IJCAI/2023 MACAO

Local-Global Transformer Enhanced Unfolding Network for Pan-sharpening

Mingsong Li[1], Yikun Liu[1], Tao Xiao[1], Yuwen Huang[2], and Gongping Yang[1]*

[1]School of Software, Shandong University, Jinan, China

[2]School of Computer, Heze University, Heze, China

August 23, 2023

Codebase

https://github.com/lms-07/LGTEUN

Homepage of Presenter

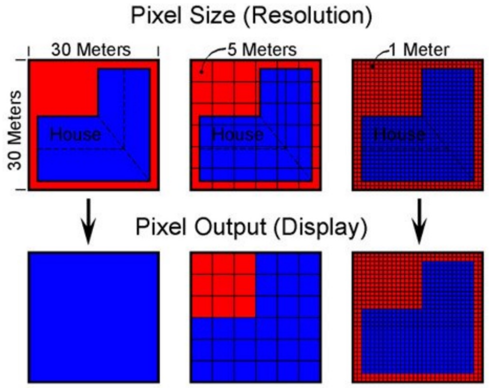
https://lms-07.github.io/

Outline

- Background
- Problem Analysis
- Method
- Experiments
- Conclusion and Discussion

Background: Vital Resolutions in Remote Sensing Image

- **Spatial resolution**



High Spatial Resolution | Medium Spatial Resolution | Low Spatial Resolution

- **Spectral resolution**

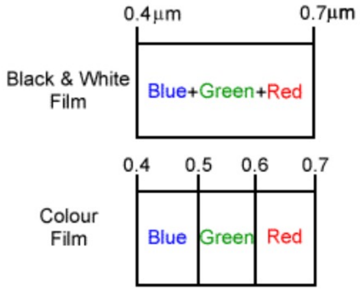
↑Bands

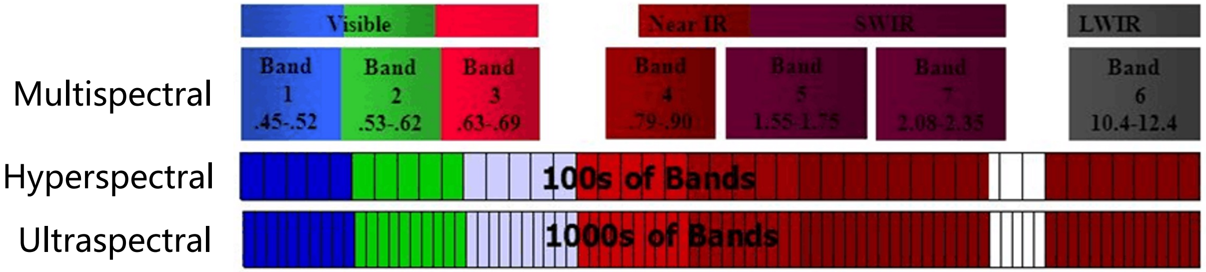
$$B = \frac{\lambda}{\Delta\lambda}$$

Wavelength

Spectral resolution↑

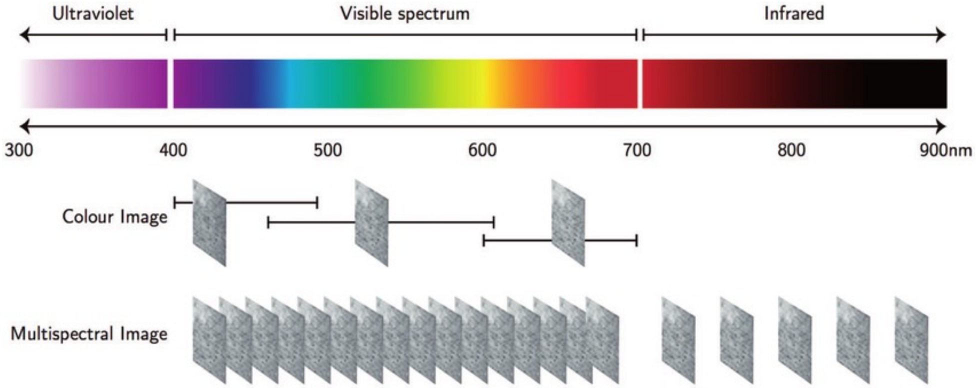
Panchromatic Image

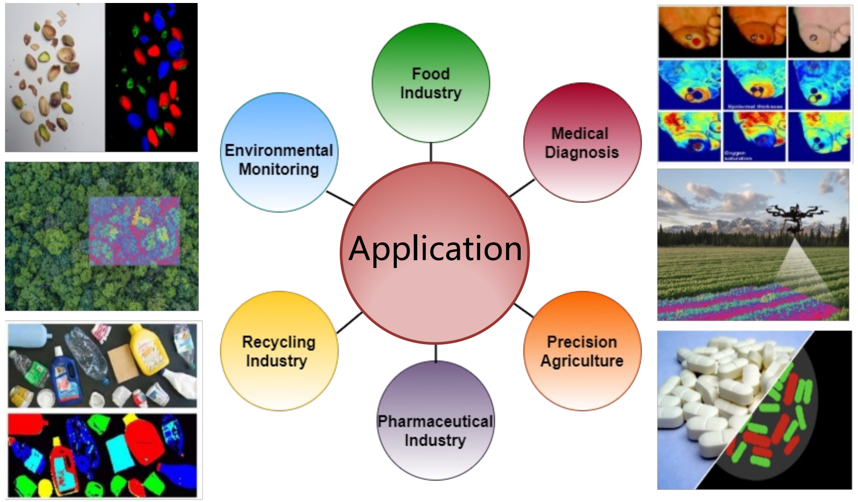




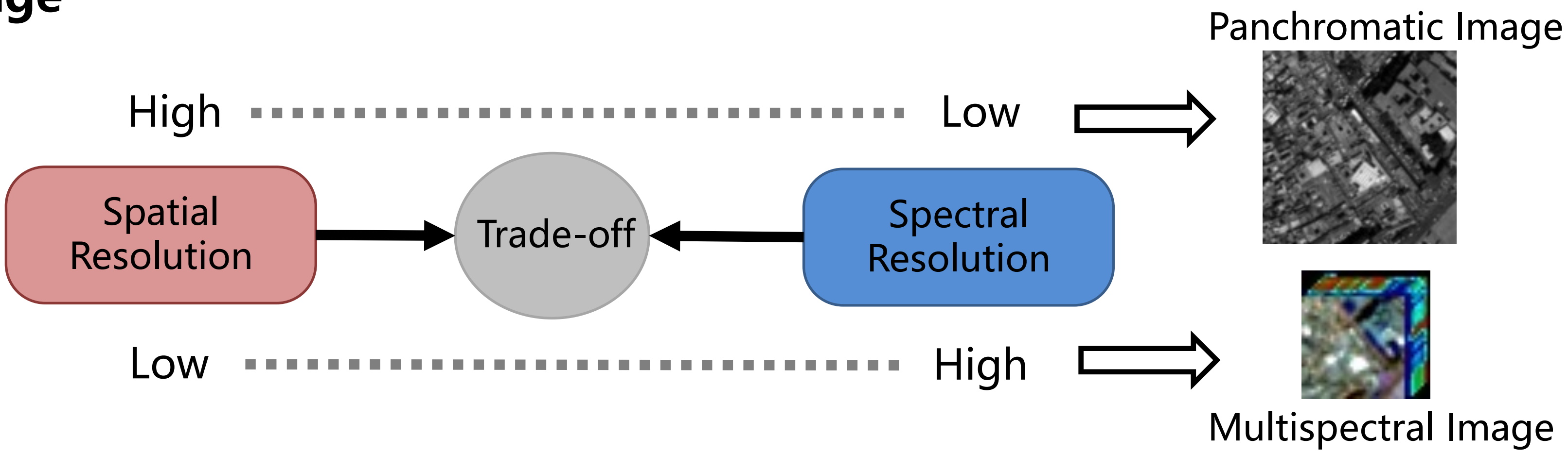
Background: Characteristic, Applications, and Challenge

➢ Wider wavelength range





➢ Challenge



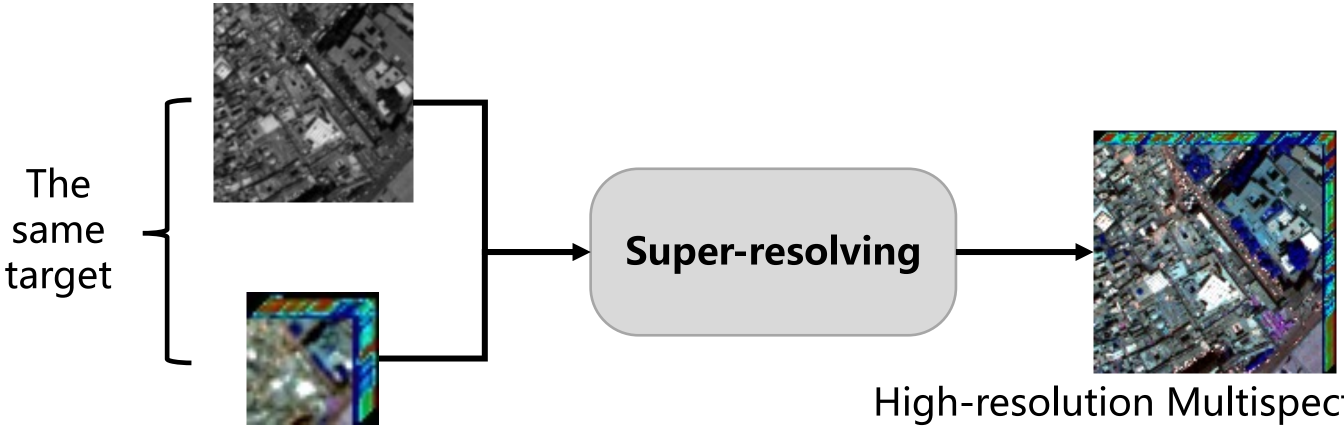
"Hyperspectral Image Classification—Traditional to Deep Models: A Survey for Future Prospects," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022.

Problem Analysis: Task Description and Literature Classification

☐ *Pan-sharpening*

Panchromatic (PAN) Image

$\mathbf{Y} \in \mathbb{R}^{HW \times 1}$



Low-resolution Multispectral (LrMS) Image

$\mathbf{X} \in \mathbb{R}^{hw \times B}$

High-resolution Multispectral (HrMS) Image

$\mathbf{Z} \in \mathbb{R}^{HW \times B}$

➢ **Categories**

Model-based

- Limited generalization ability
- Great model interpretability

VS

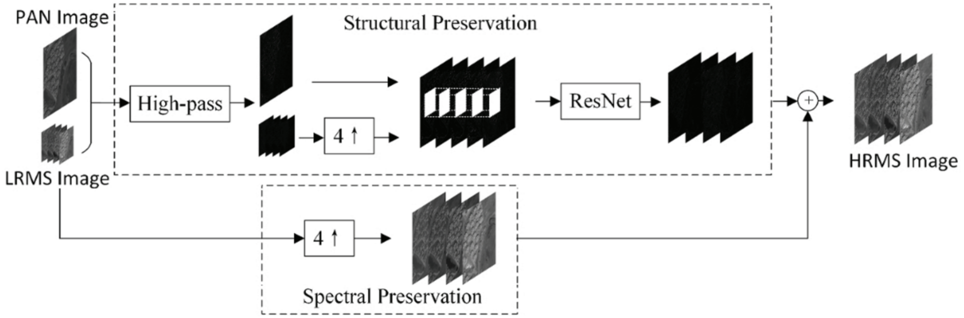
Deep Learning (DL)-based

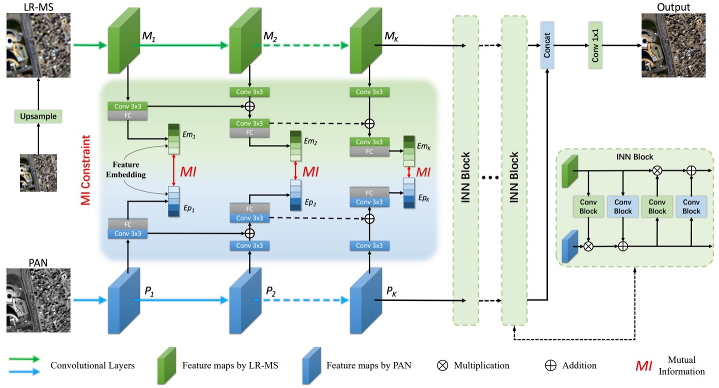
- Strong feature representation
- Weak model interpretability

Problem Analysis

1) Model Interpretability

Black box principle of DL-based methods





2) Local and Global Dependencies

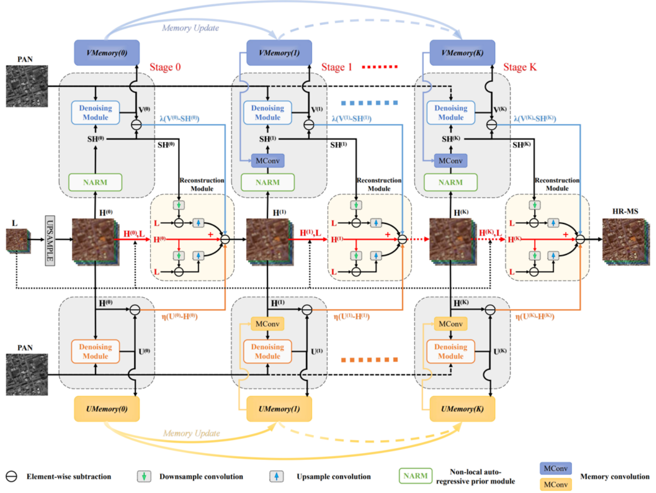
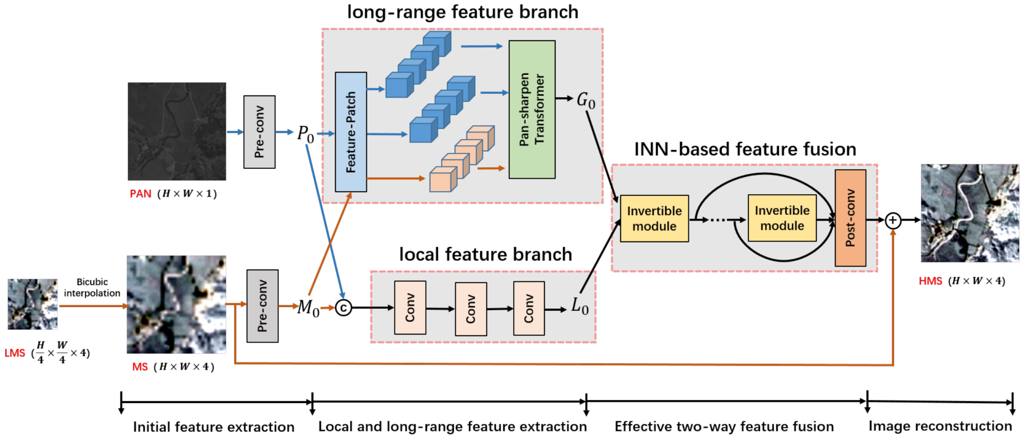


Figure 2. The overall architecture of **MDCUN**.



"PanNet: A Deep Network Architecture for Pan-Sharpening," in ICCV, 2017.
"Mutual Information-driven Pan-sharpening," in CVPR, 2022.
"Memory-Augmented Deep Conditional Unfolding Network for Pan-Sharpening," in CVPR, 2022.
"Pan-Sharpening with Customized Transformer and Invertible Neural Network," in AAAI, 2022.

Method: Problems to Solutions

1) Model Interpretability

- ***Deep unfolding methods*** ***combine***
  - Great model interpretability — **Model-based**
  - Strong feature representation — **Deep Learning (DL)-based**

☐ The degradation process of the HrMS image **Z**

$$\mathbf{X} = \mathbf{S}\mathbf{Z} + \mathbf{N}_x, \quad \mathbf{Y} = \mathbf{Z}\mathbf{R} + \mathbf{N}_y, \tag{1}$$

$\mathbb{R}^{hw\times HW}$ $\mathbb{R}^{hw\times B}$ $\mathbb{R}^{HW\times B}$ $\mathbb{R}^{B\times 1}$ $\mathbb{R}^{HW\times 1}$

$$\bar{\mathbf{Z}} = \underset{\mathbf{Z}}{argmin}\frac{1}{2}\| \mathbf{X} - \mathbf{S}\mathbf{Z} \|^2 + \frac{1}{2}\| \mathbf{Y} - \mathbf{Z}\mathbf{R} \|^2 + \lambda J(\mathbf{Z}), \tag{2}$$

➢ Proximal Gradient Descent (PGD) Alg

$$\bar{\mathbf{Z}}_k = \underset{\mathbf{Z}}{argmin}\frac{1}{2}\| \mathbf{Z} - (\bar{\mathbf{Z}}_{k-1} - \eta\nabla_f(\bar{\mathbf{Z}}_{k-1})) \|^2 + \lambda J(\mathbf{Z}), \tag{3}$$

$$\nabla_f(\bar{\mathbf{Z}}_{k-1}) = \mathbf{S}^T(\mathbf{S}\bar{\mathbf{Z}}_{k-1} - \mathbf{X}) + (\bar{\mathbf{Z}}_{k-1}\mathbf{R} - \mathbf{Y})\mathbf{R}^T. \tag{4}$$

Data Subproblem

$$\bar{\mathbf{Z}}_{k-\frac{1}{2}} = \bar{\mathbf{Z}}_{k-1} - \eta\nabla(\bar{\mathbf{Z}}_{k-1}), \tag{5}$$

$$\bar{\mathbf{Z}}_k = prox_{\eta,J}(\bar{\mathbf{Z}}_{k-\frac{1}{2}}), \tag{6}$$

Prior Subproblem

⇩

Do we need a powerful image denoiser for prior subproblem?

⇩

Efficiently model local and global feature dependencies in the same layer.

2) Local and Global Dependencies ⇦

Method: Local-Global Transformer Enhanced Unfolding Network (LGTEUN)

- Data Module D:

$$\bar{\mathbf{Z}}_{k-\frac{1}{2}} = \mathcal{D}(\bar{\mathbf{Z}}_{k-1}, \mathbf{X}, \mathbf{Y}, \eta_{k-1}).$$

- Prior Module P:

$$\bar{\mathbf{Z}}_{k} = \mathcal{P}(\bar{\mathbf{Z}}_{k-\frac{1}{2}}).$$

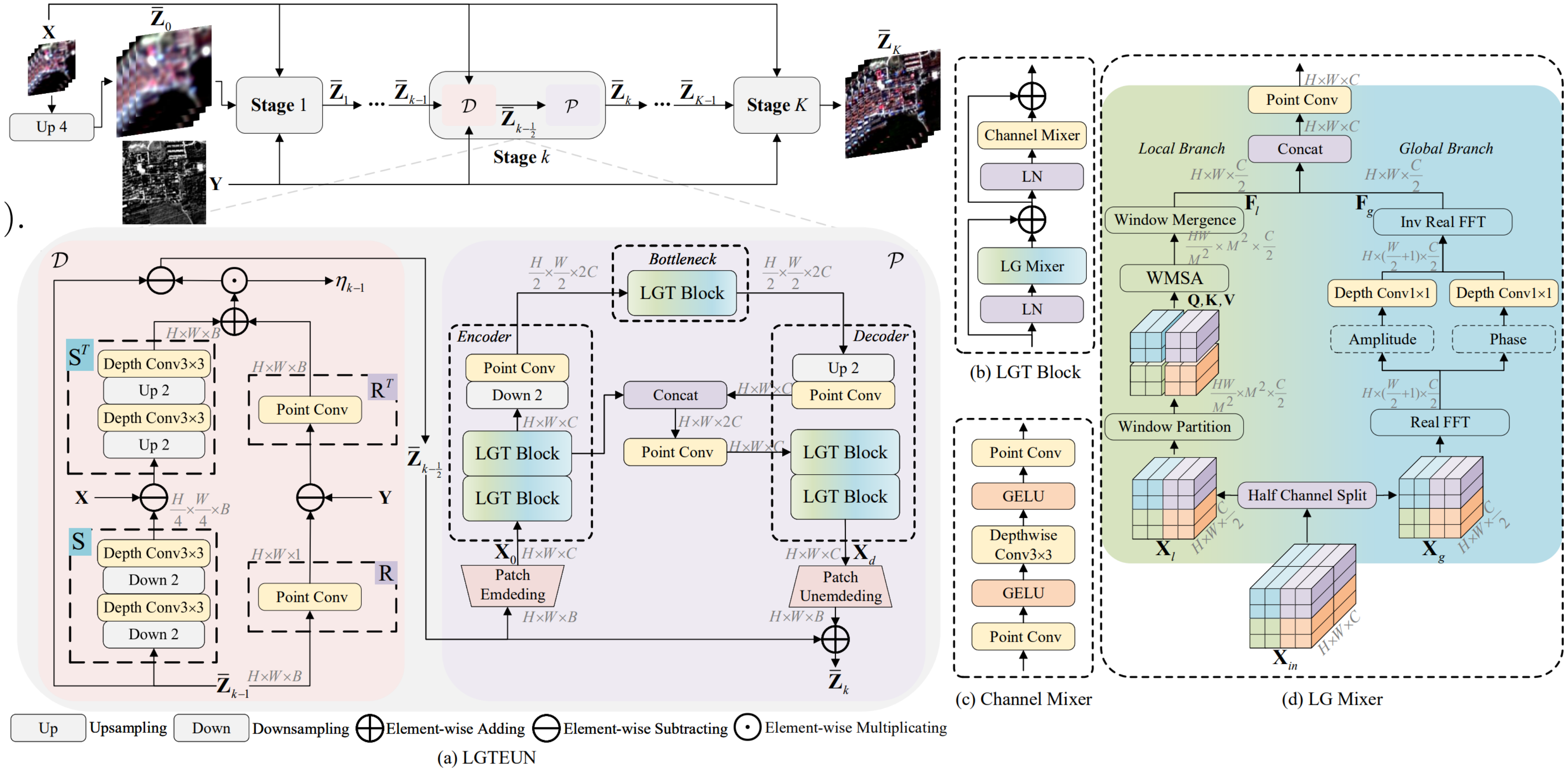
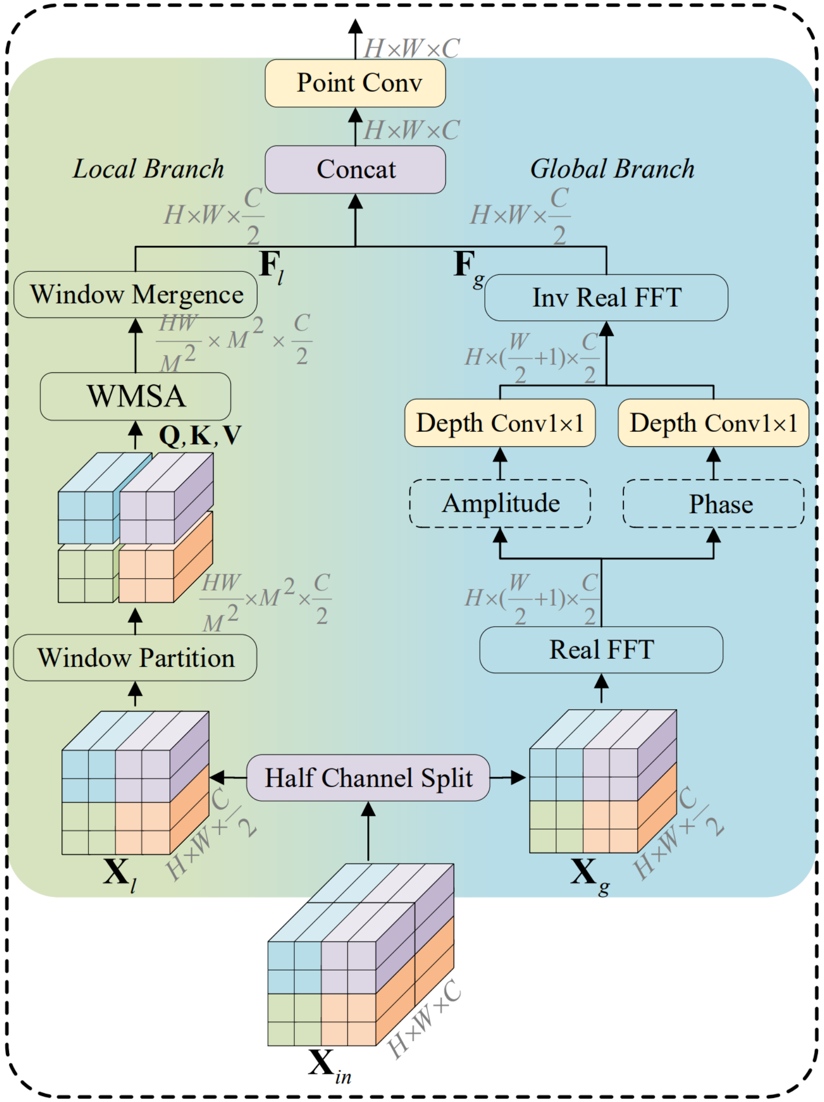


Figure 2: Illustration of the proposed LGTEUN. (a) The overall architecture of LGTEUN with K stages and details of the k-th stage. The lightweight CNN-based data module $\mathcal{D}$ and the powerful transformer-based prior module $\mathcal{P}$ in each stage correspond to the data and prior subproblems in an iteration of the PGD algorithm. (b) Components of an LGT block. (c) The adopted channel mixer. (d) The key LG Mixer is comprised of a *local branch* and a *global branch*.

Method: Local-Global Mixer



(d) LG Mixer

- ***Local Branch*** Calculating local window based self-attention in spatial domain

$$\mathbf{F}_a^i = Softmax(\frac{\mathbf{Q}^i\mathbf{K}^{iT}}{\sqrt{d}} + \mathbf{P}^i)\mathbf{V}^i, \quad i = 1, ..., h, \tag{9}$$

- ***Global Branch*** Extracting global contextual feature representation in frequency domain

$$\mathcal{F}(\mathbf{X}_g)(u,v) = \frac{1}{\sqrt{HW}}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathbf{X}_g(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \tag{10}$$

$$\mathcal{A}(\mathbf{X}_g)(u,v) = \sqrt{R^2(\mathbf{X}_g)(u,v) + I^2(\mathbf{X}_g)(u,v)}, \tag{11}$$

$$\mathcal{P}(\mathbf{X}_g)(u,v) = arctan[\frac{I(\mathbf{X}_g)(u,v)}{R(\mathbf{X}_g)(u,v)}]. \tag{12}$$

$$\mathbf{F}_g = \mathcal{F}^{-1}(DConv(\mathcal{A}(\mathbf{X}_g)), DConv(\mathcal{P}(\mathbf{X}_g))), \tag{13}$$

Experiments:

☐ Three satellite data sets:

- An 8-band MS data set (WorldView-3)
- Two 4-band MS data set (WorldView-2 and GaoFen-2)

☐ Image quality assessment:

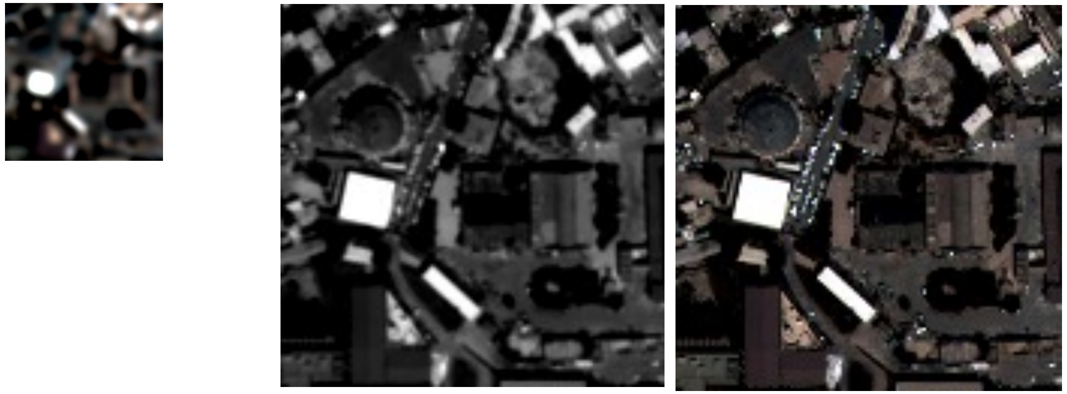
- Five reference metrics:
  PSNR, SSIM, Q-index, SAM, and ERGAS
- Three non-reference metrics:
  D_λ, D_S, and QNR

Data set displaying

- WorldView-3

LrMS PAN HrMS

- A pair of samples



Experiments: Setting Experiment-The Number of Stages

Data Set	Metric	Stage 1	Stage 2	Stage 3	Stage 4
WorldView-3	PSNR↑	32.0339	32.2188	32.068	32.0042
	SSIM↑	0.9532	0.9545	0.9535	0.9527
	Q8↑	0.9481	0.9494	0.9487	0.9480
	SAM↓	0.0605	0.0605	0.0603	0.0612
	ERGAS↓	2.6765	2.6286	2.6678	2.6898
	Time (s/img)	0.0070	0.0133	0.0205	0.0262
	Params (KB)	270.2	540.0	809.9	1079.7
	FLOPs (GB)	9.52	19.04	28.56	38.08
WorldView-2	PSNR↑	42.600	42.6837	42.4771	42.1634
	SSIM↑	0.9784	0.9786	0.9781	0.9767
	Q4↑	0.8398	0.8415	0.8383	0.8329
	SAM↓	0.0209	0.0208	0.0213	0.0222
	ERGAS↓	0.9358	0.928	0.9573	0.9787
	Time (s/img)	0.0065	0.0137	0.0204	0.0254
	Params (KB)	101.2	202.2	303.2	404.2
	FLOPs (GB)	2.57	5.14	7.71	10.28

Table 1: Performance and efficiency of LGTEUN with different numbers of stages K on WorldView-3 and WorldView-2 satellite data sets.

Experiments: Quantitative and Qualitative Comparisons

Method	WorldView-3					WorldView-2					GaoFen-2				
	PSNR↑	SSIM↑	Q8↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	Q4↑	SAM↓	ERGAS↓	PSNR↑	SSIM↑	Q4↑	SAM↓	ERGAS↓
GSA	22.5164	0.6343	0.5742	0.1106	7.8267	33.5975	0.8899	0.5681	0.0573	2.5402	36.0557	0.8838	0.5517	0.0641	3.5758
SFIM	21.4154	0.5415	0.4525	0.1147	8.8553	32.6334	0.8728	0.5159	0.0597	3.1919	34.7715	0.8572	0.4584	0.0657	4.2073
Wavelet	21.4464	0.5656	0.5271	0.1503	9.1545	32.1992	0.8500	0.4577	0.0638	3.3799	33.9208	0.8197	0.4033	0.0695	4.6445
PanFormer	30.4772	0.9368	0.9316	0.0672	3.1830	41.3581	0.9731	0.8236	0.0241	1.0617	44.8540	0.9805	0.8865	0.0271	1.3334
CTINN	31.8564	0.9518	0.9460	0.0660	2.7421	41.2015	0.9735	0.8149	0.0246	1.0880	44.2942	0.9784	0.8716	0.0293	1.4148
LightNet	32.0018	0.9525	0.9472	0.0639	2.6853	41.5589	0.9739	0.8220	0.0237	1.0382	44.6876	0.9787	0.8741	0.0279	1.3510
SFIIN	31.6587	0.9492	0.9435	0.0652	2.8016	41.9489	0.9752	0.8108	0.0229	1.0084	44.7248	0.9802	0.8721	0.0280	1.3361
MutInf	31.8298	0.9523	0.9469	0.0636	2.7526	41.9522	0.9760	0.8258	0.0227	1.0153	44.8305	0.9800	0.8836	0.0277	1.3394
MDCUN	31.2978	0.9429	0.9363	0.0661	2.9295	42.3351	0.9772	0.8370	0.0216	0.9638	45.5677	0.9825	0.8915	0.0252	1.2249
LGTEUN	32.2188	0.9545	0.9494	0.0605	2.6286	42.6837	0.9786	0.8415	0.0208	0.9280	45.8364	0.9840	0.8973	0.0247	1.1824

Table 2: Quantitative comparison of different methods on WorldView-3, WorldView-2, and GaoFen-2 satellite data sets.

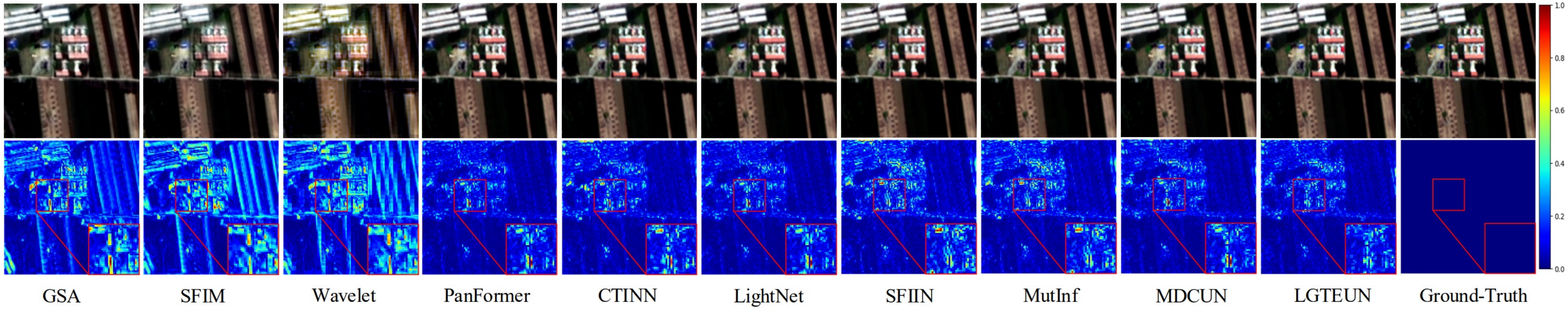


Figure 3: Qualitative comparison of different methods on the WorldView-2 satellite data set.

Experiments: Full-resolution Test and Efficiency Comparisons

Method	Full-resolution Test $D_\lambda\downarrow$	$D_S\downarrow$	QNR$\uparrow$
GSA	0.0094	0.1076	0.8839
SFIM	0.0094	0.1061	0.8854
Wavelet	0.0552	0.1330	0.8193
PanFormer	0.0191	0.0416	0.9400
CTINN	0.0123	0.0442	0.9440
LightNet	0.0185	0.0282	0.9539
SFIIN	0.0198	0.0352	0.9457
MutInf	0.0163	0.0420	0.9423
MDCUN	0.0747	0.1673	0.7708
LGTEUN	0.0162	0.0310	0.9532

Table 3: Full-resolution test of different methods on the WorldView-3 satellite data set.

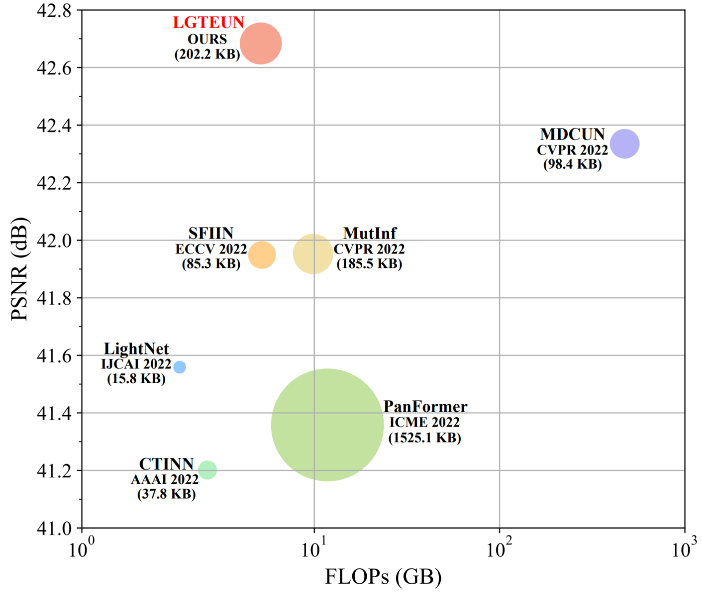


Figure 1: PSNR-Params-FLOPs comparisons between six SOTA DL-based pan-sharpening methods and our LGTEUN on the WorldView-2 satellite data set. The vertical axis is PSNR (model performance), the horizontal axis is FLOPs (computational cost), and the circle radius is Params (model complexity).

Data Set	Metric	GSA	SFIM	Wavelet	PanFormer	CTINN	LightNet	SFIIN	MutInf	MDCUN	LGTEUN
WorldView-3	Time (s/img)	0.0482	0.0591	0.0562	0.0160	0.0426	0.0019	0.0529	0.1083	0.1747	0.0133
	Params (KB)	–	–	–	1532.8	38.3	16.3	85.8	185.8	140.9	540.0
	FLOPs (GB)	–	–	–	11.92	2.68	2.02	5.25	9.87	479.54	19.04
GaoFen-2	Time (s/img)	0.0216	0.0301	0.0271	0.0257	0.0431	0.0017	0.0528	0.1141	0.1017	0.0129
	Params (KB)	–	–	–	1530.3	37.8	15.8	85.3	185.5	98.3	202.2
	FLOPs (GB)	–	–	–	11.77	2.65	1.95	5.22	9.85	473.19	5.14

Table 4: Efficiency comparison of different methods on WorldView-3 and GaoFen-2 satellite data sets.

Experiments: Ablation Study and Further Visualization

Setting		Reduced-resolution Test					Full-resolution Test		
Local Branch	*Global Branch*	PSNR↑	SSIM↑	Q8↑	SAM↓	ERGAS↓	D_λ↓	D_S↓	QNR↑
✗	✓	31.9309	0.9519	0.9468	0.0636	2.7102	0.0177	0.0364	0.9465
✓	✗	31.9742	0.9525	0.9468	0.0618	2.7029	0.0170	0.0349	0.9486
✓	✓	32.2188	0.9545	0.9494	0.0605	2.6286	0.0162	0.0310	0.9532

Table 5: Ablation study on the WorldView-3 satellite data set.

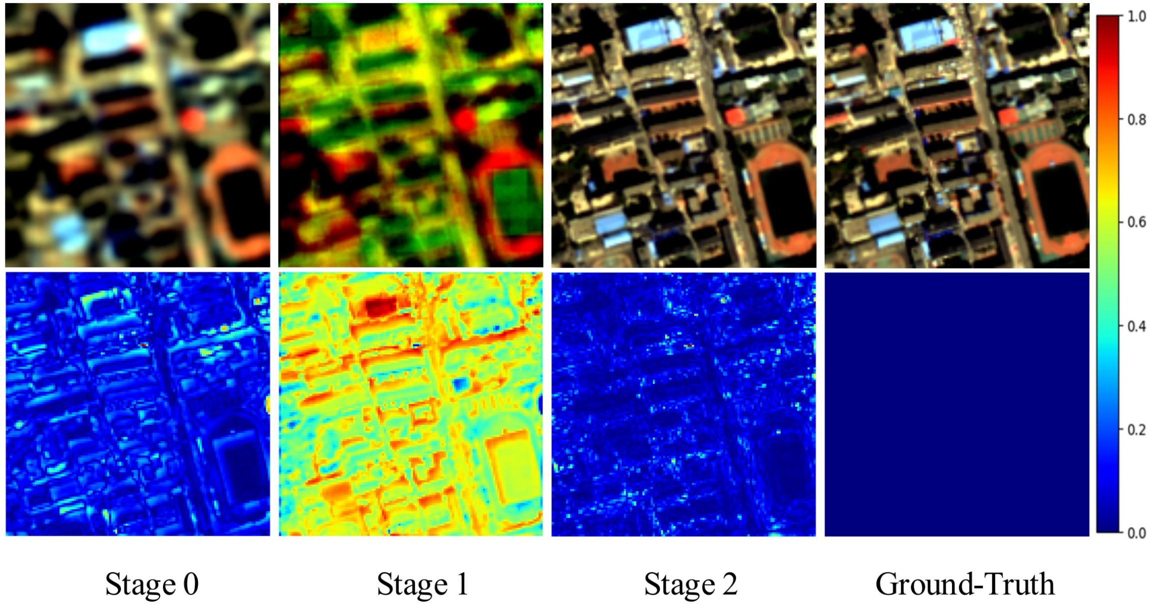


Figure 4: Stage-wise visualization on the GaoFen-2 satellite scene.

Conclusion and Discussion

Conclusion:

For the MS pan-sharpening, to address two longstanding issues, i.e., *model interpretability* and *local and global dependencies*, we unfold the iterative PGD algorithm into a stage-wise unfolding network, LGTEUN.

- The first transformer-based deep unfolding network.
- The first transformer module to perform spatial and frequency dual-domain learning.

Limitations:

- Further performance boosting on full-resolution scene.
- Further enhancements on model efficiency.

IJCAI/2023 MACAO

Thanks

Codebase

https://github.com/lms-07/LGTEUN

Contact：msli@mail.sdu.edu.cn