MFTN: A Multi-scale Feature Transfer Network Based on IMatchFormer for Hyperspectral Image Super-Resolution

Shuying Huang [1] **Mingyang Ren** [2] **Yong Yang** [2] **Xiaozheng Wang** [3] **Yingzhi Wei** [2]

Abstract

Hyperspectral image super-resolution (HISR) aims to fuse a low-resolution hyperspectral image (LR-HSI) with a high-resolution multispectral image (HR-MSI) to obtain a high-resolution hyperspectral image (HR-HSI). Due to some existing HISR methods ignoring the significant feature difference between LR-HSI and HR-MSI, the reconstructed HR-HSI typically exhibits spectral distortion and blurring of spatial texture. To solve this issue, we propose a multi-scale feature transfer network (MFTN) for HISR. Firstly, three multi-scale feature extractors are constructed to extract features of different scales from the input images. Then, a multi-scale feature transfer module (MFTM) consisting of three improved feature matching Transformers (IMatchFormers) is designed to learn the detail features of different scales from HR-MSI by establishing the cross-model feature correlation between LR-HSI and degraded HR-MSI. Finally, a multiscale dynamic aggregation module (MDAM) containing three spectral aware aggregation modules (SAAMs) is constructed to reconstruct the final HR-HSI by gradually aggregating features of different scales. Extensive experimental results on three commonly used datasets demonstrate that the proposed model achieves better performance compared to state- of-the-art (SOTA) methods.

1. INTRODUCTION

Hyperspectral remote sensing can obtain hyperspectral images (HSIs) with continuous and narrow spectral resolution due to its ability to simultaneously capture two-dimensional spatial information and three-dimensional spectral information of the targets. Because HSIs contain rich spatial and spectral information, they have become a primary data source in many critical domains such as mineral exploration (Sabins, 1999) and plant detection (Lowe et al., 2017). However, due to the inherent physical limitations of single sensors, they cannot provide high-resolution data in both spectral and spatial (Dian et al., 2021), resulting in the inability to meet the requirements of precise applications. Consequently, an increasing number of remote sensing satellites are now equipped with multiple spectral imaging instruments to overcome these constraints by simultaneously obtaining multimodal remote sensing images, such as multispectral images (MSIs) and HSIs. Compared with HSIs, MSIs have higher spatial resolution but lower spectral resolution, which can also affect the accuracy of subsequent tasks. To obtain HSIs with both high spatial and spectral resolutions, some researchers have proposed the hyperspectral image super-resolution (HISR) methods (Thomas et al., 2008; Aiazzi et al., 2011; Vivone et al., 2015), also known as the pansharpening methods, which fuse a low-spatial-resolution HSI (LR-HSI) with a high-spectral-resolution multispectral image (HR-MSI) in the same scene to obtain a HSI with high spatial and spectral resolution (HR-HSI). HISR methods can significantly improve the interpretability of remote sensing information, which helps to improve the performance of subsequent applications, such as land cover analysis (Solomon & Agnes, 2023).

Early HISR methods (Akgun et al., 2005; Wang et al., 2017; Irmak et al., 2018; Huang et al., 2014) mostly utilized single LR-HSI to reconstruct the corresponding HR-HSI. However, due to the limited spatial information provided by a single LR-HSI, such methods cannot reconstruct more spatial information, resulting in edge blurring and spectral distortion issues in the reconstructed HR-HSI. Therefore, considering the imaging characteristics of different satellite sensors, some researchers have studied multi-frame HISR reconstruction methods (Gillespie et al., 1987; Laben & Brower, 2000), which use MSIs or PAN images with the same scene as reference images to reconstruct HR-HSIs. At present, multi-frame HISR methods are mainly divided into two categories: traditional methods and deep learning-based

[1]School of Software, Tiangong University, Tianjin, China [2]School of Computer Science and Technology, Tiangong University, Tianjin, China [3]School of Control Science and Engineering, Tiangong University, Tianjin, China. Correspondence to: Yong Yang <greatyangy@126.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

methods. Traditional methods are easy to implement and have physical interpretability. For example, Akhtar et al. (2014) proposed a sparse representation-based approach for HISR, which performs sparse encoding on images with high spatial but low spectral resolution, and uses the encoding together with the scene spectrum to estimate HR-HSIs. Alparone et al. (2007) discussed the fusion results of two different sensor data in detail at the 2006 Denver International Symposium on Geoscience and Remote Sensing, and confirmed the advantages of multisolution analysis (MRA) and detail injection methods. These traditional methods typically rely on manually defined prior knowledge or regularization terms (Ma et al., 2021), whose accuracy directly affects the performance of reconstructed images. Therefore, these methods typically suffer from spatial and spectral distortions in HR-HSI images, but compared to single frame HISR methods, they can reconstruct HR-HSIs with richer details.

In recent years, to reduce human intervention in feature extraction and prior knowledge definition, deep learning has been successfully applied in HISR tasks. Initially, convolutional neural networks (CNNs) were the primary choice for implementing HISR. Dian et al. (2018) presented a deep HSI sharpening method, which realizes the fusion of an LR-HSI with an HR-MSI by directly learning the image priors via deep CNN-based residual learning. Liu et al. (2020) proposed an MSI pansharpening method by combining a shallow-deep convolutional network (SDCN) and a spectral discrimination-based detail injection (SDDI) model. Lee et al. (2021) proposed a shift-invariant pansharpening with moving object alignment (SIPSA-Net), which is the first approach in PAN sharpening tasks to consider the large misalignment of moving object regions. Hu et al. (2022c) proposed a simple and efficient CNN to fuse LR-HSIs and HR-MSIs, which can better preserve both spatial and spectral information. Due to the limitation of the receptive field of convolutional kernels, these methods mainly learn local features in convolutional operations, often leading to spatial and spectral information distortion in the reconstruction results. Later, due to the successful application of the Transformer architecture (Vaswani et al., 2017) in visual tasks, researchers have attempted to improve the performance of HISR by establishing a relationship between two modal features. Hu et al. (2022b) designed a Transformer-based architecture that can globally explore the internal relationships within features. Chen et al. (2023) proposed a spectral–spatial transformer (SST) by exploring the spectral and spatial long-range dependence to show the potentiality of transformers for HSI and MSI fusion. Although these methods focus on learning long-distance features, they do not fully consider the modal differences between the two image features. In addition, most methods directly learn the mapping relationship between an LR-HSI and its corresponding HR-HSI in LR space or HR space, leading to inaccurate spatial feature reconstruction due to the lack of establishing matching relationships between intermediate scale features.

To address the above issues, we propose a multi-scale feature transfer network (MFTN) based on the improved feature matching Transformer (IMatchFormer) for HISR. First, a multi-scale feature extraction module (MFEM) is constructed to extract features of different scales from the input LR-HSI and HR-MSI. Then, a multi-scale feature transfer module (MFTM) composed of three IMatchFormers is designed to transfer features of different scales from HR-MSI to LR-HSI by establishing matching relationships between two modal features at different scales. Finally, the obtained transfer features of different scales are fed into the constructed multi-scale dynamic aggregation module (MDAM) to achieve progressive integration of features and output the reconstructed HR-HSI. The contributions of this paper are as follows:

- An MFTN consisting of three modules, namely MFEM, MFTM, and MDAM, is proposed for HISR, aiming to reconstruct a HR-HSI with high spectral and spatial resolution by transferring the detail features from an HR-MSI to an LR-HSI.
- In MFTM, three IMatchFormers are designed to learn the transfer detail features from HR-MSI by establishing the cross-modal feature correlations between the LR-HSI and degraded HR-MSI.
- In MDAM, three spectral aware aggregation modules (SAAMs) are constructed to gradually integrate the transfer features and shallow features of LR-HSI to obtain the reconstructed HR-HSI. One spectral aware module (SAM) in each SAAM is designed to directly utilize LR-HSI features to correct the reconstructed features at different scales, so as to suppress spectral distortion in the reconstructed results.
- Numerous experiments have been conducted on different datasets to verify the performance of the proposed model. Compared with some SOTA methods, the proposed model achieves better results both subjectively and objectively.

2. Proposed Method

Although LR-HSI captures rich spectral features, the narrow range of each imaging band results in insufficient spatial information in each channel. In HISR task, reconstructing rich spatial features proves challenging with just a single LR-HSI. Therefore, we propose an MFTN based on IMatchFormer, which utilizes a HR-MSI as a reference image to achieve SR reconstruction of LR-HSI. MFTN shown in

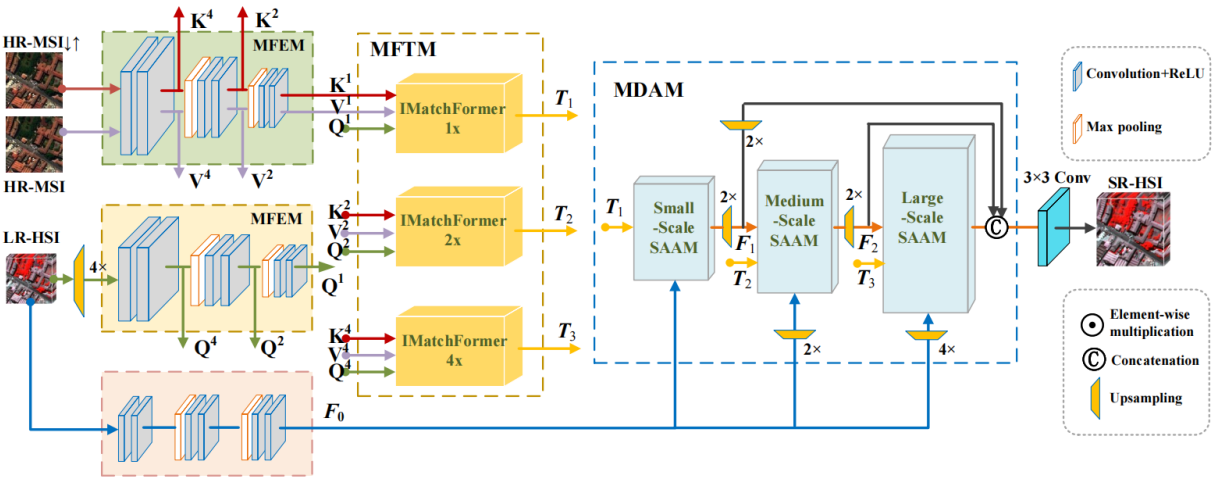

Figure 1. The overall structure of MFTN.

Figure 1 consists of three modules: MFEM, MFTM, and MDAM.

In MFEM, one multi-scale feature extractor is constructed to extract multi-scale features from the input images, which are used to achieve the transfer of spatial features from HR-MSIs to LR-HSIs. However, due to the fact that HR-MSIs and HR-HSIs come from different sensors, these two types of features do not fully match in both spatial and spectral dimensions. To make LR-HSI and HR-MSI have consistent spatial domains, 4× downsampling and 4× upsampling operations are performed on HR-MSI sequentially to obtain the degraded HR-MSI, namely HR-MSI↓↑, and 4× upsampling operation is performed on LR-HSI to obtain the upsampled LR-HSI. To extract multi-scale features of HR-MSI, HR-MSI↓↑, and LR-HSI, MFEM uses three multi-scale feature extractors, each containing three convolutional layers. Taking the multi-scale feature extraction of HR-MSI as an example, one multi-scale feature extractor in MFEM can be represented as follows:

$$V^4 = ConL(I_{HM}) = Con_{3\times3}(Con_{3\times3}(I_{HM})) \quad (1)$$

$$V^2 = ConL(MP(V^4)) \quad (2)$$

$$V^1 = ConL(MP(V^2)) \quad (3)$$

where I_{HM} represents the input HR-MSI, and $ConL(\cdot)$ represents the convolutional layers with two 3×3 convolutional operations $Con_{3\times3}(\cdot)$, and $MP(\cdot)$ represents the maximum pooling layer with a window size of 2×2 and a step size of 2. V^4, V^2 and V^1 represent features at three scales, respectively. Similarly, using the HR-MSI↓↑ as input of the feature extractor can obtain the corresponding multi-scale features K^4, K^2 and K^1, and using LR-HSI as input can obtain the multi-scale features Q^4, Q^2 and Q^1.

In MFTM, different scale features from MFEM are sent to three IMatchFormers, which are designed to obtain transfer features at different scales by learning feature correlations between HR-MSI↓↑ and LR-HSI. The operation of MFTM can be represented as:

$$T_1 = IMF_{1\times}(K^1, V^1, Q^1) \quad (4)$$

$$T_2 = IMF_{2\times}(K^2, V^2, Q^2) \quad (5)$$

$$T_3 = IMF_{4\times}(K^4, V^4, Q^4) \quad (6)$$

where $IMF_{1\times}(\cdot)$, $IMF_{2\times}(\cdot)$, and $IMF_{4\times}(\cdot)$ represent three IMatchFormers that receive features of different scales, and T_1, T_2, and T_3 represent the obtained transfer features of three scales.

In MDAM, three SAAMs are constructed to gradually achieve the fusion of features at different scale, and multi-scale fused features are concatenated and mapped into the final SR-HSI I_{SR} through a 3×3 convolutional layer. The operation of MDAM can be represented as:

$$F_1 = SAAM(F_0, T_1) \quad (7)$$

$$F_2 = SAAM(F_1, T_2) \quad (8)$$

$$F_3 = SAAM(F_2, T_3) \quad (9)$$

$$I_{SR} = Con_{3\times3}(Concat(Up(F_1), Up(F_2), F_3)) \quad (10)$$

where $SAAM(\cdot)$ represent the operation of SAAM for integrating features at different scales. F_0 represents the shallow features extracted from LR-HSI through three convolutional layers, and F_1, F_2 and F_3 represent the feature maps output by three SAAMs. $Concat(\cdot)$ represents the concatenation operation. $Up(\cdot)$ represents the upsampling operation.

Below, we provide a detailed introduction to the construction of other components in the network, such as IMatchFormer and SAAM.

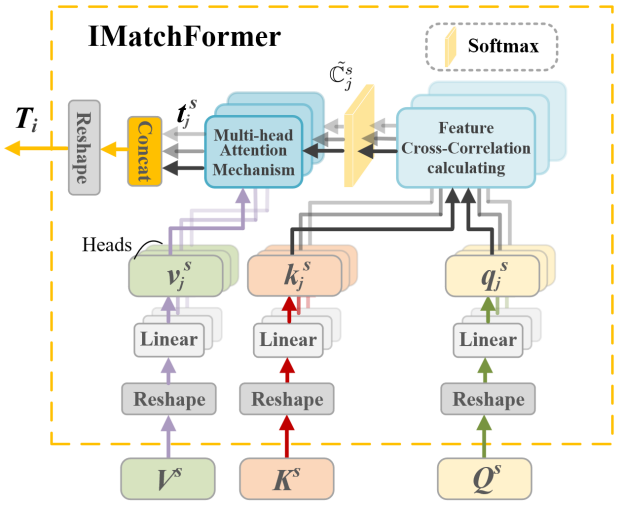

Figure 2. Architecture of IMatchFormer.

2.1. IMatchFormer

To focus on the long-distance features of images and learn the correlation between different modal features, we construct an IMatchFormer based on the transformer structure, which transfers the detail features from HR-MSI to LR-HSI by establishing the correlation between LR-HSI and HR-MSI↓↑ features. To achieve precise matching of two modal features, the multi-head attention mechanism is used in IMatchFormer. Compared with the the single-head attention mechanism, the multi-head attention mechanism can provide multiple different representation subspaces for attention through different linear transformations. Therefore, the multi-head attention mechanism is more suitable for matching multi-channel information in HISR tasks. The structure of IMatchFormer is shown in Figure 2, which takes features of the same scale Q^s, K^s and V^s (s=4, 2, 1) from MFEM as input to learn the transfer features at each scale. The operation of each IMatchFormer is described as follows. First, the feature maps Q^s, K^s and V^s generate multiple feature vectors (heads) through a linear projection layer $Linear(\cdot)$, which are defined as follows:

$$\begin{aligned} q_j^s &= Linear(Reshape(Q^s)) \\ k_j^s &= Linear(Reshape(K^s)) \\ v_j^s &= Linear(Reshape(V^s)) \end{aligned} \tag{11}$$

where $Reshape(\cdot)$ represents the operation of changing the matrix dimension.

Then, to learn the transferred detail features from HR-MSI, it is necessary to establish the feature cross-correlation between HR-MSI↓↑ and LR-HSI. Therefore, the feature cross-correlation matrices $\tilde{\mathbb{C}}_j^s$ between multiple heads of Q^s and K^s are calculated and used to perform the feature transfer operation, that is, $\tilde{\mathbb{C}}_j^s$ is used to weight the multiple heads of V^s to obtain the transfer features. The specific operations are defined as follows:

$$\tilde{\mathbb{C}}_j^s = \text{softmax}(\frac{q_j^s (k_j^s)^\top}{\sqrt{d_k}}) \tag{12}$$

$$t_j^s = MatMul(\tilde{\mathbb{C}}_j^s, v_j^s) \tag{13}$$

where d_k represents the matrix dimensions of q_j^s and k_j^s, which can address the vanishing gradients that occur when performing the softmax operation, thereby facilitating effective backpropagation during the training process. $\top$ denotes the transpose operation of the matrix, and $MatMul(\cdot)$ denotes the multiplication operation of two matrices. t_j^s represents the learned transfer features.

Finally, these transfer featuers t_j^s are concatenated and transformed back into the high-dimensional feature space through a reshaping operation, which is defined as follows:

$$T_i = Reshape(Concat(t_j^s)) \tag{14}$$

where T_i is the transfer feature maps in high-dimensional space output by the $i-th$ IMatchFormer, which contains the detail feature representations from different subspaces. The proposed IMatchFormer allows the model to jointly focus on information at different positions in different representation subspaces, and enables the model to selectively focus on highly correlated spectral and texture information within LR-HSI and HR-MSI. The transfer features obtained from three IMatchFormer are fed into MDAM to gradually achieve the fusion of features at different scales.

2.2. Multi-Scale Dynamic Aggregation Module

To fully integrate multiple scale transfer features, a MDAM containing three SAAMs is constructed, as shown in Figure 1 to gradually fuse spectral and spatial features at different scales. To improve the spectral fidelity of reconstructed features, the feature maps F0 from LR-HSI are transferred to three SAAMs to supplement spectral features. Meanwhile, to meet the requirements of spatial resolution, the feature maps F_0 need to be upsampled when they are used as input for the last two SAAMs. Here, each SAAM generates fused features at one scale by integrating features from an IMatchformer, fusion features from the previous SAAM, and shallow features from LR-HSI. Note that the input features in the first SAAM do not include the features of the previous SAAM. The construction of SAAM is described as follows.

2.2.1. Spectral Aware Aggregation Module

As shown in Figure 3, SAAM consists of a series of convolutional layers, deformable convolution (DCN) layers (Dai et al., 2017), LeakyRelu, spectral aware modulation (SAM), and residual blocks (RBs). First, the features F_{i-1} (i= 2,3) from the previous SAAM and the corresponding scale features T_i from an IMatchformer are concatenated, and passed

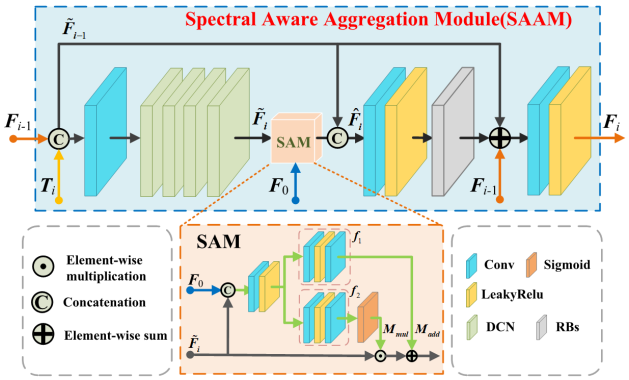

Figure 3. Architecture of SAAM.

through a 3×3 convolution layer and a group of DCN layers to achieve feature integration and extraction and output the feature maps $\tilde{F}_i$. Here, due to the insufficient use of structured information in the feature space caused by fixed grid kernels in ordinary convolutions, we adopt DCNs, which can expand the receptive fields with adaptive shapes to more accurately correct the misalignment of features. The above operations can be formulated as:

$$\tilde{F}_i = DCNs(Con_{3\times3}(Concat(T_i, F_{i-1}))) \tag{15}$$

where $DCNs$ represents 4 deformable convolution layers.

Then, the shallow features F_0 from LR-HSI and the extracted features $\tilde{F}_i$ are sent into a SAM to compensate for the spectral information. Finally, to reduce feature loss, the previously input features are reused for feature supplementation, and convolutional layers and residual blocks are used for feature integration and dimensionality reduction. The operation process is as follows:

$$\hat{F}_i = Concat(SAM(\tilde{F}_i, F_0), \tilde{F}_{i-1}) \tag{16}$$

$$F_i = ConA(RBs(Con_{3\times3}(\hat{F}_i)) + F_{i-1} + \tilde{F}_{i-1}) \tag{17}$$

where SAM represents the operation of SAM, RBs represent the residual blocks, and $ConA(\cdot)$ denotes the feature adjustment layer consisting of one 3×3 convolutional layer and a LeakyRelu.

2.2.2. Spectral Aware Module

To ensure the fidelity of spectral and spatial features, a SAM is designed, as shown in Figure 3, to utilize shallow features from LR-HSI for correction and supplementation of reconstructed features. The implementation process of SAM is described below. The input features F_0 and $\tilde{F}_i$ are first concatenated and integrated through a convolutional layer and a LeakyRelu function. Then, two branches are constructed to learn modulation coefficients and supplementary features, respectively. The modulation coefficients are used to weight the input features $\tilde{F}_i$ to obtain calibration features. Supplementary features are added to the calibration features for feature supplementation. The operation of SAM can be defined as follows:

$$M_{add} = f_1(Con_{3\times3}(Concat(F_0, \tilde{F}_i))) \tag{18}$$

$$M_{mul} = sigmoid(f_2(Con_{3\times3}(Concat(F_0, \tilde{F}_i)))) \tag{19}$$

$$F_{SAM} = \tilde{F}_i \odot M_{mul} + M_{add} \tag{20}$$

where f_1 and f_2 are non-linear mapping functions constructed by two 3×3 convolutional layers, where the first convolutional layer are followed by an LeakyReLU activation. M_{mul} and M_{add} are the modulation coefficients learned by SAM. Different channels of the modulation coefficients have different responses for the same object in the features, and these channels are collectively utilized as scaling factors to modulate the output of $DCNs$.

In summary, MDAM achieves SR reconstruction of different scale features from coarse to fine by fusing spectral and spatial features in multiple scale spaces. And it can improve some key issues in HISR tasks, such as spatial detail misalignment caused by narrow spectral coverage in MSI and imprecision in multi-source remote sensing image registration, and provide rich detail features for reconstructing HR-HSI.

2.3. Joint Loss Function

To better guide network training, a joint loss function is defined as follows:

$$\mathcal{L}_{overall} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{t-per}\mathcal{L}_{t-per} \tag{21}$$

where $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{t-per}$ represent the reconstruction loss, perceptual loss, and transfer perception loss, respectively. λ_{rec}, λ_{per}, and λ_{t-per} denote the tradeoff parameters of loss terms, which are empirically set to 1.0, 0.1, and 0.05 based on experience.

Reconstruction loss $\mathcal{L}_{rec}$ is defined as $L1$ loss, which measures the pixel distances between the SR result and corresponding HR image (i.e., ground truth, GT). This loss term can reduce false details generated by deep networks by calculating the difference between two images pixel by pixel and can be formulated as:

$$\mathcal{L}_{rec} = \frac{1}{CWH} \left\| I_{HR} - I_{SR} \right\|_1 \tag{22}$$

where I_{HR} represents GT images, I_{SR} is the generated SR-HSI, and C, H, and W represent the channel number, width, and height of HR-HSI, respectively.

Perceptual loss $\mathcal{L}_{per}$ is used to minimize the distance between high-level features of two images, which emphasizes the perceptual quality of the image and is more in line with the human eye's perception of image quality. The underlying idea of perceptual loss is to enhance the similarity in feature space between the prediction image and the target

image. The R, G, and B bands in the two images are selected to calculate the perceptual loss:

$$\mathcal{L}_{per} = \frac{1}{C_i W_i H_i} \left\| f_i^{vgg}(I_{HR}^{rgb}) - f_i^{vgg}(I_{SR}^{rgb}) \right\|_2 \quad (23)$$

where $f_i^{vgg}(\cdot)$ denotes the i-th layer's feature maps obtained by VGG-19. I_{HR}^{rgb} and I_{SR}^{rgb} are the RGB images synthesized from the R, G, and B bands in GT images and generated SR-HSI, respectively.

Transfer perception loss is designed to constrain the similarity between the features of SR-HSI and the transfer features T_s from IMatchFormers, and it can be formulated as:

$$\mathcal{L}_{t-per} = \sum_{s=1,2,4} \frac{1}{C_s W_s H_s} \left\| f_s^{MFEM}(I_{SR}) - T_s \right\|_2 \quad (24)$$

where T_s denotes the transfer feature maps at the $s-th$ spatial scale, $f_s^{MFEM}(I_{SR})$ is the feature maps of the $s-th$ spatial scale of SR-HSI obtained by the extractor in MFEM.

3. Experiments

In this section, to validate the effectiveness of the proposed MFTN, a large number of experiments are conducted on three publicly available benchmark hyperspectral datasets, including Pavia Center (Plaza et al., 2009), Botswana (Ungar et al., 2003), and Chikusei (Yokoya & Iwasaki, 2016). And the proposed MFTN is compared with eight SOTA methods, including two traditional methods, namely PCA (Chavez & Kwarteng, 1989) and GFPCA (Liao et al., 2014), and six deep learning-based methods, namely DARN (Zheng et al., 2020), HSRNet (Hu et al., 2022a), HyperPNN (He et al., 2019), SSFCNN (Han et al., 2018), SSRNet (Zhang et al., 2021), HyperDSNet (Zhuo et al., 2022), and HyperRefiner (Zhou et al., 2023). To objectively evaluate the performance of all comparison methods, six widely used reference metrics are adopted, including Spectral Cross Correlation (SCC), Spectral Angle Mapping (SAM), Erreur Relative Globale Adimensionnelle Desynthese (ERGAS), Universal Image Quality Index (UIQI), Structure Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR).

For network training, the datasets are processed following Wald's protocol (Wald, 2000). For fair comparisons, all deep learning-based methods are retrained using Python 3.9 with Pytorch 1.13 on the Linux operating system. The computing equipment contains the 64-GB CPU memory and NVIDIA GPU GeForce GTX 3090. The MFTN is trained for 10000 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}, which is decayed by a factor of 1×10^{-5} after 8000 epochs.

3.1. Experiment Results

Figures 4, 5, and 6 respectively show the reconstructed results obtained by comparison deep learning-based methods on the Pavia center, Botswana, and Chikusei datasets, and the 60th, 30th, 10th spectral bands in the reconstructed results are selected as the RGB bands. To observe the differences between the reconstruction results more clearly, the mean absolute error (MAE) maps between the reconstruction results and GTs are calculated and displayed, and the local areas are selected for magnification and displayed below the corresponding results. From these figures, it can be observed that our results are visually closest to GTs, and the corresponding MAE maps contain the least residual information compared to the results of other deep learning-based methods. Therefore, this also indicates that the proposed MFTN is superior to other deep learning-based methods, and the obtained results have better spatial and spectral fidelity.

Table 1 shows the average objective indicators of the reconstruction results obtained by all comparison methods on three datasets. The best results are highlighted in bold, while the second-best results are highlighted in underline. From the table, it can be seen that the indicator values obtained by traditional methods are much lower than those obtained by deep learning-based methods. The proposed method achieved the highest indicator values on the Pavia Center and Chikusei datasets, while on the Botswana dataset, the UIQI value ranks second, and all other indicator values are also the highest. In terms of runtime, although our method is not the least time-consuming, it is within an acceptable range and has lower runtime than HyperRefiner on all three datasets. Therefore, the proposed method has better performance compared to the comparision methods.

3.2. Ablation Studies

In this section, we conduct several ablation experiments on the Pavia Center dataset to demonstrate the effectiveness of IMatchFormer and SAAM in the proposed MFTN.

3.2.1. Effect of IMatchFormer

In MFTN, three IMatchFormers are adopted to learn the transfer features at three scales from the features of HR-MSI. The experimental results of using IMatchFormers on three scale features are presented in Table 2. $1\times$ represents the minimum scale, $2\times$ represents the intermediate scale, which is twice the minimum scale, and $4\times$ represents the maximum scale, which is four times the minimum scale and consistent with the scale of HR-MSI. In addition, $1\times$, $2\times$, and $4\times$ also refer to using only one IMatchFormer corresponding to one scale features in MFTM, respectively, and features of other scales are directly transmitted to the corresponding SAAM. $1\times2\times4\times$ indicates using IMatch-

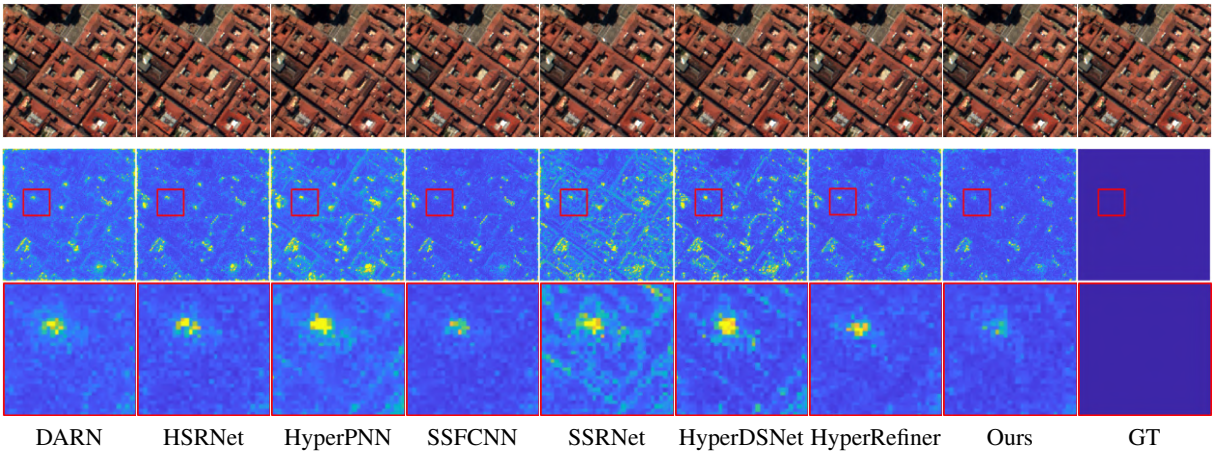

Figure 4. Comparison of visualization results obtained by deep learning-based methods on Pavia Center dataset.

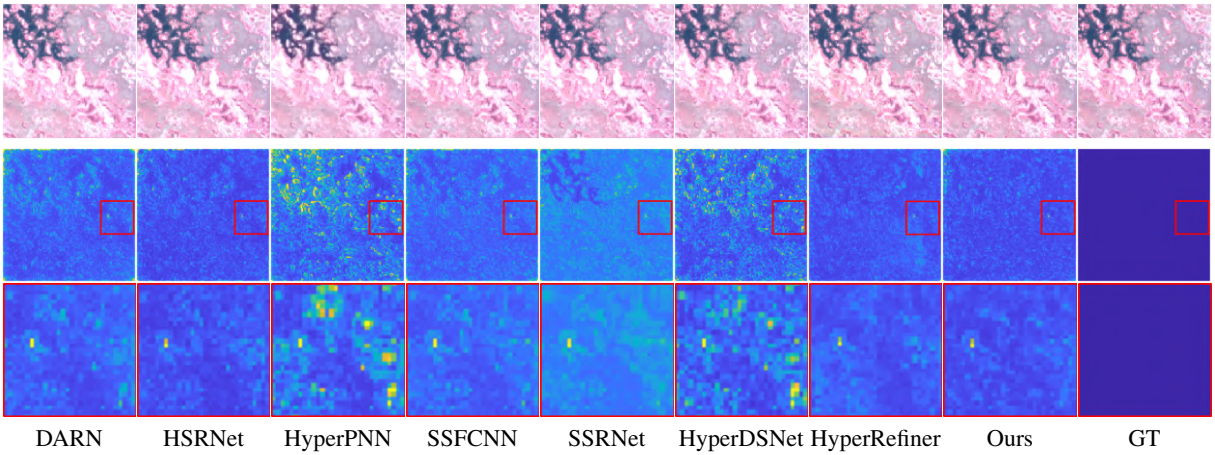

Figure 5. Comparison of visualization results obtained by deep learning-based methods on Botswana dataset.

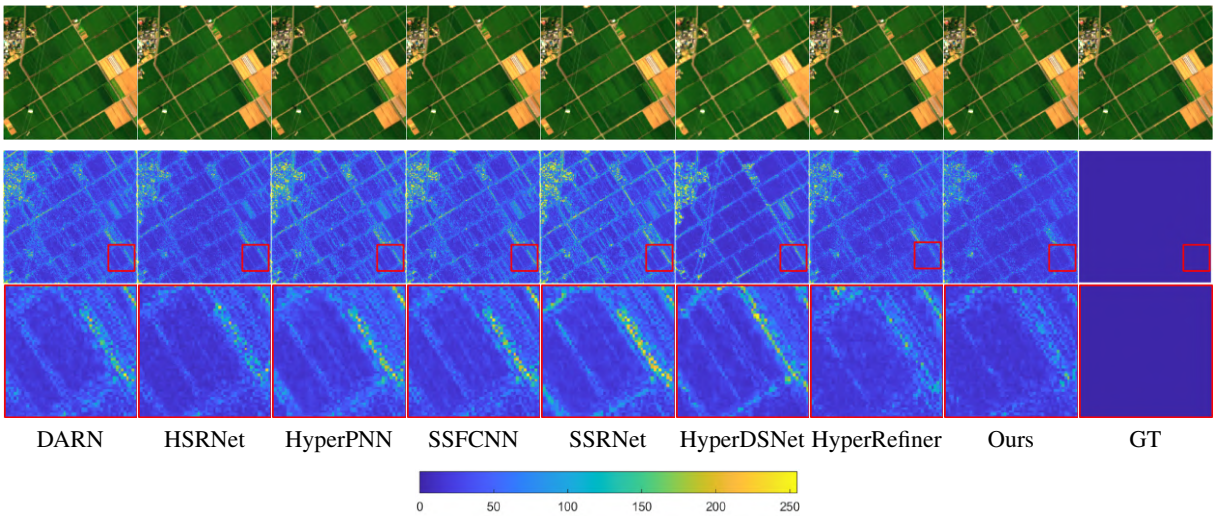

Figure 6. Comparison of visualization results obtained by deep learning-based methods on Chikusei dataset.

Former on all three scale features. From the table, it can be seen that the indicator values obtained by the model using IMatchFormer only at the minimum scale are the lowest, while the indicator values obtained by using the IMatchFormer on all three scale features are the highest. Therefore, this also demonstrates that the constructed IMatchFormer is

Table 1. The average quantitative results on the Pavia Center, Botswana, and Chikusei datasets

Dataset	Methods	SCC(↑)	SAM(↓)	ERGAS(↓)	UIQI(↑)	SSIM(↑)	PSNR(↑)	Test Time(Ms)
Pavia Center	PCA	0.780	8.01	8.86	0.457	0.458	26.01	—
	GFPCA	0.825	8.16	8.62	0.550	0.592	21.18	—
	DARN	0.980	5.13	2.58	0.960	0.951	37.61	2.153
	HSRNet	0.983	4.49	4.30	0.957	0.953	37.78	52.099
	HyperPNN	0.967	5.99	3.82	0.935	0.917	36.70	6.473
	SSFCNN	0.982	4.33	2.74	0.966	0.962	38.50	16.339
	SSRNet	0.958	5.43	3.90	0.946	0.930	37.48	11.255
	HyperDSNet	0.982	5.15	4.67	0.951	0.944	36.45	3.824
	HyperRefiner	0.984	4.23	2.44	0.967	0.965	39.61	10.178
	Ours	**0.985**	**4.20**	**2.32**	**0.969**	**0.965**	**40.98**	5.704
Botswana	PCA	0.943	2.38	1.98	0.792	0.793	40.03	—
	GFPCA	0.901	2.66	2.75	0.531	0.498	37.83	—
	DARN	0.992	1.57	2.92	0.920	0.968	38.04	1.675
	HSRNet	0.995	1.26	2.35	**0.954**	0.975	38.84	7.011
	HyperPNN	0.981	2.05	2.39	0.932	0.956	38.06	1.786
	SSFCNN	0.994	1.86	8.00	0.868	0.913	37.76	2.101
	SSRNet	0.993	2.02	2.80	0.764	0.819	38.96	1.634
	HyperDSNet	0.985	1.70	2.15	0.940	0.957	38.90	1.862
	HyperRefiner	0.994	1.30	1.96	0.945	0.976	39.79	9.135
	Ours	**0.995**	**1.24**	**1.83**	0.949	**0.978**	**41.97**	4.127
Chikusei	PCA	0.887	6.99	7.71	0.466	0.598	30.98	—
	GFPCA	0.883	4.76	7.00	0.619	0.689	30.96	—
	DARN	0.965	1.88	4.06	0.944	0.943	38.39	4.086
	HSRNet	0.974	2.08	3.60	0.950	0.953	36.95	13.569
	HyperPNN	0.953	2.04	5.64	0.926	0.927	41.57	6.041
	SSFCNN	0.959	1.91	4.56	0.940	0.936	35.74	3.893
	SSRNet	0.935	2.20	4.44	0.928	0.920	36.11	3.219
	HyperDSNet	0.977	2.42	3.87	0.943	0.959	37.76	4.109
	HyperRefiner	0.977	1.69	3.22	0.954	0.957	41.43	12.275
	Ours	**0.978**	**1.57**	**2.88**	**0.963**	**0.965**	**42.00**	12.220

Table 2. Ablation study on utilizing IMatchFormer at multiple scales.

IMatchFormer	SCC↑	SAM↓	ERGAS↓	UIQI↑	SSIM↑	PSNR↑
None	0.866	6.41	10.47	0.829	0.752	24.82
1×	0.869	6.42	9.93	0.835	0.759	25.36
2×	0.976	4.79	3.69	0.955	0.944	34.81
4×	0.983	4.30	2.37	0.968	0.964	40.52
1×2×4×	**0.985**	**4.20**	**2.32**	**0.969**	**0.965**	**40.98**

Table 3. Ablation study on the number of heads (N) in multi-head attention mechanism.

N	SCC↑	SAM↓	ERGAS↓	UIQI↑	SSIM↑	PSNR↑
N=1	0.982	4.57	2.71	0.966	0.959	38.49
N=2	0.980	4.43	2.52	0.966	0.960	39.58
N=4	0.982	4.37	2.48	0.967	0.962	40.16
N=8	**0.985**	**4.20**	**2.32**	**0.969**	**0.965**	**40.98**
N=16	0.983	4.39	2.46	0.968	0.962	39.95

effective.

In addition, IMatchFormer adopts a multi-head attention mechanism, in which the number of heads needs to be determined. Table 3 shows the objective results of IMatchFormer containing different numbers of headers in the model. N represents the number of headers, which increases exponentially from 1 to 16. From the table, when N=8, the proposed model achieves the highest objective results. Therefore, the number of heads in the multi-head attention mechanism in IMatchFormer is set to 8.

Table 4. Ablation study on the SAAM and SAM.

Method	SCC↑	SAM↓	ERGAS↓	UIQI↑	SSIM↑	PSNR↑
w/o SAAM	0.982	4.26	2.48	0.967	0.962	40.63
w/o SAM	0.983	4.23	2.47	**0.969**	0.964	40.65
MFNet	**0.985**	**4.20**	**2.32**	**0.969**	**0.965**	**40.98**

3.2.2. Effect of SAAM and SAM

To verify the effectiveness of the SAAM, we remove it from the network and replace it with a sum operation and 4 residual blocks. In addition, we also conduct the ablation experiment on the SAM in SAAM. The objective indicator results obtained by MFTN without SAAM or SAM in SAAM are presented in Table 4, respectively. The results show that the proposed MFTN with SAAM and with SAM achieves better performance compared with the other two modified structures. This also indicates that SAAM and SAM can better integrate the shallow features of LR-HSI and the transfer features at three scales.

4. Conclusion

In this paper, we propose an MFTN based on IMatchFormer for HISR task. First, HR-MSI, HR-MSI↓↑ and LR-HSI are sent to an MFEM to extract multi-scale features. Then, the features of each scale are fed into an IMatchFormer, which is constructed to learn transfer features from HR-

MSI by establishing correlations between HR-MSI↓↑ and LR-HSI features. Finally, the learned transfer features at three scales are sent together with the shallow features of LR-HSI to an MDAM containing three SAAMs, gradually achieving feature integration to obtain the reconstructed HR-HSI. Numerous experiments are conducted on three widely used HSI datasets, and the results demonstrate that the proposed MFTN is superior to some SOTA methods in both quantitative and qualitative evaluations.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62072218).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., and Selva, M. 25 years of pansharpening: a critical review and new developments. In Chen, C. H. (ed.), *Signal Image Processing for Remote Sensing*, chapter 28, pp. 533–548. CRC Press Publication, 2011.

Akgun, T., Altunbasak, Y., and Mersereau, R. M. Super-resolution reconstruction of hyperspectral images. *IEEE Transactions on Image Processing*, 14(11):1860–1875, 2005.

Akhtar, N., Shafait, F., and Mian, A. Sparse spatio-spectral representation for hyperspectral image super-resolution. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision – ECCV 2014*, pp. 63–78, Zurich, Switzerland, 2014. Springer International Publishing.

Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., and Bruce, L. M. Comparison of pansharpening algorithms: Outcome of the 2006 grs-s data-fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3012–3021, 2007.

Chavez, P. J. and Kwarteng, A. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogrammetric Engineering and Remote Sensing*, 55:339–348, 1989.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764–773, 2017.

Dian, R., Li, S., Guo, A., and Fang, L. Deep hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5345–5355, 2018.

Dian, R., Li, S., Sun, B., and Guo, A. Recent advances and new guidelines on hyperspectral and multispectral image fusion. *Information Fusion*, 69:40–51, 2021.

Gillespie, A. R., Kahle, A. B., and Walker, R. E. Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 22(3):343–365, 1987.

Han, X., Shi, B., and Zheng, Y. Ssf-cnn: Spatial and spectral fusion with cnn for hyperspectral image super-resolution. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2506–2510, 2018.

He, L., Zhu, J., Li, J., Plaza, A., Chanussot, J., and Li, B. Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(8):3092–3100, 2019.

Hu, J., Huang, T., Deng, L., Jiang, T., Vivone, G., and Chanussot, J. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7251–7265, 2022a.

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

Hu, J. F., Huang, T. Z., Deng, L. J., Jiang, T. X., Vivone, G., and Chanussot, J. Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):7251–7265, 2022c.

Huang, H., Yu, J., and Sun, W. Super-resolution mapping via multi-dictionary based sparse representation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3523–3527, Florence, Italy, 2014.

Irmak, H., Akar, G. B., and Yuksel, S. E. A map-based approach for hyperspectral imagery super-resolution. *IEEE Transactions on Image Processing*, 27(6):2942–2951, 2018.

L. Chen, G. Vivone, J. Q. J. C. and Yang, X. Spectral–spatial transformer for hyperspectral image sharpening. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2023.

Laben, C. A. and Brower, B. V. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. Technical Report 6011875, United States Patent, 2000.

Lee, J., Seo, S., and Kim, M. Sipsa-net: Shift-invariant pan sharpening with moving object alignment for satellite imagery. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10161–10169, Nashville, TN, USA, 2021.

Liao, W., Frieke, V. C., BELLENS, R., GAUTAMA, S., Pizurica, A., and PHILIPS, W. Fusion of thermal infrared hyperspectral and vis rgb data using guided filter and supervised fusion graph. *winning paper of IEEE Data Fusion Best Paper Contest 2014*, 2014.

Liu, L., Wang, J., Zhang, E., Li, B., Zhu, X., Zhang, Y., and Peng, J. Shallow–deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:1772–1783, 2020.

Lowe, A., Harrison, N., and French, A. P. Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods*, 13(80), 2017.

Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129(1):23–79, 2021.

Plaza, A., Benediktsson, J. A., Boardman, J. W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J. C., and Trianni, G. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, 2009.

Sabins, F. F. Remote sensing for mineral exploration. *Ore Geology Reviews*, 14:157–183, 1999.

Solomon, A. A. and Agnes, S. A. Land-cover classification with hyperspectral remote sensing image using cnn and spectral band selection. *Remote Sensing Applications: Society and Environment*, 31, 2023.

Thomas, C., Ranchin, T., Wald, L., and Chanussot, J. Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1301–1312, 2008.

Ungar, S. G., Pearlman, J. S., Mendenhall, J. A., and Reuter, D. Overview of the earth observing one (eo-1) mission. *IEEE Transactions on Geoscience and Remote Sensing*, 41(6):1149–1159, 2003.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000—-6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

Vivone, G., Alparone, L., Chanussot, J., Mura, M. D., Garzelli, A., Licciardi, G., Restaino, R., and Wald, L. A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5):2565–2586, 2015.

Wald, L. Quality of high resolution synthesised images: Is there a simple criterion? In *Third conference "Fusion of Earth data: merging point measurements, raster maps and remotely sensed images"*, pp. 99—-103, 2000.

Wang, Y., Chen, X., Han, Z., and He, S. Hyperspectral image superresolution via nonlocal low-rank tensor approximation and total variation regularization. *Remote Sensing*, 9(12):1286, 2017.

Yokoya, N. and Iwasaki, A. Airborne hyperspectral data over chikusei. space application laboratory. Technical report, University of Tokyo, 2016.

Zhang, X., Huang, W., Wang, Q., and Li, X. Ssr-net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5953–5965, 2021.

Zheng, Y., Li, J., Li, Y., Guo, J., Wu, X., and Chanussot, J. Hyperspectral pansharpening using deep prior and dual attention residual network. *IEEE Transactions on Geoscience and Remote Sensing*, 58:8059–8076, 2020.

Zhou, B., Zhang, X., Chen, X., Ren, M., and Feng, Z. Hyperrefiner: a refined hyperspectral pansharpening network based on the autoencoder and self-attention. *International Journal of Digital Earth*, 16(1):3268—-3294, 2023.

Zhuo, Y., Zhang, T., Hu, J., Dou, H., Huang, T., and Deng, L. A deep-shallow fusion network with multidetail extractor and spectral attention for hyperspectral pansharpening. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:7539–7555, 2022.