Learning Multi-Modal Cross-Scale Deformable Transformer Network for Unregistered Hyperspectral Image Super-resolution

Wenqian Dong*, Yang Xu*, Jiahui Qu†, Shaoxiong Hou

State Key Laboratory of Integrated Service Network, Xidian University, Xi'an 710071, China
wqdong@xidian.edu.cn, xuy@stu.xidian.edu.cn, jhqu@xidian.edu.cn, sxhou@stu.xidian.edu.cn

Abstract

Hyperspectral image super-resolution (HSI-SR) is a technology to improve the spatial resolution of HSI. Existing fusion-based SR methods have shown great performance, but still have some problems as follows: 1) existing methods assume that the auxiliary image providing spatial information is strictly registered with the HSI, but images are difficult to be registered finely due to the shooting platforms, shooting viewpoints and the influence of atmospheric turbulence; 2) most of the methods are based on convolutional neural networks (CNNs), which is effective for local features but cannot utilize the global features. To this end, we propose a multi-modal cross-scale deformable transformer network (M^2DTN) to achieve unregistered HSI-SR. Specifically, we formulate a spectrum-preserving based spatial-guided registration-SR unified model (SSRU) from the view of the realistic degradation scenarios. According to SSRU, we propose the multi-modal registration deformable module (MMRD) to align features between different modalities by deformation field. In order to efficiently utilize the unique information between different modals, we design the multi-scale feature transformer (MSFT) to emphasize the spatial-spectral features at different scales. In addition, we propose the cross-scale feature aggregation module (CSFA) to accurately reconstruct the HSI by aggregating feature information at different scales. Experiments show that M^2DTN outperforms the-state-of-the-art HSI-SR methods. Code is obtainable at https://github.com/Jiahuiqu/M2DTN.

Introduction

Hyperspectral imaging can obtain hyperspectral images (HSI) with hundreds of spectral bands. Since the intensity of electromagnetic waves reflected by various ground objects is varying in different bands, HSI provides detailed information of different ground objects. However, due to the limitations of imaging systems, HSI has low spatial resolution, and remote sensing tasks such as change detection (Qu et al. 2023), classification (Zhang et al. 2022a), and semantic segmentation (Zeng et al. 2023) require high accuracy images to obtain meaningful results. In order to obtain HSI with high spatial resolution, hyperspectral image super-resolution (HSI-SR) has appeared and gained widely concern in recent years. HSI-SR can be divided into two categories: single image super-resolution (SISR) and fusion-based super-resolution (fusion-based SR) according to whether high spatial resolution auxiliary images are used. Compared with SISR, the fusion-based SR methods provide the high spatial resolution hyperspectral images (HrHSI) with higher spatial and spectral fidelity.

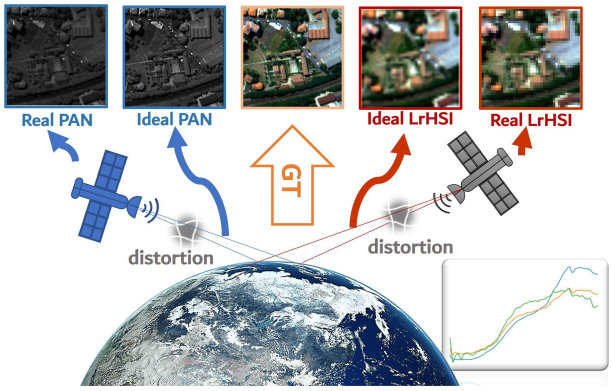

Figure 1: The figure shows the panchromatic and low-resolution hyperspectral images for ideal and realistic conditions, as well as ground truth. Spectral curves for the three HSI are also shown, with data from the Pavia Center dataset.

In recent years, with the rapid development of deep learning (DL), a series of fusion-based SR models based on HSI such as HyperPNN (He et al. 2019), LPPNet (Dong et al. 2021b), and PS-GDANet (Dong et al. 2021a) have emerged, and shown excellent performance. However, these models fuse low spatial resolution hyperspectral images (LrHSI) and high spatial resolution multispectral image (HrMSI) on the premise that the two images are strictly registered. In reality, images are difficult to be registered finely due to the imaging platforms, imaging viewpoints and the influence of atmospheric turbulence as shown in Figure 1, and how to achieve image registration in the fusion process is a more realistic problem. Meanwhile, most of the previous fusion-based SR methods are based on convolutional neural networks (CNNs) such as WSRCNN (Aburaed et al. 2020). Although the CNNs with deep layers can properly fuse the

*These authors contributed equally.
†The corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

spatial-spectral features to a certain extent, the limited receptive field of the convolution kernel cannot utilize the global information of the feature map, resulting in the performance degradation of HSI-SR methods.

In this paper, we propose a multi-modal cross-scale deformable transformer network (M^2DTN) to solve above problems. We formulate a spectral-preserving based spatial-guided registration-SR unified model (SSRU), and expand it to M^2DTN. We take the degenerate scenario in reality as the starting point, deducing and demonstrating the feasibility of SSRU. M^2DTN can be considered as the mapping of each part of SSRU in the deep network, which means that SSRU unitizes each module of M^2DTN as a whole. In M^2DTN, the multi-modal registration deformable module (MMRD) is proposed to generate the deformation field by learning the difference between the spatial distribution of features of HrMSI and LrHSI and then register the LrHSI. For both HrMSI and LrHSI, multi-scale feature transformer (MSFT) is designed to focus on spatial-spectral features at multi-scales. Meanwhile, for better image reconstruction, we propose cross-scale feature aggregation module (CSFA) to aggregate the spatial-spectral features of multi-scale to learn the relationship between the input and the desired HrHSI.

The main contributions of this paper are summarized as follows:

- Based on the realistic degradation scenarios, we formulate a novel model called SSRU, and under the guidance of it, we propose M^2DTN to achieve excellent results on the task of unregistered HSI-SR.
- We propose MMRD for registering HSI, which makes the network learn the spatial distribution of each modality by spatial relation matrix to generate the deformation field.
- Our designed MSFT and CSFA make full use of the multi-scale spatial-spectral information of multi-modal images, which helps generate more refined HrHSI.

Related Work

HSI-SR

Classical HSI-SR methods can be divided into four categories: component substitution (CS), multiresolution analysis (MRA), Bayesian and CNMF. CS-based methods first decompose LrHSI into spectral and spatial components. Subsequently, the spatial component is substituted with the PAN image and both components are converted to the original space. The MRA methods inject spatial features to LrHSI by employing a spatial filter. The limitations of single CS or MRA have led to the emergence of hybrid methods such as guided filtered PCA (GFPCA) (Kang, Li, and Benediktsson 2013). The above classcial methods are prone to spatial blur or spectral distortion due to the mismatch problem when injecting information. Bayesian methods formulate the fusion problem in a Bayesian inference framework, such as Bayesian Fusion (BF) (Wei, Dobigeon, and Tourneret 2015), sparse BF (BSF) (Wei et al. 2015). Bayesian methods like BF and BSF rely on the posterior distribution of LrHSI and HrMSI to estimate the desired image. The Bayesian-based methods consider the problem comprehensively from the model constructed by it, but it usually has large computational costs and requires certain prior conditions. CNMF (Yokoya, Yairi, and Iwasaki 2011) reconstructs the image with the abundance matrix of HrMSI and the endmember matrix of LrHSI.

For the past few years, deep learning (DL) has shown great potential in the field of image processing. A large amount of DL methods have been proposed for HSI-SR, including CNN-based methods, GAN-based methods, and have achieved satisfactory results. (He et al. 2019) proposed an HSI-SR framework named HyperPNN, which designed a spectral prediction capability to spectral difference between HSI and panchromatic images. Following the idea of MRA, (Dong et al. 2021b) proposed LPPNet to reconstruct HrHSI by using Laplacian pyramid to extract multi-scale information. (Mei et al. 2017) utilized 3D-Conv to achieve HSI-SR because 3D-Conv is easier to exploit the properties of HSI. (Dong et al. 2021a) proposed a dual-branch discriminator to compare spectral information with HSI and spatial information with multispectral images (MSI) to achieve HSI-SR. (Yao et al. 2020) designed a coupled convolution autoencoder network (CUCaNet), which attempted to solve the problem of unregistered HSI-SR. Other methods are also proposed for unregistered HSI-SR such as NonRegSRNet (Zheng et al. 2021), etc.

Image Registration

In recent years, image registration has become one of the hot issues in the field of HSI processing. These traditional registration algorithms commonly use an iterative approach to progressively optimize the problem under constraint. Some feature-based methods usually pick some feature points through a scale-invariant Fourier transform or Harris corner detector, and approximate the nonrigid transform with the thin-plate spline or Gaussian radial basis function. With the development of DL, image registration has made significant progress. (Jaderberg et al. 2015) designed the spatial transformer network (STN) to solve the problem of affine transformation between image pairs, which can be inserted anywhere in the framework to learn the transformation parameters of the input feature map. (Li and Fan 2018) used a fully convolutional network (FCN) to execute non-rigid registration of 3D brain magnetic resonance (MR) images by self-supervised. (Shu et al. 2018) proposed a coarse-to-fine unsupervised deformable registration method, in which the mean squared error (MSE) was used as the loss function between the fixed and warped moving images. (Balakrishnan et al. 2019) proposed an unsupervised learning-based method for medical image registration VoxelMorph that could be used to predict a dense deformation field. (Zhao et al. 2019) designed a recursive cascaded network for performing progressive deformation for the registration of deformable images.

Transformer-based Image SR

Transformer architectures have found great success across various computer vision tasks such as image recognition, object detection, and semantic segmentation. Owing to their strong feature representation capability, they are extended

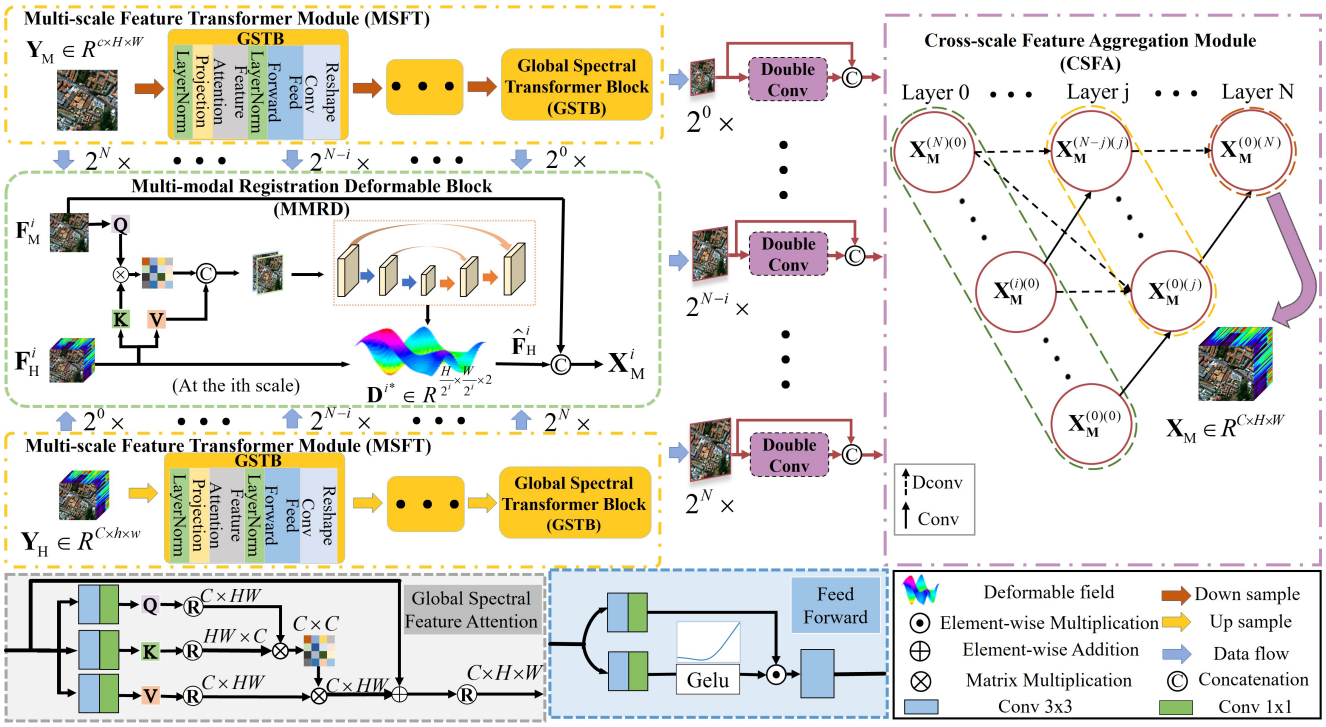

Figure 2: The overview of the M^{2}DTN. In the figure, $i \in [0, N]$ indicates the scale size, and the scale corresponding to the number 0 is the largest scale.

to HSI-SR tasks. However, navie self-attention (Dosovitskiy et al. 2020) is not suitable for HSI, because it has quadratic complexity, and HSI usually contains hundreds of bands. In order to improve the computational efficiency, some efficient transfomer architectures have been proposed. (Liang et al. 2021) used window self-attention blocks along with convolutional layers to improve the efficiency of the model. (Zhang et al. 2022b) divided the input features into five groups equally and used shift convolution to shift the first four groups of features along different spatial dimensions, including left, right, up, and down, leaving the last group unchanged. (Yang et al. 2020) added soft attention map and hard attention map to attention block to transfer texture information from the reference image. (Bandara and Patel 2022) utilized the soft attention of multiple features to obtain better spatial and spectral properties.

The above methods rarely consider the data characteristics of HSI and the problem of image misregistration of different modals simultaneously. We explore this issue next in this paper.

Method

Promblem Formulation

Given two images, LrHSI $\mathbf{Y}_\mathrm{H} \in \mathrm{R}^{C\times h\times w}$ and HrMSI $\mathbf{Y}_\mathrm{M} \in \mathrm{R}^{c\times H\times W}$, their mathematical relationship to HrHSI $\mathbf{X} \in \mathrm{R}^{C\times H\times W}$ under ideal circumstances is shown below

$$\mathbf{Y}_\mathrm{H} \approx (\mathbf{X})\downarrow_{spa} = \mathbf{XB} \tag{1}$$

$$\mathbf{Y}_\mathrm{M} \approx (\mathbf{X})\downarrow_{spe} = \mathbf{RX} \tag{2}$$

where $\mathbf{X}$ represents the HrHSI, $\mathbf{B} \in R^{HW\times hw}$ represents the spatial downsampling matrix containing a blurring operator and a downsampling operator, $\mathbf{R} \in R^{c\times C}$ represents the spectral response matrix of the multispectral image sensor, and H, W, C denote the height, width, and number of channels of the HSI, which are much larger than h, w, c, respectively.

In practice, due to the limitations of imaging systems and imaging conditions, the obtained images are often distorted. Eqs. (1), (2) can be rewritten as

$$\mathbf{Y}_\mathrm{H} = (\mathbf{X} \otimes \mathbf{D}_\mathrm{H})\,\mathbf{B} \tag{3}$$

$$\mathbf{Y}_\mathrm{M} = \mathbf{R}\,(\mathbf{X} \otimes \mathbf{D}_\mathrm{M}) \tag{4}$$

where $\mathbf{D}_\mathrm{H}$ and $\mathbf{D}_\mathrm{M} \in \mathrm{R}^{H\times W\times 2}$ denote the distorted deformation fields of HSI and MSI with respect to $\mathbf{X}$, and $\otimes$ denotes that deformation field acts on $\mathbf{X}$. Thus, we could define $\mathbf{X}_\mathrm{M} = \mathbf{X} \otimes \mathbf{D}_\mathrm{M}$ as the ground truth corresponding to $\mathbf{Y}_\mathrm{M}$. In the same way, $\mathbf{X}_\mathrm{H} = \mathbf{X} \otimes \mathbf{D}_\mathrm{H}$ as the ground truth corresponding to $\mathbf{Y}_\mathrm{H}$. The relation among $\mathbf{X}$, $\mathbf{X}_\mathrm{M}$, and $\mathbf{X}_\mathrm{H}$ can be expressed as

$$\mathbf{X} = \mathbf{X}_\mathrm{M} \otimes \mathbf{D}_\mathrm{M}^* = \mathbf{X}_\mathrm{H} \otimes \mathbf{D}_\mathrm{H}^* \tag{5}$$

where $\mathbf{D}_\mathrm{M}^*$ and $\mathbf{D}_\mathrm{H}^*$ denote the inverse operations of $\mathbf{D}_\mathrm{H}$ and $\mathbf{D}_\mathrm{M}$, respectively.

Usually, $\mathbf{X}_\mathrm{M}$ is considered closer to $\mathbf{X}$ because HrMSI receives less interference during imaging and has higher spatial resolution. Therefore, we choose $\mathbf{X}_\mathrm{M}$ as the desired HrHSI, and the relationship between $\mathbf{X}_\mathrm{M}$ and $\mathbf{X}_\mathrm{H}$ can be expressed as

$$\mathbf{X}_\mathrm{H} = \mathbf{X} \otimes \mathbf{D}_\mathrm{H} = \mathbf{X}_\mathrm{M} \otimes \mathbf{D}_\mathrm{M}^* \otimes \mathbf{D}_\mathrm{H} = \mathbf{X}_\mathrm{M} \otimes \mathbf{D} \tag{6}$$

where $\mathbf{D}$ represents the combined deformation field of $\mathbf{D}_{\mathrm{M}}^{*}$ and $\mathbf{D}_{\mathrm{H}}$. According to the above theoretical derivation, Eqs. (3), (4) can be further written as

$$\mathbf{Y}_{\mathrm{H}} = (\mathbf{X}_{\mathrm{M}} \otimes \mathbf{D})\mathbf{B} \tag{7}$$

$$\mathbf{Y}_{\mathrm{M}} = \mathbf{R}\mathbf{X}_{\mathrm{M}} \tag{8}$$

Our goal is to solve for $\mathbf{X}_{\mathrm{M}}$, which reconstructs the HrHSI by fully considering the spatial details of $\mathbf{Y}_{\mathrm{M}}$ and spectral information of $\mathbf{Y}_{\mathrm{H}}$ in the case of the unregistered images. We formulate SSRU, which is defined as follows

$$\mathbf{X}_{\mathrm{M}} = \alpha\mathbf{R}^{*}\mathbf{Y}_{\mathrm{M}} + \beta(\mathbf{Y}_{\mathrm{H}}\mathbf{B}^{*}) \otimes \mathbf{D}^{*} \tag{9}$$

where $\mathbf{R}^{*}$, $\mathbf{B}^{*}$, and $\mathbf{D}^{*}$ denote the inverse operations of $\mathbf{R}$, $\mathbf{B}$, and $\mathbf{D}$, respectively, and $\alpha, \beta \in [0, 1]$ satisfy $\alpha + \beta = 1$. Unlike Eq. (9), there is more than one $\mathbf{X}_{\mathrm{M}}$ if only Eq. (7) or Eq. (8) is satisfied, which obviously incompatible with the ground truth. In other words, the real $\mathbf{X}_{\mathrm{M}}$ is a superposition of two modal images, which conforms to the spatial prior of $\mathbf{Y}_{\mathrm{M}}$ and the spectral characteristics of $\mathbf{Y}_{\mathrm{H}}$.

Overall Pipeline

In the previous section, we construct the SSRU according to the realistic degradation scenarios. The M^{2}DTN is proposed to map the parts of SSRU. The overall pipeline of M^{2}DTN is shown in Figure 2.

According to SSRU, it is crucial to learn the spatial distribution and positional differences between HrMSI and LrHSI to generate the deformation field. To this end, we propose MMRD to learn the spatial distribution by directly modeling the relationship of the spatial dimension between the multi-modal images on the spatial dimension of the two modals. Each layer of MMRD can be expressed as follows

$$\hat{\mathbf{F}}_{\mathrm{H}}^{i} = MMRD(\mathbf{F}_{\mathrm{H}}^{i}, \mathbf{F}_{\mathrm{M}}^{i}) = \mathbf{F}_{\mathrm{H}}^{i} \otimes \mathbf{D}^{i*}\left(\mathbf{F}_{\mathrm{H}}^{i}, \mathbf{F}_{\mathrm{M}}^{i}\right) \tag{10}$$

where $\mathbf{F}_{\mathrm{H}}^{i}$, $\mathbf{F}_{\mathrm{M}}^{i}$, and $\hat{\mathbf{F}}_{\mathrm{H}}^{i}$ denote multi-scale unregistered LrHSI features, multi-scale HrMSI features, and multi-scale registered LrHSI features, respectively, $MMRD(\cdot)$ denotes MMRD, and $\mathbf{D}^{i*}$ denotes the deformation field between $\mathbf{F}_{\mathrm{H}}^{i}$ and $\mathbf{F}_{\mathrm{M}}^{i}$.

In M^{2}DTN, our proposed CSFA completes the process of image reconstruction. Before that, in order to improve the performance of M^{2}DTN, we also design MSFT to effectively utilize the information of both modals.

MSFT is a pyramid structure consisting of N global spectral transformer blocks (GSTB), which can focus on spatial features at different scales. Meanwhile, in GSTB, we apply global spectral feature attention to focus on spectral information. The MSFT module is adopted to parallel extract multi-scale features of HSI and MSI in M^{2}DTN, in which HrMSI is progressively downsampled and LrHSI is progressively upsampled to obtain feature maps of the same size. Each layer of MSFT can be expressed as follows

$$\mathbf{F}_{\mathrm{M}}^{i} = \begin{cases} GSTB\left(\mathbf{F}_{\mathrm{M}}^{i-1}\right) & , otherwise \\ \mathbf{Y}_{\mathrm{M}} & , i = 0 \end{cases} \tag{11}$$

$$\mathbf{F}_{\mathrm{H}}^{i} = \begin{cases} GSTB\left(\mathbf{F}_{\mathrm{H}}^{i+1}\right) & , otherwise \\ \mathbf{Y}_{\mathrm{H}} & , i = N \end{cases} \tag{12}$$

where $\mathbf{F}_{\mathrm{M}}^{i}$ and $\mathbf{F}_{\mathrm{H}}^{i}$ denote the features with serial number $i \in [0, N]$. The scale size of feature with the serial number i to be $R^{n \times \frac{H}{2^i} \times \frac{W}{2^i}}$, and n denotes the number of channels of the corresponding modal.

Finally, the registered multi-scale features are processed with double-layer convolution to generate reconstructed features $\mathbf{X}_{\mathrm{M}}^{i}$. The multi-scale reconstructed features are fed into CSFA to reconstruct HrHSI, which aggregates information across scales to obtain more powerful feature representations. The process of CSFA to reconstruct HrHSI can be mathematized as follows

$$\mathbf{X}_{\mathrm{M}} = CSFA(\mathbf{X}_{\mathrm{M}}^{0}, \cdots, \mathbf{X}_{\mathrm{M}}^{i}, \cdots, \mathbf{X}_{\mathrm{M}}^{N}) \tag{13}$$

where $\mathbf{X}_{\mathrm{M}}^{i}$ represents the reconstructed feature with scale index i, $\mathbf{X}_{\mathrm{M}}$ represents the final reconstruction result of M^{2}DTN, and $CSFA(\cdot)$ represents the CSFA. In the following sections, we detail MSFT, MMRD, and CSFA.

Multi-Scale Feature Attention Transformer

Due to the inconsistency of spatial resolution between LrHSI and HrMSI, it is difficult to obtain the size matched features. To this end, MSFT, a pyramid structure consisting of a series of GSTB, is proposed to generate multiple image pairs from both modals with consistent scales. For the spectral information, GSTB applies global spectral feature attention across channel dimensions to focus on the channel information of HSI. Global spectral feature attention use 1×1 convolution in projection layers to aggregate the information of different channels, while deep convolution is used to calculate contextual features channel by channel. On the other hand, MSFT employs a series of GSTB designed for hierarchical multi-scale feature extraction for the spatial information, thereby efficiently capturing the intricate spatial details of HrMSI and adeptly accommodating features across varying scales. Taking the generation of $\mathbf{F}_{\mathrm{H}}^{i}$ on the HSI branch as an example, the computational procedure of GSTB is expressed as follows

$$\mathbf{Q}_{\mathrm{H}}^{i}, \mathbf{K}_{\mathrm{H}}^{i}, \mathbf{V}_{\mathrm{H}}^{i} = W_q\left(\mathbf{F}_{\mathrm{H}}^{i+1}\right), W_k\left(\mathbf{F}_{\mathrm{H}}^{i+1}\right), W_v\left(\mathbf{F}_{\mathrm{H}}^{i+1}\right) \tag{14}$$

$$\mathbf{Attn}_{\mathrm{H}}^{i} = Softmax\left(\frac{\mathbf{Q}_{\mathrm{H}}^{i}(\mathbf{K}_{\mathrm{H}}^{i})^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\mathrm{H}}^{i} \tag{15}$$

$$\mathbf{F}_{\mathrm{H}}^{\prime i} = \mathbf{Attn}_{\mathrm{H}}^{i} + \mathbf{F}_{\mathrm{H}}^{i+1} \tag{16}$$

$$\mathbf{F}_{\mathrm{H}}^{i} = ReLU\left(Conv\left(\mathbf{F}_{\mathrm{H}}^{\prime i}\right)\right) \cdot Conv\left(\mathbf{F}_{\mathrm{H}}^{\prime i}\right) + \mathbf{F}_{\mathrm{H}}^{\prime i} \tag{17}$$

where $\mathbf{Q}_{\mathrm{H}}^{i}$, $\mathbf{K}_{\mathrm{H}}^{i}$, $\mathbf{V}_{\mathrm{H}}^{i}$, $\mathbf{Attn}_{\mathrm{H}}^{i}$ represent the qurey, key, value and attention map at the i-th scale in the HSI branch, respectively. Similar to the HSI branch, we adopt a series of GSTB to generate the feature $\mathbf{F}_{\mathbf{M}}^{i}$ corresponding to the feature $\mathbf{F}_{\mathrm{H}}^{i}$ in the MSI branch.

Multi-modal Registration Deformable Block

As mentioned before, the spatial distribution of the feature at the same scale on the two modals is not completely consistent. Therefore, we propose MMRD to align the features of LrHSI under the guidance of the spatial information of HrMSI. Inspired by the cross-attention mechanism, we compute the spatial relationship matrix by multiplying the query

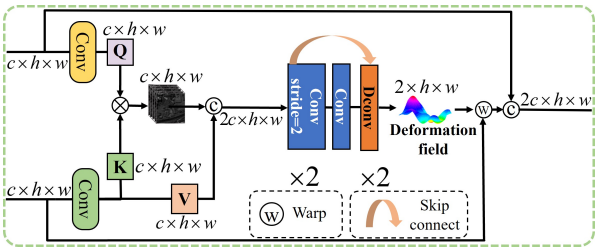

Figure 3: The specific process of MMRD.

matrix $\mathbf{Q}^i$, derived from the projection of $\mathbf{F}_{\mathrm{M}}^i$, with the key matrix $\mathbf{K}^i$, obtained from the projection of $\mathbf{F}_{\mathrm{H}}^i$. This matrix is then concatenated with the value matrix $\mathbf{V}^i$, which also comes from the projection of $\mathbf{F}_{\mathrm{H}}^i$. In this process, instead of reshaping the projection, we directly compute the spatial relationship matrix in the spatial dimension. We fit deformation field using a U-net structure with a symmetric structure containing one encoder and decoder as shown in Figure 3. It can be considered that the common effect produced by multi-scale deformation fields at all scales is the same as $\mathbf{D}^*$ in the ideal state in Eq. (9). The process of registering $\mathbf{F}_{\mathrm{H}}^i$ and $\mathbf{F}_{\mathrm{M}}^i$ by MMRD can be expressed as follows

$$\mathbf{Q}^i, \mathbf{K}^i, \mathbf{V}^i = W_q\left(\mathbf{F}_{\mathrm{M}}^i\right), W_k\left(\mathbf{F}_{\mathrm{H}}^i\right), W_v\left(\mathbf{F}_{\mathrm{H}}^i\right) \tag{18}$$

$$\mathbf{Spa}^i = \mathbf{Q}^i(\mathbf{K}^i)^T \tag{19}$$

$$\mathbf{D}^{i*} = f_D(concat(\mathbf{Spa}^i, \mathbf{V}^i)) \tag{20}$$

$$\hat{\mathbf{F}}_{\mathrm{H}}^i = \mathbf{F}_{\mathrm{H}}^i \otimes \mathbf{D}^{i*} \tag{21}$$

where $\mathbf{Spa}^i$ and $\mathbf{D}^{i*}$ represent the spatial relationship matrix and the deformation field at the i-th scale, respectively, and $f_D(\cdot)$ represents the function of U-net.

Cross-scale Feature Aggregation Module

We design CSFA for HrHSI reconstruction as shown in Figure 2. According to SSRU, the process of reconstructing HrHSI is to solve for Eq. (9). CSFA is capable of aggregating feature information at different scales, which contains N feature transfer structures. Our goal is to get the HSI with the largest scale, so CSFA does not have a branch that transfers features from large scale to small scale. Taking a node in CSFA as an example, it can be expressed as

$$\mathbf{X}_{\mathrm{M}}^{(i)(j)} = g(\hat{\mathbf{F}}_{\mathrm{H}}^i, \mathbf{F}_{\mathrm{M}}^i), j = 0, i \in [0, N] \tag{22}$$

$$\begin{aligned}\mathbf{X}_{\mathrm{M}}^{(i)(j)} = f_R(&\mathbf{X}_{\mathrm{M}}^{(i)(j-1)}, \mathbf{X}_{\mathrm{M}}^{(i+1)(j-1)}, \cdots, \\ &\mathbf{X}_{\mathrm{M}}^{(j-1)(j-1)}), j \in [1, N], i \in [0, N-j]\end{aligned} \tag{23}$$

where $\mathbf{X}_{\mathrm{M}}^{(i)(j)}$ denotes the feature node at the i-th scale of layer j in CSFA, $g(\cdot)$ denotes the double-layer convolution operation used to generate the reconstructed features at the first layer, and in the following three layers, $f_R(\cdot)$ stands for feature transfer structure consisting of a sequence of convolutions.

Loss Function

In order to optimize the parameters of the proposed M^{2}DTN, we choose l_1 norm as the loss function to constrain the reconstructed HSI. The loss function can be shown as

$$Loss = \left\|\mathbf{X}_{\mathrm{M}} - \tilde{\mathbf{X}}_{\mathrm{M}}\right\|_1 \tag{24}$$

where $\tilde{\mathbf{X}}_{\mathrm{M}}$ and $\mathbf{X}_{\mathrm{M}}$ represent the reference HrHSI and the reconstruction result of the M^{2}DTN respectively.

Experiment

Dataset

We conduct experiments on three publicly available datasets: Pavia Center (PaviaC), Houston, and Harvard.

The PaviaC dataset is acquired by ROSIS sensor over the city of Pavia, Italy, in the wavelength range of 430–860 nm. After removing 13 absorbent and noise bands, 102 bands are used for the dataset. We intercept a part of the whole graph of size $960 \times 640 \times 102$. To increase the number of samples, we further divide the image into 252 patches of size $160 \times 160 \times 102$, of which 189 patches are used for training and 63 patches for testing.

The Houston dataset is acquired by ITRES-CASI 1500 HS sensor covering the campus of the University of Houston and its neighboring urban areas in the wavelength range of 380–1050 nm. The size of the entire Houston image is 349 $\times$ 1905 $\times$ 144. An image of size 320×1280×144 is used as the ground truth image. We also divide it into 8 parts, and each part is then split into 41 patches of size 160×160×144.

Harvard dataset (Chakrabarti and Zickler 2011) contains a total of 50 indoor and outdoor images acquired using a commercial hyperspectral camera (Nuance FX, CRI Inc.). The pictures contain 31 bands with a wavelength interval of 10 nm from 420 to 720 nm. We randomly select 30 images of size 960×960×31 as training set and split it into 160×160×31 patches to increase the number of samples. Among the remaining 20 images, we randomly select 10 images of size 960×960×31 as the testing set.

To evaluate the performance of the proposed method, the LrHSI and HrMSI in the above three datasets are obtained by Wald's protocol. Moreover, we spatially distort the LrHSI using an elastic transformation to simulate Eq. (6). Elastic transformation is used for nonlinear distortion of camera lens, which is equivalent to the distortion and blur produced by different imaging devices when the sensor has jitter.

Competing Methods and Quality Assessment

In this section, we present the competing methods of this paper as well as the quality assessment metrics.

1) Competing Methods: We select fusion-based HSI-SR methods for unregistration images, including NonReg (Zheng et al. 2021), PixAwa (Wei et al. 2020), CUCaNet (Yao et al. 2020), and u2MDN (Qu et al. 2021) as competing methods. The fusion-based HSI-SR method designed for registration images named Hyperformer (Bandara and Patel 2022) is also used for comparison. The traditional fusion-based HSI-SR methods including GSA (Aiazzi, Baronti, and Selva 2007), Hysure (Lanaras, Baltsavias, and Schindler

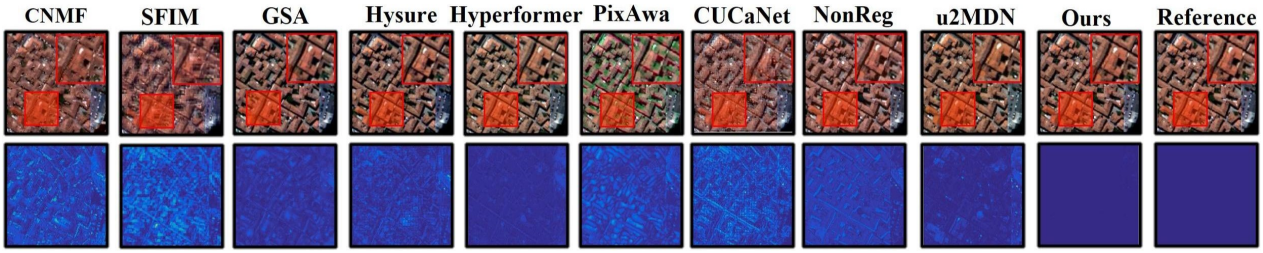

Figure 4: Visual results and the MAE images generated by different HSI-SR algorithms for unregistered Pavia Center Dataset.

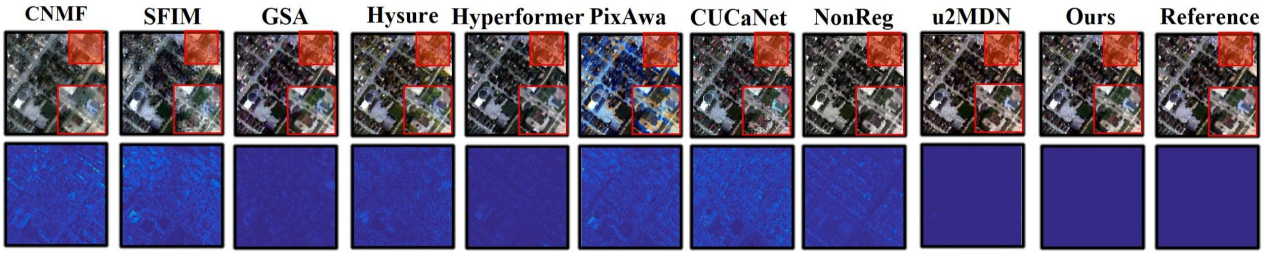

Figure 5: Visual results and the MAE images generated by different HSI-SR algorithms for unregistered Houston Dataset.

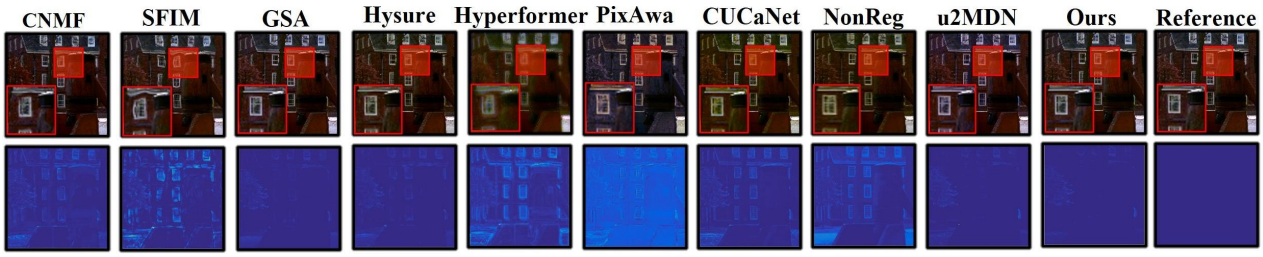

Figure 6: Visual results and the MAE images generated by different HSI-SR algorithms for unregistered Harvard Dataset.

2015), CNMF (Yokoya, Yairi, and Iwasaki 2011), and SFIM (Liu 2000) are also selected to evaluate the effectiveness of the proposed method.

2) Quality Assessment: To evaluate the similarity between the reconstructed image and the ground truth image, we selected five widely used quality assessment metrics, namely the saptial measure cross correlation (CC), the spectral measure spectral angle mapper (SAM) and the global measure peak signal-to-noise ratio (PSNR), root mean squared error (RMSE) and erreur relative globale adimensionnelle de synthese (ERGAS). The lower the SAM, RMSE, and ERGAS, and the higher the PSNR, and CC, the better the performance of the reconstructed images.

Experimental Detail

The proposed method was implemented on NVIDIA GTX 3090 GPUs based on PyTorch framework and trained using Adam optimizer with learning rate set to 0.001. In the experiment, we designed the scale layers as 4.

Experimental Result

Table 1 shows the five metrics of each method on the three datasets. As shown in Table 1, the proposed M^2DTN outperforms the existing SOTA methods on each dataset. According to the experimental results, when dealing with the task of unregistered HSI-SR, the traditional methods such as CNMF, SFIM, GSA, and Hysure perform better on the Harvard dataset than the other datasets, which means the traditional methods are greatly affected by the data imaging conditions. Unsupervised unregistered HSI-SR methods such as PixAwa, CUCaNet, and NonReg are generally better than traditional methods, but they still fall short of the current baseline of HSI-SR tasks. Only u2MDN performs well but is still not comparable to our method. Hyperformer has shown excellent performance on registered datasets, but it can be seen from the experimental results that it cannot deal with the problem of unregistered HSI-SR.

Figures 4, 5, and 6 show the visual evaluation results on the Pavia Center, Houston, and Harvard datasets, respectively. We randomly select an image from the testing set of each dataset to display the pseudo-color images of the SR results and the mean absolute error (MAE) images between the reference images and SR results. The more obvious the texture in the MAE image, the greater its difference from the reference image. It can be seen from the pseudo-color images of the experimental results that the reconstructed image of the proposed method is closest to the reference image. The pseudo-color images and MAE images together reflect

Method	PaviaC Dataset					Houston Dataset					Harvard Dataset				
	CC	SAM	RMSE	ERGAS	PSNR	CC	SAM	RMSE	ERGAS	PSNR	CC	SAM	RMSE	ERGAS	PSNR
CNMF	0.7918	8.2464	0.0523	8.8173	25.6260	0.8553	8.2503	0.0184	5.8324	26.3058	0.9709	3.6563	0.0110	5.3531	38.7914
SFIM	0.7538	7.1549	0.0572	8.7713	24.6293	0.8287	6.9124	0.0200	5.8130	25.6253	0.9378	3.3160	0.0168	7.3863	35.4870
GSA	0.9754	11.5657	0.0242	3.8124	32.0559	0.9710	3.7143	0.0076	2.316	33.8073	0.9907	2.8778	0.0060	3.7092	43.1363
Hysure	0.9452	14.0991	0.0447	7.0190	28.3515	0.9491	5.7847	0.0122	3.9145	29.5606	0.9916	3.4182	0.0055	2.9784	44.4775
Hyperformer	0.9787	7.3411	0.0185	3.5237	34.1979	0.9848	3.1672	0.0085	2.6772	36.2336	0.9128	6.1338	0.0190	12.0280	33.7227
PixAwa	0.8677	12.2810	0.0406	7.2424	28.0620	0.8608	7.5308	0.0198	5.7020	29.3526	0.9679	5.5013	0.0311	5.8107	29.5829
CUCaNet	0.9557	6.5692	0.0179	5.3505	29.9352	0.9926	5.9353	0.0256	5.7003	29.1805	0.9485	4.812 3	0.0137	6.6638	38.1548
NonReg	0.9743	5.5621	0.0223	3.8024	33.2368	0.9792	4.2441	0.0134	3.1238	34.4541	0.9915	4.6948	0.1240	5.3448	37.6673
u2MDN	0.9902	4.9743	0.0093	2.4019	41.0935	0.9913	1.4987	0.0037	1.2662	42.6521	0.9903	4.0214	0.0055	2.8396	44.5377
Ours	**0.9917**	**2.7471**	**0.0057**	**1.5308**	**45.3475**	**0.9932**	**1.0295**	**0.0021**	**1.0228**	**48.8570**	**0.9928**	**2.9371**	**0.0041**	**2.3960**	**45.8886**

Table 1: Unregistered HSI-SR results obtained by the proposed method and its competing methods on the Pavia Center, Houston and Harvard dataset. Bold font is best.

Method	Dataset	CC	SAM	RMSE	ERGAS	PSNR
MSA	PaviaC	0.9778	4.1225	0.0142	2.7990	36.4423
	Houston	0.9489	4.0193	0.0134	3.7225	32.5845
	Harvard	0.9001	7.6965	0.0275	5.5798	23.2044
Proposed	PaviaC	0.9917	2.7471	0.0057	1.5308	45.3475
	Houston	0.9932	1.0295	0.0021	1.0228	48.8570
	Harvard	0.9928	2.9371	0.0041	2.3960	45.8886

Table 2: Experimental results of multi-head self-attention and our model.

Method	Dataset	CC	SAM	RMSE	ERGAS	PSNR
without MMRD	PaviaC	0.9893	3.0732	0.0068	1.7713	43.6667
	Houston	0.9908	1.3384	0.0030	1.2600	45.2705
	Harvard	0.9889	3.8842	0.0073	3.5382	43.0694
Proposed	PaviaC	0.9917	2.7471	0.0057	1.5308	45.3475
	Houston	0.9932	1.0295	0.0021	1.0228	48.8570
	Harvard	0.9928	2.9371	0.0041	2.3960	45.8886

Table 3: Experimental results with MMRD and without MMRD.

the effectiveness of the proposed M^{2}DTN in dealing with the task of unregistered HSI-SR and the ability to deal with details.

Ablation Study

Multi-scale feature attention transformer In order to verify the powerful feature extraction ability of the proposed MSFT, we replace the global spectral feature attention with the ordinary multi-head self attention, and to control the variables, we retain the multi-scale structure in MSFT. The experimental results are shown in Table 2, and the results show that the MSFT with global spectral feature attention designed in this paper achieves better results.

Multi-modal registration deformable module In order to reflect the fine registration effect of the proposed MMRD, in the ablation study, we delete MMRD while retaining other modules, and directly cascade the extracted features into CSFA. The experimental results are shown in Table 3, and the experimental results show that MMRD can fine register the HSI to achieve better results.

Cross-scale feature aggregation module We believe that information aggregation across scales can better exploit the complex spatial-spectral features of HSI. In the third set of ablation studies, we replace CSFA with a traditional U-net decoder whose information is transmitted only at a single scale. The experimental results are shown in Table 4. Compared with the traditional U-net decoder, CSFA achieves better results.

Method	Dataset	CC	SAM	RMSE	ERGAS	PSNR
UNet	PaviaC	0.9889	3.3087	0.0082	1.8581	41.6127
	Houston	0.9890	1.5877	0.0037	1.4582	43.1989
	Harvard	0.9800	5.9059	0.0084	4.0775	39.7439
Proposed	PaviaC	0.9917	2.7471	0.0057	1.5308	45.3475
	Houston	0.9932	1.0295	0.0021	1.0228	48.8570
	Harvard	0.9928	2.9371	0.0041	2.3960	45.8886

Table 4: Experimental results using CSFA and using U-net in the feature reconstruction part.

Conclusion

In this paper, we formulate SSRU and derive the model from the realistic degradation scenarios. According to this model, we design M^{2}DTN to solve the problem of unregistered HSI-SR. Specifically, in M^{2}DTN, we design MSFT for feature extraction, MMRD for feature registration, and CSFA for feature reconstruction. MSFT uses multi-scale structure and global feature attention to efficiently extract the spatial-spectral features of the image. MMRD registers the HSI at multiple scales according to the spatial distribution of the features. Finally, CSFA aggregates the multi-scale feature information to reconstruct the HrHSI. We conduct experiments on three public datasets, and the experimental results show that the proposed method achieves excellent results on five widely accepted objective indicators.

Acknowledgments

This work was supported in part by the the National Natural Science Foundation of China under Grant 62101414 and Grant 62201423, Young Talent Fund of Xi'an Association for Science and Technology under Grant 095920221320 and Grant 959202313052, the China Postdoctoral Science Special Foundation under Grant 2022T150508 and 2023T160502, the Youth Innovation Team of Shaanxi Universities, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117, and the China Postdoctoral Science Foundation under Grant 2021M702546 and 2021M702548.

References

Aburaed, N.; Panthakkan, A.; Al-Saad, M.; El Rai, M. C.; Al Mansoori, S.; Al-Ahmad, H.; and Marshall, S. 2020. Super-resolution of satellite imagery using a wavelet multiscale-based deep convolutional neural network model. In *Image and Signal Processing for Remote Sensing XXVI*, volume 11533, 305–311. SPIE.

Aiazzi, B.; Baronti, S.; and Selva, M. 2007. Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10): 3230–3239.

Balakrishnan, G.; Zhao, A.; Sabuncu, M. R.; Guttag, J.; and Dalca, A. V. 2019. VoxelMorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8): 1788–1800.

Bandara, W. G. C.; and Patel, V. M. 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.

Chakrabarti, A.; and Zickler, T. 2011. Statistics of Real-World Hyperspectral Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 193–200.

Dong, W.; Hou, S.; Xiao, S.; Qu, J.; Du, Q.; and Li, Y. 2021a. Generative dual-adversarial network with spectral fidelity and spatial enhancement for hyperspectral pansharpening. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12): 7303–7317.

Dong, W.; Zhang, T.; Qu, J.; Xiao, S.; Liang, J.; and Li, Y. 2021b. Laplacian pyramid dense network for hyperspectral pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, L.; Zhu, J.; Li, J.; Plaza, A.; Chanussot, J.; and Li, B. 2019. 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.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Kang, X.; Li, S.; and Benediktsson, J. A. 2013. Spectral–spatial hyperspectral image classification with edge-preserving filtering. *IEEE transactions on geoscience and remote sensing*, 52(5): 2666–2677.

Lanaras, C.; Baltsavias, E.; and Schindler, K. 2015. Hyperspectral super-resolution by coupled spectral unmixing. In *Proceedings of the IEEE international conference on computer vision*, 3586–3594.

Li, H.; and Fan, Y. 2018. Non-rigid image registration using self-supervised fully convolutional networks without training data. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 1075–1078. IEEE.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Liu, J. 2000. Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21(18): 3461–3472.

Mei, S.; Yuan, X.; Ji, J.; Zhang, Y.; Wan, S.; and Du, Q. 2017. Hyperspectral image spatial super-resolution via 3D full convolutional neural network. *Remote Sensing*, 9(11): 1139.

Qu, J.; Zhao, J.; Dong, W.; Xiao, S.; Li, Y.; and Du, Q. 2023. Feature Mutual Representation Based Graph Domain Adaptive Network for Unsupervised Hyperspectral Change Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 1–1.

Qu, Y.; Qi, H.; Kwan, C.; Yokoya, N.; and Chanussot, J. 2021. Unsupervised and unregistered hyperspectral image super-resolution with mutual Dirichlet-Net. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–18.

Shu, C.; Chen, X.; Xie, Q.; and Han, H. 2018. An unsupervised network for fast microscopic image registration. In *Medical Imaging 2018: Digital Pathology*, volume 10581, 363–370. SPIE.

Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; and Tourneret, J.-Y. 2015. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Transactions on Geoscience and Remote Sensing*, 53(7): 3658–3668.

Wei, Q.; Dobigeon, N.; and Tourneret, J.-Y. 2015. Bayesian fusion of multi-band images. *IEEE Journal of Selected Topics in Signal Processing*, 9(6): 1117–1127.

Wei, W.; Nie, J.; Zhang, L.; and Zhang, Y. 2020. Unsupervised recurrent hyperspectral imagery super-resolution using pixel-aware refinement. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–15.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning Texture Transformer Network for Image Super-Resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao, J.; Hong, D.; Chanussot, J.; Meng, D.; Zhu, X.; and Xu, Z. 2020. Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 208–224. Springer.

Yokoya, N.; Yairi, T.; and Iwasaki, A. 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2): 528–537.

Zeng, X.; Wang, T.; Dong, Z.; Zhang, X.; and Gu, Y. 2023. Superpixel Consistency Saliency Map Generation for Weakly Supervised Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Zhang, T.; Xiao, S.; Dong, W.; Qu, J.; and Yang, Y. 2022a. A Mutual Guidance Attention-Based Multi-Level Fusion Network for Hyperspectral and LiDAR Classification. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Zhang, X.; Zeng, H.; Guo, S.; and Zhang, L. 2022b. Efficient Long-Range Attention Network for Image Super-resolution. arXiv:2203.06697.

Zhao, S.; Dong, Y.; Chang, E. I.; Xu, Y.; et al. 2019. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10600–10610.

Zheng, K.; Gao, L.; Hong, D.; Zhang, B.; and Chanussot, J. 2021. NonRegSRNet: A nonrigid registration hyperspectral super-resolution network. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.