Enhanced Pansharpening via Quaternion Spatial-Spectral Interactions

Supplementary Material

1. Quaternion Representation

To adapt to quaternion operators, we need to convert images into quaternion representations. As shown in Figure 1, we use two representation methods: channel quaternion representation (CQR) and spatial quaternion representation (SQR). CQR transforms $F \in \mathbb{R}^{H\times W\times C}$ into $Q^c \in \mathbb{H}^{H\times W\times C/4}$, while SQR transforms $F \in \mathbb{R}^{H\times W\times C}$ into $Q^s \in \mathbb{H}^{H/2\times W/2\times C}$. When converting quaternion representations back into image features, we use ICQR and ISQR to denote the corresponding transformations, as illustrated in Figure 2.

The use of SQR enhances the ability of the network to preserve local structures. However, SQR is a window-based representation, and its lack of cross-window connections limits its modeling capability when used alone. To further enhance the network's modeling capacity, inspired by [2, 4], we introduce shifted window quaternion representation (SWQR) into SQR. As shown in the Figure 1, SWQR is a shifted window partitioning method that alternates between two partitioning configurations across consecutive SQRs. SWQR is also demonstrated to be effective in experiments, as shown in Table 2. ISWQR denotes the inverse operation of SWQR, which converts quaternion representations back into image features while shifting the window back to its original position, as shown in Figure 2.

2. More Ablation Experiments

In this section, we provide additional ablation experiments to further validate the effectiveness of our method.

First, we conduct further ablation studies on the quaternion spatial-spectral interaction (QSSI). QSSI consists of two parts: quaternion-based global fusion (QGF) and channel-aware quaternion spatial feature injection (CQSFI). We perform separate ablations on these two parts, as shown in Table 1. When we replace QGF with CQSFI, the model performance degrades, indicating that the enhanced correlation between the two obtained through the quaternion structure is crucial for effective fusion. Similarly, when CQSFI is removed, the evaluation metrics also decline. This demonstrates that explicitly interacting across spatial and channel dimensions is essential for the fusion process.

Table 1. Ablation studies about QGF and CQSFI on the WorldView-II. The best values are highlighted in **bold**.

Configuration	QGF	CQSFI	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
I	✗	✓	42.1267	0.9689	0.0226	0.9026
II	✓	✗	42.1935	0.9703	0.0215	0.8996
Ours	✓	✓	**42.4846**	**0.9738**	**0.0208**	**0.8554**

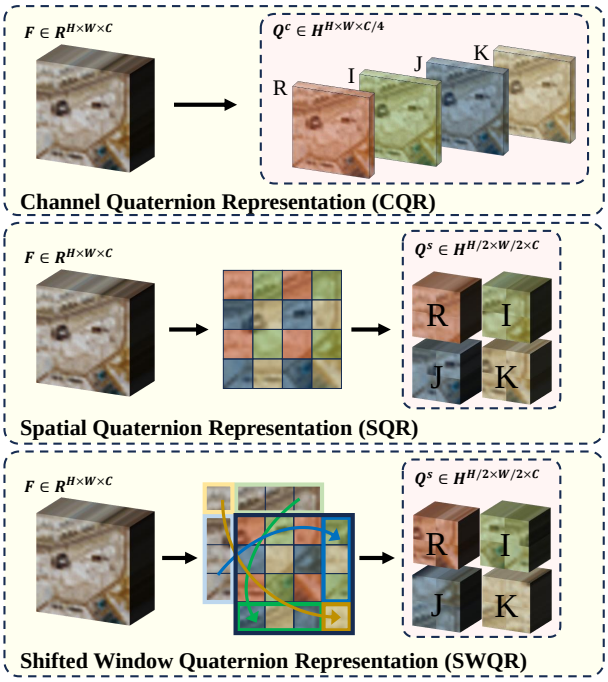

Figure 1. Illustration of quaternion representations. We design Channel Quaternion Representation (CQR), Spatial Quaternion Representation (SQR), and Shifted Window Quaternion Representation (SWQR). In the figure, R represents the real part of the quaternion, while I, J, and K represent the three imaginary parts of the quaternion respectively.

We also attempt to remove the shifted window design in the quaternion local spatial structure awareness (QLSA) branch, as shown in Table 2. It can be observed that removing the shifted window negatively impacts the model's performance. This is likely due to the lack of cross-window connections, which limits the network's modeling capability. Therefore, our method employs shifted windows to enhance the spatial quaternion representation.

Table 2. Ablation results on the WorldViewII without (w/o) and with (w) SWQR.

Configuration	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
w/o SWQR	41.9975	0.9654	0.0223	0.9267
w SWQR	42.4846	0.9738	0.0208	0.8554

3. Additional Dataset Descriptions

For visible and infrared image fusion, we use three publicly available datasets in our experiments: M3FD [3], RoadScene [9], and TNO [8]. The M3FD dataset consists of 4200 infrared and visible image pairs, with 3900 pairs used

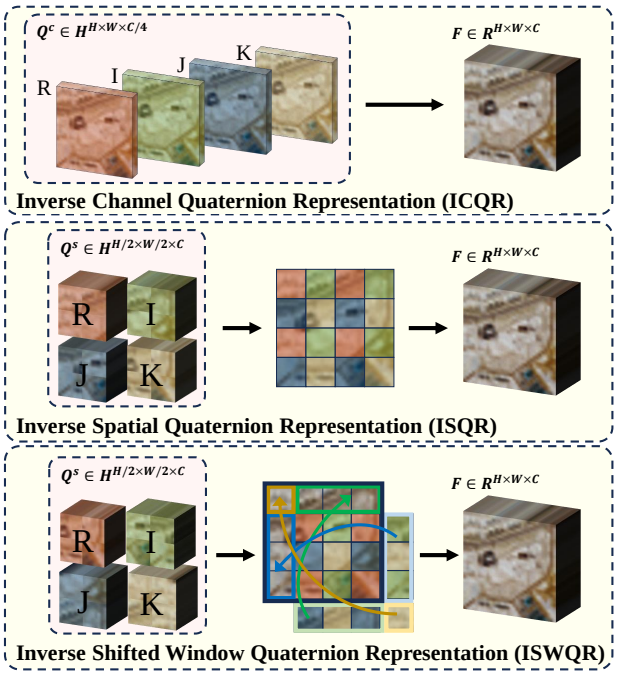

Figure 2. Illustration of inverse quaternion representations. We design Inverse Channel Quaternion Representation (ICQR), Inverse Spatial Quaternion Representation (ISQR), and Inverse Shifted Window Quaternion Representation (ISWQR). In the figure, R represents the real part of the quaternion, while I, J, and K represent the three imaginary parts of the quaternion respectively.

for training and 300 pairs for testing. To evaluate the generalizability of our method, we train our model on the M3FD dataset and test it on the RoadScene and TNO datasets.

For depth image super-resolution (SR), we use three datasets: NYU v2 [7], Middlebury [6], and Lu [5]. The NYU v2 dataset contains 1449 RGB-D image pairs, the Middlebury dataset contains 30 RGB-D image pairs, and the Lu dataset contains 6 RGB-D image pairs. We train our model on the first 1000 RGB-D image pairs from the NYU v2 dataset and evaluate the trained model on the remaining 449 pairs. To generate low-resolution depth maps, we follow the experimental protocol in [1], applying bicubic downsampling at different scales (×4, ×8, ×16). We directly test the trained model on the NYU v2 dataset as well as the additional Middlebury and Lu datasets.

4. Implementation Details

In our experiments, all deep learning models are implemented using PyTorch and trained on an NVIDIA GeForce GTX 3090 GPU. For each set, the multispectral (MS) images are cropped into patches with the size of 32×32, and the corresponding panchromatic (PAN) images are of size 128×128. During the training phase, these networks utilize the Adam optimizer with a learning rate of 1×10^{-4}. After reaching 200 epochs, the learning rate is halved. We employ common evaluation metrics, including PSNR, SSIM, SAM, and ERGAS. Additionally, we utilize three widely-used no-reference IQA metrics for real-world full-resolution scenes: D_λ,D_S and QNR.

5. Additional Visualization Results

In the manuscript, due to the limitation of space, we only present visual comparisons for a subset of methods and datasets. In this section, we provide the visualization results of all methods in each dataset. Figure 3 shows the visual results of all methods on the WorldView-II dataset, while Figure 4 presents the results on the GaoFen-2 dataset, and Figure 5 illustrates the results on the WorldView-III dataset.

6. Limitation and Discussion

While the proposed QuatPanNet achieves state-of-the-art performance on multiple benchmark datasets, certain limitations remain to be addressed. First, although our method effectively enhances spectral and spatial fidelity in pansharpening tasks, its performance is validated primarily on standard datasets under controlled conditions. Real-world satellite imagery often involves various complexities, such as atmospheric interference, sensor noise, and dynamic scene variations, which may affect the model's robustness. Extending the framework to address these challenges and validating its effectiveness on a broader range of real-world datasets is an important direction for future work.

Furthermore, the quaternion representations and interactions in QuatPanNet are specifically tailored for pansharpening tasks. While these techniques have demonstrated significant potential, their adaptability to other image fusion and restoration tasks, such as medical image super-resolution, remains unexplored. Investigating the generalizability of quaternion-based interaction mechanisms across diverse image-processing tasks could further expand the impact of this work.

Addressing these limitations through future research can further improve the applicability of the proposed QuatPanNet and pave the way for its deployment in a wide range of real-world scenarios.

7. Broader Impacts

The development of advanced pansharpening techniques, like the proposed quaternion-based spatial-spectral interaction framework QuatPanNet, has substantial implications for the fields of remote sensing, environmental monitoring, and urban planning. By leveraging the unique representational capabilities of quaternions, this method enhances the fidelity of spectral data and the richness of spatial details, significantly improving the accuracy of high-resolution multispectral (HRMS) images. These advance-

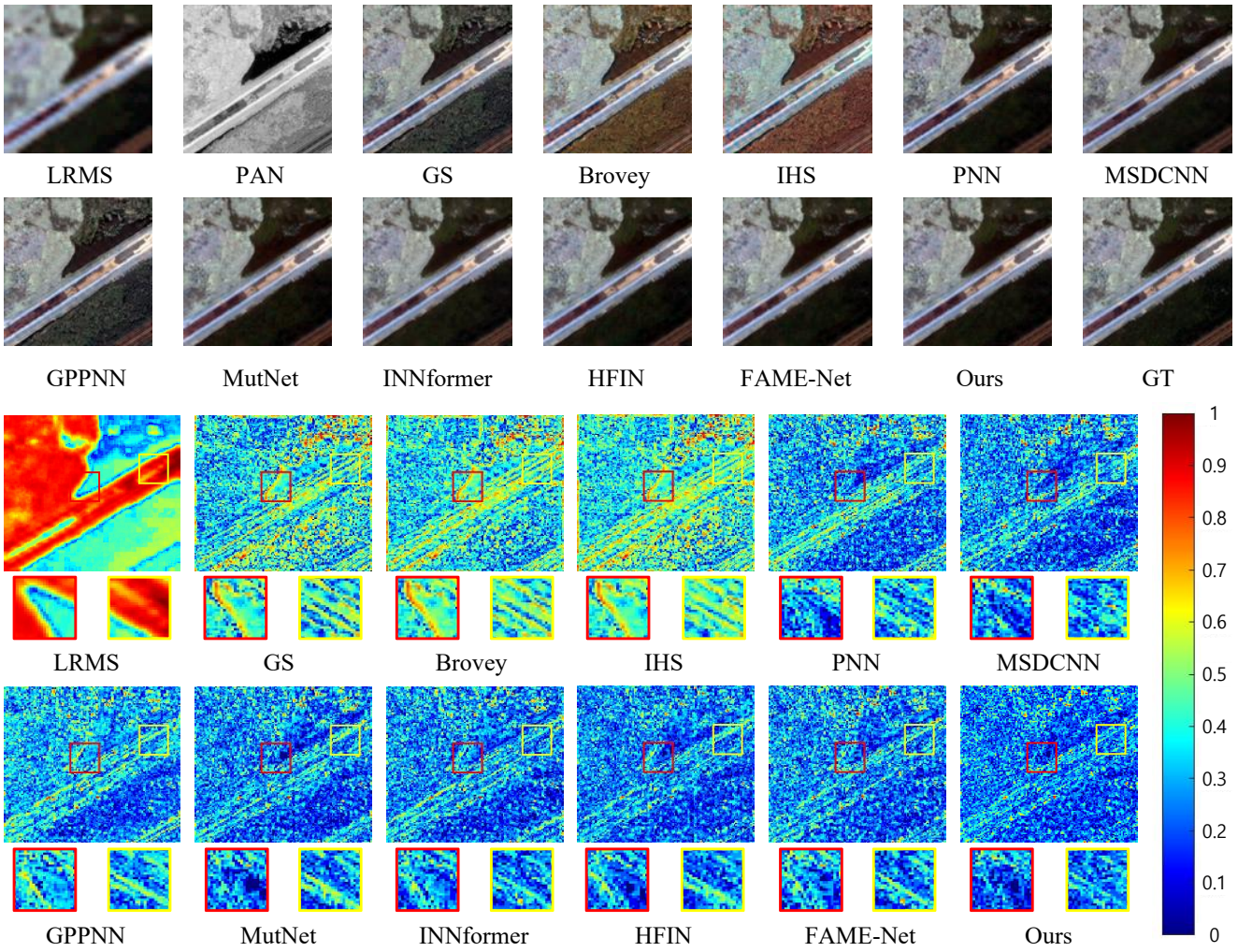

Figure 3. The visual comparisons between other pan-sharpening methods and our method on WorldView-II.

ments can lead to better resource management, disaster response, and climate monitoring by enabling more detailed and precise satellite image analysis. Furthermore, the compact nature of quaternion operations allows for a more efficient processing framework, potentially reducing the computational resources required for satellite image fusion tasks.

However, there are potential societal and ethical concerns associated with this technology. One key issue is the risk of over-reliance on the generated pan-sharpened images in critical applications such as environmental policy-making. While the improved spectral and spatial fidelity is beneficial, the generation of highly realistic but not ground-truth-verified images could lead to misinterpretations if improperly used. For instance, the integration of enhanced spatial-spectral features might inadvertently introduce artifacts that could bias analytical models or decision-making processes.

It is worth noting that the positive societal impacts of QuatPanNet far outweigh its potential issues. We advocate for the responsible use of this technology and its derived applications, ensuring that public and individual interests are not compromised.

References

[1] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021. 2

[2] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 1

[3] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 1

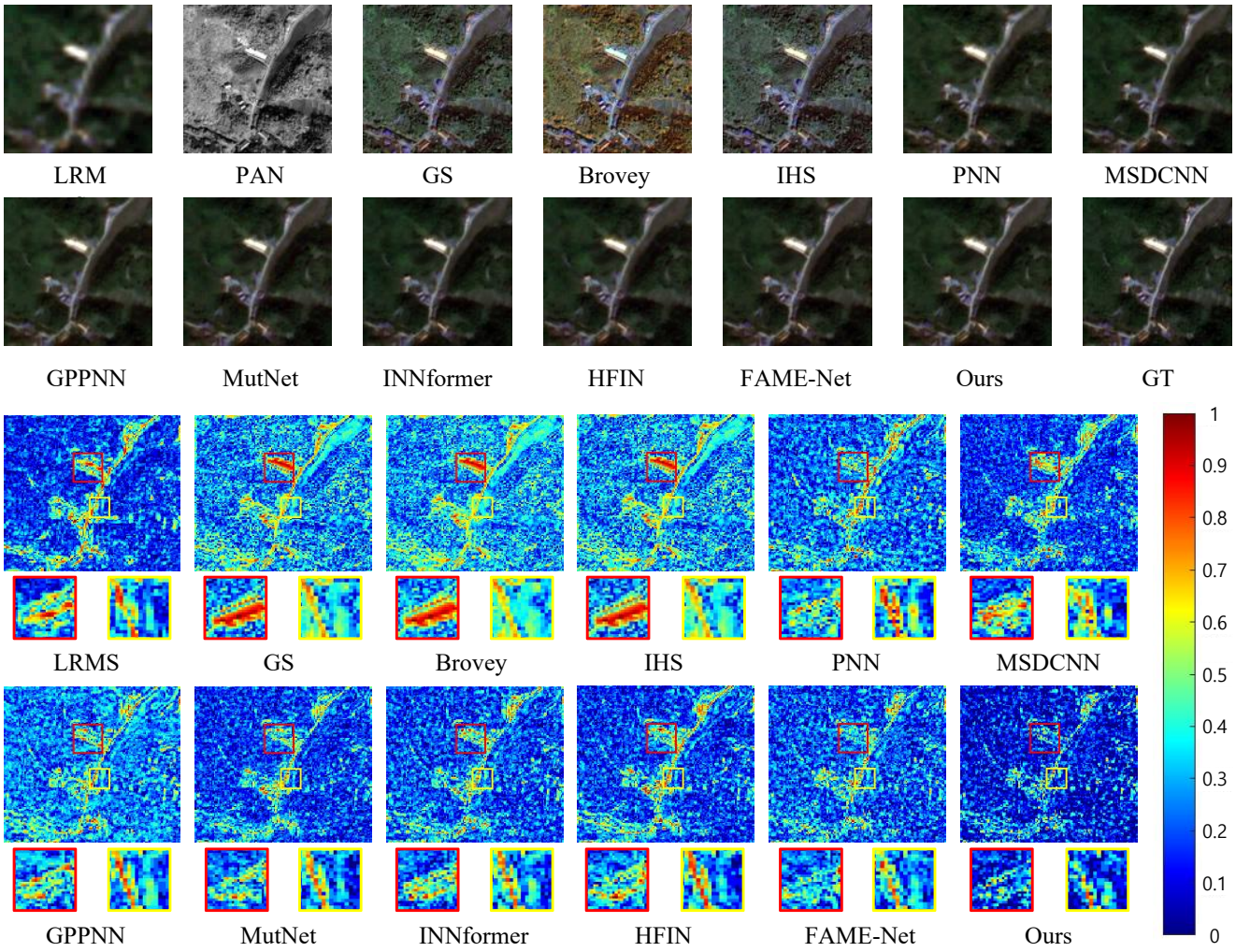

Figure 4. The visual comparisons between other pan-sharpening methods and our method on the GaoFen2.

[4] 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

[5] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3390–3397, 2014. 2

[6] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2

[8] A Toet. The tno multiband image data collection. data brief 15, 249–251 (2017). 1

[9] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020. 1

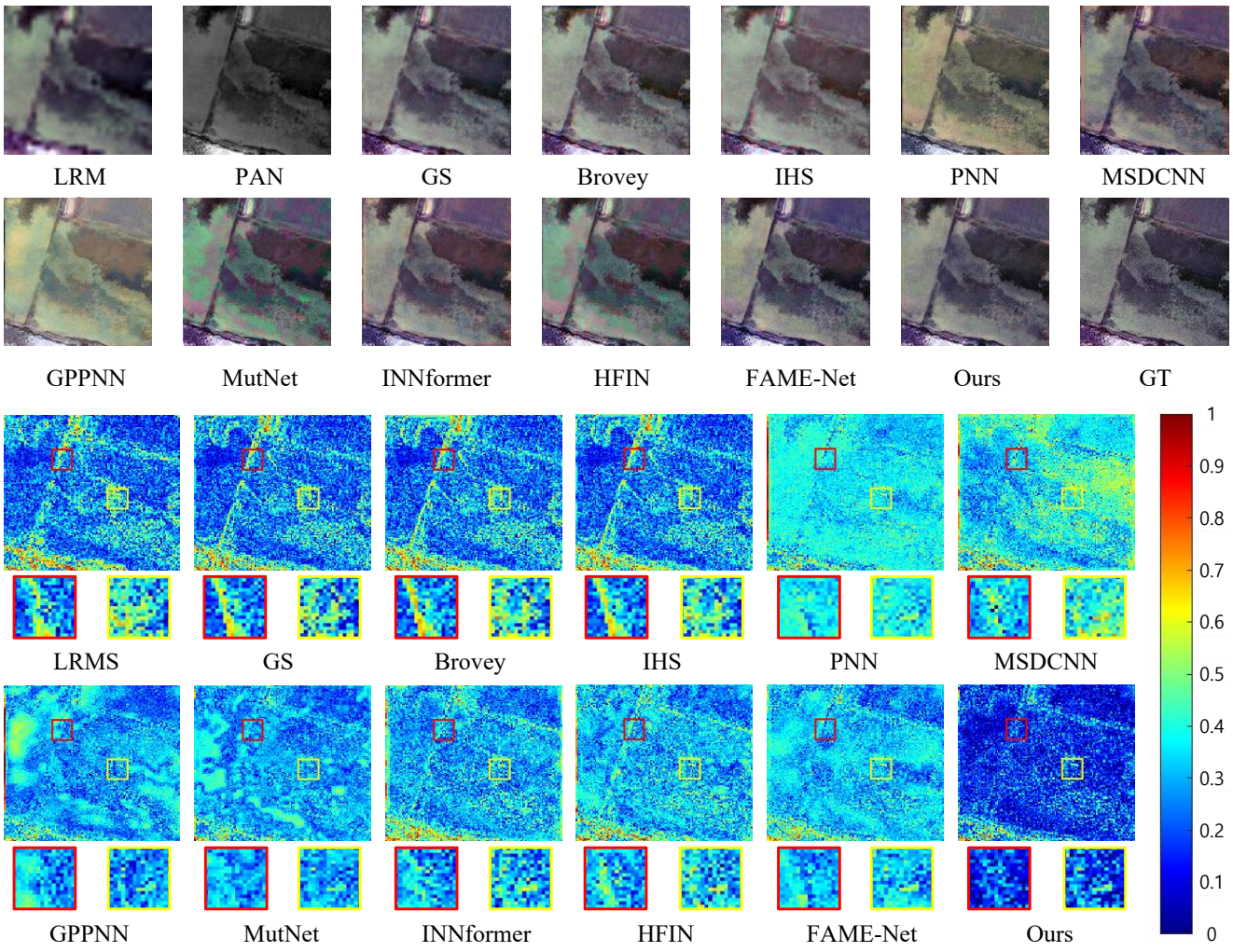

Figure 5. The visual comparisons between other pan-sharpening methods and our method on WorldView III.