https://doi.org/10.2352/EI.2026.38.10.HVEI-210

Efficient Ultra-High-Resolution Hyperspectral Reconstruction Using a Patched Input Spatial-Spectral Transformer

Jinhyeok An, Wonkyung Jung, Jintae Jang, Jeongwook Lee, Sung-Su Kim, Yitae Kim; S.LSI Division, Samsung Electronics; Hwaseong-si, Gyeonggi-do, Republic of Korea

Abstract

Transformers, which have demonstrated remarkable performance improvements in natural language processing, have been increasingly adopted in computer vision tasks since the introduction of the Vision Transformer (ViT). In hyperspectral image (HIS) reconstruction, Transformer-based models have gained popularity due to their ability to capture global dependencies. While these models alleviate the certain limitations of convolutional neural networks (CNNs), their computational complexity scales quadratically with spatial resolution, making ultra-high-resolution reconstruction infeasible. Spectral Transformer variants have been proposed to reduce the computational burden associated with high spatial resolution, yet they still face challenges in handling ultra-high-resolution imagery. In this work, we propose a "Patched Input Spatial-Spectral Transformer (PSST)" that efficiently reconstructs HSIs from ultra-high-resolution RGB images. The model integrates a spatial transformer before spectral processing, enabling global context awareness while maintaining computational efficiency through in-model patch partitioning. Although performance slightly decreases for low-resolution inputs compared to state-of-the-art (SOTA) models, our method achieves the highest reconstruction quality for ultra-high-resolution inputs, achieving higher PSNR while significantly reducing memory consumption.

Introduction

Hyperspectral imaging provides detailed information for each pixel, enabling applications in precision agriculture [4, 13], remote sensing [5, 12], and medical diagnostics [1]. The demand for ultra-high-resolution HSI has been growing, but acquiring such data is challenging due to hardware limitations, long acquisition times, and high costs. RGB-to-HSI reconstruction offers a promising alternative, but most existing methods only handle low-resolution inputs.

SOTA methods often rely on Transformer architectures [16, 20, 21], which require large amounts of GPU memory for high-resolution processing. A common workaround is to split large images into patches before reconstruction, but this approach often produces visible artifacts, such as sharp patch boundaries and brightness inconsistencies, that degrade visual quality. Overcoming these limitations is crucial for achieving ultra-high-resolution HSI reconstruction without compromising image fidelity.

We address this by introducing an in-model patch partitioning strategy that preserves global context and ensures smooth and artifact-free outputs while enabling end-to-end training on megapixel-scale images

Our work builds upon the state-of-the-art RGB-to-HSI reconstruction model MST++ [21], which employs the Spectral Multi-Scale Attention (S-MSA) [20] mechanism. When the spectral dimension is fixed, S-MSA, a spectral transformer, maintains constant computational complexity with respect to spatial resolution, thereby addressing the limitations of conventional spatial transformers. However, the feed-forward network (FFN) [6] modules within S-MSA, which provide nonlinearity to the model, significantly increase computational cost. Consequently, as input image size grows, the overall complexity escalates dramatically, making direct ultra-high-resolution reconstruction infeasible.

To mitigate this issue, we adopt in-model patch partitioning to reduce both memory consumption and computational burden. Nevertheless, naïve patch partitioning removes global context, leading to independent patch processing without considering spatial relationships. As shown in Figure 1, existing SOTA patch-based reconstructions suffer from noticeable boundary artifacts and brightness inconsistencies.

Figure 1. *Results of patch-based restoration using a SOTA model*

To address these limitations, we introduce a spatial transformer [11, 23] module prior to the spectral transformer. The spatial transformer computes a global attention map across the entire image, which is then flattened, and concatenated with each patch's features. This design allows each patch to retain global contextual information before entering the spectral transformer, thereby ensuring consistent brightness and seamless boundaries, and preserving computational efficiency. In this way, our model leverages the strengths of patch-based processing while overcoming the fundamental limitation of losing inter-patch relationships.

Related Work

Hyperspectral Imaging (HSI)

Traditional hyperspectral imaging (HSI) acquisition systems primarily rely on three types of scanners, that employ optical components to selectively transmit and detect wavelengths while sequentially scanning the scene across spatial-spectral dimensions.

The Whisk-broom scanner captures spectral information pixel by pixel using a rotating or oscillating micro-mirror. This design enables high spectral resolution and low spectral crosstalk, but its limitations include poor suitability for dynamic scenes and limited

portability due to the mechanical complexity. The Pushbroom scanner simultaneously records an entire line of pixels, providing both high spatial-spectral resolution. However, its scanning process is inherently slow, which restricts its applicability for real-time or large-scale dynamic acquisition. The Band-sequential scanner adopts a tunable wavelength filter mounted on a conventional camera to capture the full scene at each wavelength band. This approach offers simplicity and low cost, but suffers from limited sensitivity and slow scanning speed, which hinders its use in scenarios requiring rapid acquisition or high-sensitivity acquisition.

Beyond these classical scanning mechanisms, alternative HSI acquisition strategies have been developed. Multi-camera and multi-sensor fusion systems simultaneously capture the scene using cameras having different spectral sensitivities, followed by software-level data fusion to construct a virtual hyperspectral cube. This approach enhances acquisition speed but often requires complex calibration and alignment procedures. Another promising direction is the snapshot compressive imaging (SCI) [7, 19] technique, which reconstructs the 3D hyperspectral cube from a single 2D coded measurement. Among these, Coded Aperture Snapshot Spectral Imaging (CASSI) [2] has been widely studied due to its balance between acquisition efficiency and reconstruction accuracy.

These methods illustrate the trade-offs between acquisition speed, spatial-spectral resolution, hardware complexity, and portability, motivating research into data-driven reconstruction techniques that aim to overcome the physical limitations of the hardware-based systems.

Spatial and Spectral Transformer

The emergence of Vision Transformer (ViT) [11] marked a paradigm shift in computer vision, inspired by the success of Transformer architectures in natural language processing. By dividing an image into fixed-size patches and modeling their long-range dependencies through self-attention, ViT effectively addressed the limitations of convolutional neural networks (CNNs) [9], such as restricted receptive fields and difficulty in capturing global context. Consequently, spatial transformer models rapidly surpassed state-of-the-art CNN-based approaches [18, 22] in several vision benchmarks.

Despite these advances, spatial transformers have notable drawbacks. Their computational and memory complexity scales quadratically with image resolution, leading to large model footprints and a strong dependency on high-end GPUs for training. This limitation has motivated the development of more efficient architectures.

To alleviate these challenges, the spectral transformer was introduced, which shifts the self-attention mechanism from the spatial domain to the spectral (channel) domain. By modeling correlations across spectral channels instead of spatial patches, this approach significantly reduces computational burden, since its computational complexity is almost independent of spatial resolution. As a result, spectral transformers are more lightweight, require fewer computational resources, and often outperform spatial transformers in tasks such as RGB-to-HSI reconstruction, where inter-channel relationships are particularly crucial.

Nevertheless, while spectral transformers have demonstrated superior efficiency and performance in moderate-resolution HSI reconstruction, they still encounter bottlenecks when scaling to ultra-high-resolution imagery. This motivates recent research efforts toward hybrid or dual-domain transformer architectures that can simultaneously capture spatial and spectral dependencies while maintaining scalability.

Method

Network Architectures

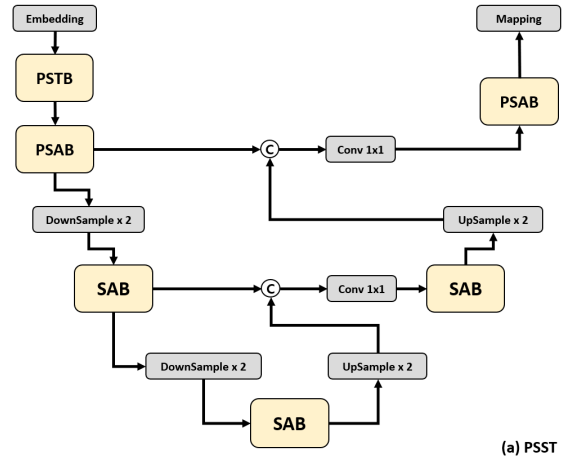


Figure 2. *Entire pipeline of the PSST model*

Figure 2(a) illustrates the overall pipeline of the proposed Patched Input Spatial-Spectral Transformer (PSST).

As shown in Figure 2(a), the PSST takes an RGB image as input, projects it into 31 spectral-band channels, and reconstructs the corresponding HSI output. The overall architecture follows a U-shaped encoder–bottleneck–decoder structure. After spatial processing, the patches are merged back to the original image resolution and passed through two down-sampling operations followed by SAB modules. The decoder adopts a symmetrical design, but no additional spatial transformer is applied. Instead, skip connections are used to propagate positional information, ensuring that global context is preserved during reconstruction.

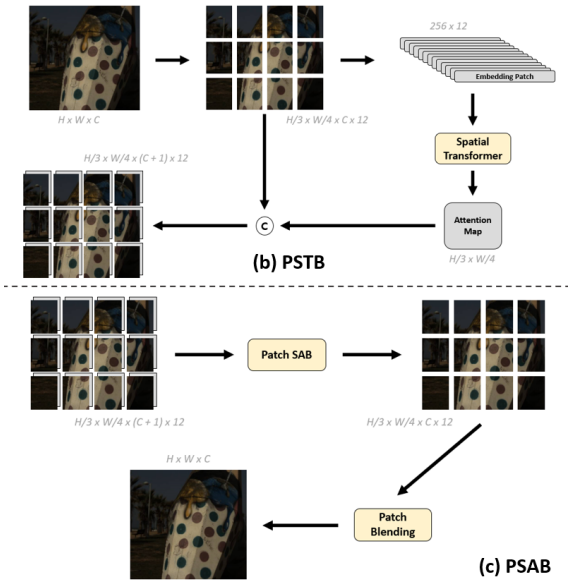


Figure 3. *Pipelines of the PSTB and PSAB model*

Figure 3(b) shows the detailed configuration of the PSTB. The model embeds patches into 256-dimensional vectors and passed through the spatial transformer to generate a unified attention map. This map is concatenated with each patch representation, thereby enabling the subsequent spectral attention to operate with global context.

Figure 3(c) illustrates the Patched Spectral Attention Block (PSAB). Each patch, enriched with global spatial information, is processed by the SAB modules, and the outputs are blended together to reconstruct a seamless image without boundary artifacts.

Finally, to achieve efficient ultra-high-resolution reconstruction, computational complexity must be minimized. Since the feed-forward network (FFN) within SABs dominates the FLOPs, our patch-based design reduces complexity approximately by the square of the patch count, yielding substantial savings in both computation cost and memory usage.

Patched Spectral-wise Attention Block

The Spectral-wise Attention Block (SAB), widely adopted in hyperspectral image reconstruction, consists of two main components: a Feed-Forward Network (FFN) and a Spectral-wise Multi-head Self-Attention (S-MSA). Given an input tensor of size $H \times W \times C$, the computational complexity of the FFN is proportional to $O(HW)$ Eq. (1), while the

$$\mathrm{MHWC^2 + M^2HWC^2 + MHWC^2 = (2M + M^2)\,HWC^2} \quad (1)$$

S-MSA also scales with $O(HW)$ Eq. (2). This quadratic dependency on spatial resolution becomes a severe bottleneck when processing ultra-high-resolution images.

$$\mathrm{2NHWC^2} \quad (2)$$

To mitigate this, we introduce a patch-based strategy. By dividing the input into patches of size $H_p \times W_p$, the effective spatial resolution per operation is reduced to $H/H_p \times W/W_p$, thereby lowering the computational complexity by a factor of $H_p \times W_p$. This significantly alleviates the computational burden, making the reconstruction of ultra-high-resolution hyperspectral images feasible. Furthermore, during the two-stage downsampling process, we restrict the channel expansion factor to only x2. This design choice prevents the computational complexity of both the FFN and S-MSA—which scale quadratically with the number of channels—from growing excessively, while still preserving sufficient representational capacity. In practice, this yields an effective balance between efficiency and reconstruction performance.

Results

Dataset

We evaluated our model using the NTIRE 2022 Spectral Reconstruction Challenge dataset. We split the dataset into training, validation, and test sets with an 18:1:1 split. Each HSI has a spatial resolution of 482 x 512 with 31 spectral channels, uniformly spanning the 400-700nm range at 10nm intervals.

Implementation Details

Training was performed with a batch size of 8 using the Adam optimizer, an initial learning rate of 0.004, and a cosine-annealing learning rate schedule for 300 epochs. All experiments were conducted on a single NVIDIA A6000 GPU, completed in 7 days.

For model optimization, we adopt a hybrid loss function consisting of L1 loss and L2 loss with a weighting ratio of 9:1. Specifically, L1 loss enforces pixel-wise accuracy by penalizing absolute differences between the predicted and ground-truth spectra, while L2 loss emphasizes larger deviations by penalizing squared differences. This combination allows the model to achieve both stable convergence and robustness to outliers.

To evaluate the reconstruction quality, we employ three standard metrics in hyperspectral image restoration. Mean Relative Absolute Error (MRAE) measures the average relative absolute error across spectral bands, which effectively captures reconstruction fidelity with respect to relative deviations. Root Mean Square Error (RMSE) quantifies the absolute reconstruction error as in terms of Euclidean distance, providing a direct measure of pixel-wise differences. Peak Signal-to-Noise Ratio (PSNR) evaluates perceptual quality by comparing the maximum possible signal intensity to the reconstruction error; a higher PSNR indicates better fidelity and less noise. By jointly analyzing these complementary metrics, we ensure a comprehensive assessment of both numerical accuracy and perceptual quality of the reconstructed hyperspectral images.

Main Results

The final training outcomes are illustrated in Figure 4, with quantitative results summarized in Table 1. All quantitative results are obtained by averaging over 50 samples from the NITRE 2022 dataset, and the numerical results for the ultra-high-resolution images are derived from the examples shown in Figure 5.

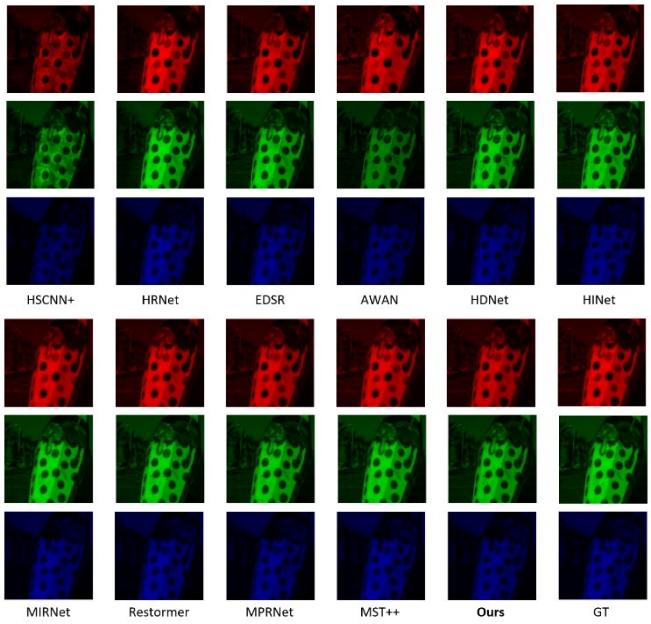


Figure 4. *Comparison of 10 model results, Ground Truth, and PSST result using Scene ARAD 1K 0930*

For low-resolution reconstruction, our method achieved a PSNR of 33.18 dB, which is slightly lower than that of leading methods such as MIRNet [14], Restormer [16], MPRNet [15], and MST++ [21]. However, when scaling to ultra-high-resolution reconstruction, our approach demonstrates a significant advantage. On the same hardware, only a limited set of models were capable of processing images larger than 12 megapixels, yet all failed to

surpass 28 dB PSNR. In contrast, our model achieved 33.02 dB PSNR on 12MP images, establishing a new benchmark in this resolution regime. This indicates that, while our method is less competitive at low resolutions, it exhibits superior scalability and performance as the image size grows.

For real 12MP images, where ground-truth HSIs could not be obtained due to hardware constraints, we reconstructed HSIs using our model and converted them back into RGB images via the RGB-channel quantum efficiency curves. The reconstructed RGB images were then compared against the captured RGB references to compute PSNR. The visual and numerical comparisons, presented in Figure 4, confirm the high fidelity and consistency of our approach in practical ultra-high-resolution scenarios.

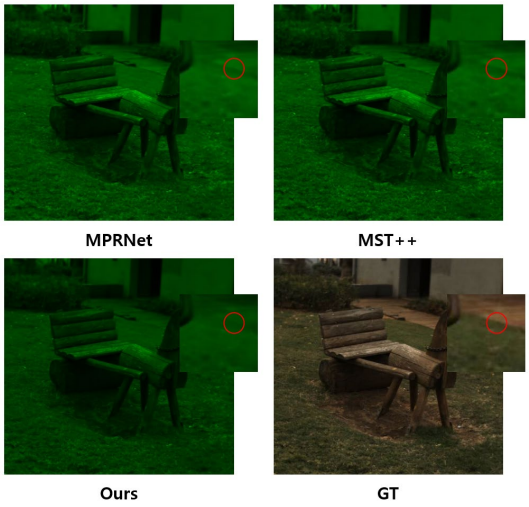


Figure 6. *Qualitative Comparison of Low-Resolution Image Reconstruction*

As shown in Figure 6, the overall reconstruction quality at low resolution exhibits slight differences in fine details when compared with MST++ [21] and MPRNet [15]. Specifically, the regions highlighted by the red circles indicate that the proposed method shows relatively less distinction at the boundary between the pavement blocks and grass. However, from a qualitative perspective, the overall reconstruction does not present significant visual differences, demonstrating that the proposed method achieves sufficiently competitive performance even under low-resolution settings.

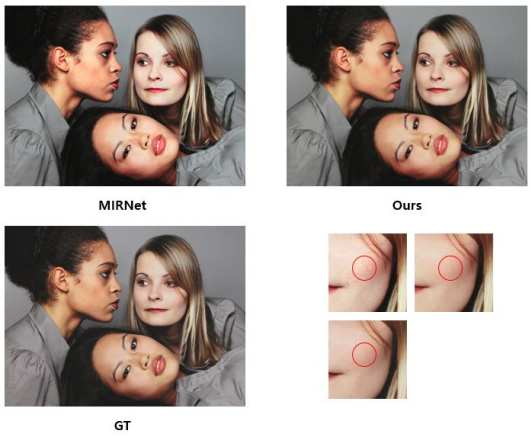


Figure 7. *Qualitative Comparison of High-Resolution Image Reconstruction*

As shown in Figure 7, a clear performance gap can be observed in the ultra-high-resolution image reconstruction results. When compared with MIRNet [14], which achieves the highest PSNR for images above 12 MP, noticeable artifacts appear in skin regions, manifesting as blotchy textures. In contrast, the proposed method produces reconstructions that are visually much closer to the ground truth (GT). These results demonstrate that the proposed approach achieves meaningful improvements not only in quantitative metrics but also in qualitative visual comparisons.

Conclusion

This work presents the first integration of in-model patch partitioning with a dual-stage spatial-spectral transformer specifically designed for ultra-high-resolution RGB-to-HSI reconstruction. Unlike previous approaches, our method preserves global context before spectral processing, thereby ensuring artifact-free boundaries and brightness consistency across patches.

Furthermore, our model demonstrates state-of-the-art performance for 12MP reconstruction, achieving a PSNR of 33.02 dB, and outperforming all existing methods in this resolution range. Importantly, the proposed framework enables end-to-end training on megapixel-scale images on a single GPU while significantly reducing memory consumption.

By addressing the fundamental limitations of prior work—namely, the loss of global context in patch-based strategies and the inability of existing architectures to handle ultra-high-resolution data—we establish a new direction for practical and scalable hyperspectral image reconstruction.

Reference

[1] Asgeir Bjorgan and Lise Lyngsnes Randeberg, “Towards real-time medical diagnostics using hyperspectral imaging technology”, In ECBO, 2015.

[2] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady, “Single disperser design for coded aperture snapshot spectral imaging”, In Applied Optics”, 2008.

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, “Enhanced deep residual networks for single image super-resolution”, In CVPRW, 2017.

[4] Bing Lu, Phuong D. Dao, Jiangui Liu, Yuhong He and Jiali Shang, “Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture”, In Remote Sens, 2020.

[5] Farid Melgani and Lorenzo Bruzzone, “Classification of hyperspectral remote sensing images with support vector machines”, In IEEE Transactions on Geoscience and Remote Sensing, 2004.

[6] George Bebis and Michael Georgiopoulos, “Feed-forward neural networks”, In IEEE Potentials, 1994.

[7] Hao Du, Xin Tong, Xun Cao, and Stephen Lin, “A prismbased system for multispectral video acquisition”, In ICCV, 2009.

[8] Jiaojiao Li, Chaoxiong Wu, Rui Song, Yunsong Li, and Fei Liu, “Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from rgb images”, In CVPRW, 2020.

[9] Keiron O'Shea and Ryan Nash, “An Introduction to Convolutional Neural Networks”, In Neural and Evolutionary Computing, 2015.

[10] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Cheng peng Chen, “Hinet: Half instance normalization network for image restoration”, In CVPRW, 2021.

[11] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng and Shuicheng Yan, “Tokens-to-Token ViT: Training Vision Transformers From Scratch on ImageNet”, In ICCV, 2021.

[12] M. Borengasser, W. S. Hungate, and R. Watkins, “Hyperspectral remote sensing: principles and applications”, In CRC press, 2007.

[13] Prachi Singh, Prem Chandra Pandey, George P. Petropoulos, Andrew Pavlides, Prashant K.Srivastava, Nikos Koutsias, Khidir Abdala Kwal Deng and Yangson Bao, “8-Hyperspectral remote sensing in precision argriculture: present status, challenges, and future trends”, In Earth Observation, 2020.

[14] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, “Learning enriched features for real image restoration and enhancement”, In ECCV, 2020.

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao, “Multi-stage progressive image restoration”, In CVPR, 2021.

[16] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, “Restormer: Efficient transformer for high-resolution image restoration”, In CVPR, 2022.

[17] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, “Restormer: Efficient transformer for high-resolution image restoration”, In CVPR, 2022.

[18] 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 CVPR, 2022.

[19] Xun Cao, Tao Yue, Xing Lin, Stephen Lin, Xin Yuan, Qionghai Dai, Lawrence Carin, and David J. Brady, “Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world”, In IEEE Signal Processing Magazine, 2016.

[20] 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 CVPR, 2022.

[21] Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte and Luc Van Gool, “MST++: Multi-stage Spectral-wise Transformer for Efficient Spectral Reconstruction”, In CVPR, 2023.

[22] Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin, “Hierarchical regression network for spectral reconstruction from rgb images”, In CVPRW, 2020.

[23] 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 ICCV, 2021.

Method	NTIRE 2022 HIS Dataset			Ultra-High-Resolution Image		
	MRAE	RMSE	PNSR	Max Image	PSNR	RMSE
HSCNN+	0.3814	0.0588	26.36	< 3MP	X	X
HRNet	0.3476	0.0550	26.89	< 5MP	21.87(3MP)	0.0806
EDSR	0.3277	0.0437	28.29	< 14MP	16.09(12MP)	0.1568
AWAN	0.2500	0.0367	31.22	< 1.2MP	X	X
HDNet	0.2048	0.0317	32.13	< 14MP	23.69(12MP)	0.0654
HINet	0.2032	0.0303	32.51	< 13MP	26.84(12MP)	0.0455
MIRNet	0.1890	0.0274	33.29	**< 15MP**	27.92(12MP)	0.0402
Restormer	0.1833	0.0274	33.40	< 3.5MP	26.86(3MP)	0.0454
MPRNet	0.1817	0.0270	33.50	< 9MP	27.76(7.8MP)	0.0409
MST++	**0.1645**	**0.0248**	**34.32**	< 5MP	30.48(3MP)	0.0299
Ours	0.1875	0.0284	33.18	< 12.5MP	**33.03(12MP)**	**0.0223**

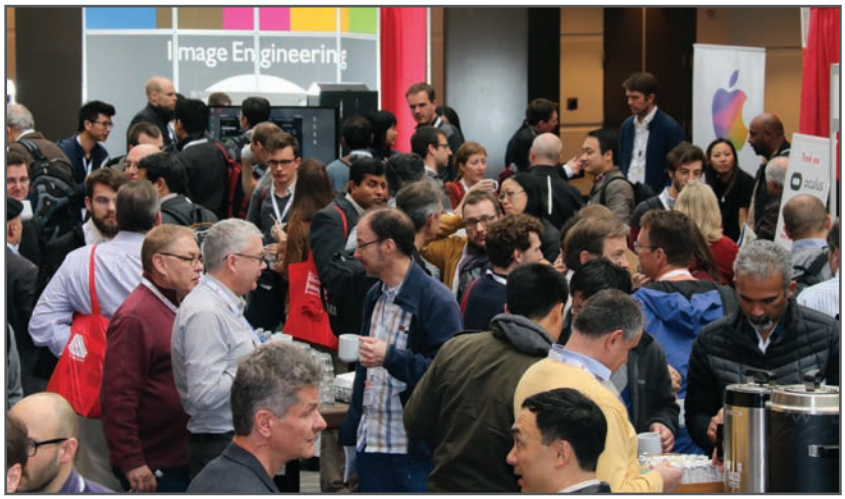
Table 1. *Comparison of NTIRE 2022 HSI Dataset valid results and max image size with results in Ultra-High-Resolution Images*

Figure 4. *Comparison of 7 model results, Ground Truth, and PSST results using Scene 3 Woman image*

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org