

Supplementary Materials: Hyperspectral and Multispectral Image Fusion in Bidirectional State Space Model

Anonymous Author(s)

1 BACKGROUND OF VISION MAMBA SERIES

In this section, we provide more details about the vision Mamba series evolving recent Vim [14], VMamba [6], LocalMamba [4], EfficientVMamba [9], RS-Mamba [12] and PanMamba [3].

CNN-based methods are adept at extracting local features, and compromise resolution, while transformer-based methods offer global perception but increase the computational burden. To address the challenges of Mamba’s unidirectional modeling and lack of positional awareness, Zhu et al. [14] proposed a universal visual backbone featuring Bidirectional Mamba Blocks (Vim), which marks the image sequences with position embeddings and compresses the visual representation with bidirectional state space models. Liu et al. [6] proposed VMamba based on the state space model, achieving linear complexity without sacrificing the global receptive field. However, the traditional Vision Mamba works directly flattening spatial tokens overlooks preserving local 2D dependencies. To address this, LocalMamba [4] introduces a local scanning strategy that divides the image into different windows, effectively capturing local dependencies while maintaining a global perspective. EfficientVMamba [9] integrates an atrous-based selective scan approach by efficient skip sampling to ensure full global receptive field coverage while minimizing computational load. However, cropping large images into small blocks can lead to a significant loss of contextual information. To capture the global context of remote sensing images with linear complexity, RS-Mamba [12] introduces an omnidirectional selective scan module, which can globally model the context of the image in multiple directions, capturing large spatial features from various directions. PanMamba [3] first utilizes the Mamba method for the pansharpening task. The proposed channel-swapping Mamba and cross-modal Mamba achieve efficient cross-modal information exchange and good fusion results. *However, it fails to realize the issue of state information loss and does not tailor state space representation to exploit the spatial and spectral domain, leading to suboptimal fusion performance.*

2 MORE DETAILS OF DATASETS

This section contains more details of the used datasets. WV3 dataset contains 9714/1080 samples for training and validation. Each sample consists of a PAN/LRMS/GT image pair of size $64 \times 64 \times 1$, $16 \times 16 \times 8$, and $64 \times 64 \times 8$, respectively. PAN image has a spatial resolution of 0.3m, whereas the LRMS image has a spatial resolution of 1.2m. GF2 dataset contains 19809/2201 samples for training and validation. Each sample consists of a PAN/LRMS/GT image pair of sizes $64 \times 64 \times 1$, $16 \times 16 \times 4$, and $64 \times 64 \times 4$, respectively. PAN images have a spatial resolution of 0.8m, while LRMS images have a spatial resolution of 3.2m. To evaluate the performance, we perform the reduced-resolution and full-resolution experiments to compute the reference and non-reference metrics, respectively.

The CAVE dataset consists of 31 indoor images captured under controlled illumination with a size of $512 \times 512 \times 31$, covering the

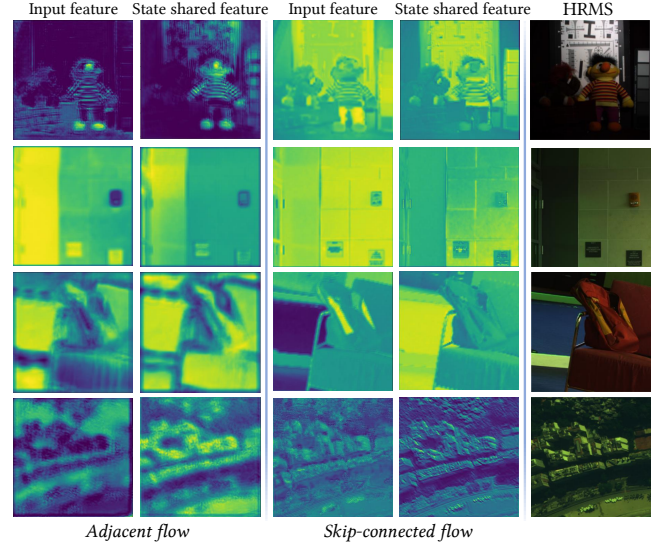


Figure 1: Visualization of features in adjacent flow and skip-connected flow in the proposed state sharing technique.

spectrum from 400nm to 700nm. We chose 20 samples to randomly extract half-sized overlapping patches and split them into training/validation pairs. The training/validation sets consist of 648/72 RGB image patches with a size of $128 \times 128 \times 3$. Both sets also include corresponding LR-HSI image patches of size $16 \times 16 \times 31$. The Harvard dataset contains 50 indoor and outdoor images recorded under daylight illumination, including 31 spectral bands, covering the visible spectrum from 420nm to 720nm. We use the same approach as for the CAVE dataset to get HR-MSIs and LR-MSIs. We chose 20 samples to randomly extract half-sized overlapping patches and split them into training/validation pairs. The training/validation sets consist of 576/144 RGB image patches with a size of $128 \times 128 \times 3$. Both sets also include corresponding LR-HSI image patches of size $16 \times 16 \times 31$.

3 STATE SHARING TECHNIQUE VISULIZATION

In this section, we provide more visual results for the proposed state sharing technique. The illustrations of the input feature and state shared feature are shown in Fig. 1. For the adjacent flow, in the first row, the strips of the toy and the contours of the objects are lighter and the network focuses on more semantic information. In the last row, the buildings are highlighted. For the skip-connected flow, we can see the state shared feature obtains more low-level information, e.g., the notes on the wall in the second row are more clear and the backpack in the third row is more likely to HRMS.

Thus, we can conclude that adjacent flow helps to learn more semantic information and skip-connected flow maintains more low-level details and helps fusion.

4 NETWORK CONFIGURATIONS & IMPLEMENTATION DETAILS

The proposed LE-Mamba configurations and training details on WV3, GF2, CAVE($\times 8$), and Harvard($\times 8$) are listed in Tab. 1. The compared methods are trained using the same AdamW [7] optimizer and training epochs to convergence for a fair comparison.

Table 1: Network and training configuration on different datasets.

Configurations	WV3&GF2	CAVE($\times 8$)&Harvard($\times 8$)
Encoder blocks	[2,1,1]	[4,3,2]
Middle blocks	1	2
Decoder blocks	[1,1,2]	[2,3,4]
Basic channel	32	32
Window size	8	8
Train epochs	800	2000
Local SSM state (N)	16	16
Global SSM state (N)	32	32
Basic lr	$1e-3 \rightarrow 1e-4$	$1e-3 \rightarrow 1e-4$
Lr scheduler	Cosine	Cosine
Batchsize	96	18
Weight decay	$1e-6$	$1e-6$

5 NETWORK GENERALIZATION

To validate the network’s generalization ability, we choose the WV2 dataset, which comprises photos of various geographic areas captured using the same equipment. The generalization performance is provided in Tab. 2 compared with previous SOTA DL-based methods. It can be observed that our architecture has a satisfactory generalization ability and outperforms most previous DL-based models.

Table 2: Generalization ability of DL-based methods. The best results are in red and the second best results are in blue.

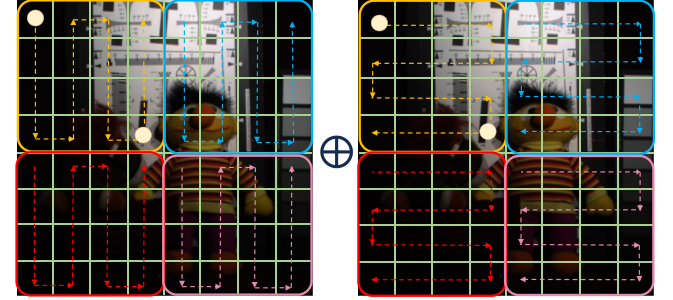
Methods	WV2 Reduced Resolution (RR): Avg \pm std			
	SAM	ERGAS	Q2n	SCC
DiCNN [2]	6.92 \pm 0.79	6.25 \pm 0.57	0.721 \pm 0.075	0.855 \pm 0.029
FusionNet [1]	6.43 \pm 0.86	5.14 \pm 0.52	0.796 \pm 0.074	0.875 \pm 0.013
LAGNet [5]	6.95 \pm 0.47	5.33 \pm 0.32	0.805 \pm 0.084	0.913 \pm 0.010
Invformer [13]	6.43 \pm 0.62	4.68 \pm 0.47	0.810 \pm 0.089	0.914 \pm 0.011
DCFNet [11]	5.62\pm0.60	4.49\pm0.38	0.829 \pm 0.082	0.915\pm0.008
HMPNet [10]	6.10 \pm 0.57	4.67 \pm 0.46	0.830\pm0.081	0.905 \pm 0.010
PanDiff [8]	5.66 \pm 0.72	5.11 \pm 0.39	0.780 \pm 0.074	0.886 \pm 0.018
Proposed	5.62\pm0.55	4.34\pm0.32	0.841\pm0.080	0.914\pm0.012

6 MORE EXPERIMENTAL ILLUSTRATIONS

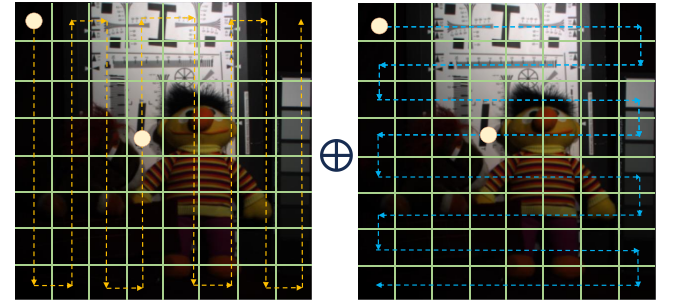
In this section, we provide more visual results in Figs. 3, 4, 5 on the used datasets. It is clear that the proposed method has fewer errors compared with GT on the reduced-resolution test set and more

fidelity with respective input LRMS and PAN on the full-resolution test set.

7 LOCAL-ENHANCED VISUAL MAMBA BLOCK ENABLES AN EFFECTIVE 2D LOCAL SCAN



(a) Local-enhanced 2D scan



(b) VMamba 2D scan

Figure 2: Local-enhanced visual Mamba block in (a) enables an effective 2D local scan. Fig. (a): The dotted line with the arrow means 2D scan and the box indicates the partitioned local window. Fig (b): VMamba 2D scan adopted in [6].

Recalling the Method section, the proposed local-enhanced visual Mamba block introduces a window partition operation (Eq. (8)). This local operation shortens the 2D image scan distance by bringing spatially distant patches closer together. For instance, within the orange window, the left upper corner patch and the right bottom patch (represented in yellow circles) are scanned at a closer distance compared to the original VMamba 2D scan. This localized approach aligns better with the inherent 2D structure of images, leading to improved performance.

8 ACKNOWLEDGEMENT

Thanks Yujie Liang for carefully revising Sect. 2.

REFERENCES

- [1] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2020. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Trans. Geosci. Remote Sens.* 59, 8 (2020), 6995–7010.
- [2] Lin He, Yizhou Rao, Jun Li, Jocelyn Chanussot, Antonio Plaza, Jiawei Zhu, and Bo Li. 2019. Pansharpening via detail injection based convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 12, 4 (2019), 1188–1204.

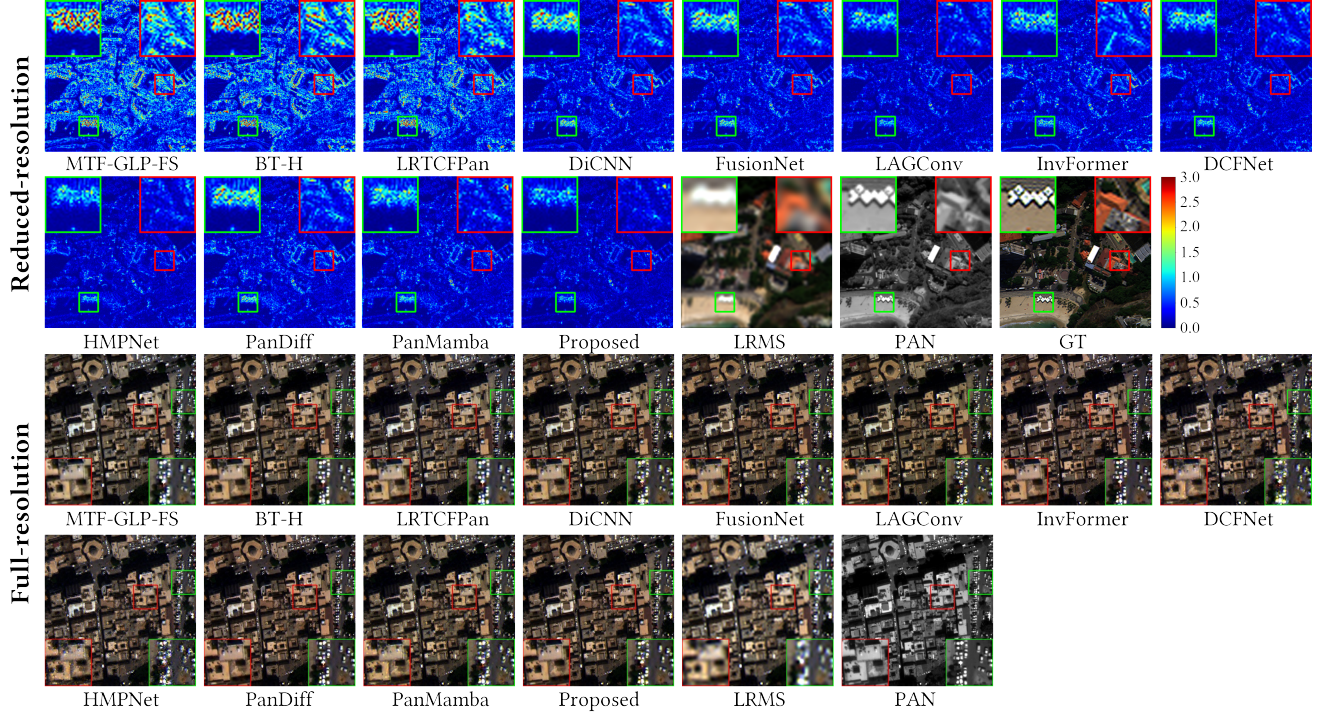


Figure 3: Visualization of fused HRMS on WV3 reduced-resolution and full-resolution test set. The first two rows are error maps on the reduced-resolution test set compared with GT. The last two rows are visual results of the full-resolution test set.

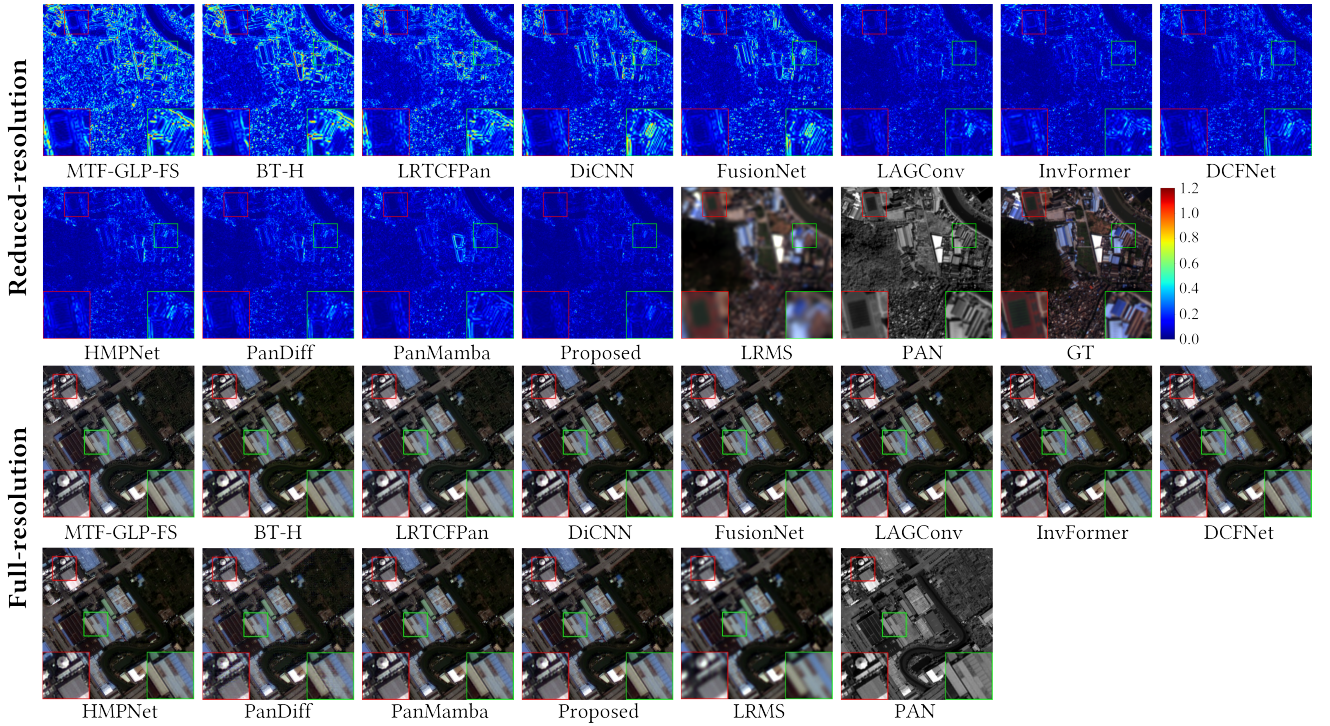


Figure 4: Visualization of fused HRMS on GF2 reduced-resolution and full-resolution test set. The first two rows are error maps on the reduced-resolution test set compared with GT. The last two rows are visual results of the full-resolution test set.

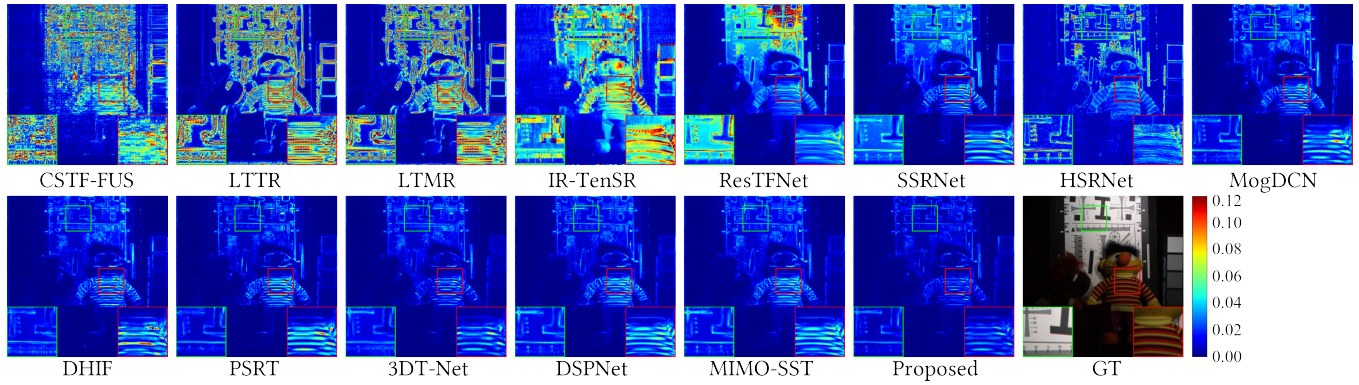


Figure 5: Visualization of fused HRHS on CAVE($\times 8$) reduced-resolution and full-resolution test set.

- [3] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. 2024. Pan-Mamba: Effective pan-sharpening with State Space Model. *arXiv preprint arXiv:2402.12192* (2024).
- [4] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2024. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338* (2024).
- [5] Zi-Rong Jin, Tian-Jing Zhang, Tai-Xiang Jiang, Gemine Vivone, and Liang-Jian Deng. 2022. LAGConv: Local-Context Adaptive Convolution Kernels with Global Harmonic Bias for Pansharpening. *AAAI* 36, 1 (Jun. 2022), 1113–1121.
- [6] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. VMamba: Visual State Space Model. *arXiv preprint arXiv:2401.10166* (2024).
- [7] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [8] Qingyan Meng, Wenxu Shi, Sijia Li, and Linlin Zhang. 2023. PanDiff: A Novel Pansharpening Method Based on Denoising Diffusion Probabilistic Model. *IEEE Trans. Geosci. Remote Sens.* 61 (2023), 1–17.
- [9] Xiaohuan Pei, Tao Huang, and Chang Xu. 2024. Efficientvmamba: Atrous selective scan for light weight visual mamba. *arXiv preprint arXiv:2403.09977* (2024).
- [10] Xin Tian, Kun Li, Wei Zhang, Zhongyuan Wang, and Jiayi Ma. 2023. Interpretable Model-Driven Deep Network for Hyperspectral, Multispectral, and Panchromatic Image Fusion. *IEEE Trans. Neural Netw. Learn. Syst.* (2023), 1–14.
- [11] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. 2021. Dynamic cross feature fusion for remote sensing pansharpening. In *ICCV*. 14687–14696.
- [12] Sijie Zhao, Hao Chen, Xuiliang Zhang, Pengfeng Xiao, Lei Bai, and Wanli Ouyang. 2024. RS-Mamba for Large Remote Sensing Image Dense Prediction. *arXiv preprint arXiv:2404.02668* (2024).
- [13] Man Zhou, Xueyang Fu, Jie Huang, Feng Zhao, Aiping Liu, and Rujing Wang. 2022. Effective Pan-Sharpener With Transformer and Invertible Neural Network. *IEEE Trans. Geosci. Remote Sens.* 60 (2022), 1–15.
- [14] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *arXiv preprint arXiv:2401.09417* (2024).