Modeling Mask Uncertainty in Hyperspectral Image Reconstruction (Supplementary Material)

Jiamian Wang[1], Yulun Zhang[2], Xin Yuan[3], Ziyi Meng[4], and Zhiqiang Tao[1]

[1]Department of Computer Science and Engineering, Santa Clara University, USA
[2]ETH Zürich, Switzerland [3]Westlake University, China
[4]Kuaishou Technology, China
{jwang16,ztao}@scu.edu, yulun100@gmail.com, xyuan@westlake.edu.cn, mengziyi64@gmail.com

1 Overview

In this Supplementary Material, we present additional results and analyses about the proposed method as follows.

- **Reconstruction Backbone SRN**: A detailed introduction to the reconstruction backbone network used in the manuscript (Section 2).
- **Alternating Reconstruction Backbones**: More analyses on alternating reconstruction backbones employed in the proposed method (Section 3).
- **Spectral Fidelity Analysis**: Evaluation on the spectral fidelity of the reconstruction results by the proposed method (Section 4).
- **Epistemic Uncertainty Analysis**: More visualization and discussion on epistemic uncertainty of the proposed method (Section 5).
- **Complementary Ablation Studies**: Ablation studies under one-to-many miscalibration and the same mask setting (one-to-one) (Section 6).
- **Self-tuning variance Analysis**: More discussions about the proposed self-tuning variance. Specifically, more results for fixed variance are provided. Also, we demonstrate the convergence of $g_\phi(m)$, variational noise distributions given distinct noise priors (Section 7).
- **Datasets**: More illustrations on the dataset, includes training data, validation data, testing data, and mask set (Section 8).
- **Reconstruction Results**: More reconstruction results under many-to-many setting (Section 9), one-to-many setting (Section 10) and the traditional setting (Section 11).

2 Reconstruction Backbone: SRN

In the manuscript, we adopt a recent deep reconstruction network, SSI-ResU-Net (SRN) [11] as the backbone $f_\theta(\cdot)$. Specifically, the network input $x_{in} \in \mathbb{R}^{H\times W\times \Lambda}$ is initialized by the measurement $y \in \mathbb{R}^{H\times(W+\Lambda-1)\times\Lambda}$ and the mask $m \in \mathbb{R}^{H\times W}$

$$x_{in}[:,:,\lambda] = \texttt{shift}(y)_\lambda \odot m, \tag{1}$$

Table 1. Averaged PSNR(dB)/SSIM of the different models. We consider the miscalibration many-to-many scenario for a fair comparison. For three types of backbones, this is implemented by training upon a mask ensemble and testing on random masks. The *mean* and *std* are obtained upon 100 testing trials.

Models	PSNR (dB)	SSIM
SRN [11]	$32.24_{\pm 0.10}$	$0.9121_{\pm 0.0010}$
Spectral ViT [1]	$31.62_{\pm 0.09}$	$0.9282_{\pm 0.0010}$
SwinIR [3]	$33.49_{\pm 0.10}$	$0.9501_{\pm 0.0010}$
SRN+GST (Ours)	$\mathbf{33.02}_{\pm 0.01}$	$\mathbf{0.9285}_{\pm 0.0001}$
Spectral ViT+GST (Ours)	$\mathbf{32.15}_{\pm 0.01}$	$\mathbf{0.9330}_{\pm 0.0001}$
SwinIR+GST (Ours)	$\mathbf{34.15}_{\pm 0.01}$	$\mathbf{0.9548}_{\pm 0.0001}$

where $\odot$ is a Hadamard product and the `shift` is the reverse operation applied in the forward process (see Eq. (1) in the manuscript for more details).

The model is composed of a 1) *main body*, which is simultaneously bridged by a global skip connection, 2) a *head* operation and 3) a *tail* operation, both of which are conducted by a `CONV-ReLU` structure. Let x_{head} and x_{body} be the output of the *head* operation and the *main body*, respectively. We have

$$x_{body} = f_{res}^{J}(f_{res}^{J-1}(...(f_{res}^{1}(x_{head}))...)), \tag{2}$$

where $J = 16$ concatenated residual blocks share the same structure, i.e., $f_{res}(x) = x + (\texttt{CONV}(\texttt{ReLU}(\texttt{CONV}(x))))$.

3 Alternating Reconstruction Backbones

The performance and the robustness toward masks of the deep reconstruction networks largely depend on their constructions. Thus, we validate the effectiveness of the proposed method upon backbones with different architectures.

SwinIR Backbone. In this supplementary material, we consider transformer architectures as the backbone. Specifically, transformer acquires modeling ability from attention mechanism [10], which has been proved to behave quite differently from the traditional ConvNets [9].

Given the initialized input x_{in} by Eq. (1), we firstly implement the backbone by Swin transformer structure [4], which computes the spatial self-attention. It is composed of three modules: (1) shallow feature extraction by a `CONV3×3` layer, i.e., $x_{\texttt{SF}} = \texttt{CONV}(x_{in})$, (2) deep feature extraction module consisting of K concatenated residual Swin transformer blocks, i.e, $x_{\texttt{DF}} = f_{\texttt{DF}}(x_{\texttt{SF}})$ where $f_{\texttt{DF}}(\cdot) = f_{\texttt{RSTB}}^{K}(f_{\texttt{RSTB}}^{K-1}(...(f_{\texttt{RSTB}}^{1}(\cdot))...))$, and (3) a reconstruction module by a `CONV3×3` layer, i.e., $\widehat{x} = \texttt{CONV}(x_{\texttt{DF}})$.

For each residual Swin transformer block $f_{\texttt{RSTB}}(\cdot)$, we have L Swin transformer layers, which conducts window-based MSA and MLP

$$x = f_{\texttt{W-MSA}}(f_{\texttt{LN}}(x)) + x, \quad x = f_{\texttt{MLP}}(f_{\texttt{LN}}(x)) + x, \tag{3}$$

where the details of the $f_{\texttt{W-MSA}}(\cdot)$, $f_{\texttt{MLP}}(\cdot)$, and $f_{\texttt{LN}}(\cdot)$ could be found in [4]. In the experiment, we set the $K = 4$, $L = 6$. For all the blocks, we let the embedding dimension to be 60 and number of heads to be 6.[1]

Spectral ViT. We also provide another type of vision transformer, which exchanges the previous spatial self-attention with the spectral self-attention. Specifically, it treats the feature map of each embedding channel as a token. Given the query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$, we have output $\mathbf{X}$

$$\mathbf{X} = \mathbf{V}\texttt{Attn}(\mathbf{K}, \mathbf{Q}), \quad \textit{where} \quad \texttt{Attn}(\mathbf{K}, \mathbf{Q}) = \texttt{softmax}(\mathbf{K}^T\mathbf{Q}/\delta), \tag{4}$$

where δ denotes a learnable scalar. For more details please refer to [1].

Comparison. We summarize the performance of different backbones under miscalibration scenario many-to-many in Table 1. For detailed illustration of the miscalibration setting, please refer to manuscript Section 4.1. Notably, the integration of the backbone into our method is implemented by training the full model upon a mask dataset $\mathcal{M}$ in a bilevel optimization framework.

By comparison, one can draw the following conclusions. (1) For the metric comparison, our method brings performance gain $\Delta\texttt{PSNR} = 0.78$dB, 0.53dB, and 0.66dB, respectively, for different backbones. (2) The proposed method enables high-fidelity reconstruction with the highest confidence. Specifically, in `many-to-many` case, $\texttt{PSNR}_{std} > 0.1$dB. Our method achieves $10\times$ the randomness control, indicating a better epistemic modeling capacity (Please see Section 5).

4 Spectral Fidelity.

In this work, we adopt two methods to demonstrate the spectral fidelity of the reconstruction results.

Firstly, given the prediction $\widehat{x} \in \mathrm{R}^{H\times W\times N_\lambda}$, we treat each spectral channel as a `R.V.` (random variable) of HW dimensions and calculate the channel-wise correlations. For each hyperspectral image with $N_\lambda = 28$, a correlation matrix of 28×28 could be visualized. We compare these matrices by the references and the predictions in Fig. 1. The more consistent they are, the higher spectral fidelity we achieve. By observation, the correlation matrices by the predictions show highly similar visual patterns as the reference, indicating that the proposed method effectively captures long-range spectra dependencies. We also notice that there might be minor differences at the centers of the matrices between the visualizations. Rectifying this part is pretty challenging as the model need to precisely distinguish the difference between each adjacent spectral pairs.

Secondly, we quantitatively compare the spectral fidelity of different methods upon density curves. As three examples demonstrated in Fig. 2, we first crop a small spatial patch from the prediction (exampled by the RGB reference on the right most column), then we draw the density curve by pixel intensities in that small patch. Finally, the correlations between the reference curve and that

[1] For implementation please refer to https://github.com/JingyunLiang/SwinIR.

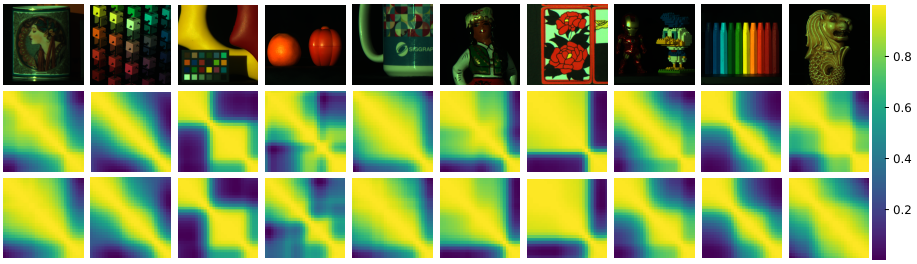

Fig. 1. RGB references of the benchmark simulation test set (top line) and spectral correlation coefficient visualizations by the reference (middle line) as well as the proposed method (bottom line). Each correlation coefficient map is of the size 28×28.

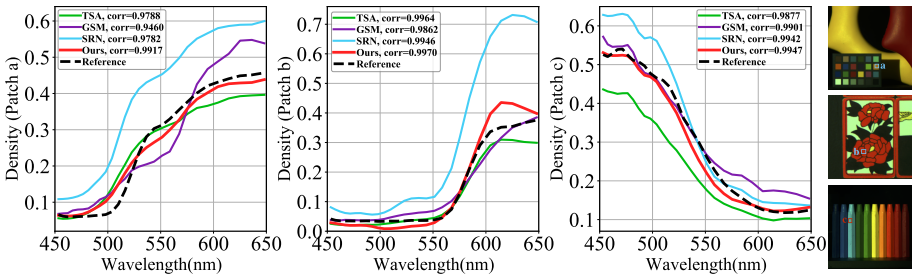

Fig. 2. Spectral correlation to the ground truth on exampled locations. Spatial patches (patch `a`,`b`,`c` as plotted in the right most column RGB references, please zoom in for better visualization) are chosen to ensure the monochromaticity. Density curves are computed upon the predictions by different methods within the chosen patch.

from the predictions are computed. Higher correlation values indicate a higher spectral fidelity for the cropped patch. The small patch is chosen to ensure the monochromaticity (wavelength). For example, if we choose the patch whose color lies in blue∼cyan range (bottom-right RGB reference in Fig. 2), the energy of the density would concentrate in the 450nm∼500nm.

Table 2. Averaged spectral correlations (↑) to the reference. For each scene, we compute the averaged correlation values upon density curves corresponding to five selected patches. Please refer to examples in Fig. 2 for a detailed computational procedure.

Methods	Scene1	Scene2	Scene3	Scene4	Scene5	Scene6	Scene7	Scene8	Scene9	Scene10	*Avg.*
GSM [2]	0.9903	0.9674	0.9751	0.9435	0.9891	0.9904	0.9951	0.9700	0.9870	0.8926	0.9701
SRN [11]	0.9910	0.9701	0.9753	0.9429	0.9898	0.9910	0.9959	0.9865	0.9870	0.9000	0.9730
Ours	**0.9956**	**0.9910**	**0.9921**	**0.9458**	**0.9925**	**0.9982**	**0.9969**	**0.9867**	**0.9903**	**0.9004**	**0.9790**

To globally compare the spectral fidelity, we randomly choose five monochromatic patch of each scene and compute an averaged correlation value upon five density curves. We report the correlation values of ten scenes in Table 2 and demonstrate the superiority of the proposed method, as compared to GSM [2] and SRN [11].

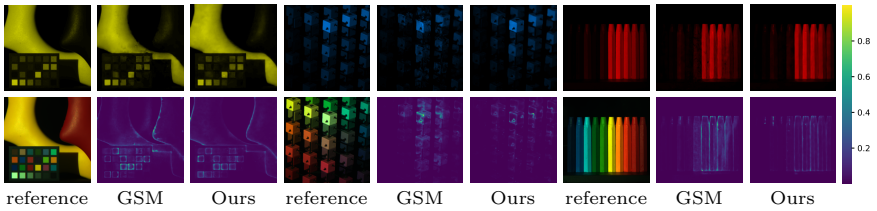

Fig. 3. Three exampled epistemic uncertainty visualizations by GSM [2] and the proposed method. For each example, we demonstrate averaged reconstruction results on selected wavelengths (i.e., 567.5nm, 471.6nm, and 614.4nm) in top line, and the epistemic uncertainty in the bottom line.

5 Epistemic Uncertainty

As illustrated in Section 4.2 of the manuscript, the proposed method demonstrates low epistemic uncertainty by approximating the mask distribution. In this section, we provide more visualizations and analyses on epistemic uncertainty in Fig. 3. Specifically, we test the well-trained models upon random real masks and repeat 100 trials. Both GSM and the proposed method are trained upon the same mask set $\mathcal{M}$ for a fair comparison. For each exampled hyperspectral images, we compare the averaged reconstruction and epistemic uncertainty on a selected spectral channel. Notably, in low-frequency regions, both methods show high confidence, while in high-frequency regions (i.e., edges), the proposed method presents a much-lower epistemic uncertainty, which would potentially benefit the down-stream applications like object detection or segmentation upon hyperspectral images.

6 Ablation Study

In this paper, three scenarios are introduced: 1) one-to-one setting, which is the traditional setting considered by previous reconstruction methods, 2) one-to-many miscalibration, 3) many-to-many miscalibration. Notably, the third scenario enables a complete mask distribution modeling, for which reason we put more emphasize on it and provide the ablation study accordingly in the manuscript. Following that, Table 3 conducts the same ablation experiments under the traditional setting (one-to-one). For miscalibration (one-to-many), we also do verification and report the performance in Table 4. The ablated models include

- `w/o GST`: we remove the graph-based self-tuning (GST) network from the proposed method. Actually it degrades into the reconstruction backbone SRN [11] applied under corresponding scenarios.
- `w/o Bi-Opt`: we simultaneously optimize all of the parameters by the lower-level loss function, i.e., Eq. (7) in the manuscript, based on the original training set.

Table 3. Ablation study of the proposed method under the traditional setting (one-to-one).

Settings	PSNR(dB)	SSIM
`w/o GST`	33.17	0.9288
`w/o Bi-Opt`	32.73	0.9193
`w/o GCN`	33.23	0.9286
Ours (full model)	33.22	0.9292

Table 4. Ablation study of the proposed method under the setting of miscalibration (one-to-many) among 100 testing trials.

Settings	PSNR(dB)	SSIM
`w/o GST`	30.17±0.63	0.8865±0.0108
`w/o Bi-Opt`	30.30±0.06	0.8843±0.0011
`w/o` GCN	30.13±0.07	0.8849±0.0011
Ours (full model)	30.60±0.08	0.8881±0.0013

– `w/o GCN`: for the self-tuning network, we exchange the GCN with a convolutional layer carrying more parameters for a fair comparison.

As shown in the Table 3 and Table 4, both the GST module and bilevel optimization strategy contribute significantly for the final performance boost. While the ablated model `w/o GCN` in self-tuning network works comparably with the `Ours (full model)` under the traditional setting by PSNR, it falls behind regarding SSIM, indicating a sub-optimal reconstruction ability.

7 Model Discussion

Fixed variance. In the manuscript, we showcase that the fixed variance with distinct values only achieves sub-optimal performance compared with the self-tuning variance. In Fig. 5 (b), we also plot the SSIM curve (green) with different values of the fixed variance. Besides, the original PSNR curve (green) is shown in Fig. 5 (a).

Self-tuning variance under different priors. Due to the limitation of time and computational resource, we only discussed the self-tuning variance under three most representative noise priors in the manuscript, i.e., $\mathcal{N}(0.006, 0.005)$, $\mathcal{N}(0.006, 0.1)$ and $\mathcal{N}(0.0, 1.0)$. In this section, we demonstrate additional results corresponding to more noise priors. Similar to the performance curves plotted in Fig. 7 (a) in the manuscript, we report the corresponding performances by red curves in Fig. 5, verifying the superiority of the self-tuning variance. In Fig. 4, we visualize both the noise prior and the subsequent variational noise distributions. The similarity between all subplots – the learned variational noise distribution gets a smaller variance than the given prior – validates the effectiveness of mask uncertainty modeling.

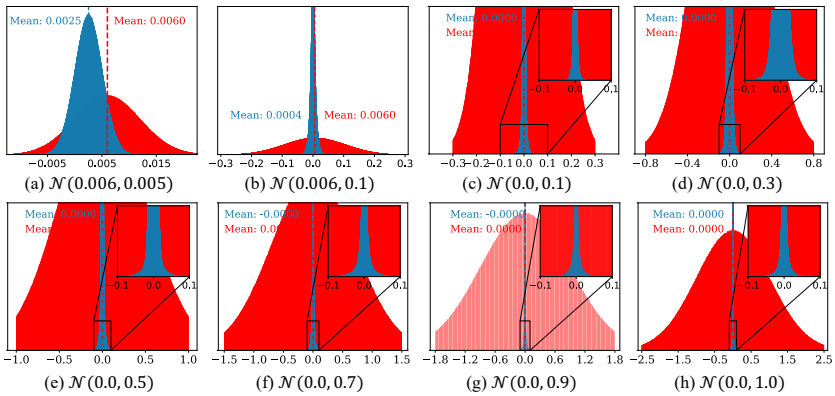

Fig. 4. Learned variational noise distribution under different priors. Eight different priors (red) are adopted in the experiment. By comparison, variational noise distributions (blue) are characterized by smaller variance. Please refer to the red curves in Fig. 5 for corresponding reconstruction performance comparison.

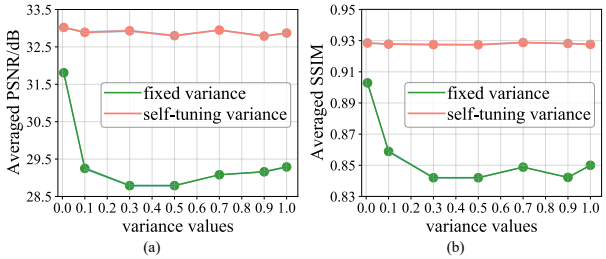

Fig. 5. Performance comparison between fixed variance (green) and self-tuning variance (red). The PSNR is compared in (a) and SSIM is compared in (b). Reconstruction using self-tuning variance outperforms that using fixed variance with different values.

In Fig. 6, we explore the convergence of the $g_\phi(m)$ during the training phase of our best model. For pre-training phase of reconstruction network (first 20 epochs as mentioned in the manuscript), the range of $g_\phi(m)$ remains invariant. An interesting observation is that the fluctuation of the $g_\phi(m)$ value is accompanied by fluctuation of the reconstruction performance, indicating the underlying impact of the self-tuning variance. During the last 200 epochs (including training and validation epochs), a converged $g_\phi(m)$ contributes to a steady performance improvement.

8 Dataset

HSI data set. We adopt the training set provided in [6] and follow the same data augmentation operations. Specifically, the training set contains 205 1024×1024×28 training samples, all of which sources from the CAVE dataset [12]. Our model is trained on 256×256×28 patches randomly cropped from these 205 samples. For a fair comparison with the other deep reconstruction networks, we create a validation data set by randomly splitting 40 hyperspectral images from the above

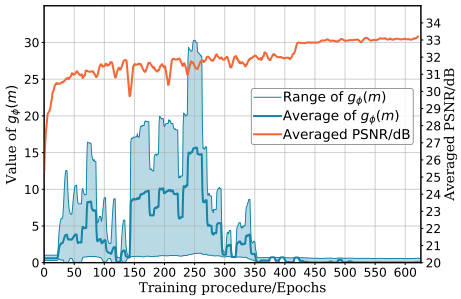

Fig. 6. Observation on $g_\phi(m)$ during training. The $g_\phi(m)$ gradually converge to a smaller range with more epochs of training. Meanwhile, a better reconstruction performance can be observed. Both training epochs and validation epochs are jointly counted.

205 samples. Therefore, *no new HSI data is introduced for our model training.* For the model testing, ten simulation hyperspectral images corresponding to 10 scenes shown in Fig. 1 are used for quantitative and perceptual comparison, following previous works [6,2,11,8,5].

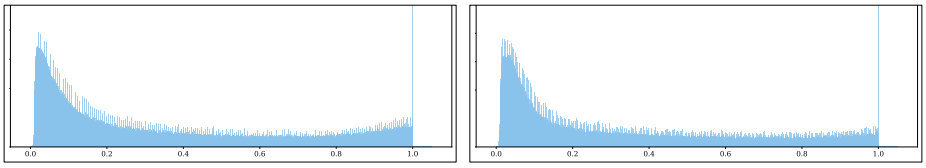

Fig. 7. Histograms of two real masks applied in this work. (a) sources from [6] and (b) sources from [7]. Both masks are produced by the same fabrication process. Bin number is set to 2000 for both histograms.

Mask set. Two 660×660 real masks following the same fabrication process are employed in this work. Fig. 7 demonstrates the histograms of both masks. As mentioned in the manuscript, the training mask set $\mathcal{M}$ is built by randomly cropping 256×256 patches from the first real mask [6]. For simulation data, testing masks are collected from both real masks. Notably, **there is no overlap between training and testing mask sets.** For real HSI reconstruction, no testing mask set is available. The second 660×660 real mask [7] is directly applied for testing purpose, indicating the miscalibration scenario.

9 Many-to-many Reconstruction

In this section, we provide two simulation example of reconstruction results by different methods, i.e., TSA-Net [6], GSM [2], SRN [11] and ours, under the scenario of miscalibration (many-to-many) in Fig. 8 and Fig. 9, respectively. Notably, we randomly select one unseen testing mask for visualization. For total ten

simulation testing hyerspectral images, please refer to the `zip` file by following the directory of `simulation results` > `many to many`.

10 One-to-many Reconstruction

In this section, we provide one real example of reconstruction result by different methods, i.e., TSA-Net [6], GSM [2], SRN [11] and ours, under the scenario of miscalibration (one-to-many) in Fig. 10. Notably, we randomly select one unseen testing mask for visualization. For total ten simulation testing hyperspectral images, please refer to the `zip` file by following the directory of `simulation results` > `one to many`. For total five real data results, please refer to the `zip` file by following the directory of `realdata results` > `one to many`.

11 Same mask Reconstruction

In this section, we provide one simulation example of reconstruction result by different methods, i.e., TSA-Net [6], GSM [2], SRN [11] and ours, under the traditional setting in Fig. 11. For total ten simulation testing hyperspectral images, please refer to the `zip` file by following the directory of `simulation results` > `one to one`. For total five real data results, please refer to the `zip` file by following the directory of `realdata results` > `one to one`.

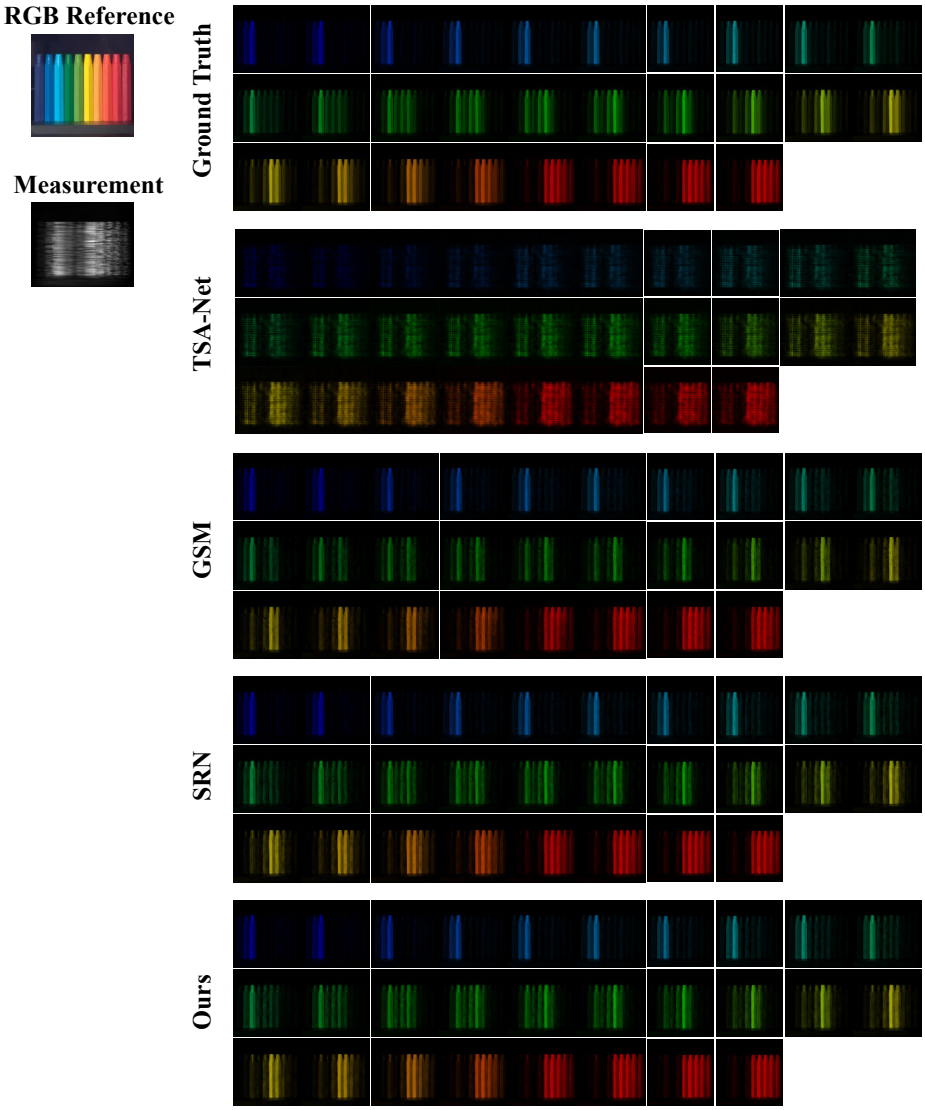

Fig. 8. Exampled simulation reconstruction result under the miscalibration setting of many-to-many. The proposed method, TSA-Net [6], GSM [2], and SRN [11] are compared. Among diverse spectral channels, the proposed method contains least blurring and distortion by comparison. Zoom in for better visualization.

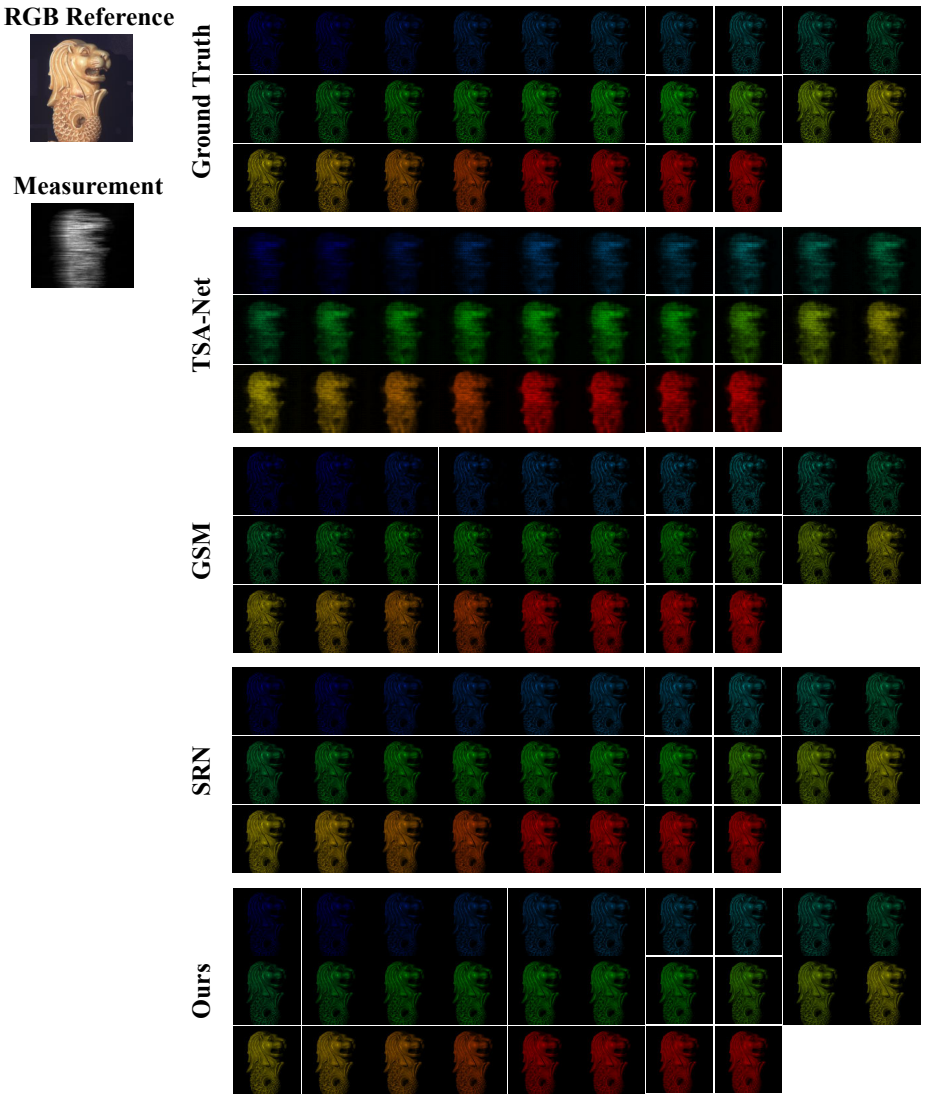

Fig. 9. Exampled simulation reconstruction result under the miscalibration setting of many-to-many. The proposed method, TSA-Net [6], GSM [2], and SRN [11] are compared. Among diverse spectral channels, the proposed method contains least blurring and distortion by comparison. Zoom in for better visualization.

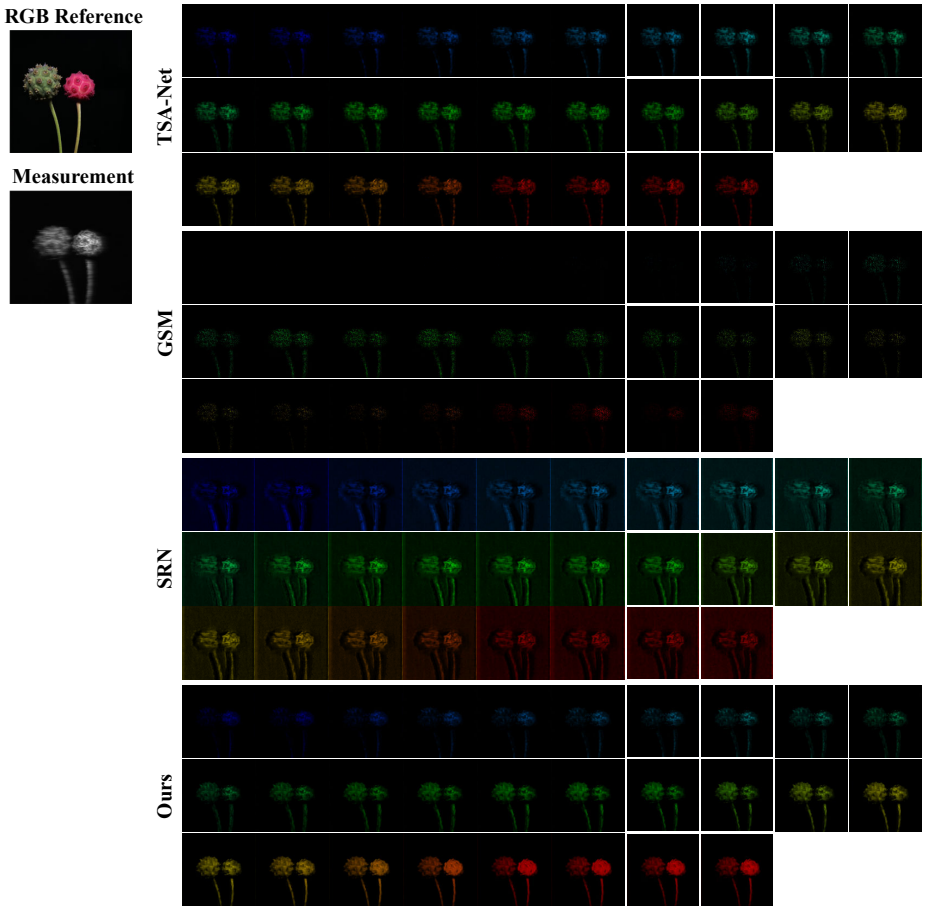

Fig. 10. Exampled real reconstruction result under miscalibration (one-to-many). The proposed method, TSA-Net [6], GSM [2], and SRN [11] are compared. Our method enables least artifact while TSA-Net suffers from the shot noise, GSM endures the low brightness, and SRN gives distorted results. Zoom in for better visualization.

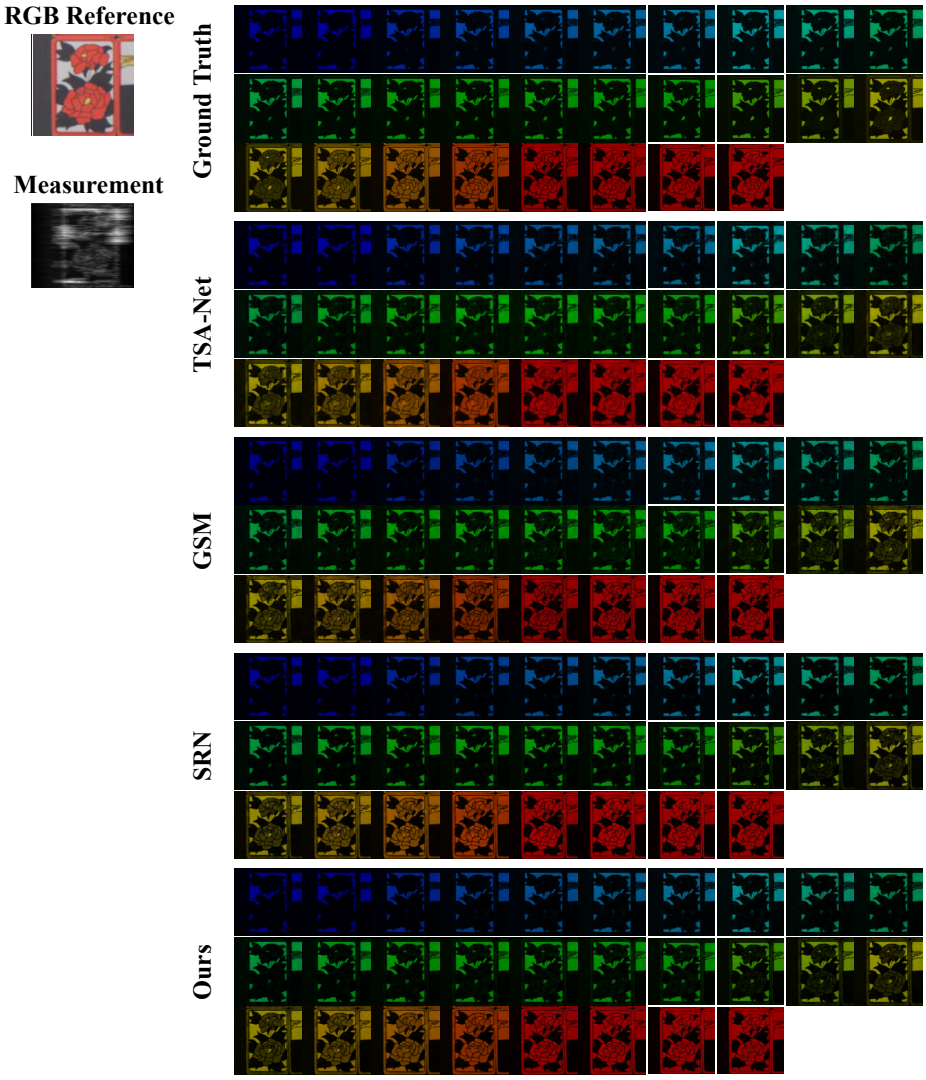

Fig. 11. Exampled simulation reconstruction result under the traditional setting. The proposed method, TSA-Net [6], GSM [2], and SRN [11] are compared. The proposed method works better in high frequency regions. Zoom in for better visualization.

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. In: NeurIPS (2021) 2, 3
2. Huang, T., Dong, W., Yuan, X., Wu, J., Shi, G.: Deep gaussian scale mixture prior for spectral compressive imaging. In: CVPR (2021) 4, 5, 8, 9, 10, 11, 12, 13
3. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: CVPR (2021) 2
4. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: CVPR (2021) 2, 3
5. Meng, Z., Jalali, S., Yuan, X.: Gap-net for snapshot compressive imaging. arXiv preprint arXiv:2012.08364 (2020) 8
6. Meng, Z., Ma, J., Yuan, X.: End-to-end low cost compressive spectral imaging with spatial-spectral self-attention. In: ECCV (2020) 7, 8, 9, 10, 11, 12, 13
7. Meng, Z., Qiao, M., Ma, J., Yu, Z., Xu, K., Yuan, X.: Snapshot multispectral endomicroscopy. Optics Letters **45**(14), 3897–3900 (2020) 8
8. Meng, Z., Yu, Z., Xu, K., Yuan, X.: Self-supervised neural networks for spectral snapshot compressive imaging. In: ICCV (2021) 8
9. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., Dosovitskiy, A.: Do vision transformers see like convolutional neural networks? In: NeurIPS (2021) 2
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 2
11. Wang, J., Zhang, Y., Yuan, X., Fu, Y., Tao, Z.: A new backbone for hyperspectral image reconstruction. arXiv preprint arXiv:2108.07739 (2021) 1, 2, 4, 5, 8, 9, 10, 11, 12, 13
12. Yasuma, F., Mitsunaga, T., Iso, D., Nayar, S.: Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum. Tech. rep. (2008) 7