

Supplementary Material: Linearly-evolved Transformer for Pan-sharpening

Junming Hou*
Southeast University
Nanjing, China
junming_hou@seu.edu.cn

Zihan Cao*
University of Electronic Science and
Technology of China
Chengdu, China
iamzihan666@gmail.com

Naishan Zheng
University of Science and Technology
of China
Hefei, China
nszheng@mail.ustc.edu.cn

Xuan Li
Southeast University
Nanjing, China
xuanli2003@seu.edu.cn

Xiaoyu Chen
Southeast University
Nanjing, China
213214058@seu.edu.cn

Xinyang Liu
Hong Kong Polytechnic University
Hong Kong, China
codex.lxy@gmail.com

Xiaofeng Cong
Southeast University
Nanjing, China
cxf_svip@163.com

Danfeng Hong[†]
Aerospace Information Research
Institute, Chinese Academy of
Sciences
Beijing, China
hongdf@aircas.ac.cn

Man Zhou[†]
University of Science and Technology
of China
Hefei, China
manman@mail.ustc.edu.cn

1 Datasets and Experimental Settings

Pan-sharpening Benchmark. We compare the quantitative and qualitative performance of our model with state-of-the-art methods on the pan-sharpening task. Three traditional methods including: BDDSD-PC [19], MTF-GLP-FS [21], BT-H [14]; and nine deep-learning based methods: PNN [15], DiCNN [7], MSDCNN [27], FusionNet [2], DCFNet [25], SFIIN [32], PanViT [16], InvFormer [30], and Fourmer [31] are selected.

Dataset Simulation. We assess our proposed methods using two popular commercial satellites over pan-sharpening task: WorldView3 (WV3) and GaoFen2 (GF2). In detail, each satellite dataset includes numerous image pairs for training, validation, and testing. The training set has a spatial resolution of 64×64 for LRMS, PAN, and GT, while 16×16 for MS. The reduced-resolution testing dataset adopts 256×256 for LRMS, PAN, and GT, and 64×64 for MS. In contrast, the full-resolution dataset employs 512×512 for LRMS and PAN, and 128×128 for MS. More details about the datasets can refer to [3].

Metrics. In our experiments, we employ the spectral angle mapper (SAM) [28], the dimensionless global error in synthesis (ERGAS) [23], the Q2n (Q8 for 8-band datasets and Q4 for 4-band datasets) [6], and the peak signal to noise ratio (PSNR) indicators for reduced-resolution evaluation. Additionally, for full-resolution assessment, we incorporate three non-reference metrics: the hybrid quality with no reference (HQNR) index, the spectral distortion D_λ index, and spatial distortion D_s index [20].

Experimental Settings. All deep learning models are implemented using PyTorch, trained on a single NVIDIA RTX 4090 GPU. We employ the Adam [12] algorithm with beta values of (0.9, 0.999) and weight decay of 0.1 for model training. The minibatch size is 32, and the initial learning rate is 3×10^{-4} . The learning rate decay is applied

by multiplying 0.1 at 300 and 500 epochs, with training concluding after 800 epochs. In all experiments, the hyperparameter λ in the loss function is fixed at 0.1, and we utilize 5 LFormer modules.

2 Hyperspectral Image Super-resolution

Dataset Simulation. To further evaluate the robustness and adaptability of our approach, we apply it to another analogous application, *i.e.*, the hyperspectral image super-resolution (HISR). We conduct the experiment on the widely used CAVE dataset¹. The raw CAVE dataset includes 32 hyperspectral images (HSIs) with a size of $512 \times 512 \times 31$. In the experiment, we follow the protocols of existing methods [4, 11, 18], thus using 20 images for training and validation, while 11 images for testing. Like the pan-sharpening task, the ground truth images (GTs), *i.e.*, HR-HSIs, are usually unavailable in HISR. Therefore, we need to simulate the LR-HSIs and high-resolution multispectral images (HR-MSIs). According to Wald's protocol [24], we view the original HSIs as HR-HSIs (*i.e.*, GTs) and simulate the LR-HSIs and HR-MSIs from them. Specifically, we first crop the original HSIs into 3920 overlapping patches with a size of $64 \times 64 \times 31$, respectively. Next, we employ a Gaussian blur with a kernel of 3×3 and a standard deviation of 0.5 to HR-HSIs and then conduct $4 \times$ bicubic downsampling to obtain the LR-HSIs with a size of $16 \times 16 \times 31$. In addition, we also generate the paired RGB images (*i.e.*, HR-MSIs) with a size of $64 \times 64 \times 3$ using a general spectral response function R of Nikon D700 camera [4, 11, 18]. Finally, we obtain 3920 training patches, including LR-HSIs, HR-HSIs (GTs), and HR-MSIs, which are divided into two parts: 1) training set (90%) and 2) validation set (10%). Note that we use the original 11 testing examples to evaluate the trained model.

HISR Benchmark. For the HISR task, we also select three traditional methods including LTMR [5], MTF-HS [22], UTV [26]; and seven deep-learning based methods: ResTFNet [13], SSRNet [29],

*Equal Contribution.

[†]Corresponding authors.

¹<https://www.cs.columbia.edu/CAVE/databases/multispectral/>.

Fusformer [8], HSRNet [9], U2Net [17], HyperTransformer [1], DHIF [10] for comparison purpose. All comparison networks are trained using the same methodology. Moreover, the related hyperparameters are selected consistent with the original papers.

3 Additional Visual Results

We provide more visual results to further demonstrate the superiority of our LFormer. Figure 1 and Figure 2 show the qualitative outcomes of all compared methods and ground truth (GT) on a reduced-resolution WV3 example and GF2 example, respectively. To be specific, the first two rows display the RGB visualization, while the last two rows illustrate the error maps between the fused images and GT. It is clearly observed that the images yielded by our model achieve smaller differences with GT as indicated by the dark blue residual maps.

Figure 3 demonstrates the visual results of all compared techniques on a real-world full-resolution GF2 sample. The products generated by our model exhibit clear edges and textures while preserving realistic spectral information, as clearly depicted in the enlarged areas. As displayed in the third row of Figure 3, moreover, our model showcases a deep hot HQNR map with sparse bright spots, consistent with the higher HQNR score reported in quantitative measurement, further substantiating its comprehensive spatial and spectral reconstruction quality.

Figure 4 gives the RGB visualization and the corresponding residual maps of all approaches on a CAVE case. Again, our model shows the desired textures and colors, and lower deviations from the GT, as illustrated by its dark blue residual map.

Figure 5 provides the feature maps of different LFormer layers on a testing example from the CAVE dataset, the WV3 dataset, and the GF2 dataset, respectively. It is evident that the feature maps display differing levels of granularity across various layers. Moreover, the detail information becomes increasingly discriminative throughout the linearly-evolved process.

Figure 6 illustrates the feature maps of the proposed LFormer (i.e., Baseline) and its two variants over the GF2 dataset, corresponding to the ablation experiment: "Effect of Attention Evolution" in section 5. Although the feature maps of the Config.I with dense cross-attention computations, resulting in significant computational overhead, can capture global information to some extent, they struggle to learn more fine-grained local details. While the feature maps of the Config.II not only fail to represent the long-range feature information but also exhibit blurry texture details. In contrast, the feature maps of the baseline can effectively model the global feature dependencies and demonstrate more intricate details compared to the two variants as the block number increases.

References

- [1] Wele Gedara Chaminda Bandara and Vishal M Patel. 2022. HyperTransformer: A textural and spectral feature fusion transformer for pansharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1767–1777.
- [2] Liang-Jian Deng, Gemine Vivone, Cheng Jin, and Jocelyn Chanussot. 2020. Detail injection-based deep convolutional neural networks for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8 (2020), 6995–7010.
- [3] Liang-jian Deng, Gemine Vivone, Mercedes E. Paoletti, Giuseppe Scarpa, Jiang He, Yongjun Zhang, Jocelyn Chanussot, and Antonio Plaza. 2022. Machine Learning in Pansharpening: A benchmark, from shallow to deep networks. *IEEE Geoscience and Remote Sensing Magazine* 10, 3 (2022), 279–315. <https://doi.org/10.1109/MGRS.2022.3187652>
- [4] Shang-Qi Deng, Liang-Jian Deng, Xiao Wu, Ran Ran, Danfeng Hong, and Gemine Vivone. 2023. PSRT: Pyramid Shuffle-and-Reshuffle Transformer for Multispectral and Hyperspectral Image Fusion. *IEEE Transactions on Geoscience and Remote Sensing* (2023).
- [5] Renwei Dian and Shutao Li. 2019. Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization. *IEEE Transactions on Image Processing* 28, 10 (2019), 5135–5146.
- [6] Andrea Garzelli and Filippo Nencini. 2009. Hypercomplex quality assessment of multi/hyperspectral images. *IEEE Geoscience and Remote Sensing Letters* 6, 4 (2009), 662–665.
- [7] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 4 (2019), 1188–1204.
- [8] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Hong-Xia Dou, Danfeng Hong, and Gemine Vivone. 2022. Fusformer: A transformer-based fusion network for hyperspectral image super-resolution. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5.
- [9] Jin-Fan Hu, Ting-Zhu Huang, Liang-Jian Deng, Tai-Xiang Jiang, Gemine Vivone, and Jocelyn Chanussot. 2021. Hyperspectral image super-resolution via deep spatio-spectral attention convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 33, 12 (2021), 7251–7265.
- [10] Tao Huang, Weisheng Dong, Jinjian Wu, Leida Li, Xin Li, and Guangming Shi. 2022. Deep hyperspectral image fusion network with iterative spatio-spectral regularization. *IEEE Transactions on Computational Imaging* 8 (2022), 201–214.
- [11] Sen Jia, Zhichao Min, and Xiyu Fu. 2023. Multiscale spatial-spectral transformer network for hyperspectral and multispectral image fusion. *Information Fusion* 96 (2023), 117–129.
- [12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [13] Xiangyu Liu, Qingjie Liu, and Yunhong Wang. 2020. Remote sensing image fusion based on two-stream fusion network. *Information Fusion* 55 (2020), 1–15.
- [14] Simone Lolli, Luciano Alparone, Andrea Garzelli, and Gemine Vivone. 2017. Haze correction for contrast-based multispectral pansharpening. *IEEE Geoscience and Remote Sensing Letters* 14, 12 (2017), 2255–2259.
- [15] Giuseppe Masi, Davide Cozzolino, Luisa Verdoliva, and Giuseppe Scarpa. 2016. Pansharpening by convolutional neural networks. *Remote Sensing* 8, 7 (2016), 594.
- [16] Xiangchao Meng, Nan Wang, Feng Shao, and Shutao Li. 2022. Vision transformer for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11.
- [17] Siran Peng, Chenhao Guo, Xiao Wu, and Liang-Jian Deng. 2023. U2Net: A General Framework with Spatial-Spectral-Integrated Double U-Net for Image Fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3219–3227.
- [18] Ran Ran, Liang-Jian Deng, Tai-Xiang Jiang, Jin-Fan Hu, Jocelyn Chanussot, and Gemine Vivone. 2023. GuidedNet: A general CNN fusion framework via high-resolution guidance for hyperspectral image super-resolution. *IEEE Transactions on Cybernetics* (2023).
- [19] Gemine Vivone. 2019. Robust band-dependent spatial-detail approaches for panchromatic sharpening. *IEEE transactions on Geoscience and Remote Sensing* 57, 9 (2019), 6421–6433.
- [20] Gemine Vivone, Luciano Alparone, Jocelyn Chanussot, Mauro Dalla Mura, Andrea Garzelli, Giorgio A Licciardi, Rocco Restaino, and Lucien Wald. 2014. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing* 53, 5 (2014), 2565–2586.
- [21] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. 2018. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing* 27, 7 (2018), 3418–3431.
- [22] Gemine Vivone, Rocco Restaino, and Jocelyn Chanussot. 2018. Full scale regression-based injection coefficients for panchromatic sharpening. *IEEE Transactions on Image Processing* 27, 7 (2018), 3418–3431.
- [23] Lucien Wald. 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.
- [24] Lucien Wald, Thierry Ranchin, and Marc Mangolini. 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. *Photogrammetric engineering and remote sensing* 63, 6 (1997), 691–699.
- [25] Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, and Tian-Jing Zhang. 2021. Dynamic cross feature fusion for remote sensing pansharpening. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14687–14696.
- [26] Ting Xu, Ting-Zhu Huang, Liang-Jian Deng, Xi-Le Zhao, and Jie Huang. 2020. Hyperspectral image superresolution using unidirectional total variation with Tucker decomposition. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 4381–4398.
- [27] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei Zhang. 2018. A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpener. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11, 3 (2018), 978–989. <https://doi.org/10.1109/JSTARS.2018.2819188>

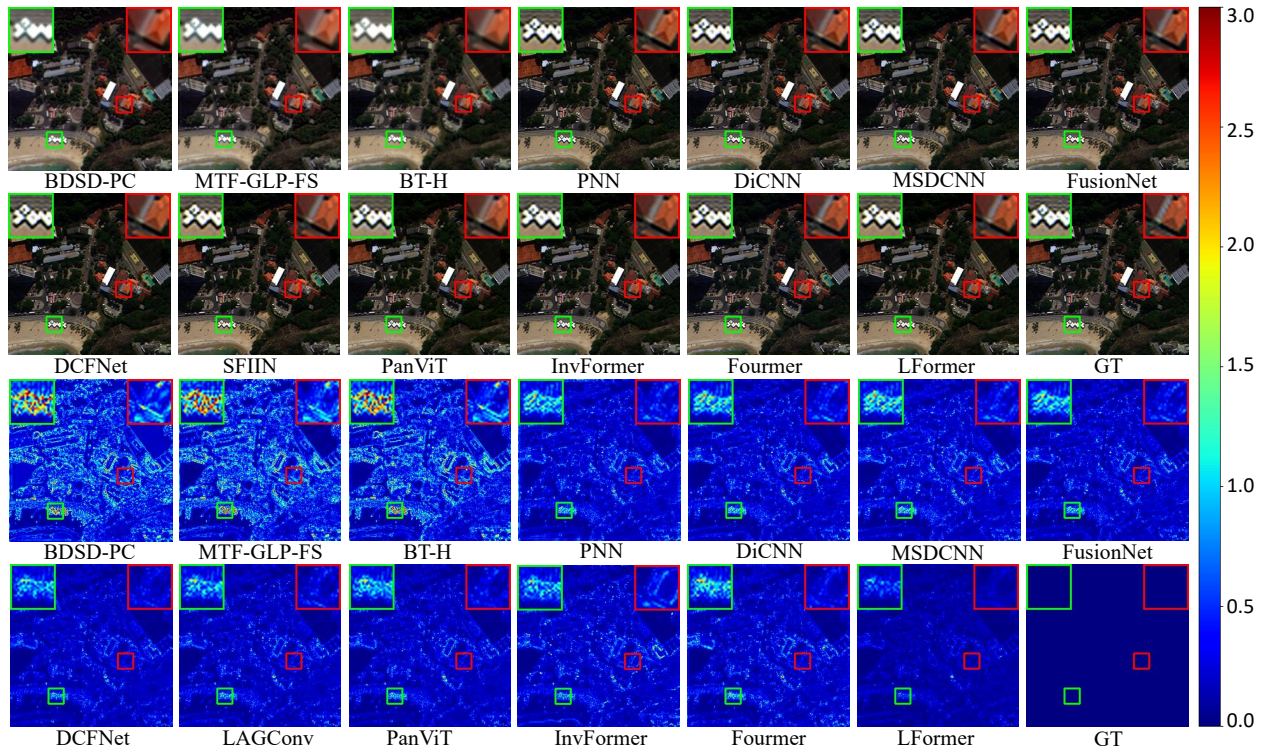


Figure 1: RGB visualization for both our model and other state-of-the-art approaches over reduced resolution WV3 Dataset.

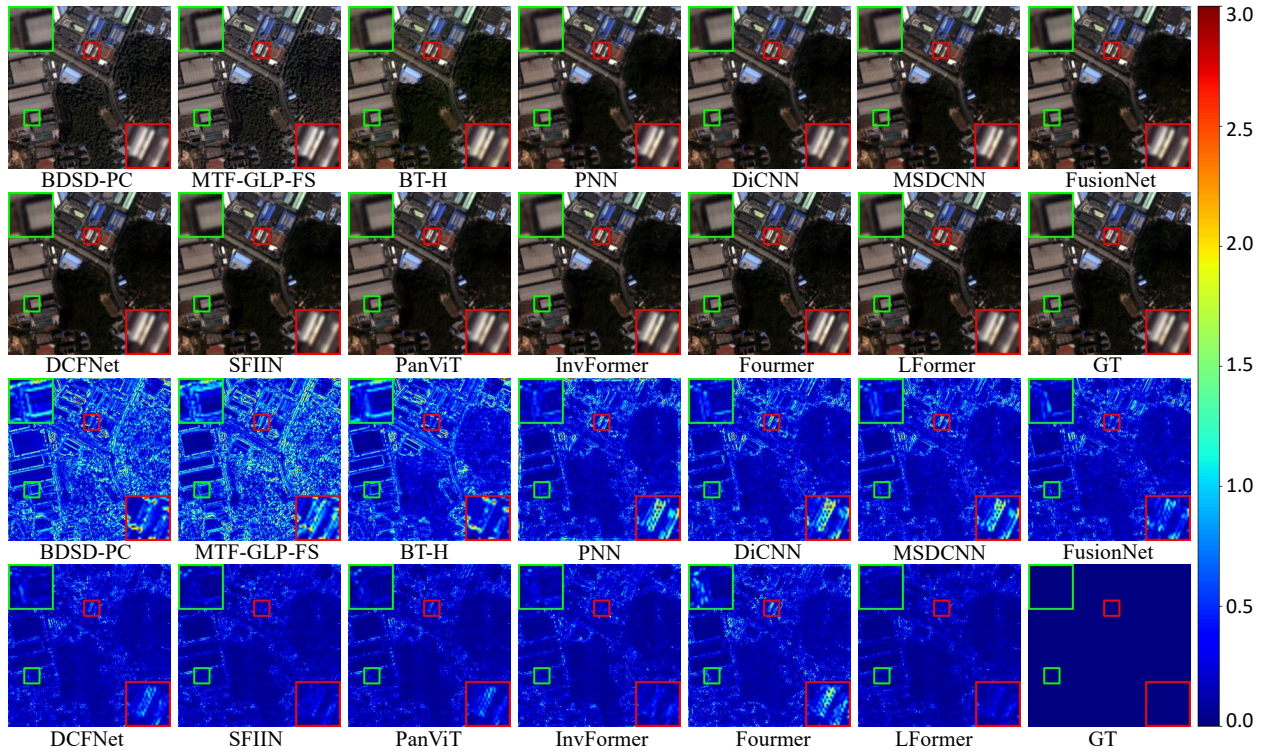


Figure 2: RGB visualization for both our model and other state-of-the-art approaches over reduced resolution GF2 Dataset.

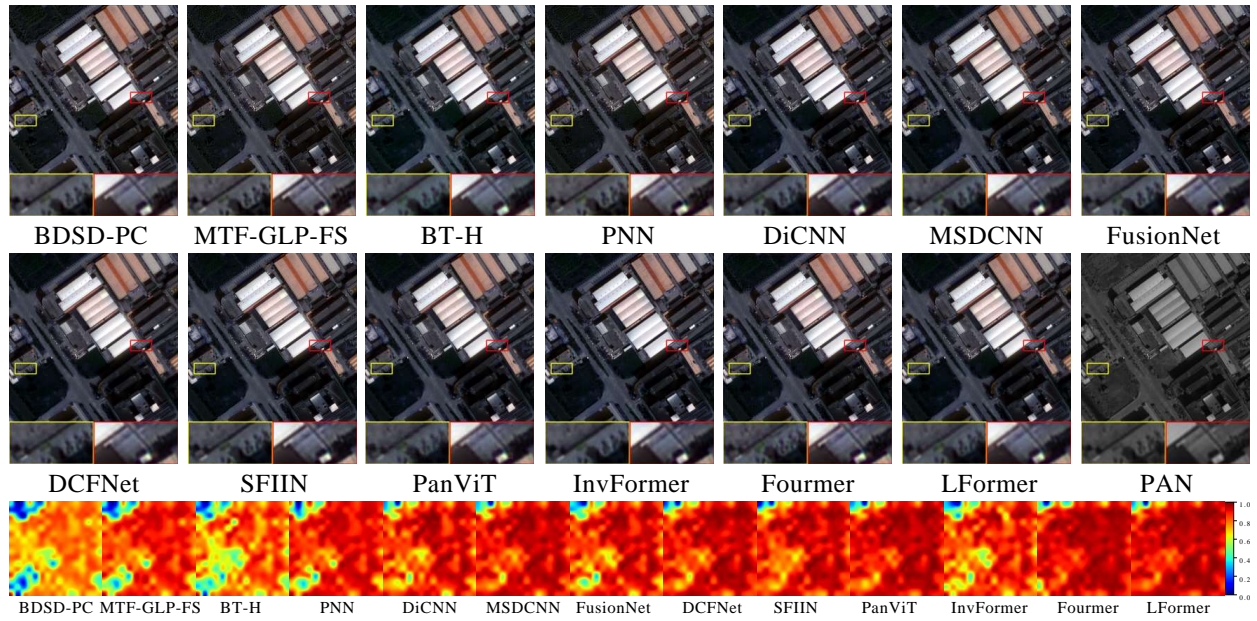


Figure 3: RGB visualization and HQNR map for both our model and other state-of-the-art approaches over real-world full resolution GF2 Dataset.

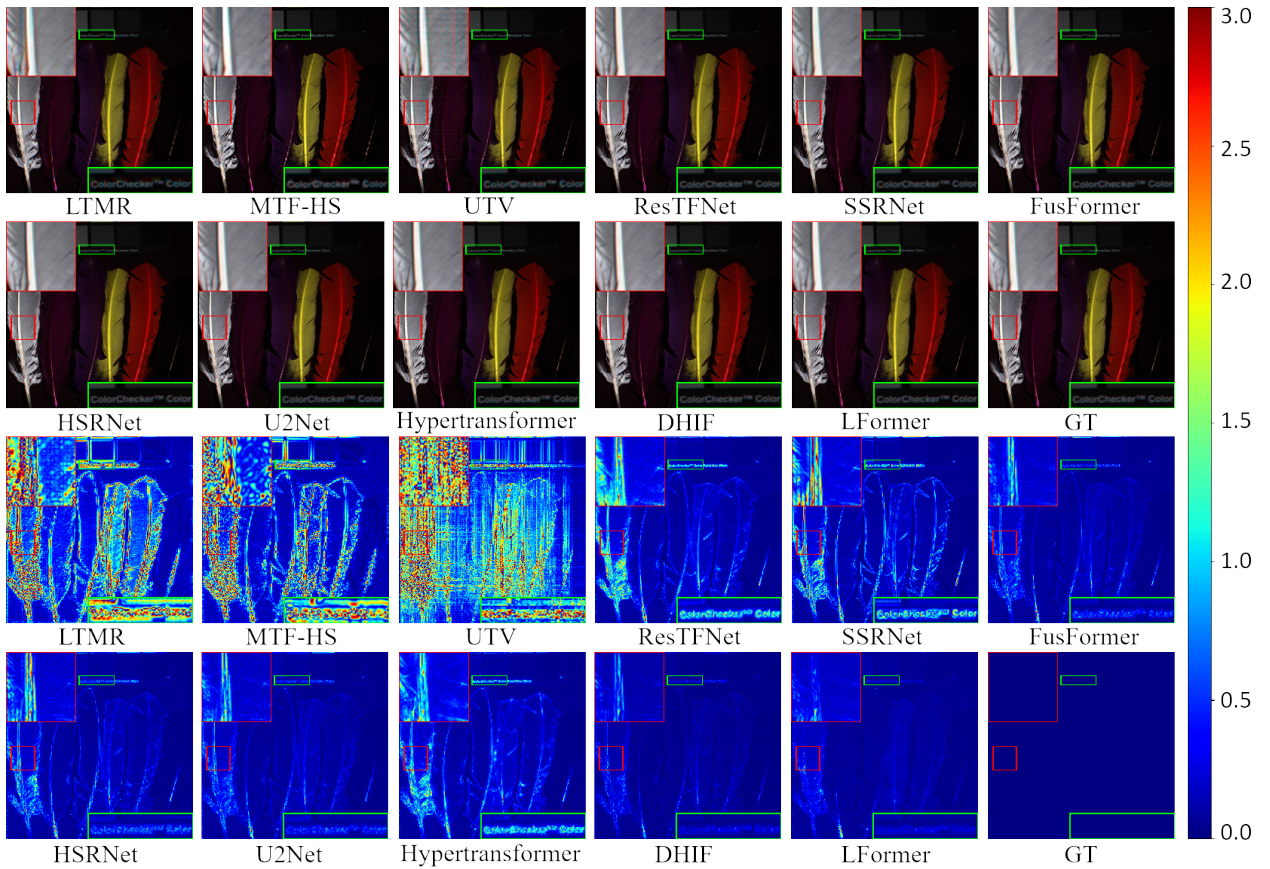


Figure 4: RGB visualization for both our model and other state-of-the-art approaches over the CAVEx4 Dataset.

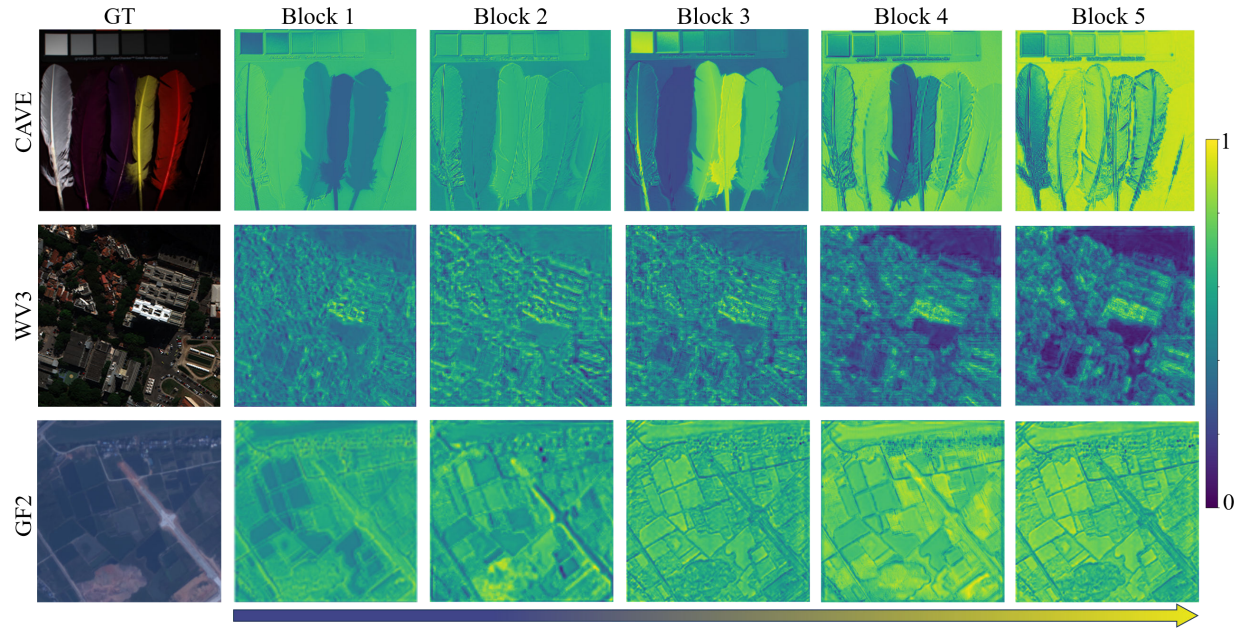


Figure 5: Additional feature map visualization. The first row, the second row and the third row depict the feature maps derived from various LFormer blocks over CAFE dataset, WV3 dataset, and GF2 dataset, respectively. It is evident that as the block number increases, all three groups of feature maps display progressively clearer contours and more intricate details, underscoring the potential effectiveness of the proposed key linear evolution design.

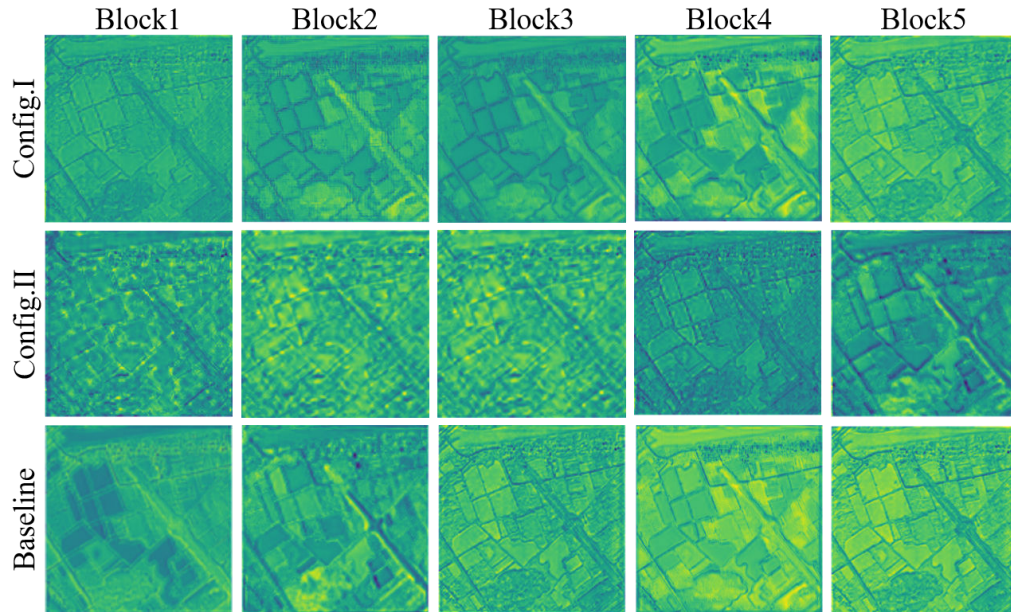


Figure 6: Feature map visualization on the "Effect of Attention Evolution" in section 5. It can be observed that the feature maps of the baseline exhibits more intricate details compared to the two variants as the block number increases.

- [//doi.org/10.1109/JSTARS.2018.2794888](https://doi.org/10.1109/JSTARS.2018.2794888)
- [28] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman. 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*.
- [29] Xueting Zhang, Wei Huang, Qi Wang, and Xuelong Li. 2020. SSR-NET: Spatial-spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2020), 5953–5965.
- [30] Man Zhou, Jie Huang, Yanchi Fang, Xueyang Fu, and Aiping Liu. 2022. Pan-sharpening with customized transformer and invertible neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3553–3561.
- [31] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. 2023. Fourmer: An efficient global modeling paradigm for image restoration. In *International Conference on Machine Learning*. PMLR, 42589–42601.
- [32] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. 2022. Spatial-frequency domain information integration for pan-sharpening. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*. Springer, 274–291.