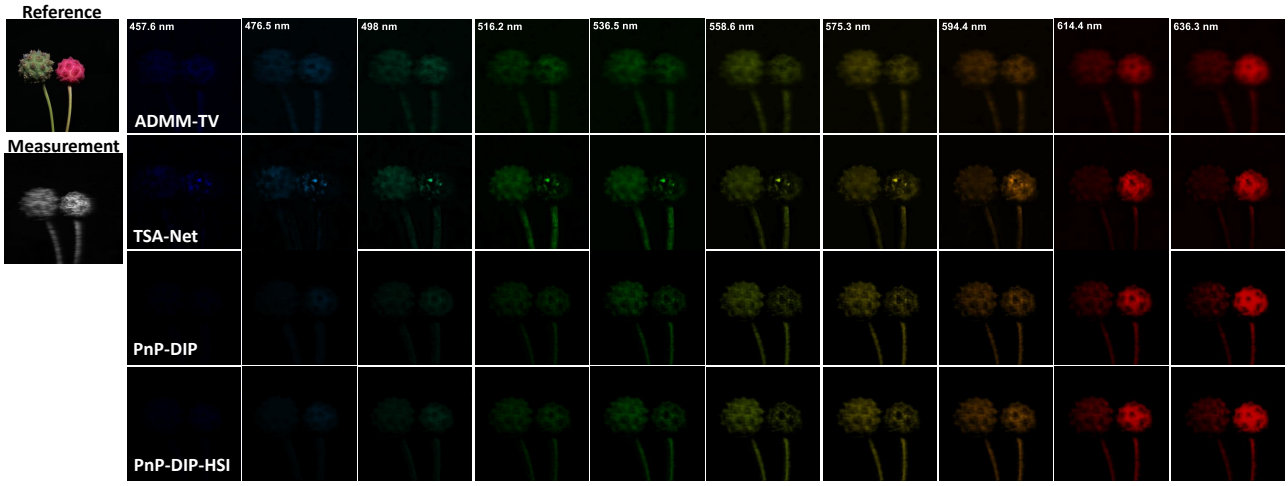

Figure M18. The results of the real data *Plant* with 10 spectral channels reconstructed by ADMM-TV, TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

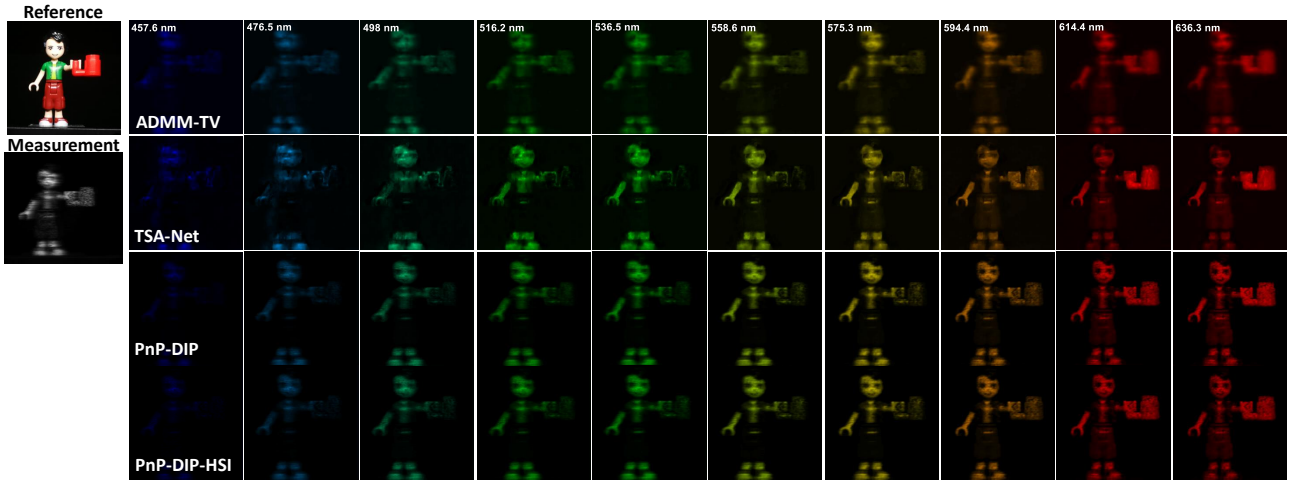

Figure M19. The results of the real data *Lego man* with 10 spectral channels reconstructed by ADMM-TV, TSA-Net and the proposed PnP-DIP and PnP-DIP-HSI.

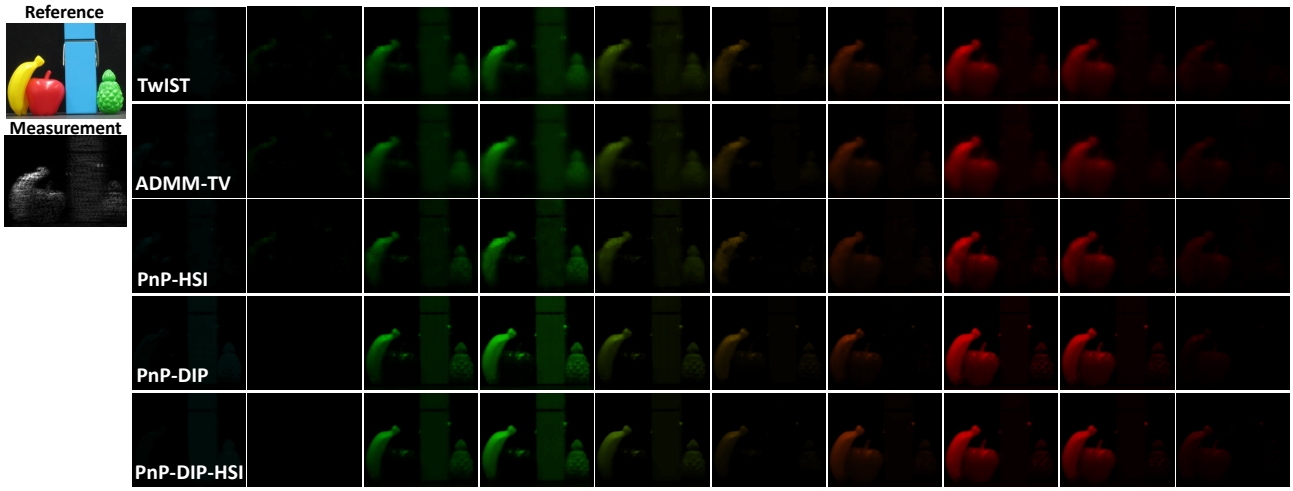

Figure M20. The results of the real data *Object* with 10 spectral channels reconstructed by TwIST, ADMM-TV, deep PnP method (PnP-HSI) and the proposed PnP-DIP and PnP-DIP-HSI.

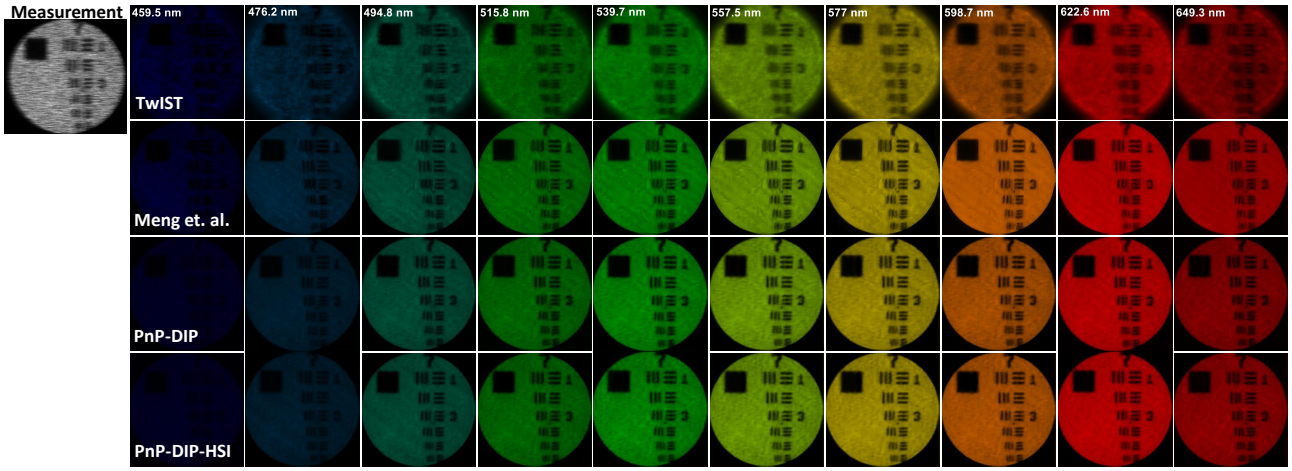

Figure M21. The results of the real data *Resolution target* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.

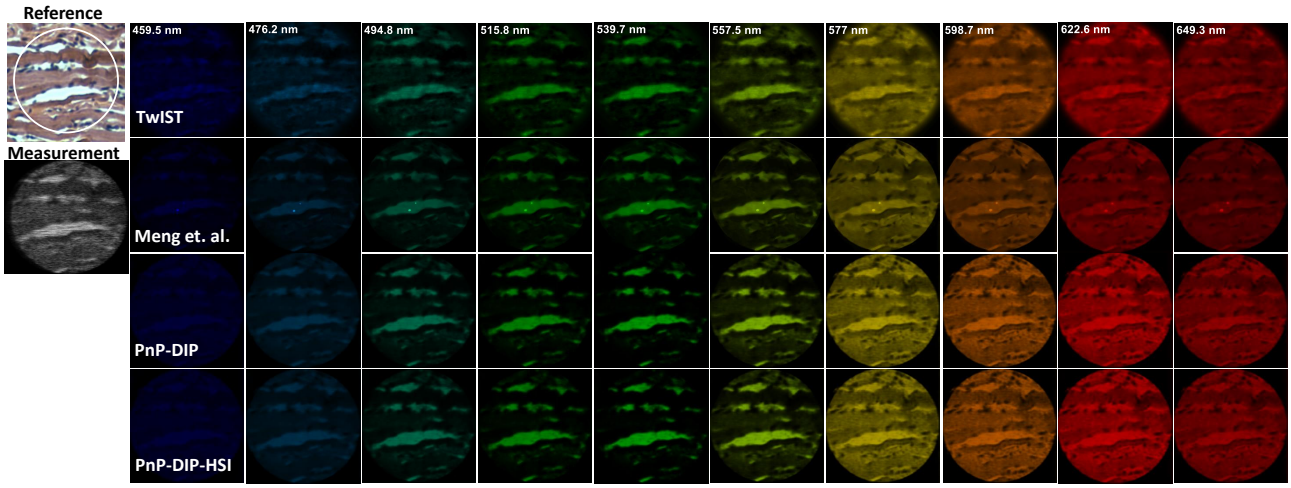

Figure M22. The results of the real data *Dog olfactory membrane section 1* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.

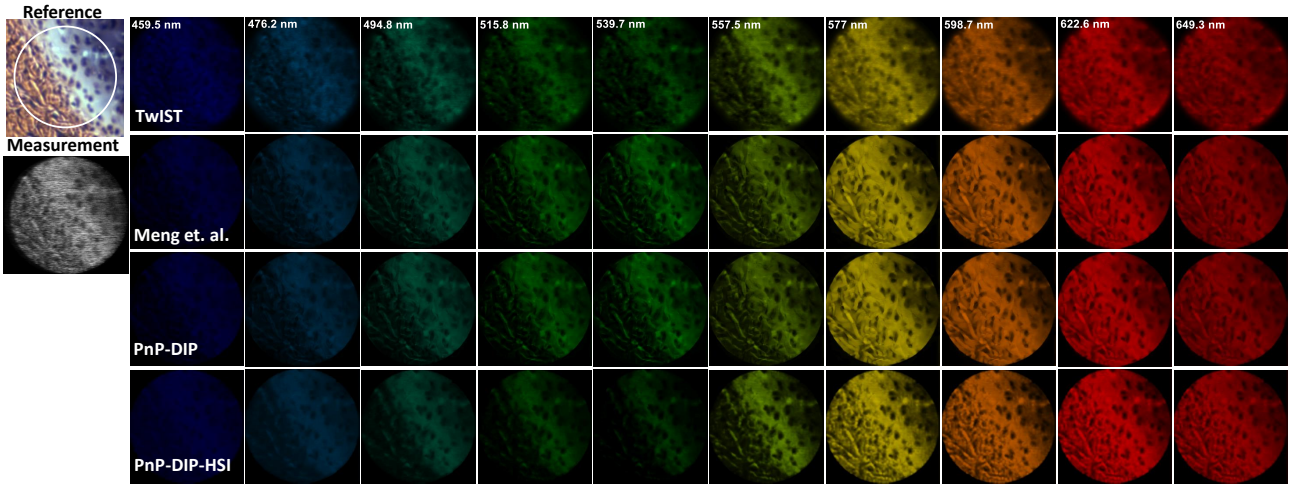

Figure M23. The results of the real data *Dog olfactory membrane section 2* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.

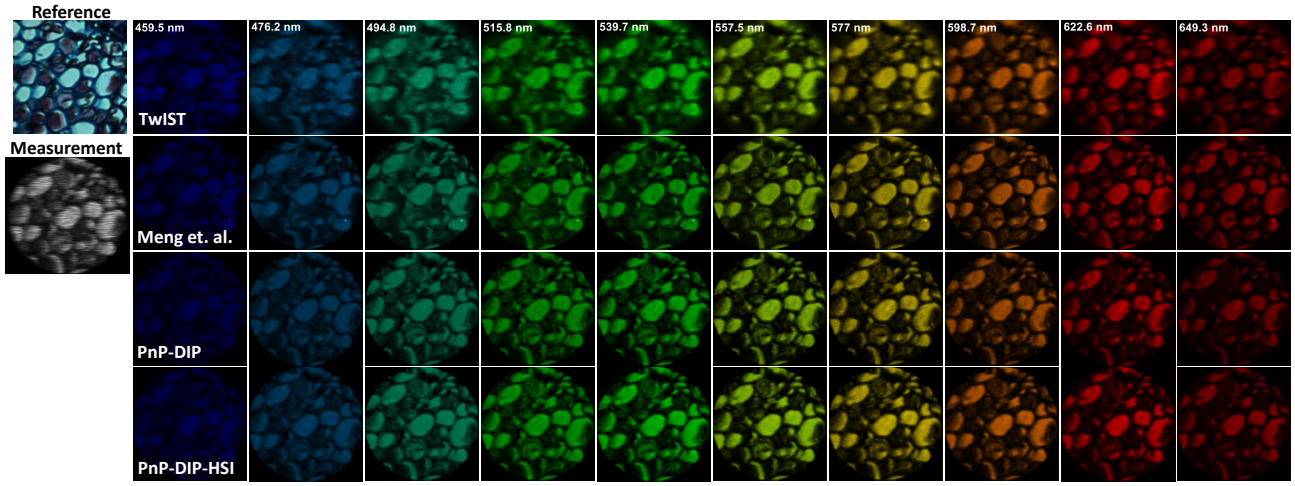

Figure M24. The results of the real data *Fern root section* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.

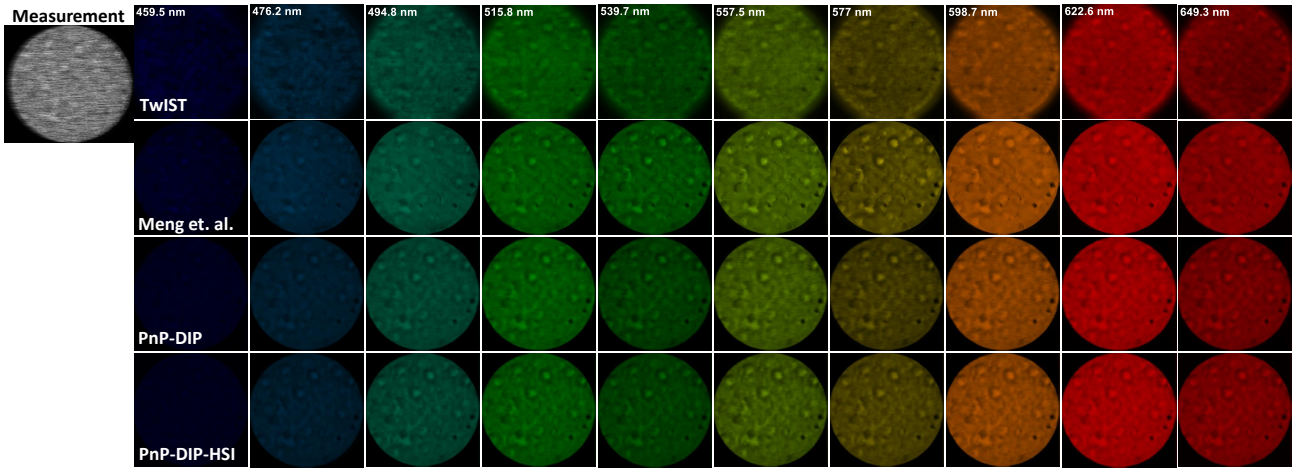

Figure M25. The results of the real data *Red blood cell 1* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.

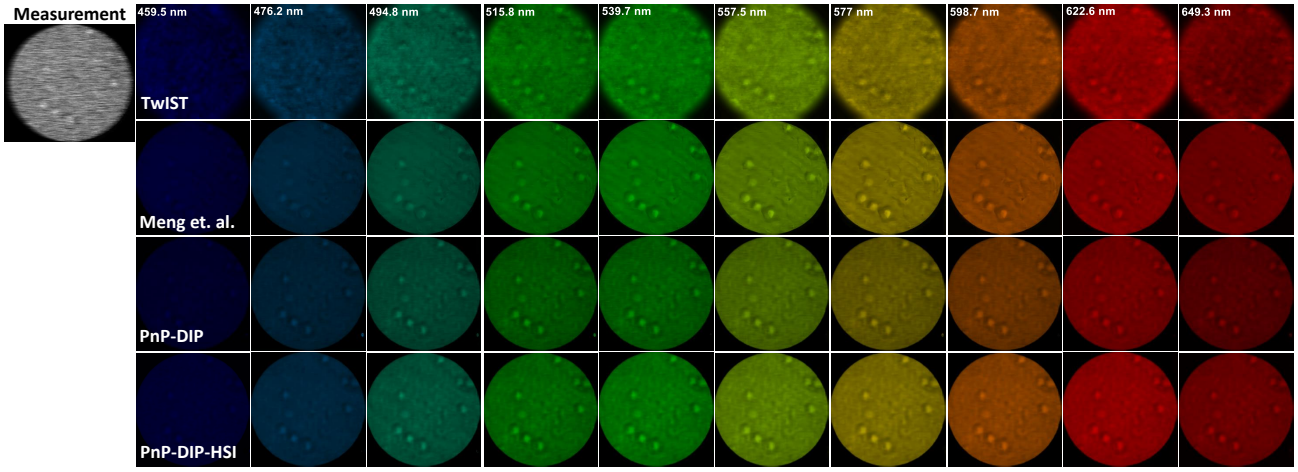

Figure M26. The results of the real data *Red blood cell 2* with 10 spectral channels reconstructed by TwIST, a supervised deep neural network and the proposed PnP-DIP and PnP-DIP-HSI.