EMR-Diff: Edge-aware Multimodal Residual Diffusion Model for Hyperspectral Image Super-resolution

Tao Zhang[1,4], Shengtao Yao[1], Rong Zeng[1], Zunjie Zhu[1,4,*], Bolun Zheng[1,4],
Yaoqi Sun[2], Ying Fu[3], Chenggang Yan[1,4]

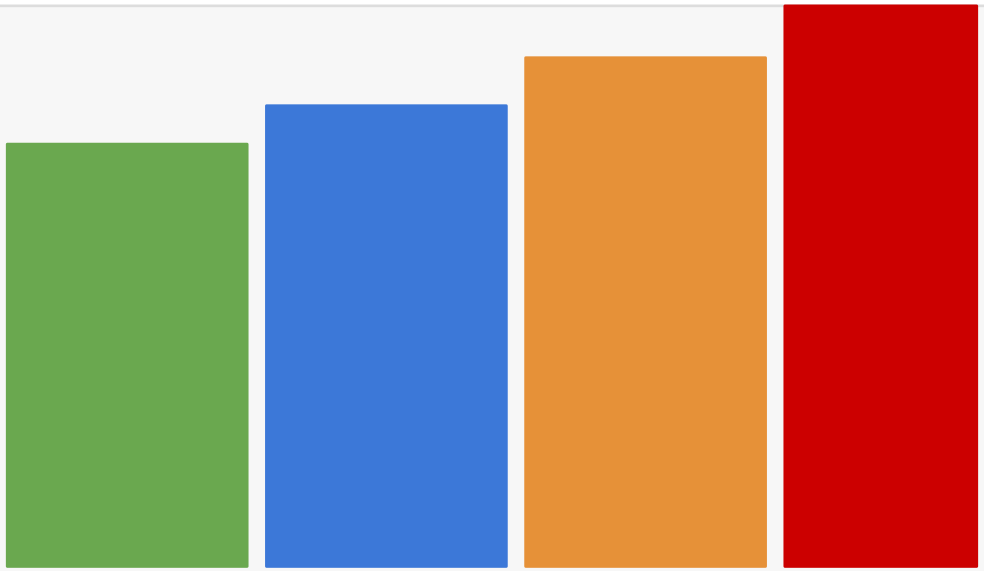
[1]Hangzhou Dianzi University, [2]Lishui University, [3]Beijing Institute of Technology
[4]Zhejiang Provincial Key Laboratory of Low Altitude Ubiquitous Networking Technology

1. Introduction

Hyperspectral Image (HSI)

- Continuously samples hundreds of spectral bands.
- Useful in environmental monitoring, geology, military and food safety.
- Reveals fine material-specific spectral features.

Fine spectral information



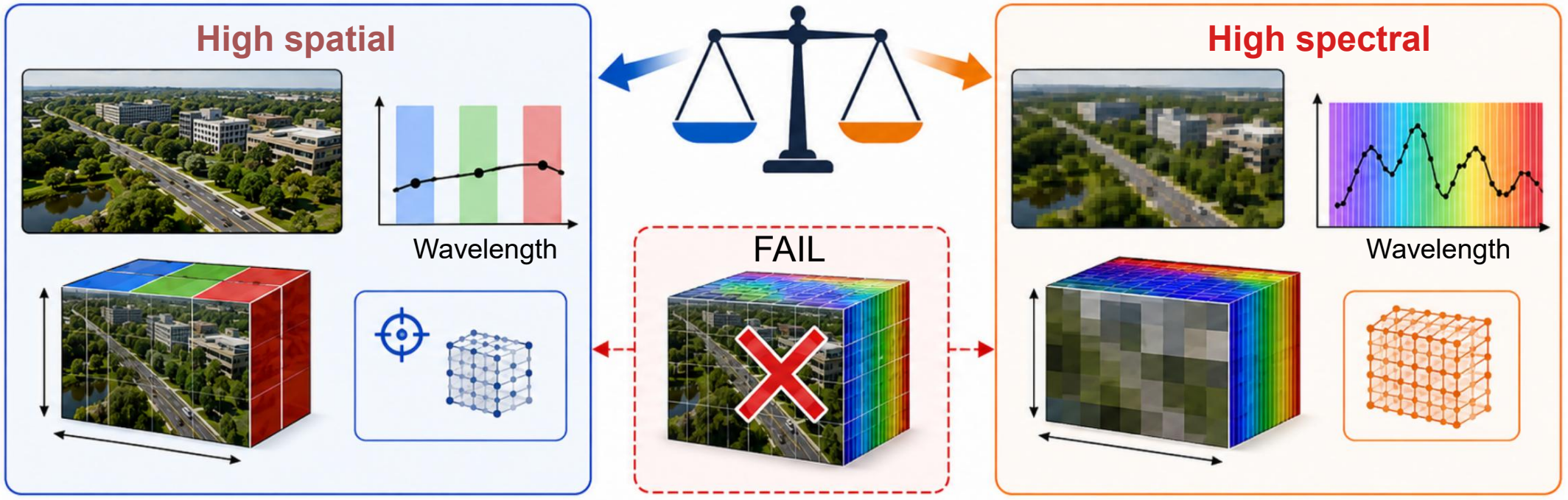
Visible → SWIR

Each pixel is a spectrum, not just RGB.

1. Introduction

Hardware constraints

- HSI acquisition struggles to obtain both **high spatial** and **high spectral** resolutions simultaneously.



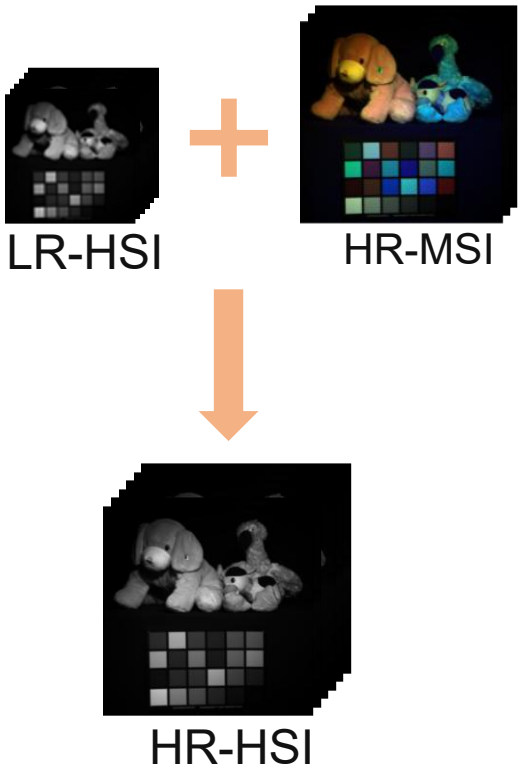
[1] G. Zhang, X. Fu, K. Huang and J. Wang, "Hyperspectral Super-Resolution: A Coupled Nonnegative Block-Term Tensor Decomposition Approach," *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Le gosier, Guadeloupe, 2019.

1. Introduction

HSI super-resolution: two mainstream directions

Single-image SR

Fusion-based SR



[2] W. Dong *et al*., "Hyperspectral Image Super-Resolution via Non-Negative Structured Sparse Representation," in *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2337-2352, May 2016.
[3] Zhang, Y., He, M. Multi-spectral and hyperspectral image fusion using 3-D wavelet transform. *J. of Electron.(China)* 24, 218–224 (2007).

2. Previous Works

Traditional HSI Super-resolution

- Zhang et al. [1] propose a coupled tensor decomposition method for image fusion by utilizing the characteristics of canonical multivariate decomposition.
- Dong et al. [2] propose a non-negative dictionary learning algorithm based on block coordinate descent to estimate sparse encoding of HSIs and obtain high-resolution results.
- Zhang et al. [3] propose a fusion method based on three-dimensional wavelets for the characteristics of MSI and HSI data volume and three-dimensional feature analysis.
- Moeller et al. [4] propose a wavelet variational method for fusing high-resolution images and HSIs with any number of bands.

They are limited by linear mathematical models and manually designed priors.

[4] Michael Moeller, Todd Wittman, and Andrea L. Bertozzi "A variational approach to hyperspectral image fusion", Proc. SPIE 7334, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV, 73341E (27 April 2009).

2. Related Works

Deep Learning HSI Super-resolution

Supervised methods

- Xie et al. [5] construct a HSI fusion model by expanding the proximal gradient algorithm.
- Sun et al. [6] solve the problem of deep HSI fusion by jointly utilizing spatial spectral regularization and physical imaging models.

Unsupervised methods

- Li et al. [7] propose an enhanced deep image prior method that accurately simulates complex HSI priors.
- Cao et al. [8] introduce an unsupervised hybrid network for blind HSI fusion.

They often fall short in generating realistic high-frequency details.

[5] Q. Xie, M. Zhou, Q. Zhao, Z. Xu and D. Meng, "MHF-Net: An Interpretable Deep Network for Multispectral and Hyperspectral Image Fusion,"in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1457-1473, 1 March 2022.

2. Related Works

HSI Super-resolution Based on Diffusion Model

- Liu et al. [9] propose a HSI fusion method that utilizes a spatial autoregressive model for internal structure guidance.
- Wang et al. [10] present an HSI fusion method called HyperGAN based on GANs.
- Rui et al. [11] employ a low rank diffusion model for HSI pan-sharpening.

Low-efficiency sampling, detail-limited generation, and insufficient denoising.

[6] Sun et al., "Dual Spatial–Spectral Pyramid Network With Transformer for Hyperspectral Image Fusion," in IEEE Transactions on Geoscience and Remote Sensing, vol. 61, pp. 1-16, 2023.
[8] X. Cao, Y. Lian, K. Wang, C. Ma and X. Xu, "Unsupervised Hybrid Network of Transformer and CNN for Blind Hyperspectral and Multispectral Image Fusion," in IEEE Transactions on Geoscience and Remote Sensing, vol. 62, pp. 1-15, 2024.

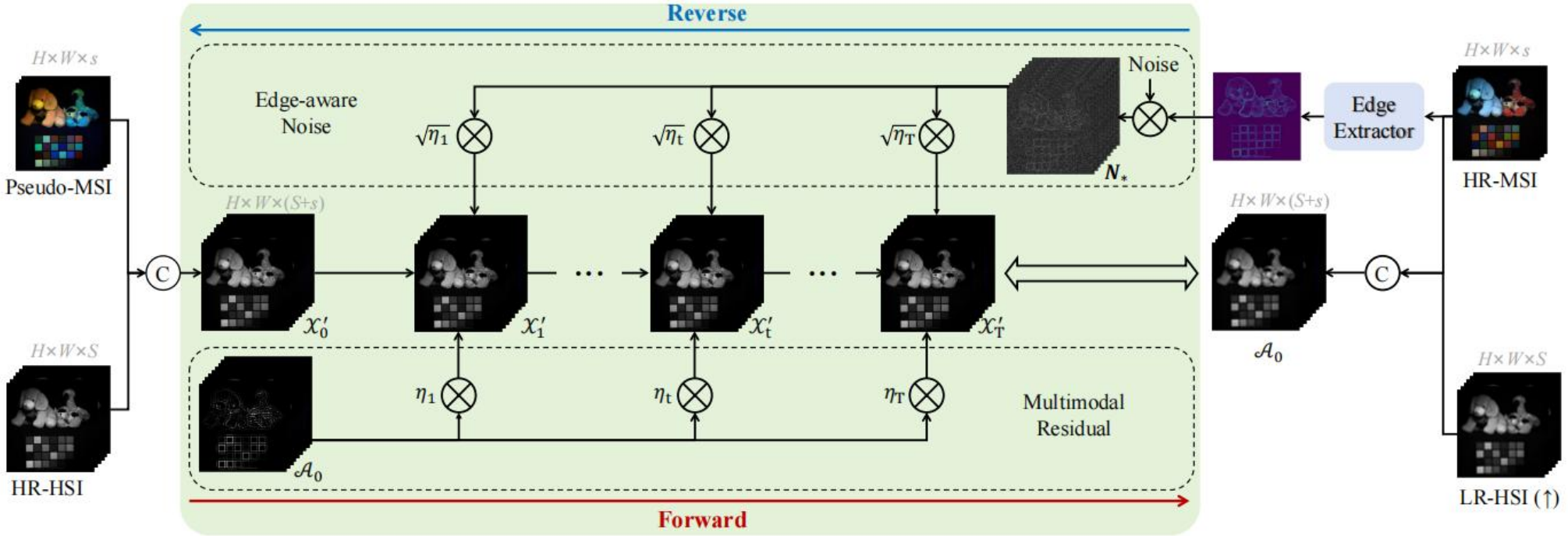
3. Method

- **In our work, we propose Edge-aware Multimodal Residual Diffusion Model for Hyperspectral Image Super-resolution.**
- **Multimodal residual mechanism**
  - Transfers multimodal residuals between HR-HSI, LR-HSI, and HR-MSI to reduce the sampling cost.
- **Edge-aware noise strategy**
  - Allows the model to focus more on the recovery of image details.
- **BAF-UNet architecture**
  - Enable progressive reconstruction and collaborative optimization of spectral and spatial features.

[7] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li and J. Chanussot, "Enhanced Deep Image Prior for Unsupervised Hyperspectral Image Super-Resolution," in IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-18, 2025, Art no. 5504218, doi: 10.1109/TGRS.2025.

3. Method

Overall architecture of EMR-Diff



[9] Liu C, Qian J, Fang F. ISGM-Fus: Internal structure-guided model for multispectral and hyperspectral image fusion[J]. Neurocomputing, 2025.
[10] Wang J, Zhu X, Jing L, et al. HyperGAN: a hyperspectral image fusion approach based on generative adversarial networks[J]. Remote Sensing, 2024.

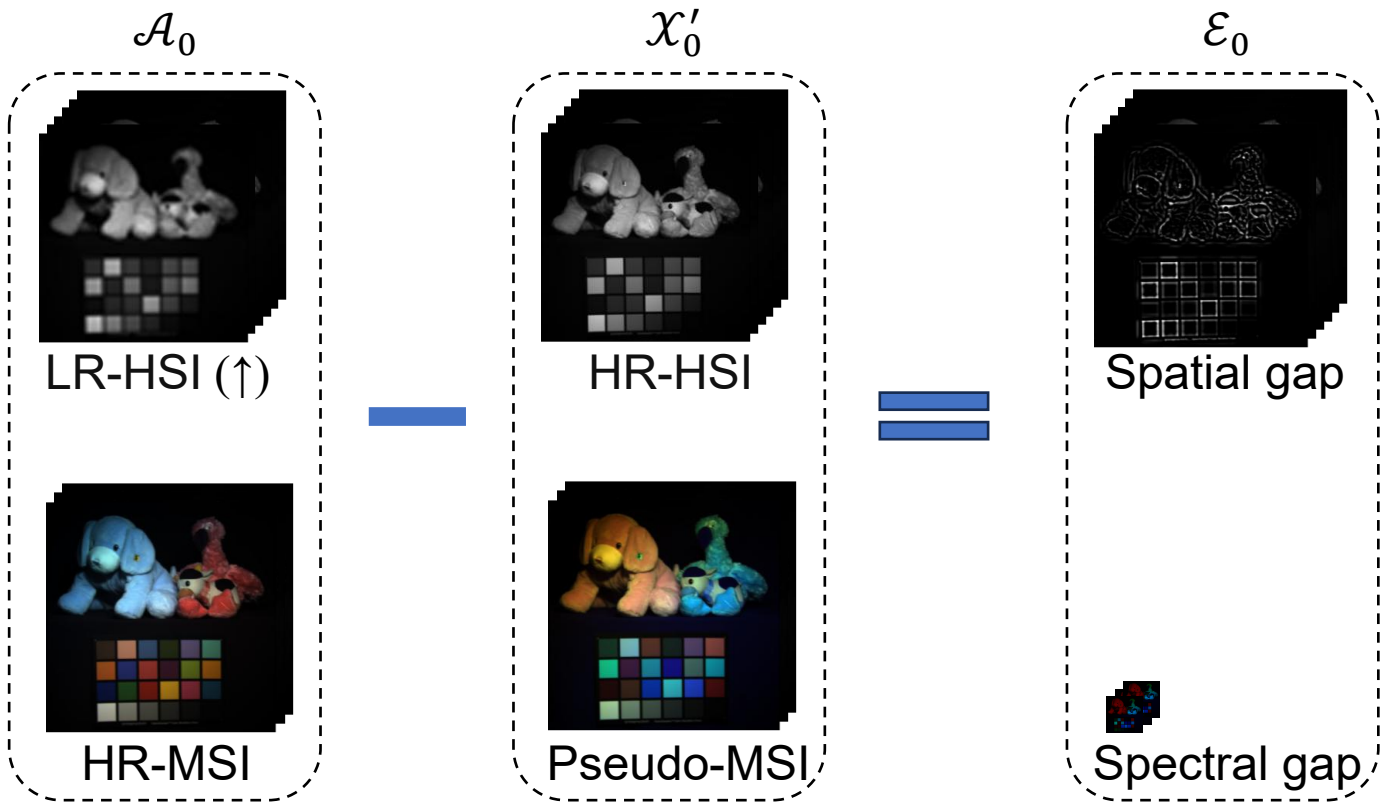
3. Method

Forward Process

- Multimodal residual

$$\mathcal{E}_0 = \mathcal{A}_0 - \mathcal{X}_0'$$

We intercept the (1,...,s) band of HR-HSI as Pseudo-MSI



[11] Rui X, Cao X, Pang L, et al. Unsupervised hyperspectral pansharpening via low-rank diffusion model[J]. Information Fusion, 2024.

3. Method

Forward Process

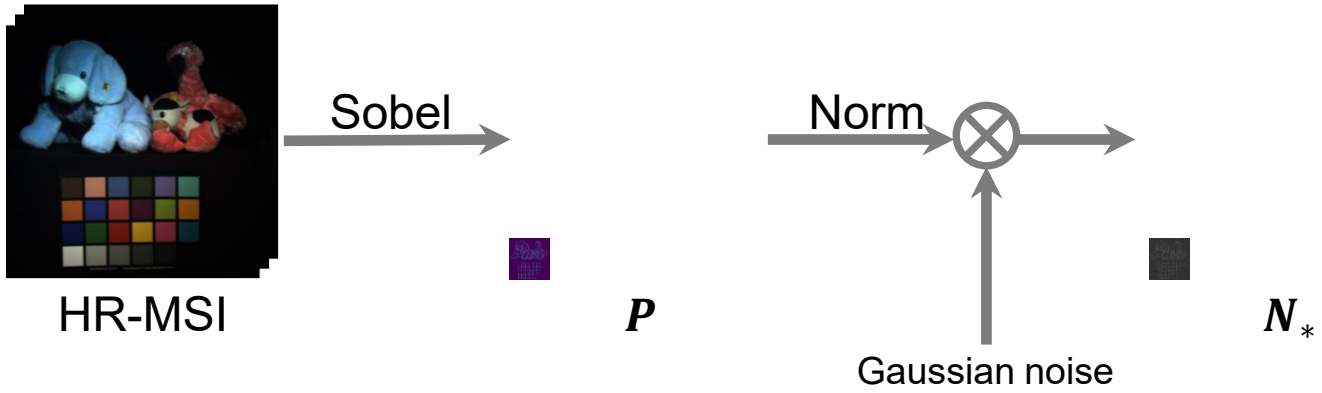
- Edge-aware noise

$$\boldsymbol{G_x} = \boldsymbol{C_x} * \boldsymbol{P}, \boldsymbol{G_y} = \boldsymbol{C_y} * \boldsymbol{P}$$

$$\boldsymbol{C_x} = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad \boldsymbol{C_y} = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}$$

$$\boldsymbol{M} = \sqrt{\boldsymbol{G_{x^2}} + \boldsymbol{G_{y^2}} + \epsilon}, \quad \epsilon = 10^{-8}$$

$$\boldsymbol{N_*} = \boldsymbol{N} \cdot \boldsymbol{W} = \boldsymbol{N} \cdot norm(\boldsymbol{M})$$Norm



3. Method

Forward Process

$$\mathcal{X}_t' = \mathcal{X}_0' + \eta_t \mathcal{E}_0 + \kappa\sqrt{\eta_t}\boldsymbol{N}_*$$

η_t is a monotonically increasing sequence

$t = 1$ ➡ η_t approaches 0 ➡ $\mathcal{X}_1' \approx \mathcal{X}_0'$

$t = \mathrm{T}$ ➡ η_t approaches 1 ➡ $\mathcal{X}_\mathrm{T}' = \mathcal{A}_0 + \kappa\boldsymbol{N}_*$

The multimodal residual is added to the Markov chain, which can reduce the number of sampling steps.

3. Method

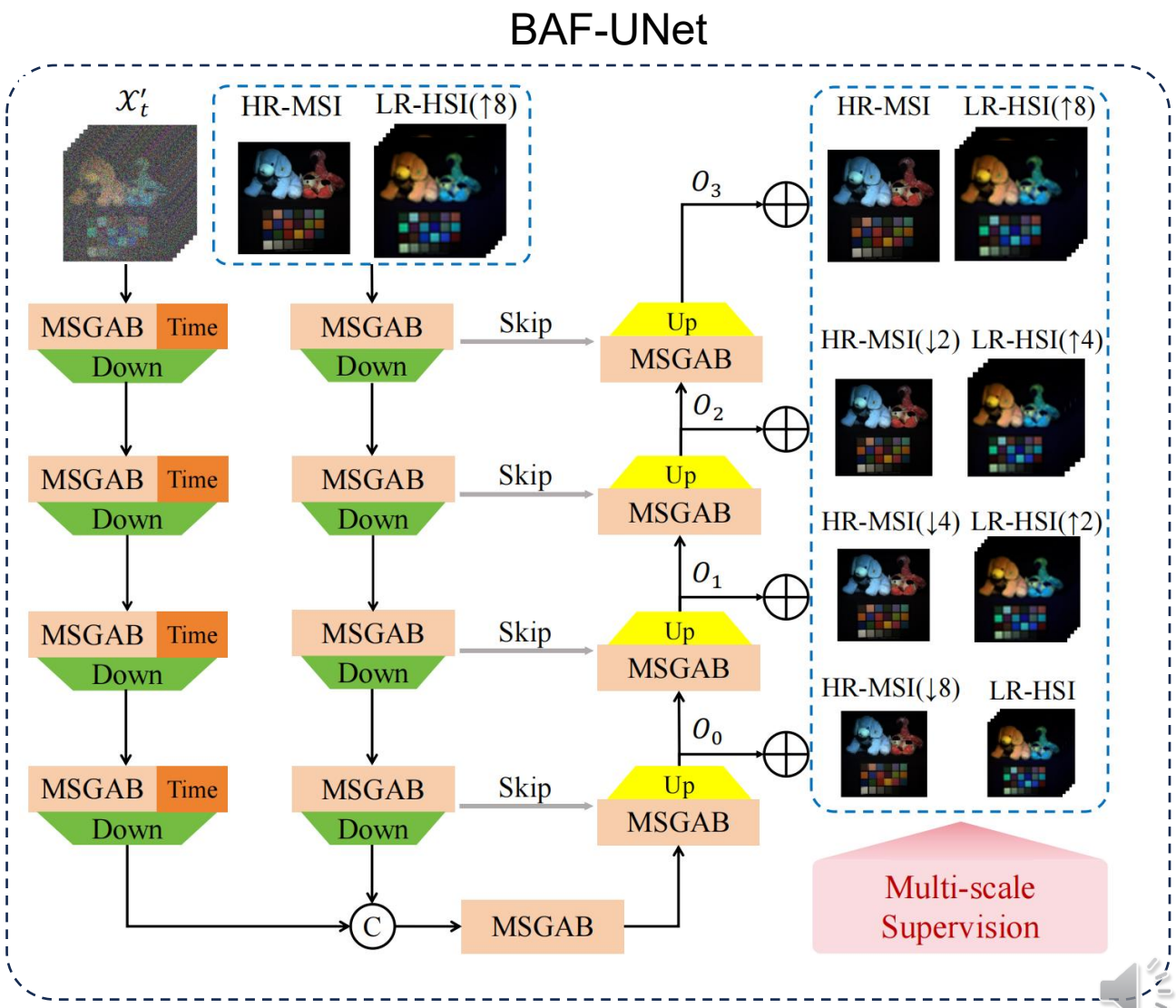
Reverse process

$$X'_{t-1} = \frac{\eta_{t-1}}{\eta_t} X'_t + \frac{\alpha_t}{\eta_t} f_{\boldsymbol{\theta}}(X'_t, \mathcal{A}_0, t) + \kappa \sqrt{\frac{\eta_{t-1}}{\eta_t} \alpha_t} \boldsymbol{N}_*$$

$$\alpha_t = \eta_t - \eta_{t-1}$$

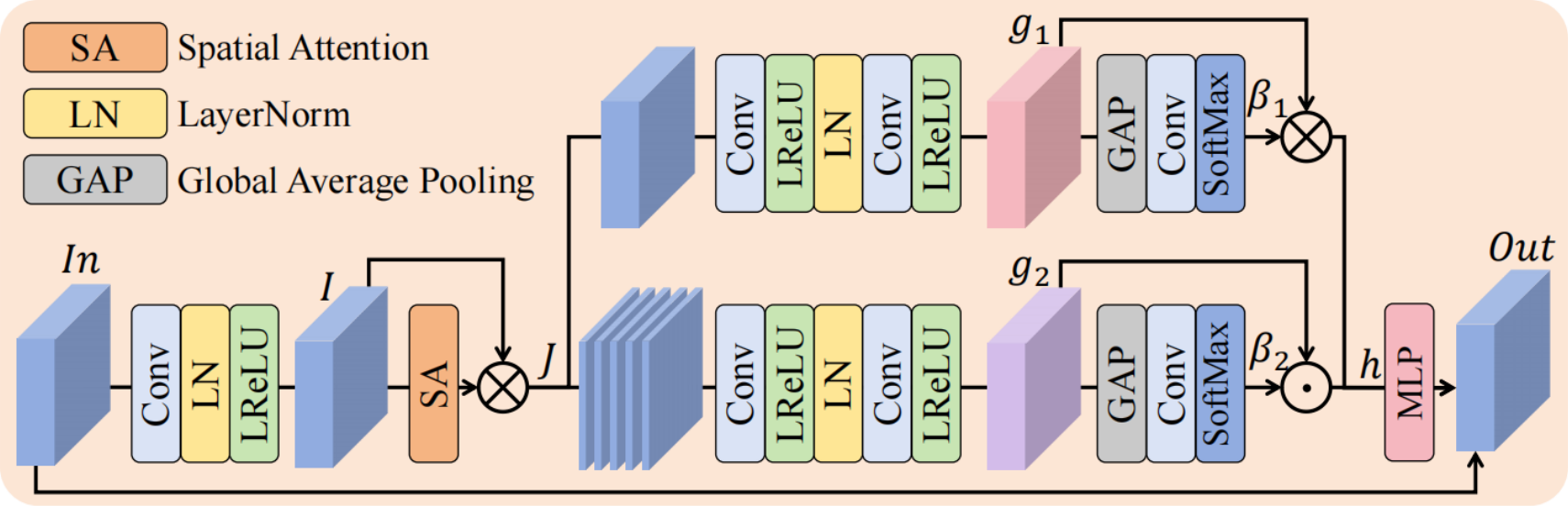
$f_{\boldsymbol{\theta}}$ is Bilateral Attention Fusion UNet (BAF-Unet)

Prevent information interference and enabling separate learning of noise distribution and structural features.



3. Method

Multi-Scale Group Attention Block (MSGAB)



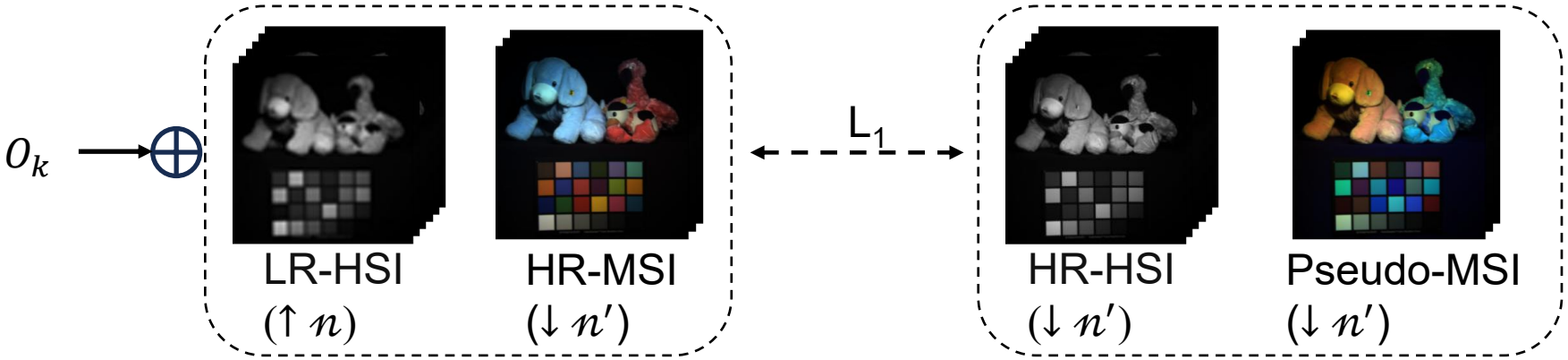
Integrate group convolution, attention, and MLP to collaboratively capture local details, channel dependencies, and spatial context.

3. Method

Multi-scale loss function

$$L_{multi} = \sum_{k=0}^{3} \left\| O_k + [Y_{\uparrow n}, Z_{\downarrow n'}] - X'_{\downarrow n'} \right\|_1$$

$[\cdot]$ is channel-wise concatenation



Clearly limit the upsampling process, LR-HSI guides spectral recovery, HR-MSI optimizes spatial structure, simulates super-resolution mapping, coordinates spectral and spatial optimization.

4. Experimental Results

Experimental Setup

- Dataset Setting

  ICVL, Harvard, Chikusei.

  LR-HSI: a 3×3 Gaussian filter (standard deviation of 2) and downsampling by 8 times bicubic interpolation.

  HR-MSI: using the spectral response function of Nikon D700 camera.

- Evaluation Metrics

  PSNR, SSIM, SAM, ERGAS.

- Comparison Methods

  Traditional methods: CNMF, Hysure.

  Unsupervised learning methods: PLR, SDP, ARGS.

  Supervised learning methods: LAGConv, DHIF, PSRT, DSPNet, SMGU, LRTN, EMR-Diff.

4. Experimental Results

Quantitative Results

- Comparison experiments

Methods	ICVL				Harvard				Chikusei			
	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	SAM↓	ERGAS↓
CNMF [46]	36.21	0.9895	0.0512	2.2128	32.94	0.9697	0.0730	4.3996	32.45	0.9798	0.2936	5.7328
Hysure [32]	38.37	0.9901	0.0457	1.9432	32.65	0.9737	0.1177	5.2153	33.59	0.9802	0.2812	5.6064
LAGConv [17]	53.53	0.9996	0.0044	0.1846	47.75	0.9984	0.0265	1.0045	43.96	0.9960	0.1160	2.1326
DHIF [16]	52.57	0.9995	0.0042	0.1938	47.78	0.9973	0.0260	0.8087	44.06	0.9959	0.1173	2.0143
PSRT [5]	52.00	0.9987	0.0061	0.3669	46.76	0.9972	0.0244	1.0380	44.74	0.9968	0.1055	1.8804
DSPNet [34]	55.19	0.9996	0.0042	0.1609	48.68	0.9986	0.0237	0.8814	46.97	0.9979	0.0977	1.5178
PLR [31]	45.94	0.9854	0.0233	1.4218	41.02	0.9920	0.0653	1.9737	38.47	0.9826	0.1288	4.0849
SDP [26]	46.11	0.9913	0.0219	1.0692	42.31	0.9931	0.0678	1.7294	39.01	0.9842	0.1243	3.4658
ARGS [60]	46.56	0.9922	0.0210	0.6728	41.97	0.9938	0.0662	1.7595	39.43	0.9850	0.1224	3.2289
SMGU [43]	51.44	0.9985	0.0062	0.3921	46.77	0.9979	0.0280	1.1435	44.53	0.9961	0.1042	1.9832
LRTN [27]	52.35	0.9992	0.0051	0.3796	46.80	0.9974	0.0277	0.9837	43.77	0.9961	0.1129	1.9872
EMR-Diff	**55.40**	**0.9997**	**0.0040**	**0.1588**	**49.28**	**0.9990**	**0.0233**	**0.7800**	**47.55**	**0.9980**	**0.0950**	**1.4943**

4. Experimental Results

Ablation study

- Ablation study of multimodal residual

Methods	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
None Residual	47.93	0.9984	0.0250	0.8994
Unimodal Residual	48.46	0.9986	0.0241	0.8025
Multimodal Residual	**49.28**	**0.9990**	**0.0233**	**0.7800**

- Ablation study of edge-aware noise

Methods	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
Pure Noise	48.36	0.9986	0.0245	0.8198
Edge-aware Noise	**49.28**	**0.9990**	**0.0233**	**0.7800**

- Ablation study of diffusion steps.

Steps	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
3	48.54	0.9987	0.0239	0.8009
4	48.70	0.9988	0.0237	0.7890
5	**49.28**	**0.9990**	**0.0233**	**0.7800**
10	48.89	0.9988	0.0236	0.7882

4. Experimental Results

Ablation study

- Ablation study on BAF-UNet

Module	UNet	B-UNet	AF-UNet	BAF-UNet(S)	BAF-UNet
Bilateral		✓		✓	✓
MSGAB			✓	✓	✓
MSS		✓	✓		✓
PSNR ↑	44.87	48.22	48.63	48.54	**49.28**
SSIM ↑	0.9972	0.9983	0.9986	0.9985	**0.9990**
SAM ↓	0.0260	0.0244	0.0237	0.0239	**0.0233**
ERGAS ↓	0.9369	0.8637	0.8024	0.8811	**0.7800**

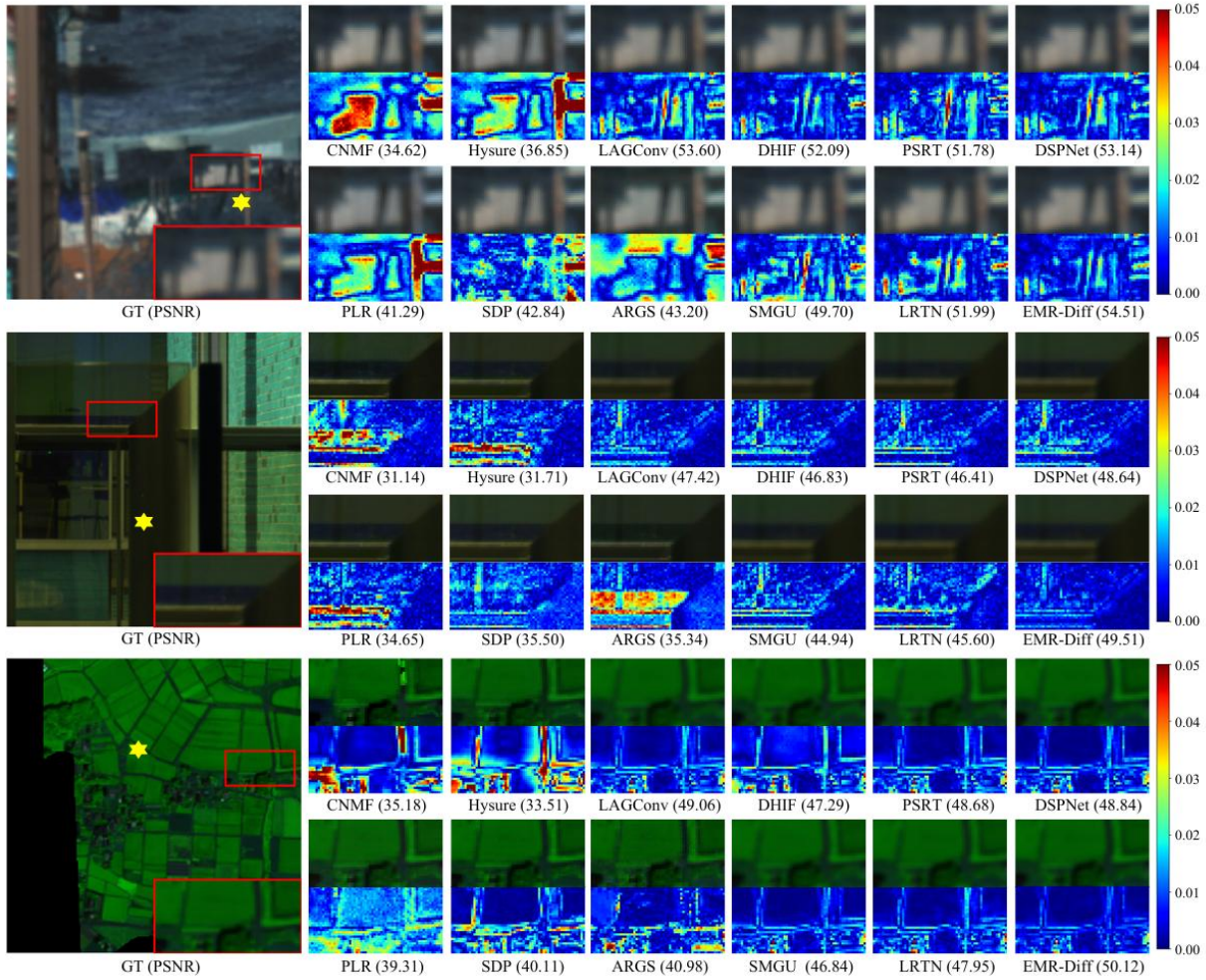
- Ablation study of pseudo-MSI synthesis

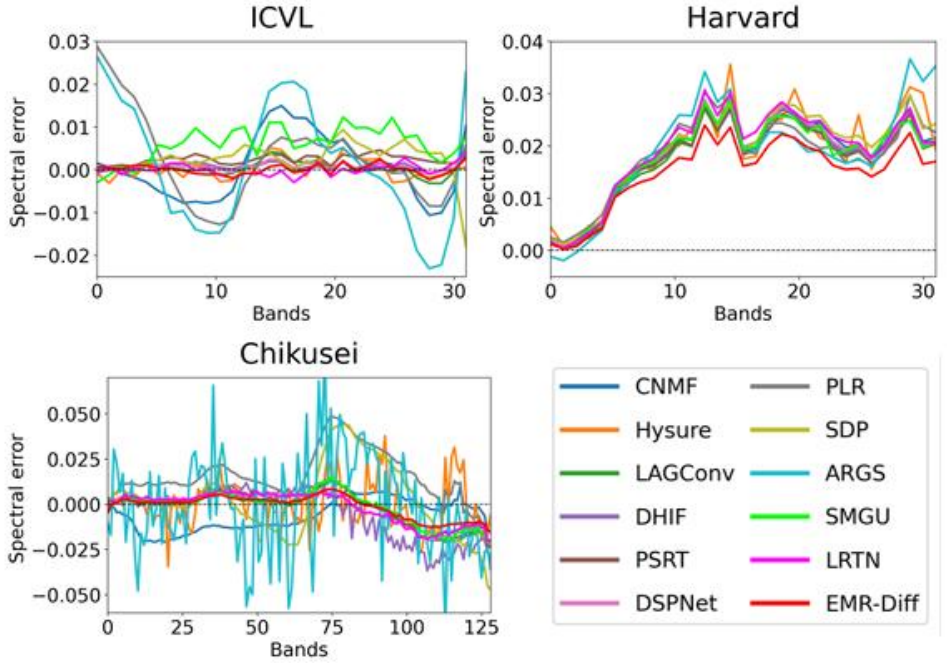
Bands	PSNR ↑	SSIM ↑	SAM ↓	ERGAS ↓
(29,30,31)	48.89	0.9988	0.0235	0.7892
(1,15,31)	49.02	0.9989	0.0236	0.7848
(1,2,3)	**49.28**	**0.9990**	**0.0233**	**0.7800**

4. Experimental Results

Qualitative Visualization

- Comparison experiments

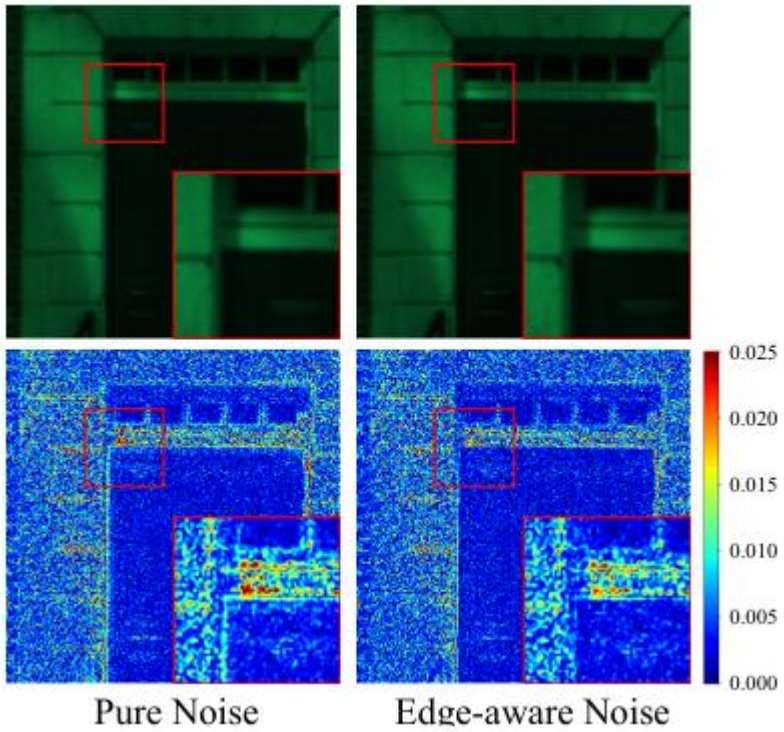




4. Experimental Results

Qualitative Visualization

- Ablation study of edge-aware noise



Using edge-aware noise exhibit significantly smaller errors in edge regions.

5. Conclusion

We propose EMR-Diff for HSI super-resolution

- Multimodal residual significantly shortens the diffusion process.
- Edge-aware noise guides denoising toward high-frequency image details.
- BAF-UNet decouples denoising and guidance while using MSGAB and multi-scale supervision.

Future work

- Explore more efficient samplers.
- Improve generalization across sensors and scenes.

Thank you