Supplemental Materials for Stability and Non-local Modeling in Hybrid Convolution–Transformer Networks for Snapshot Hyperspectral Reconstruction

Xian-Hua Han
Graduate School of Artificial Intelligence and Science, Rikkyo University
3-34-1 Nishi-Ikebukuro, Toshima-ku, Tokyo 171-8501, Japan
hanxhua@rikkyo.ac.jp

1. Overview

Due to the page limit of the main manuscript, we provided only the visualization results comparing our deep unfolding implementation with the proposed BCPTE for prior learning and the existing deep unfolding methods (DUM). However, the visualization comparison between our end-to-end (E2E) hyperspectral reconstruction network, HCT-UNet, and the recent state-of-the-art E2E approaches was omitted from the main paper. To present a more comprehensive evaluation, this supplemental material includes three additional visualization examples: (1) Two examples comparing E2E methods, including our proposed HCT-UNet and representative state-of-the-art E2E reconstruction networks. (2) One additional example comparing DUM-based reconstruction methods, complementing the limited visual results included in the main text.

Moreover, we also conduct a Robustness Evaluation Under Different Noise Levels to further demonstrate the potential stability introduced by the incorporated Spatial Stability Enhanced Convolution (SSE-Conv) block used for initial prior learning in HCT-UNet.

2. Visualization Comparison of E2E Methods

We present two representative hyperspectral scenes to qualitatively assess the reconstruction performance of various end-to-end learning–based methods as shown in Figs. 1 and. 2. For each scene, we show: 1) the reconstructed HS image with various E2E methods including our BCPTE-L, TSA-Net [4], MST-L [2] and CST-L+ [1], S^2-Tran [5] and DWMT [3]; 2) the erro maps of the differenc images betwen reconstruction and the Ground-Truth. These examples demonstrate the advantage of our architecture in recovering fine spatial structures and preserving spectral fidelity, especially in regions with complex textures and abrupt spectral variations.

3. Visualization Comparison of DUM Methods

In addition, we include one more visualization example comparing our unfolding-based implementation with BCPTE-prior learning and several existing DUM baselines. This example further highlights how integrating hierarchical contextual learning and transformer-based reconstruction contributes to improved detail restoration and reduces artifacts commonly observed in traditional unfolding schemes.

4. Robustness Evaluation Under Different Noise Levels

To further assess the robustness of the proposed HCT-UNet, we introduce varying levels of additive noise into the compressive snapshot measurements, even though no explicit noise modeling or noise augmentation is incorporated during training. This evaluation setting enables an examination of the intrinsic stability of different architectures when confronted with progressively degraded measurement quality. We compare our method with a pure Transformer-based U-Net and analyze the evolution of PSNR and SSIM as the noise level increases. The corresponding quantitative trends are presented in Fig. 4. As shown in Fig. 4, HCT-UNet consistently achieves higher PSNR and SSIM across all noise levels compared with the pure Transformer U-Net. Although both models exhibit expected performance degradation under increased noise, the rate and magnitude of degradation differ substantially, underscoring the superior robustness of HCT-UNet. This enhanced robustness can be largely attributed to the Spatial Stability Enhanced Convolution (SSE-Conv) module employed at the initial stage of the network. SSE-Conv establishes a noise-resistant and structurally coherent local feature space, ensuring that shallow representations suppress noise-induced local fluctuations while preserving essential spatial regularity.

The PSNR and SSIM curves in Fig. 4 further reveal sev-

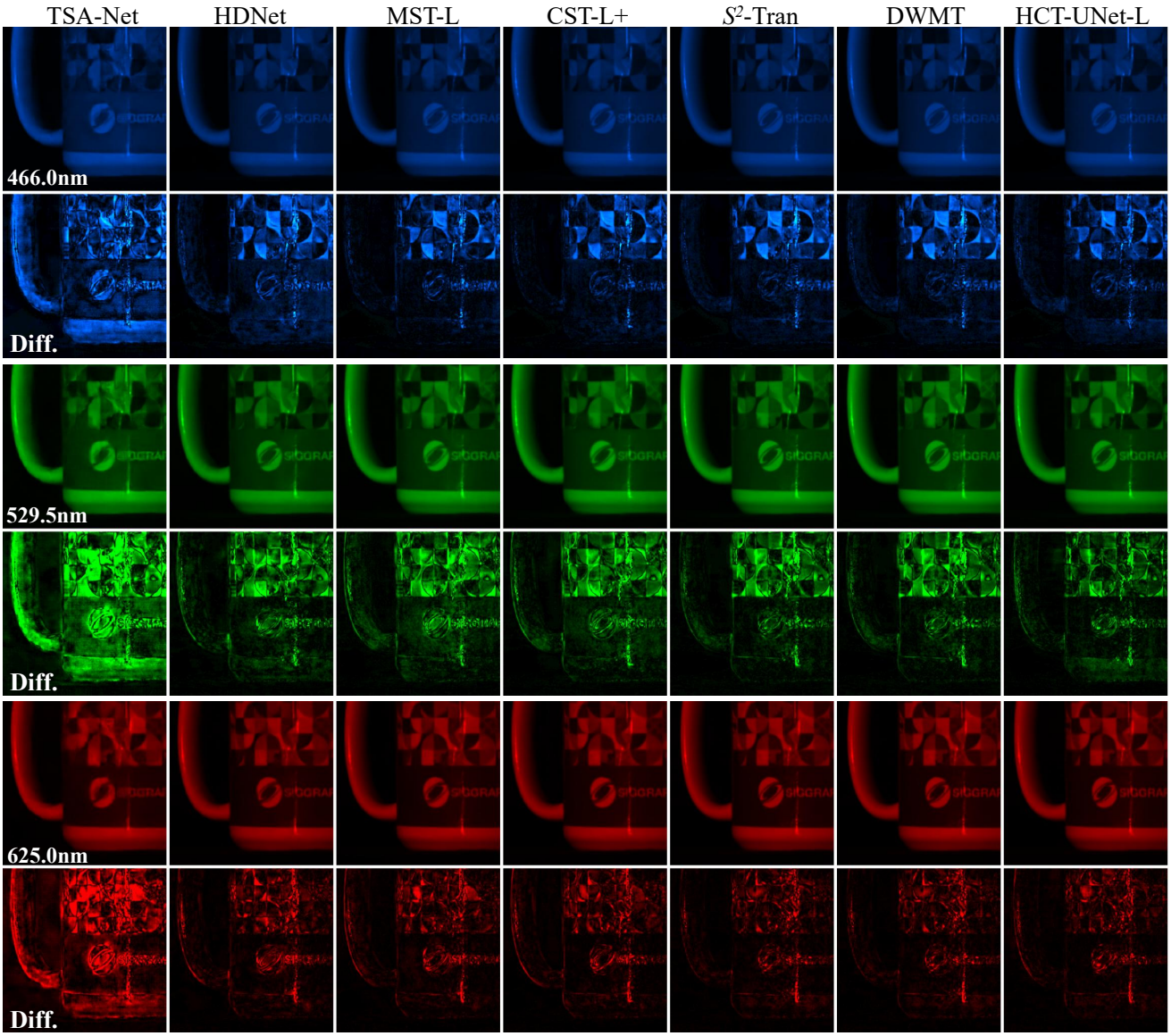

Figure 1. Qualitative reconstruction results of the proposed HCT-UNet and representative state-of-the-art E2E reconstruction networks for Scene 5 utilizing three out of 28 spectral channels, along with the corresponding difference images between the ground truth and the reconstructions.

eral important observations:

1) Slower performance degradation: The decline in HCT-UNet's performance is considerably more gradual than that of the pure Transformer U-Net. This suggests that the stable convolutional priors introduced by SSE-Conv mitigate error propagation into subsequent Transformer layers.

2) Higher resilience to spatial perturbations: Transformer architectures are known to be sensitive to local perturbations due to their global attention mechanism. In the pure Transformer U-Net, injected noise is amplified through the attention layers, leading to sharper drops in both PSNR and SSIM.

3) Superior preservation of structural similarity: HCT-UNet maintains higher SSIM even under heavy noise, indicating more faithful preservation of edges, textures, and local smoothness. This advantage arises from the spatial consistency priors reinforced by SSE-Conv, which stabilize shallow features before they are processed by deeper Conv–Transformer interactions.

Together, these observations confirm that SSE-Conv effectively enhances the network's resilience to measurement noise, enabling HCT-UNet to maintain both spectral fidelity

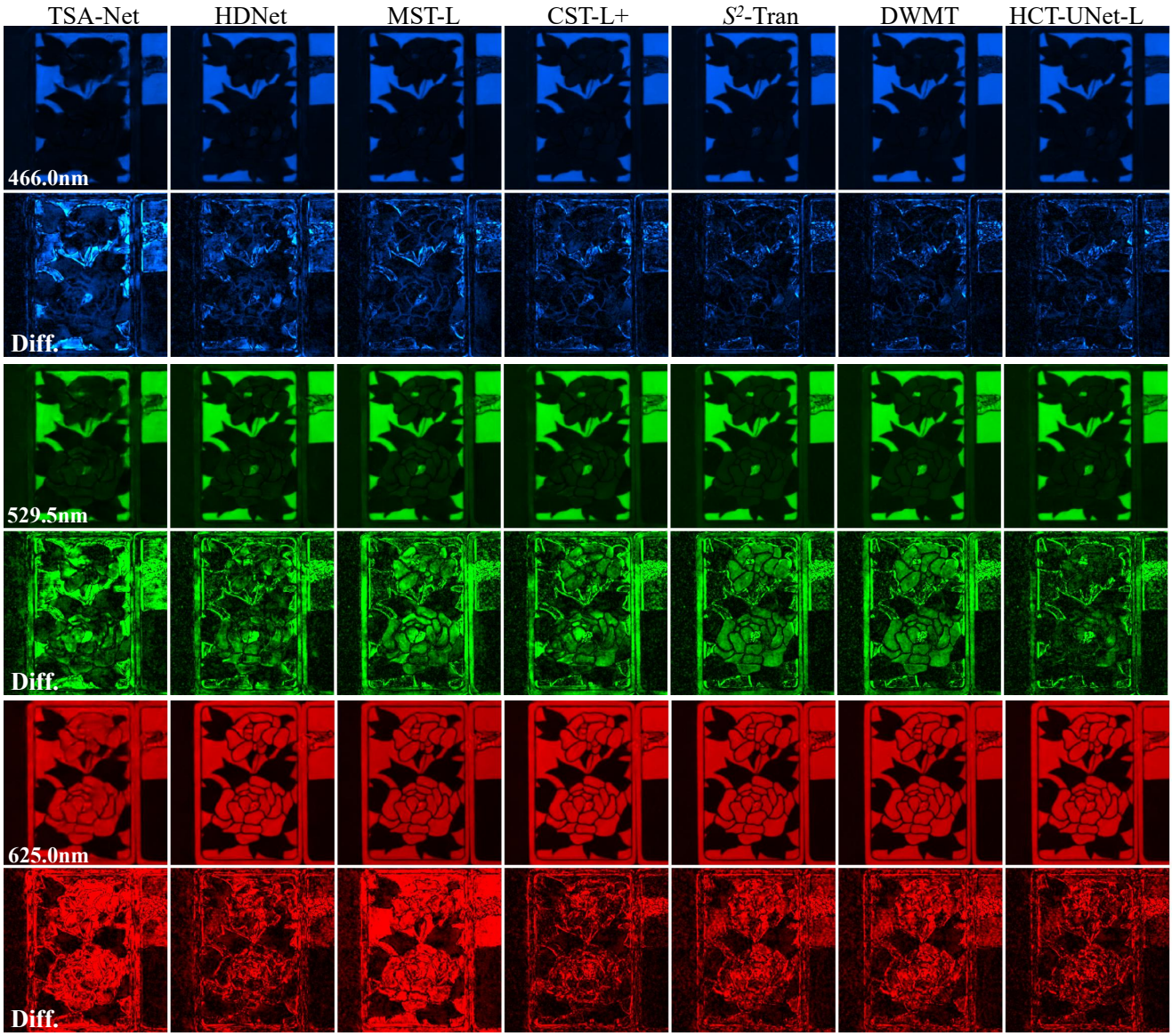

Figure 2. Qualitative reconstruction results of the proposed HCT-UNet and representative state-of-the-art E2E reconstruction networks for Scene 7 utilizing three out of 28 spectral channels, along with the corresponding difference images between the ground truth and the reconstructions.

and spatial consistency under increasingly challenging conditions.

References

[1] Y. Cai, J. Lin, X. Hu, H. Wang, X. Yuan, Y. Zhang, R. Timofte, and L. Van Gool. Coarse-to-fine sparse transformer for hyperspectral image reconstruction. *ECCV*, 2022. 1

[2] 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. *CVPR*, pages 17502–17511, 2022. 1

[3] F. Luo, X. Chen, X. Gong, W. Wu, and T. Guo. Dual-window multiscale transformer for hyperspectral snapshot compressive imaging. *AAAI*, 442:3972—-3980, 2024. 1

[4] Ziyi Meng, Jiawei Ma, and Xin Yuan. End-to-end low cost compressive spectral imaging with spatial-spectral self attention. *ECCV*, pages 187–204, 2020. 1

[5] Jiamian Wang, Kunpeng Li, Yulun Zhang, Xin Yuan, and Zhiqiang Tao. s^2-transformer for mask-aware hyperspectral image reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 1

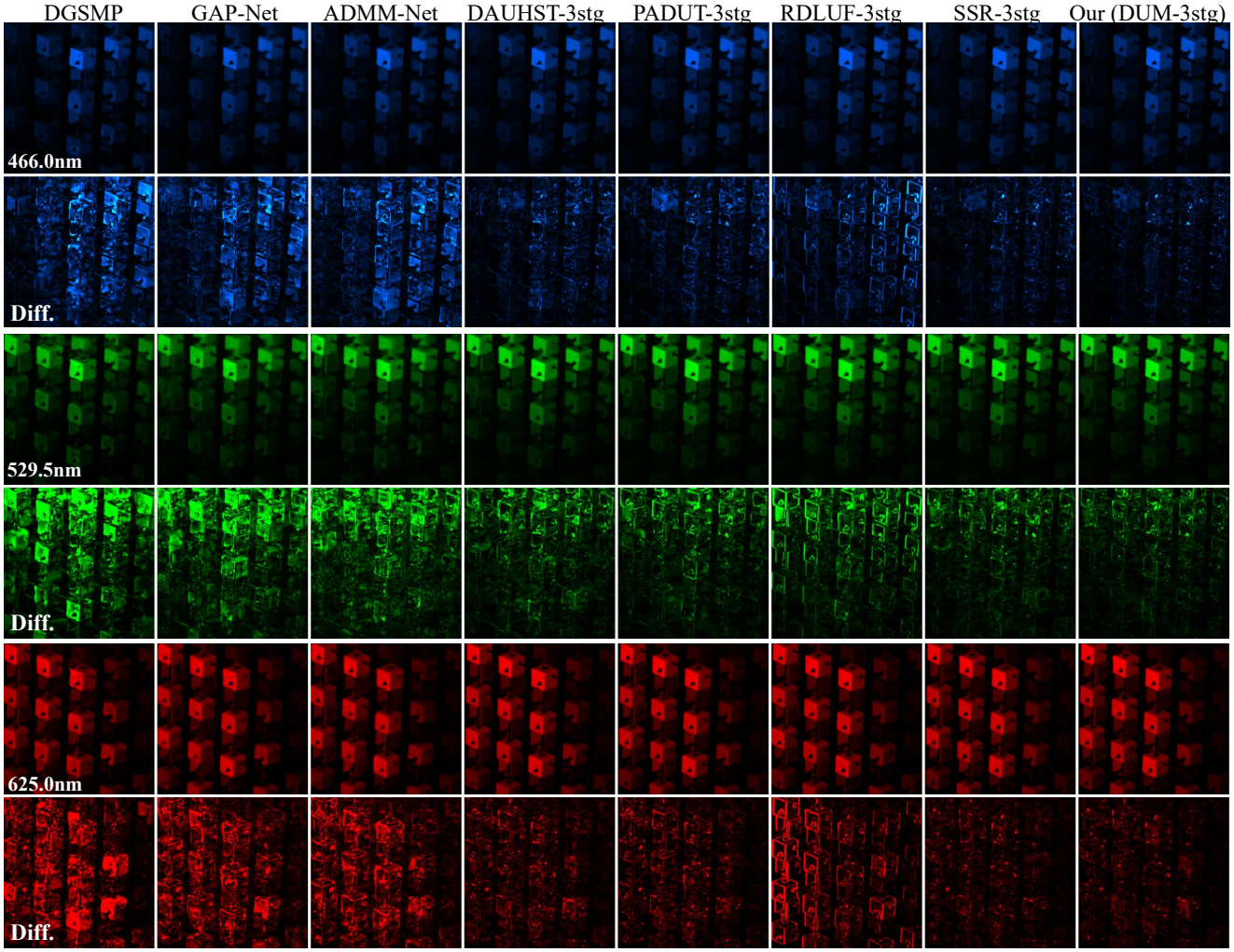

Figure 3. Additional qualitative comparison among the proposed deep-unfolding implementation with HCT-UNet for prior learning and existing DUM-based approaches for Scene 2 utilizing three out of 28 spectral channels, along with the corresponding difference images between the ground truth and the reconstructions.

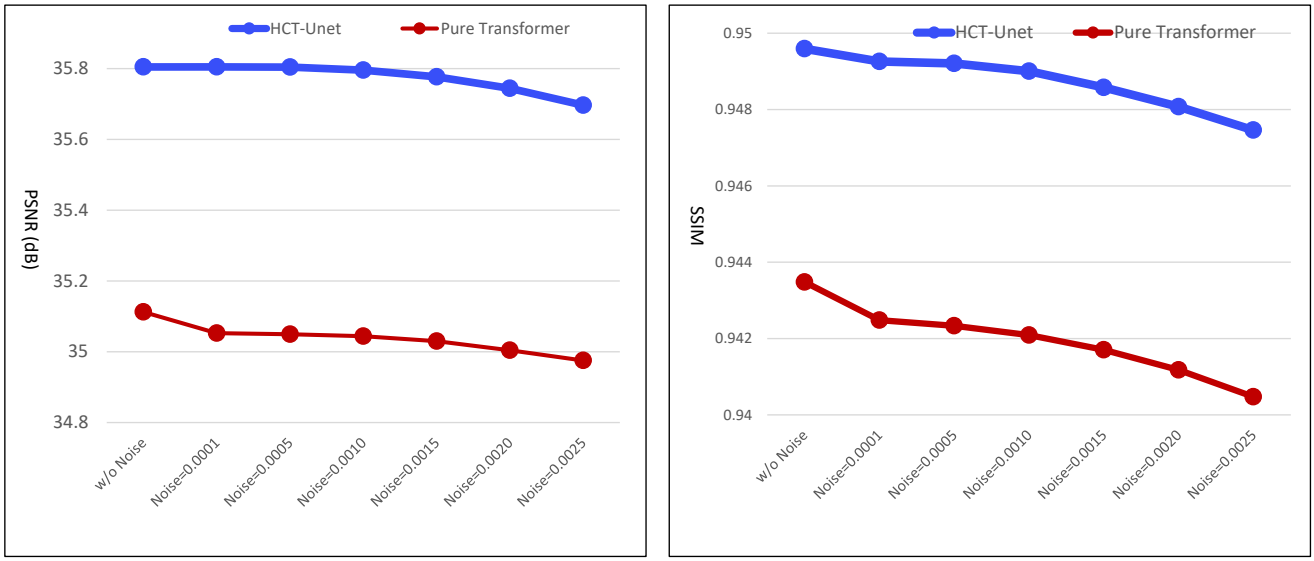

Figure 4. PSNR and SSIM evolution curves of the proposed HCT-UNet and a pure Transformer-based UNet under increasing Gaussian noise levels added to the compressive snapshot measurements. Despite no noise modeling during training, HCT-UNet consistently achieves higher reconstruction quality and slower degradation. This robustness stems from the Spatial Stability Enhanced Convolution (SSE-Conv) block, which enables the model to learn a stability prior and maintain spatial coherence under noisy conditions..