UMAG-Net: A New Unsupervised Multiattention-Guided Network for Hyperspectral and Multispectral Image Fusion

Shuaiqi Liu, Siyu Miao, Jian Su, Bing Li, Weiming Hu, and Yu-Dong Zhang

***Abstract*—To reconstruct images with high spatial resolution and high spectral resolution, one of the most common methods is to fuse a low-resolution hyperspectral image (HSI) with a high-resolution (HR) multispectral image (MSI) of the same scene. Deep learning has been widely applied in the field of HSI-MSI fusion, which is limited with hardware. In order to break the limits, we construct an unsupervised multiattention-guided network named UMAG-Net without training data to better accomplish HSI-MSI fusion. UMAG-Net first extracts deep multiscale features of MSI by using a multiattention encoding network. Then, a loss function containing a pair of HSI and MSI is used to iteratively update parameters of UMAG-Net and learn prior knowledge of the fused image. Finally, a multiscale feature-guided network is constructed to generate an HR-HSI. The experimental results show the visual and quantitative superiority of the proposed method compared to other methods.**

***Index Terms*—Deep learning, hyperspectral images (HSIs), image fusion, multispectral images (MSIs).**

I. Introduction

REMOTE sensing hyperspectral images (HSIs) are images of high spectral dimensions consisting of hundreds or even thousands of narrow bands [1]. Benefiting from the high spectral resolution, they are more sensitive to subtle changes in reflected energy and can be used for material identification. Images produced by hyperspectral sensors contain more information than those produced by multispectral sensors that have material identification properties. As a result, the formers are widely used in environmental monitoring, military, industrial and agricultural applications [2]–[10]. However, due to the physical constraints of the imaging equipment, the radiant light energy received by the hyperspectral imaging sensor is divided among many bands, which results in a lower spatial resolution for HSI than multispectral image (MSI). The low spatial resolution severely affects the use of HSI in computer vision-related tasks [11]. In order to improve the spatial resolution of HSI, we usually fuse high-resolution (HR) MSI and low-resolution (LR) HSI to HR-HSI. This process is also known as HSI super-resolution reconstruction. The reconstructed HR-HSI can better perform computer vision tasks such as anomaly detection [12] and change detection [13]. Generally, the HSI-MSI fusion algorithms can be classified into four kinds, such as extensions fusion method based on pansharpening, fusion methods based on matrix factorization, fusion methods based on tensor representation, and fusion methods based on deep learning.

Early methods to fusing spatial and spectral information, which aimed to fuse LR-MSI with HR panchromatic (PAN) images to enhance the spatial and spectral resolution of the fused image, are known as pansharpening image fusion methods. Subsequently, pansharpening image fusion algorithms have been gradually extended to HSI-MSI fusion. For example, Aiazzi *et al.* [14] proposed an HSI-MSI fusion method by using the spectral response function (SRF). Zhang and He [15] proposed an HSI-MSI fusion method through three-dimensional (3-D) wavelet transform. Chen *et al.* [16] proposed a fusion framework of HSI and MSI based on region chunking. Selva *et al.* [17] applied HR MSI to construct high spatial resolution images of each spectral band of HSI by linear regression, and then we can get the final fused image by HSI and the synthesized images. In general, extended fusion methods based on pansharpening are simple and efficient, but the quality of the fusion needs to be improved.

The degradation from high spatial resolution to low spatial resolution can be regarded as the process of image element blending, whereas image fusion is the inverse process, which can be regarded as the process of unmixing. Therefore, a matrix factorization-based approach can be used for MSI fusion. In recent years, image fusion algorithms based on matrix

Manuscript received June 9, 2021; revised July 2, 2021; accepted July 10, 2021. Date of publication July 14, 2021; date of current version August 2, 2021. This work was supported in part by the Natural Science Foundation of Hebei Province under Grant F2020201025, Grant F2019201151, and Grant F2018210148; and in part by the Science Research Project of Hebei Province under Grant BJ2020030 and Grant QN2017306, and the National Natural Science Foundation of China under Grant 61572063 and Grant 62172003. *(Corresponding authors: Jian Su; Yu-Dong Zhang.)*

Shuaiqi Liu is with the College of Electronic and Information Engineering, Machine Vision Technological Innovation Center of Hebei, Hebei University, Baoding 071002, China, and also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: shdkj-1918@163.com).

Siyu Miao is with the Key Laboratory of Digital Medical Engineering of Hebei Province, College of Electronic and Information Engineering, Hebei University, Baoding 071002, China (e-mail: siyumiao_hbu@163.com).

Jian Su is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210094, China (e-mail: sj890718@gmail.com).

Bing Li and Weiming Hu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: bli@nlpr.ia.ac.cn; wmhu@nlpr.ia.ac.cn).

Yu-Dong Zhang is with the Department of Informatics, University of Leicester, LE1 7RH Leicester, U.K. (e-mail: yudongzhang@ieee.org).

Digital Object Identifier 10.1109/JSTARS.2021.3097178

factorization have been widely developed for their intuitive interpretation of fusion results. For example, Yokoya *et al.* [18] gave an HSI-MSI image fusion algorithm by coupled nonnegative matrix decomposition. In this algorithm, they used the vertex component analysis algorithm to extract the initial endmember features in LR-HSI. The endmember matrix of LR-HSI and the abundance matrix of HR-MSI were then iteratively derived by using the sensor's sensing model and the matrix decomposition algorithm. The two matrices were multiplied together to obtain an HR fusion result. Dong *et al.* [19] first used LR HSI to learn overcomplete dictionaries and then proposed an image fusion method by using nonlocal (NL) similarity and sparse prior, which can effectively improve the spatial resolution of the fused images. And in [20], He *et al.* gave an HSI-MSI image fusion algorithm based on robust nonnegative matrix factorization with sparse noise regularizers. Han *et al.* [21] modeled the global similarity of HR-HSI by grouping similar blocks and combined it with a constrained sparse representation for HSI super-resolution reconstruction. Wei *et al.* [22] proposed a fast multiband HSI-MSI image fusion method. The method constructed a closed-form solution of the corresponding Sylvester equation by using a circular matrix and downsampling matrix, which greatly saved the running time of this algorithm. Furthermore, combined with the alternating iteration algorithm and the block coordinate descent algorithm, it can be easily extended to fusion methods based on Bayesian estimation, which can lead to better fusion results.

In recent years, tensor analysis is widely used in the field of HSI. Dian *et al.* [23] proposed an HSI super-resolution algorithm based on NL sparse tensor decomposition. The algorithm treated the HSI as a 3-D tensor. And the sparse Tucker factorization was used to decompose the HSI into a 3-D core tensor and a 2-D dictionary of three modes. Then, the input HR-MSI is divided into several image blocks, and it is considered that the set of image blocks belonging to the same class share a common spatial and spectral dictionary based on the NL self-similarity of images. The spectral dictionary is learned from LR-HSI image block set, and spatial dictionary is learned from HR-MSI image block set, whereas core tensor is extracted from HR-MSI image block by tensor sparse coding. Finally, the HSI fusion image is obtained by multiplying the core tensor and the dictionary of the three modes. In [24], Li *et al.* proposed an HSI and MSI image fusion method based on coupled sparse tensor factorization. Dian *et al.* [25] factorized the HSI into smaller full-band blocks, and the HSI-MSI fusion problem can be turned to an optimization problem of a sparse core tensor and three dictionaries estimating for each full-band block. Wang *et al.* [26] proposed a pansharpening method based on sparse tensor neighbor embedding. In this method, each tensor constructed by MSI can be sparsely coded based on its neighbor tensor and the joint sparse coding assumption was constructed on bands. Finally, an HR multispectral tensor was obtained by weighting the sparse tensor coefficients on the PAN image. Xu *et al.* [27] proposed an HSI-MSI fusion method based on NL tensor factorization. This method first constructed an NL similar block tensor of HSI according to MSI. Then, HSI-MSI fusion was performed by coupling tensor canonical polymorphism decomposition, which achieved a good fusion effect. Dian and Li [28] proposed an HSI-MSI fusion method base on low tensor multi rank regularized, which also achieved a good fusion effect.

HSI-MSI fusion methods based on matrix factorization and HSI-MSI fusion methods based on tensor factorization are collectively referred to as image fusion methods based on factorization. The representation of image fusion methods based on factorization is consistent with the representation of the imaging model and is an intuitive description of imaging. Image fusion methods based on factorization typically assume that the HR-HSI share the same endmember to the corresponding LR-HSI scene. Therefore, the spectral properties of the fusion image are extracted from the LR-HSI. Although LR-HSI and HR-HSI in the same scene have the same spectral information in a physical sense, in terms of realistic imaging processes, LR-HSI is the spatial degradation result of HR-HSI in the same scene. During the degradation process, HR-HSI loses some details and produces a certain degree of distortion in the spectral domain. So, there are differences between LR-HSI and HR-HSI in the spectral domain, and it has some errors between the fusion image based on factorization and the ground truth (GT) image. In addition, image fusion methods based on factorization are difficult to obtain stable and accurate fusion results because they are usually sensitive to initial values. With the development of deep learning [29], HSI-MSI fusion methods based on deep learning come out on top, which have high reconstruction accuracy and fast computational speed. For example, Rao *et al.* [30] proposed an HSI-MSI fusion method based on a residual convolutional neural network constructed by the sparse residuals between multispectral and PAN images, which is helpful to solve the spectral distortion problem of traditional methods. Dian *et al.* [31] proposed a pansharpening method based on deep learning. The final HR-HSI was reconstructed by combining the trained network model and the image prior information. Xie *et al.* [32] proposed an HSI-MSI fusion method by combining model-based HSI-MSI fusion method and deep learning. Experimental results show that the superiority of the method.

The good performance of deep learning based methods often needs a lot of paired training data. However, hyperspectral datasets are often difficult to acquire due to imaging conditions and hardware limitations [33]. As a result, synthetic data are often used for network training, which reduces model flexibility and affects the performance of the network. Therefore, HSI-MSI image fusion methods based on deep learning without training are a hot research issue. For example, Zhang *et al.* [34] proposed an unsupervised HSI-MSI image fusion method that did not require paired datasets. In this method, LR images were generated in an unsupervised manner via a generative adversarial network. The generated images were then used for supervised training of the HR fusion images. Uezato *et al.* [35] proposed a guided deep decoder (GDD) network, which can be applied for image denoising and image fusion without training. Although HSI-MSI fusion image can be obtained based on the above methods without needed supervised learning, the structure of these networks does not make full use of the semantic features and detailed information of the images.

For fusing LR-HSI and HR-MSI, we proposed an unsupervised multiattentive guidance network to get better HSI super-resolution reconstruction. Specifically, a new network is constructed as a regularizer. The network is initialized by random noise and does not require any training, and the final fusion HR-HSI can be gained by only a pair of MSIs. The proposed network consists of two parts: multiattention encoding (MAE) network and multiscale feature-guided (MSFG) network. The semantic features of the MSI are extracted by the MAE network in the proposed network. Then, in order to guide the output of the MSFG network, multiscale features of the image can be extracted by attention mechanisms and NL similarity blocks in the MAE network. Our contributions are as follows.

1) An unsupervised end-to-end network, namely UMAG-Net, for HIS-MSI image fusion is proposed. UMAG-Net can generate corresponding HR-HSI images by using only one pair of images without any training data.
2) An MAE network is constructed. The NL block and spatial cross attention (SCA) block in the encoding network enable the full extraction of the semantic information and image details of the HSI.
3) A novel network structure is proposed as a regularizer for the unsupervised HSI-MSI fusion problem instead of using a handcrafted regularizer. The Laplacian guide (LG) block and upsampling guided (UG) block in the novel network generate fuse images by exploiting the multiscale semantic features of LR-MSI.

Experimental results show that the proposed UMAG-Net can achieve superior performance on unsupervised HSI-MSI fusion problems.

II. Related Work

In recent years, deep learning based HSI and MSI fusion methods have made great progress [36]. These kinds of methods are mainly based on learning the correspondence among LR-HSI, HR-MSI, and the corresponding HR-HSI through multilayer deep neural networks. The learned correspondences will be used as *a priori* knowledge to construct the missing spatial and spectral information in the new input sources images to complete the target HSI-MSI fusion work. It is generally accepted that such methods make use of not only the information carried by the input sources images itself but also the mapping relationships learned by relying on a library of training samples. Thus, better performance can be obtained than with HSI-MSI fusion methods based on manually defined prior information. However, Ulyanov *et al.* [37] proposed the deep image prior (DIP) algorithm, which argued that human-designed network structures are inherently capable of capturing a large amount of low-level statistical prior information about an image. Thus, targets like denoising and super-resolution can be achieved by iteratively learning the prior knowledge of an image. It is also believed that neural networks with random initialization can be able to extract better hand-designed prior distribution features. Gandelsman *et al.* [38] proposed the double-DIP algorithm to segment an image into its basic components by coupling multiple DIP networks, which makes it suitable for various tasks. Sidorov and Hardeberg [39] extended the DIP algorithm to the field of hyperspectral imaging.

Attention mechanisms enable networks to ignore irrelevant features and focus on important features. Therefore, attention mechanisms are widely used in various deep network constructions. Wang *et al.* [40] proposed an NL attention mechanism that can well capture the relational weights of any pixels in an image with respect to the current pixel. Good results were achieved in the fields of target detection, instance segmentation, and key point detection. Gu *et al.* [41] proposed a channel attention module based on a local cross-channel interaction strategy without dimensionality reduction, which effectively improves the computational efficiency of the network. Yao *et al.* [42] first introduced cross-attention to an HSI super-resolution task by multiplying two original features (LR-HSI and HR-MSI) with the attention map of another image, respectively, to transmit important information thus obtaining better image super-resolution results.

Inspired by the above works, an unsupervised multiattention-guided network (named UMAG-Net) is proposed in this article, which consists of an MAE and an MSFG network, with random noise and HR-MSI as inputs to iteratively learn a prior. Specifically, two attention mechanisms in the encoder–decoder are incorporated. NL blocks are used to better retain spectral and spatial details of the image. Spatial cross-attention blocks are utilized instead of traditional cascades to better achieve spatial-spectral information transfer by highlighting useful information in the image while suppressing irrelevant information in the encoder–decoder network. An LG block is used to connect the MAE network with the MSFG network to achieve better fusion results while ensuring feature alignment.

III. Proposed Method

A. Fusion Model

The HSI-MSI fusion problem is actually a problem of estimating HR-HSI by using HR-MSI and LR-HSI. Let $X \in \mathbb{R}^{W\times H\times L}$ denote HR-HSI. Let W, H, and L denote the width, height, and the count of bands, separately. HR-MSI is denoted by $Y \in \mathbb{R}^{W\times H\times l}$. $Z \in \mathbb{R}^{w\times h\times L}$ denotes LR-HSI, where l is the count of bands of $Y(l \ll L)$, and w and h are the width and height of $X(m \ll M, n \ll N)$.

In general, the HSI fusion problem based on deep neural networks can be formulated as

$$\min_X \mathcal{L}(X, Y, Z) + \mathcal{R}(X) \tag{1}$$

where $\mathcal{L}$ denotes the loss function and $\mathcal{R}$ is the handcrafted regularizer.

Instead of using a handcrafted regularizer, a new convolutional neural network was applied to estimate the generated HR-HSI, which is shown as

$$X = g_\theta(N) \tag{2}$$

where g_θ denotes a mapping function with network parameters of θ and N denotes the random noise of the input network. It is worth noting that the size of N may vary when conducting

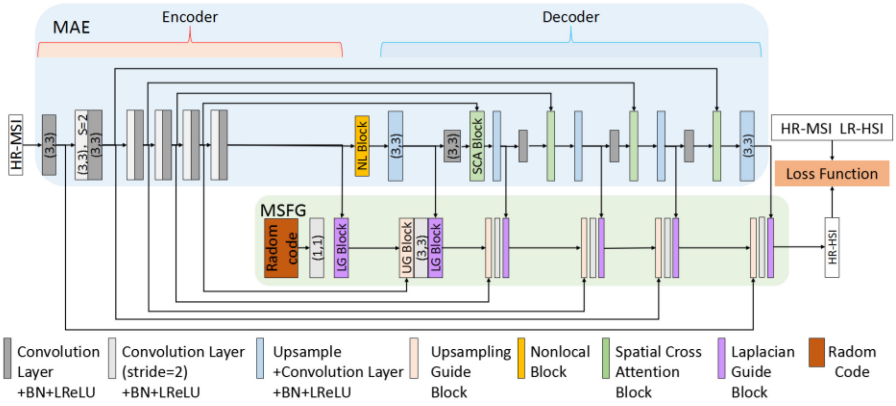

Fig. 1. Structure of the UMAG-net.

experiments on different datasets. Its size needs to be consistent with the size of Z.

Instead of $\mathcal{R}(\cdot)$, the prior distribution is obtained by the neural network. In addition, X is represented by a neural network mapping. Thus, manual errors are avoided and the flexibility of the network has been increased. Formula (1) can be rewritten as

$$\min_{\theta} \mathcal{L}\left(g_{\theta}\left(N\right), Y, Z\right). \tag{3}$$

From formula (3), we use the implicit prior regularization X of the neural network to generate HR-HSI with only LR-HSI and HR-MSI, and the whole network does not need to be trained. Specifically, the network is first randomly initialized. Then, the network parameters are iteratively updated by using a loss function that contains a pair of LR-HSI and HR-MSI. The parameter update process is similar to the traditional neural network training process. The number of iterations is set manually. When the number of iterations reaches the set value, the iteration stops and the fusion is completed. The structure of the proposed network is described in the following.

B. Network Architecture

In this article, an UMAG-Net is proposed. Fig. 1 shows the structure of UMAG-Net, which consists of an MAE network and an MSFG network. As shown in Fig. 1, MAE incorporates two attention mechanisms to the traditional encoder–decoder: NL attention and spatial cross-attention, which can extract multiscale image features of MSI well.

MSFG takes random noise as input for generating HR-HSI images and uses LG blocks in the guided network to connect the MAE network with the MSFG network to achieve better fusion while ensuring feature alignment. The construction of MAE and MSFG is described in the following, respectively.

1) Multiattention Encoder: The structure of the encoder in UMAG-Net is like that of U-Net, as shown in the blue background section in Fig. 1. For UMAG-Net, the number of layers of the encoder–decoder is determined by the downsampling factor. With a downsampling factor of 32, the number of layers is 5 ($\log_2 32$). U-Net [43], which is widely used in image processing, consists of a downsampling section and an upsampling section. The former is designed to gradually highlight background information, whereas the upsampling process combines information from the downsampling layers with input information from the upsampling to recover image details and recover the image step by step. MAE incorporates NL blocks between the downsampling network and the upsampling network and uses SCA to fuse features at the same scale. Thus, MAE can adequately exploit the semantic features of the MSI and preserve details in each band.

In Fig. 1, the orange frame section is the encoder. The first convolutional layer of the encoder is a 3×3 convolution with a step size of 1. It is used for shallow feature extraction. The remaining four convolution blocks are identical in structure, each consisting of two convolution layers. The first convolution layer has a convolution kernel of size $3 \times 3$with stride 2. The second convolution layer has a convolution kernel of size 3×3 and a step size of 1.

As can be seen from Fig. 1, the decoder of the MAE consists of convolutional layers and spatial cross-attention modules. All modules contain upsampling and batch normalization operations. Incorporating batch normalization between the convolution and the rectified linear unit (ReLU) prevents gradient disappearance and explosion while accelerating network convergence. In the proposed UMAG-net, instead of using the encoder and decoder features as the output directly to obtain HSI, the different scale features are input to the guided network for further processing to obtain a better HSI reconstruction. SCA block is used to fuse all features of the encoder and decoder, highlighting effective salient features that contribute to HSI-MSI fusion while suppressing irrelevant information in the input image. The structure of the SCA block is shown in Fig. 2.

The spatial attention (SA) block [41] is shown in the red frame in Fig. 2. Let X_l denote the input encoder low-level features and X_h denote the decoder high-level features. For SA, the feature map size is assumed to be $c \times h \times w$, where c is the number of input channels, and h and w are the width and height of the feature map, respectively. SA first reduces the dimensionality of X by using parallel point convolution φ, and we can obtain the compressed features X_l' and X_h' of the encoder and decoder. Then the compressed features are summed. After that, the feature with a channel number of 1 is then obtained by

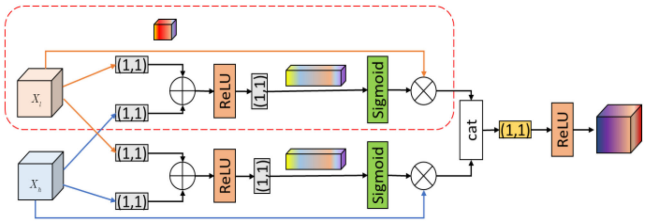

Fig. 2. Cross-space attention blocks.

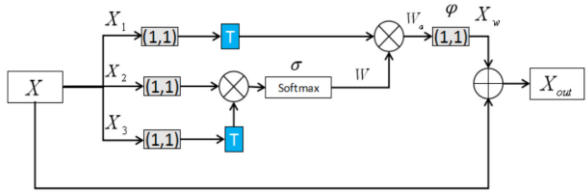

Fig. 3. NL blocks.

the ReLU function [denoted by $ReLU(\cdot)$] followed by another point convolution [denoted by $\varphi_1(\cdot)$]. Finally, the pixel attention weight W_p is obtained by the Sigmoid function, which is shown as

$$W_p = \sigma\left(\varphi_1\left(ReLU\left(X'_l + X'_h\right)\right)\right). \tag{4}$$

Recalibration of the encoder low-level features X_l by using weight W_p, i.e.,

$$X_{al} = W_p X_l. \tag{5}$$

Although SA blocks can effectively suppress irrelevant regions in the input image and do a better job of highlighting salient features that are useful for a particular task, their results may contain noise. Therefore, SCA blocks are used for feature enhancement of images. The SCA block consists of two SA blocks, as shown in Fig. 2. Let outputs of the two SA blocks be X_{al} and X_{ah}, respectively. Cascade the two (the cascade is denoted by ©) and obtain the feature with channel number 1 by point convolution φ. Finally, we can obtain the final feature by the ReLU function as follows:

$$X_{lh} = ReLU\left(\varphi\left(W_{p1}X_l \copyright W_{p2}X_h\right)\right). \tag{6}$$

As can be seen from Fig. 1, the MAE constructed in this article is different from other end-to-end encoder–decoder networks in which its encoder features will also enter the guided network through the upsampling block. Therefore, it is not appropriate to use all encoder features as input of the decoder.

An NL block is incorporated between the encoder and decoder to capture the interaction information between all pixels, which helps the network better extract the contextual semantic information. Fig. 3 shows the structure of the NL block.

For the input feature map, let X denote the feature map with a size of $C \times h \times w$, where C represents the count of channels, and h and w denote the height and width of X, separately. The feature enhancement block first uses point convolution φ on X to produce features X_1, X_2, and X_3 with a three-way channel count halved and the size is $c \times h \times w$ ($c = 0.5C$). The three-way features can be reshaped into a 2-D matrix of size $c \times hw$. The

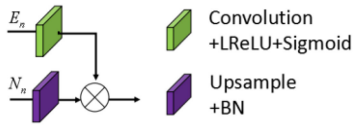

Fig. 4. Structure of the UG block.

first two features are multiplied by the Softmax function to obtain the weight of each pixel, which can be expressed as

$$W = \sigma(X_1^T X_2) \tag{7}$$

where $W \in \mathbb{R}^{hw \times hw}$, σ is the Softmax function, and T denotes the transpose. The third way feature X_3 is multiplied by the weighting factor W to obtain the weighted feature $W_a (W_a \in \mathbb{R}^{0.5c \times h \times w})$. By applying the point convolution function and batch normalization to the weighted feature W_a, we can recover the number of channels to get a feature F_w of the same size as the input features. Finally, the output features X_{out} are obtained through jump connections that can facilitate information dissemination and recovery, which is shown as follows:

$$X_{\text{out}} = \varphi(W_a) + X. \tag{8}$$

The NL feature enhancement block enhances the feature map representation by encoding a wider range of semantic information into the local receptive field. It is worth noting that this block has no restriction on the input size of the feature map and is less computationally intensive.

2) Multiscale Guided Network: In image fusion, it will inevitably lose some of the effective information only by using the coder and decoder result as the output. And a single coder-decoder structure cannot retain both shallow and deep features of the image. Therefore, an MSFG network combined with the output of MAE is constructed for image fusion.

MSFG contains three structures: convolutional layers, UG blocks, and LG blocks. In MSFG, the MAE encoder extracts multiscale features of MSI as the input to the UG block, whereas the MAE decoder features are used as the input to the LG block. So that the MSFG can fully extract the multiscale features of the MSI image to achieve the final image fusion to accurately recover the HSI image detail information and finish the HSI-MSI fusion.

Bilinear interpolation upsampling may result in an image too smooth to effectively recover the boundaries and fine structures of the image. Therefore, we use the same structure in [35] for the upsampling operation with the upsampling module, whose structure is shown in Fig. 4.

As shown in Fig. 4, the UG block consists of SA gates and channel normalizations. The SA gate consists of point convolutions, LReLU, and sigmoid function, which is effective in ensuring the spatial localization of MSIs when processing MSI features. Notably, the two input features of the upsampling module have the same scale.

The Laplacian attention block consists of an adaptive mean pooling layer, convolutional layers, activation functions, and a cascade function, the structure of which is shown in Fig. 5.

Let the $Average(\cdot)$ function denote the adaptive average pooling layer in the LG block, $\phi_k(\cdot)(k = 3, 5, 7)$ denote the

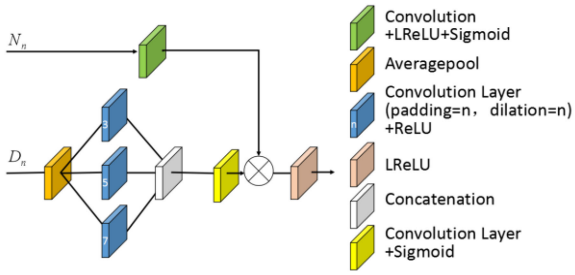

Fig. 5. LG block structure.

convolution kernel size of 3×3, the convolution operation with padding and dilation of k, respectively. And let $\sigma(\cdot)$ denote the activation function. The features from the encoder are pooled and then convolved by three convolutions, and we can obtain multiscale features f_3, f_5, and f_7, which is shown as

$$\begin{cases} f_3 = \phi_3(Average(E_n)) \\ f_5 = \phi_5(Average(E_n)) \\ f_7 = \phi_7(Average(E_n)) \end{cases}. \tag{9}$$

Subsequently, the cascaded multiscale feature ($f_3 © f_5 © f_7$) is passed through a convolution layer and an activation function to obtain the following features:

$$D_{\phi_n} = \sigma(\varphi(f_3 © f_5 © f_7)). \tag{10}$$

Feature N_n in the guided network is obtained as feature weights W_f through an attention gate, which consists of point convolution, LReLU, and sigmoid functions. W_f can be expressed as

$$W_f = \sigma(LReLU(\varphi(N_n))). \tag{11}$$

Laplacian attention ensures that the input features are aligned with the features of the MSI while extracting image features at multiple scales. The whole weighing process can be expressed as follows:

$$N_{outn} = LReLU(W_f \otimes D_{\phi_n}). \tag{12}$$

C. Loss Function

Let $X \in \mathbb{R}^{W \times H \times L}$, $Y \in \mathbb{R}^{W \times H \times l}$, and $Z \in \mathbb{R}^{W \times H \times l}$ denote HR-HSI, HR-MSI, and LR-HSI, respectively. If $\tilde{X}$ denotes the estimated HR-HIS, the loss function can be defined as

$$\mathcal{L}(\tilde{X}, Y, Z) = \left\| R\tilde{X} - Z \right\|_F^2 + \lambda \left\| \tilde{X}D - Y \right\|_F^2 \tag{13}$$

where R is the SRF and D is the spatial downsampling function. The first term measures the spectral similarity between X and Z, and the second term measures the spatial similarity between X and Y. λ is the balance term, which is used to adjust the balance between the two, which is set to 0.1.

IV. Experimental Results

A. Datasets

To test the effectiveness of UMAG-Net, two public datasets, the Columbia Computer Vision Laboratory dataset (CAVE) [44] and the Harvard database [45], are chosen for the experiments.

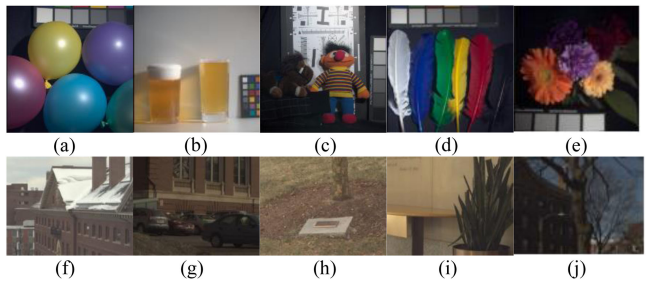

Fig. 6. Examples of RGB images from the CAVE dataset (first row) and the Harvard dataset (second row). (a) Balloons. (b) Beer. (c) Toys. (d) Feathers. (e) Flowers. (f) Img 1. (g) Img b2. (h) Img b6. (i) Img d9. (j) Img f2.

The first dataset contains 32 high-quality indoor HSI of size 512×512captured by a universal classification pixel camera. Each HSI in the CAVE dataset has 31 bands starting at 400 nm and covering a wavelength range of 300 nm with 10 nm intervals. The Harvard dataset contains 50 images of size 1024×1024. Each HSI has 31 bands covering the wavelength range from 420 to 720 nm at 10 nm intervals. Fig. 6 shows the RGB images from CAVE and Harvard datasets.

B. Experimental Setup

1) Compare Methods: To verify the reliability and validity of UMAG-Net, we compared our method with seven MSI-HSI fusion algorithms. The first two methods among the seven methods are deep learning based methods, whereas the last five are traditional MSI-HSI fusion methods. Deep learning based MSI-HSI fusion methods include: 1) MSI-HSI fusion method based on MS/HS fusion networks (supervised deep learning model, MHF) proposed in [32]; 2) MSI-HSI fusion method based on GDD networks (unsupervised deep learning model, GDD) proposed in [35]. Traditional MSI-HSI fusion methods include the following.

1) MSI-HSI fusion method based on coupled nonnegative matrix factorization (CNMF) proposed in [18].
2) MSI-HSI fusion method based on nonlocal sparse tensor factor decomposition (NLSTF) proposed in [23].
3) Semiblind MSI-HSI fusion method based on nonlocal sparse tensor factor factorization (NSTF) proposed in [25].
4) MSI-HSI fusion method based on local low-rank coupled spectral factorization (LRCS) proposed in [46].
5) MSI-HSI fusion method based on low-rank tensor training rank representation (LTTR) proposed in [47].

Only MHF needs to be trained in the traditional method. We use the code published by the author in the corresponding paper for the test data. Our method is implemented under Pytorch 1.5.1 framework running in the Windows 10 environment with Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz and NVIDIA GeForce RTX 2080 Ti GPU.

2) Evaluation Metrics: Together with evaluating the performance of each fusion method by subjective visualization, seven objective evaluation metrics are presented for better evaluation The objective evaluation indicators adopted in this article are: correlation coefficient (CC), the mean of absolute error (MoAE), relative dimensionless global error in synthesis

(ERGAS) [48], peak signal-to-noise ratio (PSNR), root-mean-square error (RMSE), spectral angle mapper (SAM) [49], structural similarity (SSIM) [50], and universal image quality index (UIQI) [51], which are described in the following.

The similarity of the content of an image is determined by the score, CC is mainly used to score the similarity of the content between two images, which is defined as

$$\mathrm{CC} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(\tilde{X}(i,j)-\bar{\tilde{X}})(X(i,j)-\bar{X})}{\sqrt{\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(\tilde{X}(i,j)-\bar{\tilde{X}})^2\right)\left(\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\bar{X})^2\right)}} \tag{14}$$

where X denotes the GT, $\tilde{X}$ denotes the estimated HSI, and $M \times N$ denotes the image size. CC in HSI fusion is calculated as averaged over all bands. The larger the CC is, the nicer the fusion image can be.

MoAE is the mean of the absolute error. And it is often used to indicate the magnitude of the difference between two images. The definition of MoAE is as follows:

$$\mathrm{MoAE} = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|X(i,j)-\tilde{X}(i,j)\right| \tag{15}$$

where X denotes GT, $\tilde{X}$ denotes the estimated HSI, and m denotes the pixels number in data X. A smaller MoAE indicates that the error between the fused image and the GT is smaller.

RMSE is often used to indicate the similarity between two images. The smaller its value is, the better the image is. Let the size of X and $\tilde{X}$ be $M \times N$. The mean square error is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|X(i,j)-\tilde{X}(i,j)\right|^2} \tag{16}$$

where $X(i,j)$ indicates the pixel value of image X at position (i,j) and $\tilde{X}(i,j)$ indicates the pixel value of image $\tilde{X}$ at position (i,j).

PSNR is a full reference image quality evaluation metric, which is often a representation of the degree of difference between two images. It is often defined through the mean squared error and can be expressed as

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\mathrm{MAX}}{\mathrm{RMSE}}\right)^2 = 20\log_{10}\left(\frac{\mathrm{MAX}}{\mathrm{RMSE}}\right) \tag{17}$$

where MAX denotes the maximum value of the image color. PSNR of HSI is defined as the average of all bands. A higher PSNR value indicates that the difference between the fused image and the original image is less, and more detail is preserved.

ERGAS is expressed as a synthetic error for all bands, i.e.,

$$\mathrm{ERGAS}(X,\tilde{X}) = \frac{100}{d}\sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{\mathrm{RMSE}(X_i,\tilde{X}_i)}{\mu(\tilde{X}_i i)}\right)^2} \tag{18}$$

where d is the spatial subsampling factor, μ denotes the mean value of the image, and S denotes the band number. A smaller ERGAS indicates a higher spectral agreement between the two images.

As an indispensable index for evaluating spectral distortion, SAM is defined as

$$\mathrm{SAM}(X,\tilde{X}) = \frac{1}{M}\sum_{j=1}^{M}\arccos\frac{\langle X,\tilde{X}\rangle}{\|X^j\|_2\left\|\tilde{X}^j\right\|_2} \tag{19}$$

where M is the number of spectral pixels, and $\langle\rangle$ is the inner product of the two vectors. Smaller SAM indicates less spectral distortion.

SSIM measures the similarity of two images. The value of SSIM would be 1 when the two images are identical. As an implementation of structural similarity theory, SSIM defines information about the structure in terms of image composition as a property that reflects the structure of objects in a scene, irrelevant of luminance and contrast. Distortion is expressed as a combination of structure, luminance and contrast. The mean is used as an estimate of brightness, the standard deviation as an estimate of contrast, and the SSIM is measured by covariance

$$\mathrm{SSIM} = \frac{(2\mu_X\mu_{\tilde{X}}+c_1)(\sigma_{X\tilde{X}}+c_2)}{\left(\mu_X^2+\mu_{\tilde{X}}^2+c_1\right)\left(\sigma_X^2+\sigma_{\tilde{X}}^2+c_2\right)} \tag{20}$$

where μ_X and $\mu_{\tilde{X}}$ represent the mean of X and $\tilde{X}$, respectively, σ_X and $\sigma_{\tilde{X}}$ represent the variance of them, respectively, and $\sigma_{X\tilde{X}}$ denotes the covariance of them. To maintain stability, constants $c_1 = (pD)^2$ and $c_2 = (qD)^2$ are employed, where $p = 0.01$ and $q = 0.03$. The dynamic range of the pixel values, denoted by D , is usually set to 255. The smaller the gap between SSIM and 1, the better the fusion result will be.

UIQI evaluates the effect of the fusion image by measuring the correlation loss, brightness distortion, and contrast distortion between the fusion image and the source image. The UIQI of two images X and $\tilde{X}$ is defined as

$$\mathrm{UIQI}(X,\tilde{X}) = \frac{\sigma_{X,\tilde{X}}}{\sigma_X\sigma_{\tilde{X}}}\frac{2\bar{\mathrm{X}}\bar{\tilde{\mathrm{X}}}}{\bar{\mathrm{X}}^2+\bar{\tilde{\mathrm{X}}}^2}\frac{2\sigma_X\sigma_{\tilde{X}}}{\sigma_X^2+\sigma_{\tilde{X}}^2} \tag{21}$$

where σ and μ represent the variance and mean, separately. The CC of X and $\tilde{X}$, given in the first term, is a measure of the linear correlation between the two images and has an optimum value of 1. Linear correlation does not suggest that there is no relative distortion between images. Thus, the second and third terms are used to assess the relative distortion. The second term is used to measure the proximity of the average brightness between the two images, which obtains the best value of 1 when the average brightness is equal. σ_X and $\sigma_{\tilde{X}}$ can be regarded as the contrast estimation of X and $\tilde{X}$. The third term can measure the degree of contrast similarity between images. When and only when σ_X is equal to $\sigma_{\tilde{X}}$, the third term obtains the optimum value of 1. A larger UIQI value indicates a better effect of image fusion.

C. Experimental Results

1) Experiments Based on CAVE Dataset: When these methods are tested on the CAVE dataset, the HR-HSI is given 32-fold downsampling to obtain the LR-HSI. The HR-HSI is then

TABLE I
OBJECTIVE EVALUATION METRICS FOR EACH FUSED IMAGE ON THE CAVE DATASET

Method	GDD	MHF	CNMF	NLSTF	NSTF	LRCS	LTTR	Ours
CC ↑	0.9958	0.9960	0.9951	0.9935	0.9924	0.9921	0.9932	**0.9962**
MoAE ↓	0.9575	0.9017	2.8232	1.0482	1.4105	1.2709	1.3416	**0.8251**
ERGAS ↓	0.3209	**0.2950**	0.3848	0.3844	0.4544	0.4598	0.4261	0.3058
PSNR ↑	43.5939	**44.6670**	39.3969	43.2614	41.2821	40.6672	41.7723	44.5414
RMSE ↓	2.1669	1.9963	3.3624	2.6038	3.1356	2.9465	2.9910	**1.9923**
SAM ↓	5.9215	5.7585	6.2703	6.5876	11.5292	7.3159	7.1752	**5.2420**
SSIM ↑	0.9852	0.9880	0.9754	0.9824	0.9742	0.9814	0.9773	**0.9881**
UIQI ↑	0.8834	**0.8960**	0.8784	0.8945	0.8686	0.8738	0.8797	0.8931

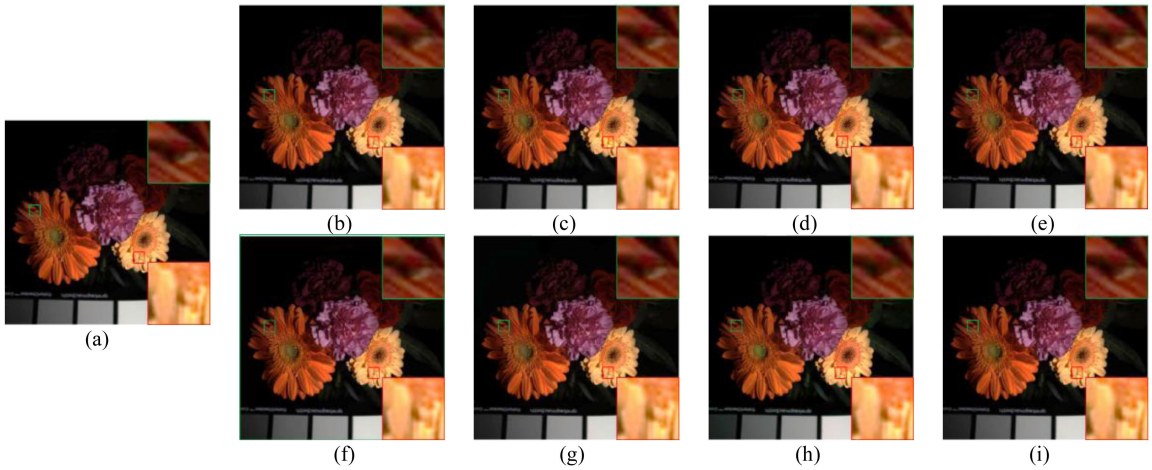

Fig. 7. RGB image of each fusion algorithm for the flower image. (a) GT. (b) GDD. (c) LTTR. (d) MHF. (e) NLSTF. (f) NSTF. (g) LRCS. (h) CNMF. (i) Our.

spectrally downsampled with the Nikon D700 SRF to generate an HR-MSI. Although the proposed method does not require training data, there are deep learning methods that need to be trained in the comparison algorithm. The CAVE dataset is divided into a training set and a test set. 12 images are selected for testing and the remaining 12 images are utilized for model training of the comparison methods.

12 randomly selected images on the CAVE dataset are tested and the objective evaluation metrics obtained from 12 images are averaged to obtain the final objective evaluation metrics of each fusion method. Table I presents the test results. In Table I, ↓ indicates that a smaller value for an objective indicator is better and ↑ indicates that a larger value for an objective indicator is better. The best results of the objective indicator are bolded, whereas the second best results are underlined.

In Table I, obviously, the objective evaluation metrics of UMAG-Net and MHF are far superior to other methods. However, MHF needs 20 images for model training in advance, whereas the proposed method can obtain better image fusion results without training data. Notably, compared to MHF, the SAM value of the proposed algorithm is reduced by 8.7%, which shows that the spectral distortion of the HSI fusion image obtained by UMAG-Net is much smaller than MHF. Although the overall performance of GDD is not as good as MHF and the proposed method, it still yields better results than NLSTF and LTTR in terms of detail preservation and spectrum preservation. NSTF has the most serious spectral distortion because of its semiblind fusion. CNMF and LRCS have a relatively large spectral distortion due to the consideration of coupling and unmixing mechanisms.

To intuitively observe the performance of each method, a set of test images is randomly selected to show. Fig. 7 shows RGB images generated from the fused image obtained by each image fusion method. The petal areas with rich features (green box and red box) are selected to enlarge. Fig. 7 shows that GDD, MHF, and UMAG-Net can better restore image details, and the color fidelity is better. The color of CNMF is distorted in some areas and the image details cannot be recovered well. LRCS, LTTR, NLSTF, and NSTF have problems such as blurring and loss of fine structures at the petal boundaries.

To provide a visual representation of the effect of each image fusion method, the error maps of the fusion results and the GT at each band are plotted. Fig. 8 shows the error map for the 10th band of the flower image in the CAVE dataset, with Reference indicating the HSI in the 10th band, whereas Fig. 9 shows the error map for the 28th band of the flower image in the CAVE dataset, with Reference indicating the HSI in the 28th band.

From the results in Fig. 8, it is noticeable that the errors produced by UMAG-Net and NLSTF are smaller, but UMAG-Net has a better fusion effect at the left rear flower. GDD and LTTR have obvious errors at the lower left square. MHF with excellent quantitative performance has a serious error at the center of the

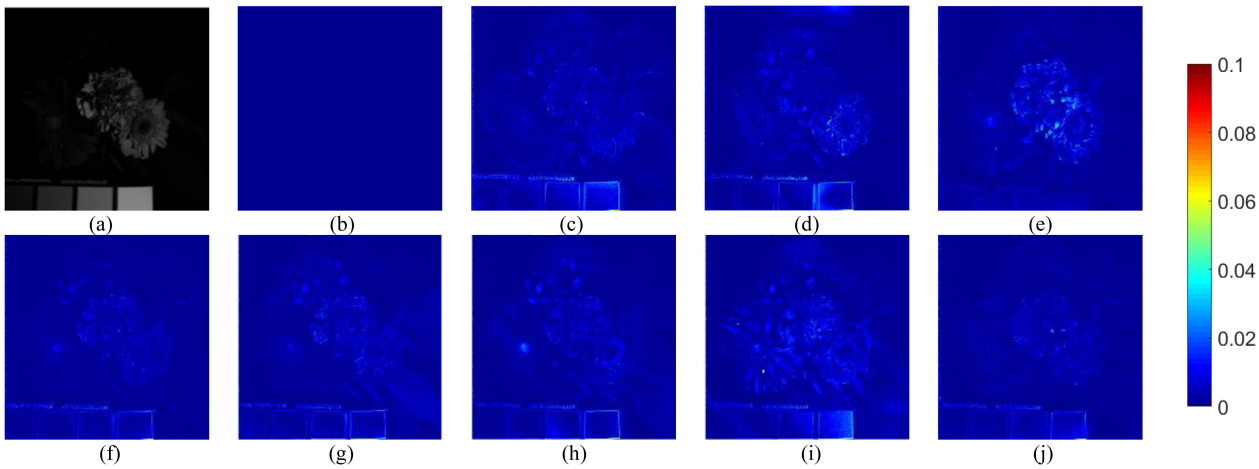

Fig. 8. Error map for the 10th band of each fusion algorithm for the flower image. (a) Reference. (b) GT. (c) GDD. (d) LTTR. (e) MHF. (f) NLSTF. (g) NSTF. (h) LRCS. (i) CNMF. (j) Our.

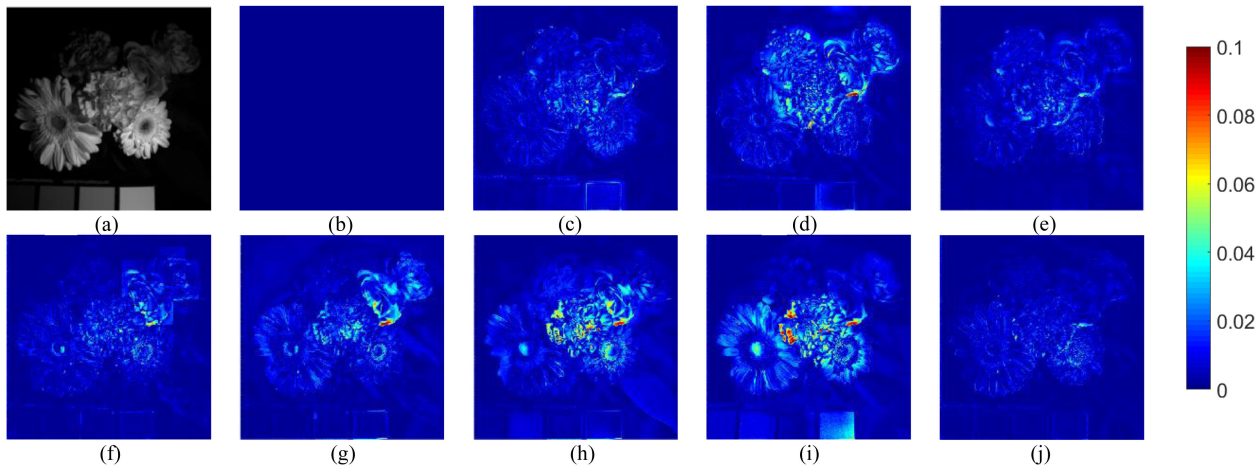

Fig. 9. Error map for the 28th band of each fusion algorithm for the flower image. (a) Reference. (b) GT. (c) GDD. (d) LTTR. (e) MHF. (f) NLSTF. (g) NSTF. (h) LRCS. (i) CNMF. (j) Our.

image. LRCS has more errors at the pistil. NSTF has more errors at the rear leaf. And CNMF has obvious error in a large area and performs the worst.

In Fig. 9, UMAG-Net still maintains a good fusion effect in the 28th band. In contrast, the fused images of the remaining two deep learning methods (GDD and MHF) have severe loss of detail at the edges of the flowers and the fused image of MHF even has bad artificial texture in the background. The fusion results of the traditional methods generally performed poorly in the bands at the back. NLSTF, which performs better in band 10, shows significant block errors in band 28. And the fused images of LTTR, NSTF, LRCS, and CNMF show significant errors at the petals and leaves.

2) Experiments Based on Harvard Dataset: The HSI in the Harvard dataset is cropped into blocks of size 1024×1024. Downsampling by a factor of 32 used HR-HSI to obtain LR-HSI. And HR-MSI is obtained by down-sampling HR-HSI spectrum using the Nikon D700 SRF. Since MHF requires training data, the Harvard dataset is divided into a training set and a test set. In total, 20 images are selected as test images, whereas the remaining 30 HSIs are used for training.

In all, 20 randomly selected images on the Harvard dataset are tested and then the objective evaluation metrics obtained by each method on the 20 images are averaged to obtain the final objective evaluation metrics for each fusion algorithm, as presented in Table II. In Table II, $\downarrow$ indicates that a smaller value for an objective metric is better and $\uparrow$ indicates that a larger value for an objective metric is better. The best results for objective metrics are bolded, whereas the second best results are underlined.

In Table II, all the metrics of the UMAG-Net are optimal except for the UIQI metric on the Harvard dataset. Even on the UIQI metric, a very small gap of 0.0061 appears between UMAG-Net and the optimal value, which fully illustrates the advantage of the proposed method in HSI-MSI fusion.

Compared to the CAVE dataset, fusion of HSIs on the Harvard dataset is relatively less difficult. In order to visualize the effect of fusion of each method, a random set of images from the Harvard dataset is selected to show the fused images of each algorithm. Fig. 10 shows the RGB images obtained by spectral downsampling of the fused images from Imgf2. Two regions,

TABLE II
OBJECTIVE EVALUATION METRICS FOR EACH FUSED IMAGE ON THE HARVARD DATASET

Method	GDD	MHF	CNMF	NLSTF	NSTF	LRCS	LTTR	Ours
CC↑	0.9945	0.9943	0.9916	0.9941	0.9925	0.9909	0.9884	**0.9948**
MoAE↓	0.9269	0.9260	1.3074	0.9299	1.0785	0.9886	1.4795	**0.9146**
ERGAS↓	0.2357	0.2510	0.3078	0.2389	0.2537	0.3005	0.3262	**0.2350**
PSNR↑	46.1360	46.1385	43.3310	46.3596	42.9593	44.6014	41.2363	**46.3652**
RMSE↓	1.5974	1.6252	2.2775	1.7022	1.8980	1.9245	3.3414	**1.5870**
SAM↓	3.2442	3.3372	3.5565	3.3319	3.7188	3.3930	5.1474	**3.2267**
SSIM↑	0.9834	0.9840	0.9797	0.9836	0.9326	0.9827	0.9202	**0.9846**
UIQI↑	0.8524	0.8616	0.8441	**0.8639**	0.8085	0.8570	0.7868	0.8578

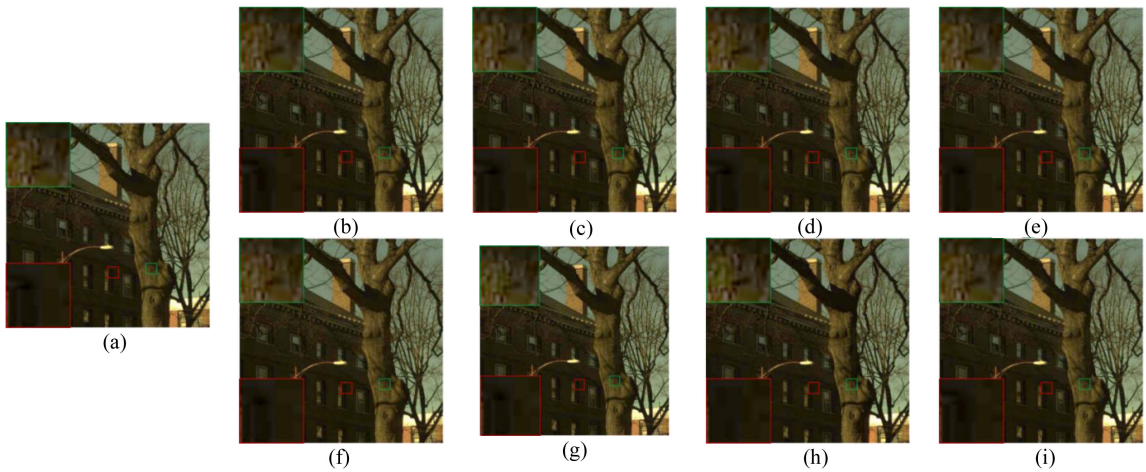

Fig. 10. Imgf2 image of each fusion algorithm RGB image. (a) GT. (b) GDD. (c) LTTR. (d) MHF. (e) NLSTF. (f) NSTF. (g) LRCS. (h) CNMF. (i) Our.

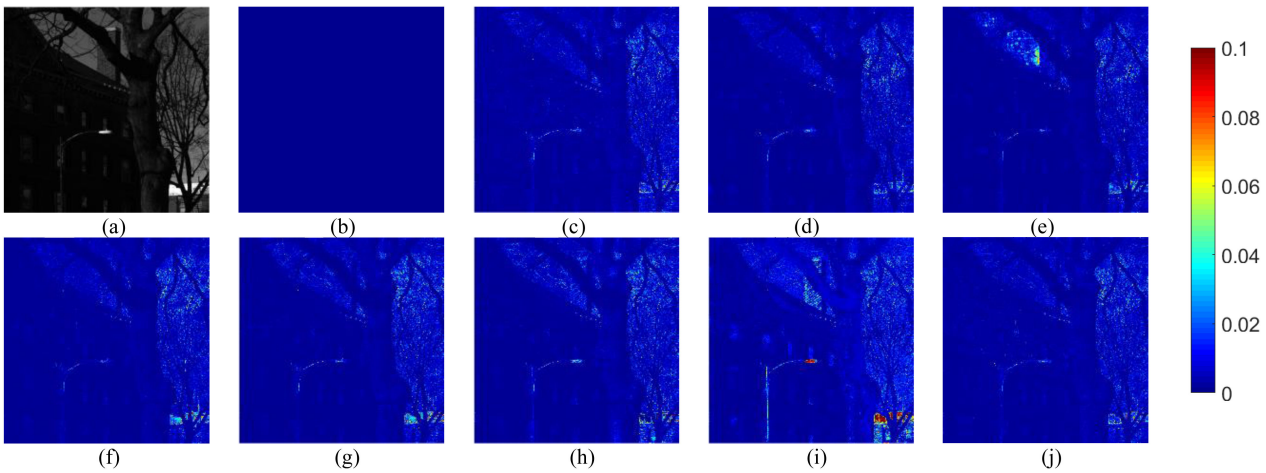

Fig. 11. Error map for band 12 of the Harvard dataset Imgf2. (a) Reference. (b) GT. (c) GDD. (d) LTTR. (e) MHF. (f) NLSTF. (g) NSTF. (h) LRCS. (i) CNMF. (j) Our.

the window edge (green box) and the tree trunk (red box), are selected for magnification.

For an intuitive understanding of the effectiveness of each image fusion method, the error maps of the fused image and the GT in each band are plotted. Fig. 11 shows the error map for the 12th band of the Imgf2 image in the Harvard dataset, with Reference indicating the HSI in the 12th band, whereas Fig. 12 shows the error map for the 26th band of the Imgf2 image in the Harvard dataset, with Reference indicating the HSI in the 26th band.

Fig. 11 shows that the visual effect of the fusion results from the proposed method, GDD, and the traditional method LTTR (which relies on complex *a priori* knowledge) is better. Compared to GDD, the reconstruction error of our method is smaller at the light pole and window of the house on the left. The reconstruction error of our method at the tree trunk on the left

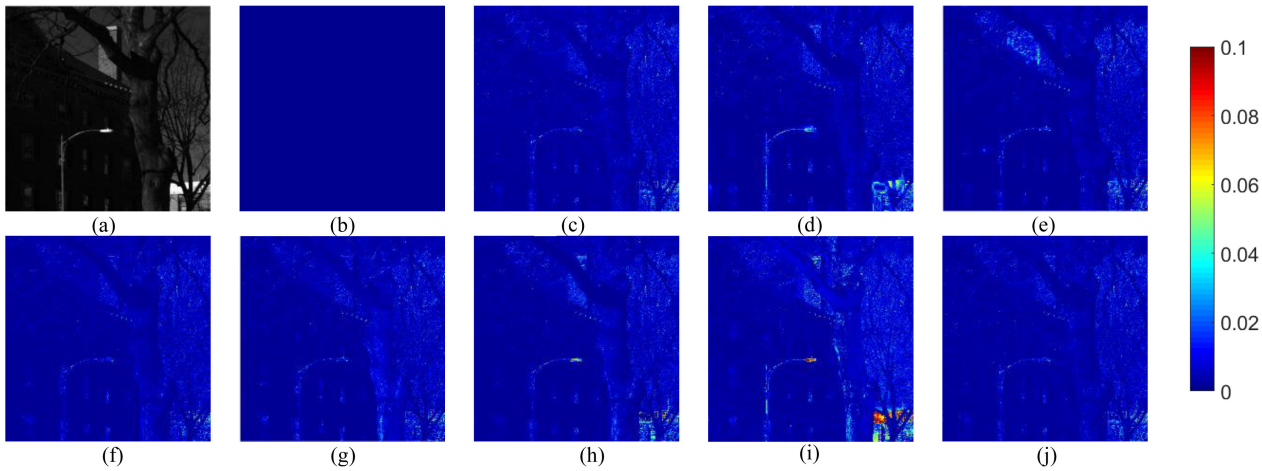

Fig. 12. Error map for band 26 of the Harvard dataset Imgf2. (a) Reference. (b) GT. (c) GDD. (d) LTTR. (e) MHF. (f) NLSTF. (g) NSTF. (h) LRCS. (i) CNMF. (j) Our.

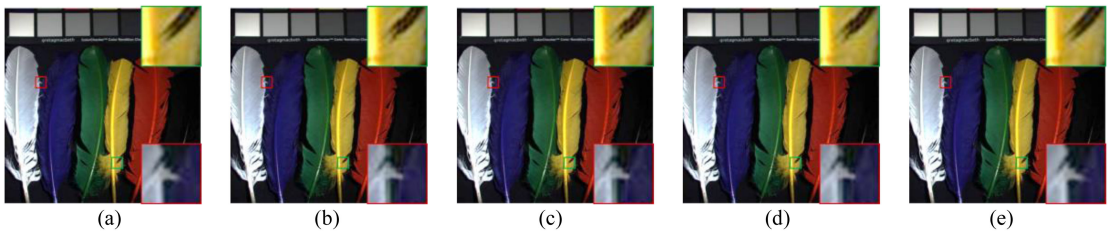

Fig. 13. RGB images generated from the fusion results of each model. (a) GT. (b) w/o–SCA. (c) w/o–LG. (d). w/o–NL. (e) UMAG.

side is smaller compared to LTTR. Other methods have obvious local errors, such as the left border of the chimney in the fusion image of MHF, the chimney and street light in the fusion image of CNMF and the right rear house in NLSTF, NSTF, and LRCS fusion results.

Fig. 12 shows the error map for band 26 of the Harvard dataset. In band 26, deep learning methods that do not require pre-trained models (GDD and the proposed method) still maintain good reconstructions, whereas fused images from deep learning methods that require training (MHF) and traditional algorithms (CNMF, NLSTF, NSTF, LRCS, and LTTR) have obvious local reconstruction errors.

TABLE III
COMPARISON OF THE RESULTS OF THE REMOVAL OF A MODULE WITH THE UMAG RESULTS

Module	w/o-SCA	w/o-LG	w/o-NL	UMAG-Net
CC↑	0.9955	0.9960	0.9953	**0.9962**
MoAE↓	0.9215	0.8759	0.9291	**0.8244**
ERGAS↓	0.3317	0.3142	0.3313	**0.3064**
PSNR↑	43.5642	44.0735	43.6329	**44.5576**
RMSE↓	2.1901	2.0651	2.2143	**1.9912**
SAM↓	5.8499	5.5313	5.6590	**5.2446**
SSIM↑	0.9858	0.9872	0.9859	**0.9881**
UIQI↑	0.8811	0.8906	0.8862	**0.8929**

D. Ablation Experiments

To better illustrate the validity of the proposed model, the role of each part of the proposed network structure is analyzed. Table III gives the effects of the three components (the spatial cross-attention block, the LG block, and the NL block) on the fusion results of CAVE dataset. w/o-SCA denotes the network model of UMAG-Net without the spatial cross-attention block, w/o-LG denotes the network model of UMAG-Net without the LG block, and w/o-NL denotes the network model of UMAG-Net without the NL block.

In Table III, UMAG shows the best performance. And by reducing anyone block, the network performance will be degraded to some extent. After removing the SCA block, the indicators become significantly worse. Among them, the SAM value is reduced by 0.6053 and the PSNR value is reduced by 0.9934, which indicates that the SCA block can better recover the spatial details and maintain the spectral properties compared with the traditional cascade method. After removing the NL or LG blocks, there are different degrees of deterioration in the indicators. For example, the PSNR values of w/o-NL and w/o-LG are reduced by 0.9247 and 0.4841, respectively, which indicates that the NL and LG blocks can better extract MSI features and thus are more conducive to the retention of image details.

For an intuitive view of the fusion effect of each algorithm, a set of randomly selected images from the CAVE dataset is shown to display the RGB images generated by the fused images of each method, as shown in Fig. 13. The top right corner of the white feather (red box) and the bottom right corner of the

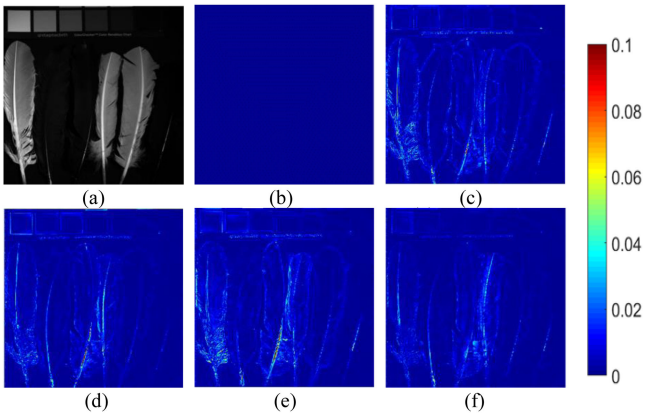

Fig. 14. Ablation experiment error map. (a) Reference. (b) GT. (c) w/o–SCA. (d) w/o–LG. (e) w/o–NL. (f) UMAG.

yellow feather (green box) are zoomed in. In Fig. 13, the fused images from w/o-SCA, w/o-LG, and w/o-NL are presented. The white feather in the red box has blurred edges and artifacts in the background, the yellow feather in the green box has incorrect texture recovery at the notch, whereas the fused image obtained by UMAG-Net has clearer and more accurate feather boundaries.

Similarly, an error map of the fused image and the GT image at band 28 is presented in Fig. 14.

In Fig. 14, Reference denotes the 28th band HSI image. It can be clearly seen that w/o-SCA is poorly fused at the feather border, losing a lot of detailed texture, whereas w/o-LG and w/o-NL are poorly fused in the lower part of the green feather, losing a lot of detailed texture; UMAG-Net maintains the feather edge detail and the edges between the squares at the top of the picture.

V. Conclusion

In this article, an UMAG-Net without training data is proposed for the HSI-MSI fusion. In the proposed UMAG-Net, the MAE network is used for deep extraction of multiscale image features from MSIs and the UG block is used to generate HR-HSIs. Features at different scales in encoder and decoder are injected into the upsampling attention network via LG blocks, and the network takes random noise as input. Spatial detail and spectral features of HSI and MSI are fully leveraged by UMAG-Net. Compared with other HSI-MSI fusion methods, the proposed method achieves optimum image fusion results.

References

[1] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.

[2] Q. Guo, B. Zhang, Q. Ran, L. Gao, J. Li, and A. Plaza, "Weighted-RXD and linear filter-based RXD: Improving background statistics estimation for anomaly detection in hyperspectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2351–2366, Jun. 2014.

[3] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. X. Zhu, "Learning to propagate labels on graphs: An iterative multitask regression framework for semi-supervised hyperspectral dimensionality reduction," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 35–49, 2019.

[4] C. Li, L. Gao, Y. Wu, B. Zhang, J. Plaza, and A. Plaza, "A real-time unsupervised background extraction-based target detection method for hyperspectral imagery," *J. Real Time Image Process*, vol. 15, no. 3, pp. 597–615, 2018.

[5] D. Hong, X. Wu, P. Ghamisi, J. Chanussot, N. Yokoya, and X. X. Zhu, "Invariant attribute profiles: A spatial-frequency joint feature extractor for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 3791–3808, Jun. 2020.

[6] X. Cao, J. Yao, Z. Xu, and D. Meng, "Hyperspectral image classification with convolutional neural network and active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4604–4616, Jul. 2020.

[7] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognit.*, 2010, pp. 44–51.

[8] D. Hong, N. Yokoya, J. Chanussot, and X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[9] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "CoSpace: Common subspace learning from hyperspectral-multispectral correspondences,"*IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4349–4359, Jul. 2019.

[10] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 147, pp. 193–205, 2019.

[11] D. Hong, N. Yokoya, J. Chanussot, J. Xu, and X. Zhu, "Joint and progressive subspace analysis (JPSA) with spatial-spectral manifold alignment for semisupervised hyperspectral dimensionality reduction,"*IEEE Trans. Cybern.*, vol. 51, no. 7, pp. 3602–3615, Jul. 2021.

[12] Y. Qu, H. Qi, B. Ayhan, C. Kwan, and R. Kidd, "Does multispectral/hyperspectral pansharpening improve the performance of anomaly detection?" in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 6130–6133.

[13] V. Ferraris, N. Dobigeon, Q. Wei, and M. Chabert, "Robust fusion of multiband images with diffffferent spatial and spectral resolutions for change detection," *IEEE Trans. Comput. Imag.*, vol. 3, no. 2, pp. 175–186, Jun. 2017.

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

[15] Y. Zhang and M. He, "Multi-spectral and hyperspectral image fusion using 3-D wavelet transform," *J. Electron.*, vol. 24, no. 2, pp. 218–224, 2007.

[16] Z. Chen, H. Pu, B. Wang, and G.-M. Jiang, "Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pansharpening methods," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1418–1422, Aug. 2014.

[17] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hypersharpening: A first approach on Sim-GA data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 3008–3024, Jun. 2015.

[18] N. Yokoya, S. Member, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.

[19] W. Dong, F. Fu, G. Shi, X. Cao, G. Li, and X. Li, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, May 2016.

[20] W. He, H. Zhang, and L. Zhang, "Sparsity-regularized robust non-negative matrix factorization for hyperspectral unmixing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 9, pp. 4267–4279, Sep. 2016.

[21] X. Han, B. Shi, and Y. Zheng, "Self-similarity constrained sparse representation for hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5625–5637, Nov. 2018.

[22] Q. Wei, N. Dobigeon, and J. Tourneret, "Fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4109–4121, Nov. 2015.

[23] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3862–3871.

[24] S. Li, R. Dian, L. Fang, and J. M. Bioucas-Dias, "Fusing hyperspectral and multispectral images via coupled sparse tensor factorization," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4118–4130, Aug. 2018.

[25] R. Dian, S. Li, L. Fang, T. Lu, and J. M. Bioucas-Dias, "Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion," *IEEE Trans. Syst., Man, Cybern.*, vol. 50, no. 10, pp. 4469–4480, Oct. 2020.

[26] M. Wang, K. Zhang, X. Pan, and S. Yang, "Sparse tensor neighbor embedding based pan-sharpening via N-way block pursuit," *Knowl.-Based Syst.*, vol. 149, pp. 18–33, 2018.
[27] Y. Xu, Z. Wu, J. Chanussot, P. Comon, and Z. Wei, "Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 348–362, Jan. 2020.
[28] R. Dian and S. Li, "Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 5135–5146, Oct. 2019.
[29] D. Hong, N. Yokoya, G. Xia, J. Chanussot, and X. X. Zhu, " X-ModalNet: A semi-supervised deep cross-modal network for classification of remote sensing data," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 12–23, 2020.
[30] Y. Rao, H. Lin, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4.
[31] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5345–5355, Nov. 2018.
[32] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 1585–1594.
[33] D. Hong *et al.*, "Interpretable hyperspectral AI: When non-convex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.
[34] S. Zhang, G. Fu, H. Wang, and Y. Zhao, "Degradation learning for unsupervised hyperspectral image super-resolution based on generative adversarial network," *Signal, Image Video Process.*, pp. 1–9, 2021.
[35] T. Uezato, D. Hong, N. Yokoya, and W. He, "Guided deep decoder: Unsupervised image pair fusion," *Proc. Euro. Conf. Comput. Vis. (ECCV)*, pp. 87–102, 2020.
[36] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.
[37] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 9446–9454.
[38] Y. Gandelsman, A. Shocher, and M. Irani, "Double-dip: Unsupervised image decomposition via coupled deep-image-priors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 11018–11027.
[39] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Denoising, inpainting, super-resolution," in *Proc. Int. Conf. Comput. Vision*, 2019, pp. 3844–3851.
[40] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7794–7803.
[41] R. Gu *et al.*, "CA-Net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 699–711, Feb. 2021.
[42] J. Yao, D. Hong, J. Chanussot, D. Meng, X. Zhu, and Z. Xu, "Cross-Attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 208–224.
[43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
[44] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
[45] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Comput. Vision Pattern Recognit.*, 2011, pp. 193–200.
[46] Y. Zhou, L. Feng, C. Hou, and S. Kung, "Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5997–6009, Oct. 2017.
[47] R. Dian, S. Li, and L. Fang, "Learning a low tensor-train rank representation for hyperspectral image super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2672–2683, Sep. 2019.
[48] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in *Proc. 3rd Conf. Fusion Earth Data, Merging Point Meas., Raster Maps, Remotely Sensed Images*, 2000, pp. 99–103.
[49] F. A. Kruse *et al.*, "The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 283, no. 1, pp. 192–201, 1993.
[50] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[51] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.