Two-Stage Transformer-based Deep Hyperspectral and Multispectral Image Fusion Network for Hyperspectral Image Super-Resolution

Wo-Yen Li[1], Chia-Ming Lee[2], Chih-Chung Hsu[3], Volodymyr Khylenko[4], and Li-Wei Kang[1,5,*]

[1]Department of Electrical Engineering, National Taiwan Normal University, Taipei, Taiwan
[2]Institute of Data Science, and Miin Wu School of Computing, National Cheng Kung University, Tainan, Taiwan
[3]Institute of Intelligent Systems, College of Artificial Intelligence, National Yang Ming Chiao Tung University, Tainan, Taiwan
[4]Department of Informatics and Information Technologies, Slovak University of Technology in Bratislava, Bratislava, Slovak
[5]Graduate Institute of AI Interdisciplinary Applied Technology, National Taiwan Normal University, Taipei, Taiwan
[*]E-mail: lwkang@ntnu.edu.tw

***Abstract*—Hyperspectral images (HSIs), which capture detailed spectral information per pixel, are widely used in fields like forestry, satellite imaging, medicine, and hydrology. However, acquiring high-resolution HSIs (HR-HSIs) directly is limited by hardware constraints and computational demands. A common workaround is to fuse high-resolution multispectral images (HR-MSI) with low-resolution HSIs (LR-HSI) to achieve super-resolution (SR) of the LR-HSI. In this paper, we propose SwinDFN, a two-stage deep fusion network based on Swin Transformers, designed to combine HR-MSI and LR-HSI to generate high-quality HR-HSIs. The first stage performs multi-scale feature fusion, while the second stage refines the result using residual Swin Transformer blocks (RSTB). Experiments and ablation studies confirm the effectiveness of each module, and both quantitative and qualitative results show that SwinDFN outperforms existing state-of-the-art methods.**

I. INTRODUCTION

Hyperspectral imaging aims to capture dense spectral information at each pixel of a scene, playing a vital role in the field of remote sensing. Unlike traditional digital images, hyperspectral images (HSIs) provide spectral data across a wide range of wavelengths, encompassing tens to hundreds of spectral bands [1]. This rich spectral information makes HSIs suitable for a variety of applications, including vegetation monitoring, water resource management, soil surveying, and geological inspection. These applications benefit from the ability of HSIs to effectively distinguish between materials, enabling accurate image classification, target detection, and anomaly detection [2]-[4]. While high spatial resolution HSIs can further enhance the performance of these applications, their spatial resolution is often limited due to the hardware constraints of current hyperspectral sensing devices [5].

To obtain a high-resolution hyperspectral image (HR-HSI), a common approach is to separately acquire a low-resolution hyperspectral image (LR-HSI) and a high-resolution multispectral image (HR-MSI), then fuse them to generate the desired HR-HSI. Traditional methods typically employ pan-sharpening techniques, which include component substitution (CS)-based and multi-resolution analysis (MRA)-based approaches. CS-based methods (e.g., [6]) decompose the captured LR-HSI into spatial and spectral components, replacing the spatial component with the corresponding HR-MSI to reconstruct the HR-HSI. On the other hand, MRA-based methods (e.g., [7]) apply multi-scale decomposition to extract spatial details from the HR-MSI, which are then integrated into each spectral band of the LR-HSI. However, both CS and MRA approaches may introduce significant spectral distortions during the injection of HR-MSI spatial information into the LR-HSI. Moreover, model-based methods [8], [9], which are designed using mathematical models and image priors, utilize matrix factorization and tensor representations to fuse information from LR-HSI and HR-MSI. However, the fusion performance of these methods may degrade when real-world scenarios do not align well with the assumed image priors, such as sparsity or low-rank structures.

With the rapid development of deep learning techniques and their success in various perceptual tasks, such as image classification [10], object detection [11], [12], and image restoration [13], [14], numerous deep learning-based image fusion frameworks have been proposed for fusing HR-MSI and LR-HSI [15]–[26]. These methods have demonstrated significantly better performance in SR of LR-HSI compared to traditional approaches. For instance, deep learning-based pan-sharpening frameworks, such as [15], have led to notable improvements over earlier pan-sharpening-based methods.

However, most existing deep learning-based approaches utilize the concatenation of HR-MSI and LR-HSI along the spectral channel dimension as input to the deep network, which may not fully exploit the underlying spatial information. Moreover, these methods often struggle to achieve satisfactory SR performance for input LR-HSIs, particularly in error-prone

979-8-3315-7206-8/25/$31.00

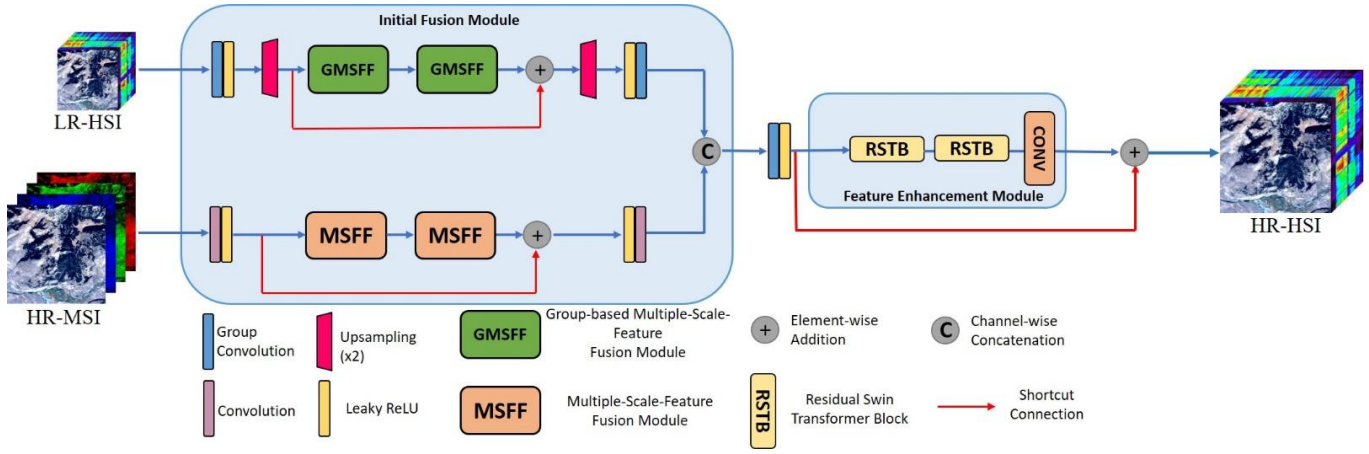

Fig. 1. Proposed SwinDFN, a two-stage deep fusion network based on Swin Transformer for HR-MSI and LR-HSI fusion.

environments, and may also fail to maintain low model complexity and run-time efficiency.

To address these challenges, this paper proposes a deep learning-based framework for fusing HR-MSI with LR-HSI to generate a SR version of the input LR-HSI. The framework, termed SwinDFN (Swin-based Deep Fusion Network), is a hybrid model that integrates a shallow convolutional neural network (CNN) with a Swin Transformer-based architecture. The proposed network operates in two stages. In the first stage, a lightweight shallow CNN performs an initial fusion of the input HR-MSI and LR-HSI. In the second stage, a transformer-based model, built on residual Swin Transformer blocks (RSTB) [27], extracts deep features to refine and enhance the fusion results, ultimately producing the final HR-HSI. The main novelties and contributions of this paper are three-fold: *i*) we propose a two-stage deep network that combines a shallow CNN with an RSTB-based transformer to address the HR-MSI/LR-HSI fusion problem; *ii*) in the first stage, a lightweight and shallow CNN performs multi-scale feature fusion to generate an effective initial result. In the second stage, a simple RSTB-based feature enhancement module further refines this result to produce the final HR-HSI output.; and *iii*) the proposed deep fusion framework demonstrates superior or competitive performance compared to state-of-the-art methods, both quantitatively and qualitatively, particularly in challenging, error-prone scenarios. Moreover, it achieves strong results without requiring extensive spectral information from the HR-MSI bands.

The remainder of this paper is organized as follows. Section II provides a brief review of state-of-the-art deep learning-based HR-MSI/LR-HSI fusion frameworks for SR of LR-HSI. Section III introduces the proposed SwinDFN framework. Section IV presents the experimental results. Finally, Section V concludes the paper.

II. Related Work

Several deep learning frameworks have been proposed for HR-MSI/LR-HSI fusion. In [16], a framework combining spatio-spectral regularization with a physical imaging model was introduced, featuring a model-guided unfolding network (DHIF-Net) to iteratively optimize spatio-spectral regularization. In [17], the progressive zero-centric residual network (PZRes-Net), a CNN-based approach, was proposed to perform multi-scale feature decomposition for information fusion. Similarly, [18] presented MSSJFL, a multi-scale spatial-spectral joint feature learning framework, to extract spatial, spectral, and joint features. Based on the UNet architecture, [19] proposed D-UNet, which incorporates multistage detail injection using a detail extraction and a spatio-spectral fusion network. A multi-resolution detail-enhanced dual-UNet was also introduced in [20], leveraging UNet's multi-scale capability for feature fusion. More recently, [21] presented Fusformer, a transformer-based fusion framework utilizing attention mechanisms for improved HR-MSI/LR-HSI integration. More advanced deep learning-based fusion frameworks are discussed in [22]-[26].

III. Proposed SwinDFN for HR-MSI/LR-HSI Fusion

A. Problem Formulation and Overview of SwinDFN

Similar to most studies on LR-HSI/HR-MSI fusion, let the target HR-HSI be denoted as Y. We assume the observed image data include an LR-HSI, represented as $X_h = \mathrm{YB}$, where B is a blurring matrix that reduces the spatial resolution of Y, and an HR-MSI, represented as $X_m = \mathrm{DY}$, where D is a down-sampling matrix that reduces the spectral resolution of Y. Therefore, the problem to be addressed in this paper can be formulated as follows: Given the observed LR-HSI $X_h \in \mathbb{R}^{h_h \times w_h \times b}$ and HR-MSI $X_m \in \mathbb{R}^{h \times w \times b_m}$, the goal is to reconstruct the HR-HSI $Y^* \in \mathbb{R}^{h \times w \times b}$, expressed as:

$$Y^* = SwinDFN(X_h, X_m, \theta), \quad (1)$$

where $h_h \times w_h$ and $h \times w$ are the spatial resolutions of a single band of LR-HSI and HR-MSI, respectively, b and b_m are the number of spectral bands (channels) in LR-HSI and HR-MSI, respectively, satisfying $h_h < h$, $w_h < w$, and $b > b_m$. $SwinDFN$ denotes the proposed LR-HSI/HR-MSI fusion network model, and θ represents its network parameters.

To address this problem, as illustrated in Fig. 1, we propose a transformer-based HR-MSI/LR-HSI fusion framework built on RSTBs [27]. Each RSTB consists of multiple Swin Transformer Layers (STLs), which are primarily responsible for computing local attention and propagating features through cross-window interactions.

B. Initial Fusion Module

The proposed SwinDFN framework comprises two main stages. In the first stage, the Initial Fusion Module (IFM) performs preliminary upsampling and feature extraction on the LR-HSI, followed by fusion with the features extracted from the HR-MSI. For the input HR-MSI (X_m), we first apply a convolution operation followed by a Leaky ReLU activation [28]. The result is then fed into our Multi-Scale Feature Fusion (MSFF) module, where MSFF extracts and fuses features at different scales using convolutions with multiple receptive fields. The features extracted by two MSFF modules are added element-wise with the features propagated through a shortcut (residual) connection. This sum is then passed through another Leaky ReLU activation and a convolution operation to obtain the final HR-MSI feature representation, denoted as Z_{X_m}.

For the input LR-HSI (X_h), considering a case where $h = 4h_h$ and $w = 4w_h$, we first apply grouped convolution (which is more efficient due to the large number of spectral bands) followed by Leaky ReLU activation. Then, a 2× upsampling is performed. The result is passed through two consecutive group-based MSFF (GMSFF) modules for feature extraction, with shortcut connections in between. This is followed by another 2× upsampling to match the spatial resolution of the HR-MSI. Finally, the Leaky ReLU activation and grouped convolution are applied to obtain the LR-HSI feature representation, denoted as Z_{X_h}. Next, Z_{X_m} and Z_{X_h} are concatenated and processed using grouped convolution followed by Leaky ReLU activation. This operation is denoted as:

$$\mathrm{IFM}(X_m, X_h) = \mathrm{Conv.}\left(\mathrm{Cat.}\left(Z_{X_m}, Z_{X_h}\right)\right), \quad (2)$$

where $\mathrm{Conv.}$ represents the grouped convolution followed by Leaky ReLU.

C. Feature Enhancement Module

The initially fused result, $\mathrm{IFM}(X_m, X_h)$, is then fed into the Feature Enhancement Module (FEM) in the second stage. FEM primarily consists of two RSTBs based on the Swin Transformer [27], which are used to further enhance the fused features from the HR-MSI and LR-HSI and generate the final HR-HSI output. The main motivation for using RSTBs is that both HR-MSI and LR-HSI contain rich inter-band correlations and intra-band local self-correlations. These correlations can significantly benefit feature extraction, fusion, and the reconstruction of the corresponding HR-HSI. Therefore, we leverage the powerful self-attention and cross-attention capabilities of the Transformer architecture. The reconstructed HR-HSI Y^* is given by:

$$Y^* = \mathrm{CONV}\left(\mathrm{RSTB}^2\left(\mathrm{IFM}(X_m, X_h)\right)\right) \oplus \mathrm{IFM}(X_m, X_h), \quad (3)$$

where CONV denotes a convolution operation, RSTB^2 represents two consecutive RSTB modules, and $\oplus$ indicates element-wise addition between the convolution output and the shortcut connection from the initial fused features.

D. Model Learning

To train the proposed SwinDFN model, the loss function we use is defined as follows:

$$L(Y^*, \mathrm{Y}) = \lambda_1 L_1(Y^*, \mathrm{Y}) + \lambda_2 L_{SAM}(Y^*, \mathrm{Y}), \quad (4)$$

where Y is the ground truth, λ_1 and λ_2 are the weighting coefficients, empirically set to 1 and 0.1, respectively, L_1 is the ℓ_1 loss, and L_{SAM} is the Spectral Angle Mapper (SAM) loss, defined as follows (SAM was originally proposed in [29]):

$$L_{SAM} = 1 - \frac{1}{hw}\sum_{n=1}^{hw}\left(\frac{\mathrm{Y}_n^{\mathrm{T}} Y_n^*}{|\mathrm{Y}_n|_2 \cdot |Y_n^*|_2 + \epsilon}\right), \quad (5)$$

where Y_n and Y_n^* denote the n-th spectral vector of the ground truth Y and the reconstructed output Y^*, respectively, and ϵ is an error term. The SAM loss is primarily based on computing the cosine similarity between the reconstructed spectral vector and its corresponding ground truth.

IV. Experimental Results

A. Parameter Settings and Network Training

The dataset used in this study was collected by the Airborne Visible Infrared Imaging Spectrometer (AVIRIS) sensor [30], comprising 2,078 HR-HSIs. These were randomly split into training (1,678 images), validation (200 images), and testing (200 images) sets. The spatial and spectral resolutions were 256×256×b_m for each HR-MSI and 64×64×172 for each LR-HSI, where b_m is either 4 or 6 in our experiments. The proposed SwinDFN was implemented using the PyTorch

Table I. Performance Evaluation of the proposed SwinDFN and other fusion models.

Model	Model Complexity		Fusion Performance for 4 MSI Bands				Fusion Performance for 6 MSI Bands			
	#Params	FLOPs	PSNR↑	SAM↓	ERGAS↓	RMSE↓	PSNR↑	SAM↓	ERGAS↓	RMSE↓
D-UNet [19]	2.97M	88.65G	35.423	1.892	1.796	33.183	38.453	1.548	1.205	26.148
PZRes-Net [17]	40.15M	5262.34G	34.963	1.934	1.935	35.498	37.427	1.478	1.538	28.234
MSSJFL [18]	16.33M	175.56G	34.966	1.792	2.245	33.636	38.006	1.390	1.535	26.893
DHIF-Net [16]	57.04M	13795.11G	34.458	1.829	2.613	34.769	39.146	1.239	1.113	25.309
FusFormer [21]	0.18M	11.74G	34.217	2.012	1.996	35.687	38.637	1.678	1.204	28.674
QRCODE [23]	41.88M	2231.19G	35.361	1.623	2.027	32.711	38.948	1.148	1.429	24.617
Proposed SwinDFN	**20.12M**	**809.8G**	**36.682**	**1.369**	**1.728**	**28.178**	**39.856**	**1.055**	**1.240**	**22.803**

Table II. Performance evaluation of the proposed SwinDFN and other fusion models on 6-band MSI with AWGN noise at varying SNR levels.

Model	SNR = 45	SNR = 35	SNR = 25
	PSNR↑	PSNR↑	PSNR↑
D-UNet [19]	33.863	26.935	17.810
PZRes-Net [17]	34.455	28.772	19.865
MSSJFL [18]	36.516	31.768	23.417
DHIF-Net [16]	35.312	27.424	18.697
FusFormer [21]	34.400	26.409	19.183
QRCODE [23]	35.468	30.157	21.757
Proposed SwinDFN	**39.850**	**39.795**	**39.238**

framework. Training was conducted with a batch size of 4 for 600 epochs across all experiments. For peer methods, training epochs followed the default settings in their original publications. The ADAM optimizer [31] was used with an initial learning rate of 0.0001, which was adjusted during training using a cosine annealing scheduler.

B. Performance Evaluation

To evaluate the performance of our SwinDFN, we compared it against six supervised hyperspectral image fusion methods: DHIF-Net [16], PZRes-Net [17], MSSJFL [18], D-UNet [19], FusFormer [21], and QRCODE [23]. The evaluation utilized four metrics: PSNR (Peak Signal-to-Noise Ratio in dB), SAM, RMSE (Root Mean Squared Error), and ERGAS (Erreur Relative Globale Adimensionnelle de Synthèse) [32]. Experiments were conducted using both 4-band and 6-band MSI configurations. Table I and Figs. 2 and 3 present the quantitative and qualitative HR-MSI/LR-HSI fusion results of the proposed SwinDFN in comparison with these state-of-the-art methods. Furthermore, to assess the robustness of SwinDFN under Additive White Gaussian Noise (AWGN), Table II reports the fusion performances of all evaluated models across different Signal-to-Noise Ratios (SNRs). As shown in Tables I and II and Figs. 2 and 3, the proposed SwinDFN consistently outperforms the baseline models in both quantitative and qualitative metrics, under both ideal and noisy conditions. On the other hand, Table I also presents the model complexity of the evaluated models in terms of the number of parameters (#Params) and floating-point operations (FLOPs).

Table III. Ablation study evaluating the performance of the proposed SwinDFN with and without the FEM on 4-band MSI.

Model	PSNR↑	SAM↓	ERGAS↓	RMSE↓
SwinDFN without FEM	35.875	1.509	1.921	30.684
SwinDFN with FEM	36.682	1.369	1.728	28.178

C. Ablation Study

To evaluate the effectiveness of the proposed Feature Enhancement Module (FEM) based on the Swin Transformer, we conduct an ablation study by comparing the performance of our SwinDFN model with and without the FEM. The fusion results are presented in Table III. As shown in Table III, the integration of the proposed FEM into our SwinDFN significantly enhances the overall fusion performance.

V. CONCLUSIONS

In this paper, we have proposed a hybrid deep learning model called SwinDFN (Swin Transformer-based Deep hyperspectral and multispectral image Fusion Network) for fusing HR-MSI and LR-HSI to generate HR-HSI. SwinDFN combines a shallow CNN with multi-receptive fields for multi-scale feature fusion, and a Transformer network based on Residual Swin Transformer Blocks (RSTB) with a multi-head attention mechanism. The shallow CNN captures both local and global image features to produce an initial fused result, which is then refined by the RSTB-based Transformer using a residual structure to generate the final HR-HSI.

Based on our experiments comparing SwinDFN with state-of-the-art deep learning-based fusion methods, SwinDFN demonstrates superior qualitative performance in reconstructing both fine local details and overall global consistency. Quantitatively, it also outperforms existing methods, particularly in error-prone environments. Furthermore, SwinDFN shows notable advantages when the input HR-MSI contains a limited number of bands (e.g., 4

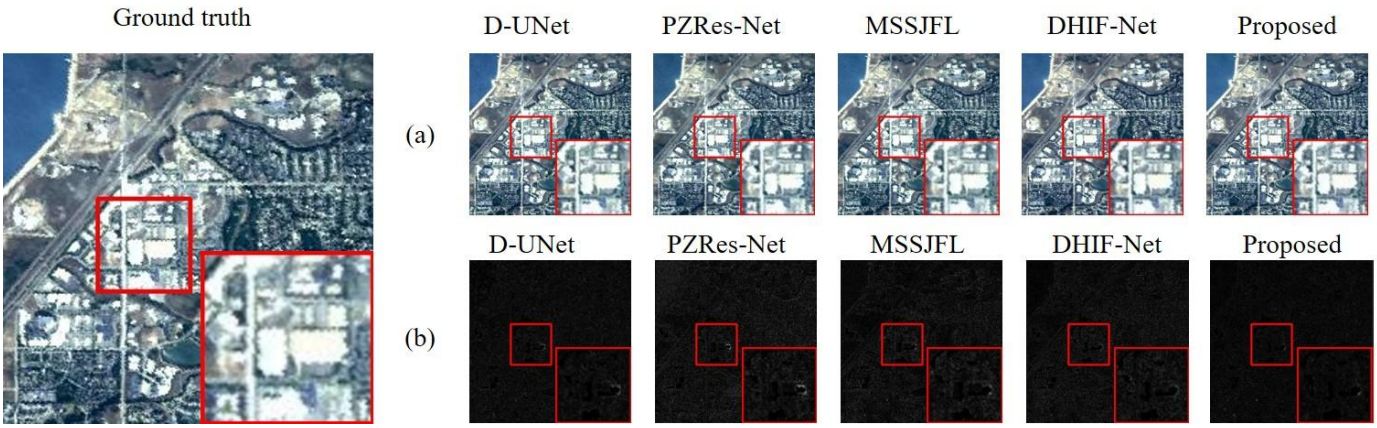

Fig. 2. Qualitative results of HR-MSI/LR-HSI image fusion: (a) SR images generated by different models, with key regions highlighted; (b) corresponding difference images between each SR result and the ground truth, illustrating reconstruction errors across models.

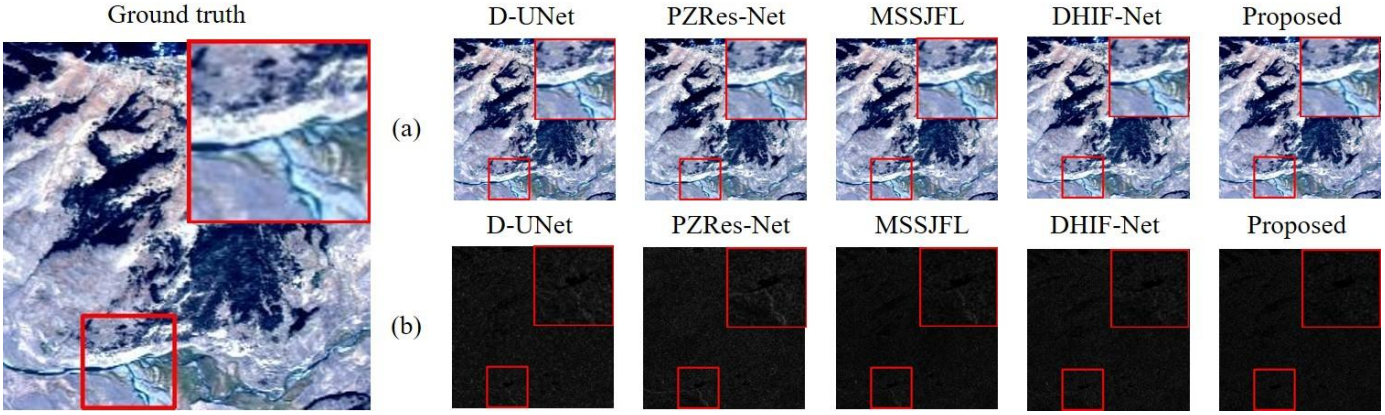

Fig. 3. Qualitative results of HR-MSI/LR-HSI image fusion: (a) SR images generated by different models, with key regions highlighted; (b) corresponding difference images between each SR result and the ground truth, illustrating reconstruction errors across models.

bands), and continues to perform well as the number of bands increases (e.g., 6 bands). This indicates that SwinDFN achieves good super-resolution performance for LR-HSI without relying on extensive spectral information from HR-MSI. Additionally, our ablation studies confirm the effectiveness of the proposed RSTB-based Feature Enhancement Module (FEM). In summary, SwinDFN is a simple yet effective model for fusing HR-MSI and LR-HSI to generate high-quality HR-HSI, and holds promise for various hyperspectral imaging applications.

VI. Acknowledgment

This work was supported in part by National Science and Technology Council (NSTC), Taiwan, under the Grant NSTC 114-2221-E-003-001-MY3.

References

[1] N. Aburaed, M. Q. Alkhatib, S. Marshall, J. Zabalza, and H. Al Ahmad, "A review of spatial enhancement of hyperspectral remote sensing imaging techniques," *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 2275-2300, 2023.

[2] C.-I. Chang, C. -C. Liang, and P. F. Hu, "Iterative random training sampling convolutional neural network for hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 61, pp. 1-26, 2023.

[3] C.-I. Chang, "Hyperspectral target detection: Hypothesis testing, signal-to-noise ratio, and spectral angle theories," *IEEE Trans. Geoscience and Remote Sensing*, vol. 60, pp. 1-23, 2022.

[4] S.-Y. Chen et al., "Automated peanut defect detection using hyperspectral imaging and deep learning: A real-time approach for smart agriculture," *Smart Agricultural Technology*, vol. 11, Aug. 2025.

[5] H.-F. Yan et al., "Hyperspectral and multispectral image fusion: When model-driven meet data-driven strategies," *Information Fusion*, vol. 116, Apr. 2025.

[6] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + Pan data," *IEEE Trans. Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230-3239, Oct. 2007.

[7] J. Cheng, H. Liu, T. Liu, F. Wang, and H. Li, "Remote sensing image fusion via wavelet transform and sparse representation," *ISPRS J. Photogrammetry and Remote Sensing*, vol. 104, pp. 158-173, 2015.

[8] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658-3668, July 2015.

[9] Y. Xu, Z. Wu, J. Chanussot, and Z. Wei, "Hyperspectral images super-resolution via learning high-order coupled tensor ring representation," *IEEE Trans. Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4747-4760, Nov. 2020.

[10] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690-6709, Sept. 2019.

[11] L. Yan, M. Zhao, X. Wang, Y. Zhang, and J. Chen, "Object detection in hyperspectral images," *IEEE Signal Processing Letters*, vol. 28, pp. 508-512, 2021.

[12] C.-H. Yeh et al., "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Trans. Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129-6143, Nov. 2022.

[13] Y. Chang, L. Yan, H. Fang, S. Zhong, and W. Liao, "HSI-DeNet: Hyperspectral image restoration via convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 667-682, Feb. 2019.

[14] C.-H. Yeh, C.-H. Huang, and L.-W. Kang, "Multi-scale deep residual learning-based single image haze removal via image decomposition," *IEEE Trans. Image Processing*, vol. 29, pp. 3153-3167, 2020.

[15] M. Ciotola, S. Vitale, A. Mazza, G. Poggi, and G. Scarpa, "Pansharpening by convolutional neural networks in the full resolution framework," *IEEE Trans. Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2022.

[16] T. Huang, W. Dong, J. Wu, L. Li, X. Li, and G. Shi, "Deep hyperspectral image fusion network with iterative spatio-spectral regularization," *IEEE Trans. Computational Imaging*, vol. 8, pp. 201-214, 2022.

[17] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, "Hyperspectral image super-resolution via deep progressive zero-centric residual learning," *IEEE Trans. Image Processing*, vol. 30, pp. 1423-1438, 2021.

[18] Z. Min, Y. Wang, and S. Jia, "Multiscale spatial-spectral joint feature learning for multispectral and hyperspectral image fusion," *Proc. IEEE Int. Conf. High Performance Computing & Communications*, Haikou, Hainan, China, 2021, pp. 1265-1270.

[19] J. Xiao, J. Li, Q. Yuan, and L. Zhang, "A dual-UNet with multistage details injection for hyperspectral image fusion," *IEEE Trans. Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2021.

[20] J. Fang, J. Yang, A. Khader, and L. Xiao, "A multiresolution details enhanced attentive dual-UNet for hyperspectral and multispectral image fusion," *IEEE J. Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 638-655, 2023.

[21] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1-5, 2022.

[22] S. Peng, X. Zhu, H. Deng, L.-J. Deng, and Z. Lei, "Fusionmamba: Efficient remote sensing image fusion with state space model," *IEEE Trans. Geoscience and Remote Sensing*, vol. 62, pp. 1–16, 2024.

[23] C.-H. Lin, C.-C. Hsu, S.-S. Young, C.-Y. Hsieh, and S.-C. Tai, "Qrcode: Quasi-residual convex deep network for fusing misaligned hyperspectral and multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[24] J. Li, K. Zheng, L. Gao, Z. Han, Z. Li, and J. Chanussot, "Enhanced deep image prior for unsupervised hyperspectral image super-resolution," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 63, pp. 1-18, 2025.

[25] C.-M. Lee, Y.-F. Lin, Y.-H. Ho, L.-W. Kang, and C.-C. Hsu, "HyFusion: Enhanced reception field Transformer for hyperspectral image fusion," *Proc. IEEE Int Geoscience and Remote Sensing Symposium*, Brisbane, Australia, Aug. 2025.

[26] C.-M. Lee, Y.-F. Lin, L.-W. Kang, and C.-C. Hsu, "Robust hyperspectral image pansharpening via sparse spatial-spectral representation," *Proc. IEEE Int Geoscience and Remote Sensing Symposium*, Brisbane, Australia, Aug. 2025.

[27] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops*, Montreal, BC, Canada, 2021, pp. 1833-1844.

[28] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv:1505.00853*, May 2015.

[29] A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz, "The spectral image processing system (SIPS) – interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, pp. 145-163, 1993.

[30] G. Vane, R. O. Green, T. G. Chrien, H. T. Enmark, E. G. Hansen, and W. M. Porter, "The airborne visible/infrared imaging spectrometer (aviris)," *Remote Sens. Environ.*, vol. 44, no. 2-3, pp. 127–143, 1993.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv*:1412.6980, 2014.

[32] L. Wald, Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions. Paris, France : Les Presses de lEcole des Mines, 2002.