Deep Unsupervised Blind Hyperspectral and Multispectral Data Fusion

Jiaxin Li, Ke Zheng, Jing Yao, Lianru Gao, *Senior Member, IEEE*, and Danfeng Hong, *Senior Member, IEEE*

***Abstract*— Hyperspectral images (HSIs) usually have finer spectral resolution but coarser spatial resolution than multispectral images (MSIs). To obtain a desired HSI with higher spatial resolution, great research attention has been paid to achieving hyperspectral super-resolution by fusing the observed HSI with an auxiliary MSI of the same scene. However, most of the existing HSI-MSI fusion methods rely either on prior knowledge of the degradation model or on sufficient training data, hindering their practicality and interpretability. In this letter, we propose a novel unsupervised HSI-MSI fusion network with the ability of degradation adaptive learning, namely, UDALN. Specifically, we propose three modules to straightly encode the spatial and spectral transformations across resolutions, i.e., SpaDnet, SpeUnet, and SpeDnet. Through an elaborately designed three-stage unsupervised training strategy, the estimated network parameters can exhibit clear physical meanings of degradation processes and therefore help guarantee a faithful reconstruction of the desired HSI. The experimental results on two widely used hyperspectral datasets demonstrate the effectiveness of our method in comparison to the state-of-the-art HSI-MSI fusion models. (Code available at https://github.com/JiaxinLiCAS/UDALN_GRSL.)**

***Index Terms*— Data fusion, deep learning, hyperspectral, multispectral, unsupervised learning.**

I. Introduction

HYPERSPECTRAL imaging (HSI) technology is one of the most significant breakthroughs in the history of remote sensing by capturing contiguous spectral information in narrow spectral intervals, which produces HSIs product in the form of three-order data tensors with two spatial dimensions and one spectral dimension [1]. However, there exists an intrinsic trade-off between spectral resolution and spatial resolution due to the mechanisms in designing imaging systems. Hyperspectral and multispectral image fusion (HSI-MSI fusion) is a practically feasible way to achieve the goal of hyperspectral super-resolution and obtain a high spatial resolution HSI (HHSI).

Manuscript received November 12, 2021; revised January 15, 2022; accepted February 7, 2022. Date of publication February 15, 2022; date of current version March 8, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62161160336 and Grant 42030111 and in part by the China Postdoctoral Science Foundation under Grant 2021M693234. *(Corresponding author: Jing Yao.)*

Jiaxin Li is with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: lijiaxin203@mails.ucas.ac.cn).

Ke Zheng, Jing Yao, Lianru Gao, and Danfeng Hong are with the Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: zhengkevic@aircas.ac.cn; yaojing@aircas.ac.cn; gaolr@aircas.ac.cn; hongdf@aircas.ac.cn).

Digital Object Identifier 10.1109/LGRS.2022.3151779

The HSI-MSI fusion methods can be divided into three categories which are detail injection-based methods, optimization-based methods, and deep-learning (DL)-based methods. The first one is originally designed to deal with pan-sharpening and later adapted to tackle HSI-MSI fusion problem [2], [3]. Though these methods are computationally efficient, their fused results commonly contain non-negligible spectral distortion. The optimization-based methods, which include Bayesian representation, matrix factorization, and tensor factorization, are specifically designed for HSI-MSI fusion. The methods belonging to this category treat the fusion problem as an inverse problem, and desired results can be obtained by exploiting the degradation model and designing proper handcrafted priors. Wei *et al.* [4] transform the maximum likelihood probability problem into a closed-form solution problem for solving Sylvester equation, which can easily incorporate different priors. Point spread function (PSF) and spectral response function (SRF) are first estimated in [5], and then total variation prior is used in the subspace to regularize the ill-posed inverse problem. Besides, the unfolding matrix of HSI tensor can be expressed as the product of two matrices. For example, Yokoya *et al.* [6] and Lanaras *et al.* [7] use the coupled matrix factorization to alternatively update endmembers and abundances. Akhtar *et al.* [8] impose a local similarity prior on the sparse code solved by the greedy pursuit algorithm and use estimated sparse codes together with learned spectral dictionaries to reconstruct HHSI. The idea of coupled tensor factorization is adopted by Li *et al.* [9] and Kanatsoulis *et al.* [10] using Tucker decomposition and canonical decomposition, respectively, and the latter does not need any prior about PSF. Though optimization-based methods are mathematically and physically interpretable by considering the degradation model, the degradation models (i.e., PSF and SRF) are often given in advance, which restricts their practicality.

Recently, DL has attracted much attention in HSI-MSI fusion. This class of methods mainly focuses on a supervised manner and aims to learn the nonlinear mapping from input MSI and HSI to ground-truth HHSI via one or multiple branches' network without designing handcrafted priors [11], [12], but lacks adequate interpretability due to ignoring the degradation model. Though methods in [13]–[15] are capable of learning PSF and SRF as parameters of the network, the need of large training samples and complex network architectures make them lose practicality in real applications. Methods in an unsupervised manner are rarely studied. Qu *et al.* [16] use SRF as a known prior to constrain generated HHSI but ignore PSF. Though methods in [17] and [18] use adaptively learned degradation model to constrain the generated HHSI, some extra datasets are needed to pretrain

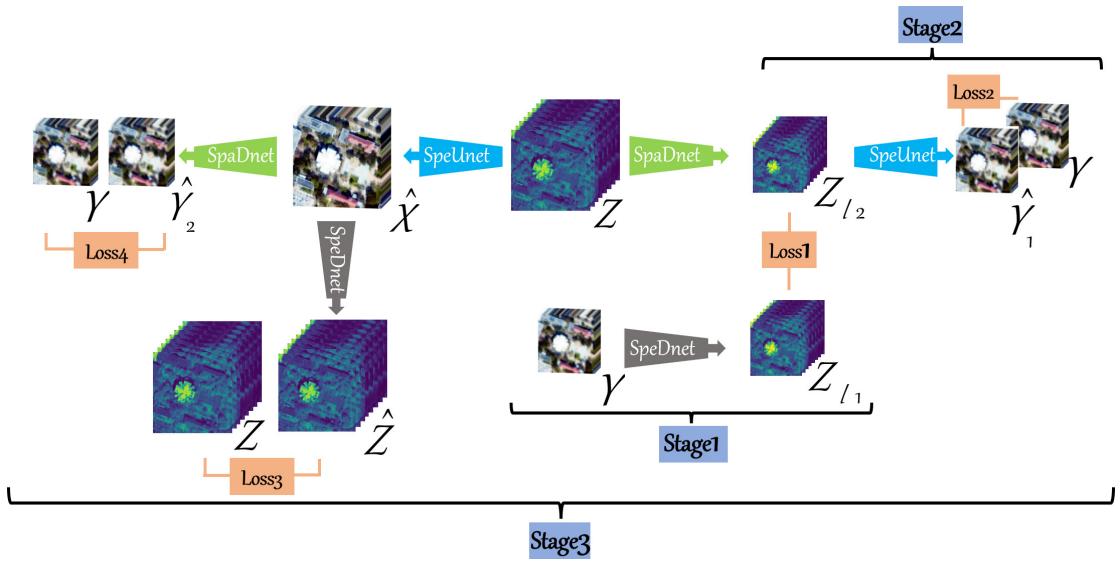

Fig. 1. Architecture of our method which contains a three-stage training procedure.

relevant modules. Very recently, Zheng *et al.* [19] model the coupled unmixing idea in DL to adaptively learn PSF and SRF, and Yao *et al.* [20] further introduce cross-modal attention mechanism, both yielding good results.

To overcome the limitations mentioned above, we propose an unsupervised degradation adaptive learning net, abbreviated as UDALN, for the task of HSI-MSI fusion. The main advantages are summarized as follows.

1) We propose a DL-based HSI-MSI fusion method in an unsupervised way which works on input HSI and MSI pair only and treats PSF and SRF as trainable parameters in the neural network to help get more accurate reconstruction results.
2) Three modules in our method are trained through a three-stage procedure instead of a plain end-to-end manner to obtain a better reconstruction result.

II. Methodology

We propose three novel modules in our UDALN method, i.e., spatial downsample network (SpaDnet), spectral downsample network (SpeDnet), and spectral upsample network (SpeUnet), which can be optimized in a three-stage training. The first two stages are designed to learn the degradation model and a coarse HHSI, and the last stage uses the learned degradation model to further refine the HHSI. Fig. 1 illustrates the architecture of our proposed UDALN.

A. Spatial Downsample Network and Spectral Downsample Network in Stage One

A low spatial resolution HSI $\mathbf{Y} \in \mathbb{R}^{wh \times C}$ and a low spectral resolution MSI $\mathbf{Z} \in \mathbb{R}^{WH \times Cm}$ can be regarded as spatially degraded and spectrally degraded version of ground-truth HHSI $\mathbf{X} \in \mathbb{R}^{WH \times C}$, where W, H, and C are the width, height, and spectral bands of $\mathbf{X}$, respectively ($w \ll W$, $h \ll H$, $C_m \ll C$). Accordingly, w and h are the width and height of $\mathbf{Y}$, and C_m is the spectral band of $\mathbf{Z}$. According to the widely used degradation model, HSI and MSI can be expressed as

$$\mathbf{Y} = \mathbf{PX} \tag{1}$$

$$\mathbf{Z} = \mathbf{XS} \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{wh \times WH}$ denotes the spatial degradation operation, which combines a convolution operation using PSF with a spatial downsampling operation, and $\mathbf{S} \in \mathbb{R}^{C \times Cm}$ is the spectral degradation operation whose column corresponds to a spectral response of one band in MSI $\mathbf{Z}$. Given $\mathbf{Y}$ and $\mathbf{Z}$, the goal of HSI-MSI fusion is to reconstruct the latent $\mathbf{X}$.

Besides (1) and (2), the spectrally degraded version of HSI $\mathbf{Y}$ should be equivalent to the spatially degraded version of MSI $\mathbf{Z}$

$$\mathbf{Z}_{l1} = \mathbf{YS} \tag{3}$$

$$\mathbf{Z}_{l2} = \mathbf{PZ} \tag{4}$$

$$\mathbf{Z}_{l1} = \mathbf{Z}_{l2} \tag{5}$$

where $\mathbf{Z}_{l1} \in \mathbb{R}^{wh \times Cm}$ and $\mathbf{Z}_{l2} \in \mathbb{R}^{wh \times Cm}$.

By designing a proper network and a loss function using (5), PSF and SRF can be estimated through back-propagation. The function of SRF is essentially the definite integral along the spectral dimension. That is to say, the spectral vector of each pixel in HSI $\mathbf{Y}$ is integrated into several scalars by each column of $\mathbf{S}$ and these scalars will form a new spectral vector of the corresponding pixel in $\mathbf{Z}_{l1}$. SpeDnet contains one convolutional layer with shape $C_m \times W \times 1 \times 1$ and stride 1 to formulate the integral process with weights acted as SRF, where C_m is both the band number of MSI $\mathbf{Z}$ and the number of convolutional kernels in SpeDnet, 1×1 is the spatial size, and for each kernel, and W is different and determined by the number of hyperspectral bands covered by the spectral response of each band in MSI $\mathbf{Z}$. The whole is represented by SpeDnet as follows:

$$\mathbf{Z}_{l1(i,j)} = \text{SpeDnet}(\mathbf{Y}, \theta) = \frac{\sum_{t \in \Theta_j} \mathbf{Y}_{i,t} \omega_j}{\sum \omega_j} \tag{6}$$

where θ denotes the weights in SpeDnet, i and j are the index of row and column, respectively, Θ_j denotes the jth support set that the band of $\mathbf{Y}$ belongs to, and ω_j is the weight of the jth $W \times 1 \times 1$ convolutional kernel.

PSF is easily implemented in the convolutional neural network by convolving each band in the spatial dimension with one same convolutional kernel of size $1 \times r \times r$ with stride r, where $r = W/w = H/h$ is both the scale factor in the spatial dimension and the size of convolutional kernels. The

whole is represented by SpaDnet as follows:

$$\mathbf{Z}_{l2} = \text{SpaDnet}(\mathbf{Z}, \beta) \tag{7}$$

where β is the weight of SpaDnet.

An $L1$-norm loss is adopted to evaluate the similarity of outputs of SpaDnet and SpeDnet in the pixelwise domain. The function is defined as follows:

$$\text{Loss1}(\mathbf{Z}_{l1}, \mathbf{Z}_{l2}) = \frac{1}{whC_m}\|\mathbf{Z}_{l1} - \mathbf{Z}_{l2}\|_1. \tag{8}$$

By back-propagation in the network training, the weights in SpaDnet and SpeDnet are automatically updated. In other words, PSF and SRF are adaptively learned, and the low spatial resolution MSI $\mathbf{Z}_{l2}$ can be obtained.

B. Spectral Upsample Network in Stage Two

Inspired by [21], a low spatial resolution MSI $\mathbf{Z}_l$ is obtained by degrading $\mathbf{Z}$ using the spatial degradation operation $\mathbf{P}$ in (1)

$$\mathbf{Z}_l = \mathbf{PZ} = \mathbf{PXS} = \mathbf{YS} \tag{9}$$

where $\mathbf{Z}_l \in \mathbb{R}^{wh \times Cm}$. As can be seen in (2) and (9), $\mathbf{Z}_l$ and $\mathbf{Z}$ are generated by the same spectral degradation operation $\mathbf{S}$ from $\mathbf{Y}$ and $\mathbf{X}$, respectively. In other words, if the spectral inverse mapping from $\mathbf{Z}_l$ to $\mathbf{Y}$ learned in the low resolution is applied to $\mathbf{Z}$, the latent HHSI can be reconstructed. Though inverse mapping is learned in low spatial resolution, it contains sufficient information to recover the target image. However, $\mathbf{P}$ in [21] is assumed as a known prior to generate $\mathbf{Z}_l$, which is not practical in real applications. Hence, we use $\mathbf{Z}_{l2}$ in (7) generated by SpaDnet in section A to replace $\mathbf{Z}_l$ and learn the spectral inverse mapping from $\mathbf{Z}_{l2}$ to $\mathbf{Y}$. Similar to SpeDnet, 1×1 convolution kernels are exploited in SpeUnet to learn mapping. The major difference between them is that multi-convolutional layers are used here to gradually recover the spectral information of $\mathbf{Y}$ instead of using one convolutional layer in SpeDnet, i.e., the number of feature maps becomes twice that of the previous layer after each convolution layer until the final output. The whole is represented by SpeUnet as follows:

$$\hat{\mathbf{Y}}_1 = \text{SpeUnet}(\mathbf{Z}_{l2}, \alpha) \tag{10}$$

where $\mathbf{Z}_{l2}$ is the input of SpeUnet and α is the weight of SpeUnet. The $L1$-norm loss is used to learn spectral inverse mapping, and the function is defined as follows:

$$\text{Loss2}(\mathbf{Y}, \hat{\mathbf{Y}}_1) = \frac{1}{whC}\left\|\mathbf{Y} - \hat{\mathbf{Y}}_1\right\|_1. \tag{11}$$

When the learned SpeUnet is applied to the original spatial resolution MSI $\mathbf{Z}$, a coarse HHSI $\hat{\mathbf{X}}_{\text{coarse}}$ can be obtained

$$\hat{\mathbf{X}}_{\text{coarse}} = \text{SpeUnet}(\mathbf{Z}, \alpha). \tag{12}$$

C. Three-Stage Training Details

HHSI $\hat{\mathbf{X}}_{\text{coarse}}$ can be reconstructed to a certain extent using the degradation model learned in Section II-A and the spectral inverse mapping learned in Section II-B, respectively. However, the result is limited by the following factors: the accuracy of the estimated degradation model and the accuracy and generalization of the estimated spectral inverse mapping learned in low resolution. To further constrain and refine the coarse HHSI, the degradation model learned in Section II-A is used to get a more accurate $\hat{\mathbf{X}}$, i.e., spatially degraded and spectrally degraded version of $\hat{\mathbf{X}}$ should be identical to $\mathbf{Y}$ and $\mathbf{Z}$, respectively

$$\hat{\mathbf{Z}} = \text{SpeDnet}(\hat{\mathbf{X}}_{\text{coarse}}, \theta) \tag{13}$$

$$\hat{\mathbf{Y}}_2 = \text{SpaDnet}(\hat{\mathbf{X}}_{\text{coarse}}, \beta). \tag{14}$$

The $L1$-norm loss can be expressed as follows:

$$\text{Loss3}(\mathbf{Z}, \hat{\mathbf{Z}}) = \frac{1}{WHC_m}\left\|\mathbf{Z} - \hat{\mathbf{Z}}\right\|_1 \tag{15}$$

$$\text{Loss4}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{whC}\left\|\mathbf{Y} - \hat{\mathbf{Y}}_2\right\|_1. \tag{16}$$

Ultimately, the loss of our model for training is given by

$$\begin{aligned}\text{Loss}(\mathbf{Y}, \mathbf{Z}) = {} & \text{Loss1}(\mathbf{Z}_{l1}, \mathbf{Z}_{l2}) + \text{Loss2}(\mathbf{Y}, \hat{\mathbf{Y}}_1) \\ & + \text{Loss3}(\mathbf{Z}, \hat{\mathbf{Z}}) + \text{Loss4}(\mathbf{Y}, \hat{\mathbf{Y}}_2).\end{aligned} \tag{17}$$

However, we do not simply use the final loss (17) to train three modules simultaneously. Instead, the whole training process is divided into three stages. The first two stages aim to initialize the weights of three modules, and the last stage is to refine HHSI $\hat{\mathbf{X}}_{\text{coarse}}$ to get a more accurate result $\hat{\mathbf{X}}$. The optimization steps are detailed in Algorithm 1.

Algorithm 1 Proposed UDALN Algorithm

Input: HR MSI $\mathbf{Z}$, LR HSI $\mathbf{Y}$
Training Procedure:
Stage1: Initialize weights β in SpaDnet and θ in SpeDnet
  a) Use $\mathbf{Z}$ and $\mathbf{Y}$ as input
  b) Back-propagation using $Loss1(\mathbf{Z}_{l1}, \mathbf{Z}_{l2})$
  c) Obtain β, θ, and $\mathbf{Z}_{l2}$
Stage2: Initialize weights α in SpeUnet
  a) Use $\mathbf{Z}_{l2}$ as input
  b) Back-propagation using $Loss2(\mathbf{Y}, \hat{\mathbf{Y}}_1)$
  c) Obtain α
Stage3: Obtain refined weights $\hat{\beta}, \hat{\theta}$, and $\hat{\alpha}$
  a) Back-propagation using $Loss(\mathbf{Y}, \mathbf{Z})$
  b) Obtain refined $\hat{\beta}, \hat{\theta}$, and $\hat{\alpha}$
End Procedure
Output: Generate the result HHSI $\hat{\mathbf{X}} = SpeUnet(\mathbf{Z}, \hat{\alpha})$

III. EXPERIMENTS

In this section, two widely used hyperspectral datasets are briefly reviewed, and the setup in our experiment is introduced. Then, the estimated degradation model is evaluated, and the ablation study of the three-stage training strategy is implemented to demonstrate the effectiveness of a three-stage training procedure. Our method is compared with eight state-of-the-art HSI-MSI fusion methods which are FUSE [4] from Bayesian representation, HySure [5], CNMF [6], CSU [7], and G-SOMP+ [8] from matrix factorization, CSTF [9] and STEREO [10] from tensor factorization, and unsupervised uSDN [16] from DL in the last.

A. Datasets and Settings

To evaluate the proposed method, two widely used public hyperspectral datasets are chosen. The first dataset is called Chikusei for short, which has 128 bands covering the wavelength from 363 to 1018 nm and consists of 2517×2335 pixels with a spatial resolution of 2.5 m. The second dataset is called Houston for short, which has 48 bands in the

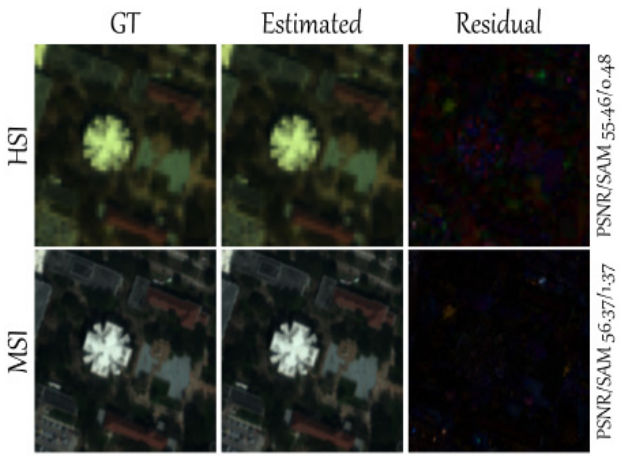

Fig. 2. Estimated HSI and MSI using the learned degradation model in the scale factor of 5 on the Houston dataset. The first line shows the results of HSI and the second line shows the results of MSI.

spectral band range from 380 to 1050 nm and contains 1202 × 4172 pixels with a spatial resolution of 1 m. Subimages of 240 × 240 × 110 on Chikusei and 240 × 240 × 46 on Houston are selected for test after discarding some noisy and water absorption bands. The two subimages are used as reference images and compared with the output of each method.

We generate simulated HSI and MSI as inputs following Wald's protocol [22]. Since the spatial degradation consists of a spatial convolution using PSF and a downsampling operation, we convolve each disjoint r by r pixel block in the reference images by an isotropic Gaussian PSF with kernel size and full width at half maxima equaling to r, where $r = W/w = H/h$ is the scale factor, to generate input HSI. We conduct two sets of experiments on each dataset by setting the scale factor to 5 and 8. The input MSI with eight bands is simulated by the spectral response of WorldView 2 multispectral imager.

B. Estimated Degradation Model

Since the degradation model is adaptively learned in our method to help achieve a faithful reconstruction result, we investigate the learned degradation model in this section. The MSI and HSI generated by the adaptively learned degradation model using the reconstructed HHSI $\hat{\mathbf{X}}$ are shown in Fig. 2. It can be found that the outputs of the estimated results are very similar to the ground truth, which is hard to tell the difference between them. Besides, the residuals of three chosen bands are so small that the original result is black, so we magnify the original residual by a factor of 40 to show the differences more clearly. In addition, two widely used metrics, namely, PSNR and SAM, are calculated to indicate that the learned degradation model can reconstruct the input MSI and HSI effectively and hence promote the reconstruction quality.

C. Ablation Study

A three-stage training strategy is used to optimize our method, and the ablation study of the training strategy is shown in Table I, where "$S1 + S2$" means the model is only trained via the first two stages without constraint and refinement from stage three, and "$S3$" means the model is directly trained via stage three without initialization from the first two stages. The optimal results are bold, which is the same for other tables. As can be seen in Table I, the three-stage training strategy achieves better results compared with other conditions. Specifically, the results of "$S1 + S2$" demonstrate the effectiveness of constraint from stage three, and the results of "$S3$" show that the first two stages provide the modules with better initialized parameters than random initialization, which will be of benefit to training.

TABLE I

ABLATION STUDY ON HOUSTON DATASET WITH DIFFERENT TRAINING STAGES

Scale factor	Stages	Metric				
		SAM	ERGAS	PSNR	RMSE	CC
5	S1+S2	**0.8654**	0.5905	43.9479	0.0046	**0.9994**
	S3	1.0162	0.6749	43.1181	0.0051	0.9992
	S1+S2+S3	0.8748	**0.5232**	**45.2627**	**0.0041**	**0.9994**
8	S1+S2	0.9238	0.3394	44.8672	0.0042	0.9993
	S3	0.9523	0.3607	44.3196	0.0045	0.9993
	S1+S2+S3	**0.9048**	**0.3185**	**45.5478**	**0.0039**	**0.9994**

TABLE II

QUANTITATIVE PERFORMANCE OF NINE METHODS ON CHIKUSEI DATASET

Methods	Scale factor									
	5					8				
	SAM	ERGAS	PSNR	RMSE	CC	SAM	ERGAS	PSNR	RMSE	CC
HySure	1.2008	1.0781	42.2838	0.0081	0.9975	1.5504	0.9064	39.9008	0.0108	0.9968
FUSE	1.3411	1.2058	41.1569	0.0094	0.9966	1.4452	0.8318	40.5546	0.0098	0.9971
G-SOMP+	1.2878	1.3192	41.2449	0.0090	0.9945	1.5257	1.0831	38.8846	0.0117	0.9913
CSU	1.3812	1.7623	39.7220	0.0097	0.9889	1.6958	1.2280	38.2603	0.0117	0.9913
CNMF	1.0747	1.0996	42.6517	0.0078	0.9963	1.2380	0.8835	40.0916	0.0103	0.9951
STEREO	0.8447	0.7745	51.4839	0.0041	0.9970	1.1967	0.6450	48.0002	0.0059	0.9953
CSTF	0.8255	0.7818	**53.0044**	0.0041	0.9970	0.8093	0.5072	49.2651	0.0042	0.9970
uSDN	1.0700	1.5033	43.2698	0.0069	0.9909	1.3457	1.018	41.5412	0.0081	0.9896
UDALN	**0.6921**	**0.5816**	52.5624	**0.0036**	**0.9983**	**0.7602**	**0.4592**	**50.7592**	**0.0040**	**0.9974**

TABLE III

QUANTITATIVE PERFORMANCE OF NINE METHODS ON HOUSTON DATASET

Methods	Scale factor									
	5					8				
	SAM	ERGAS	PSNR	RMSE	CC	SAM	ERGAS	PSNR	RMSE	CC
HySure	1.5592	0.6661	43.2444	0.0050	0.9989	2.2145	0.5778	40.2108	0.0071	0.9983
FUSE	1.6766	0.9116	40.5941	0.0072	0.9978	1.8366	0.5607	40.8336	0.0065	0.9980
G-SOMP+	1.4909	0.6883	42.8372	0.0053	0.9989	1.8977	0.5437	40.5537	0.0067	0.9984
CSU	1.5235	0.7080	42.4352	0.0054	0.9986	1.9117	0.5413	40.4194	0.0067	0.9980
CNMF	1.2008	0.5981	43.8871	0.0047	0.9991	1.5564	0.4883	41.5621	0.0061	0.9986
STEREO	1.5078	0.5698	45.0847	0.0043	0.9986	1.5497	0.3629	44.9795	0.0044	0.9985
CSTF	1.2607	0.5432	44.8660	0.0043	0.9988	1.2984	0.3475	44.7972	0.0044	0.9987
uSDN	1.9640	0.7834	42.1170	0.0060	0.9978	2.0795	0.5261	41.4574	0.0064	0.9973
UDALN	**0.8748**	**0.5232**	**45.2627**	**0.0041**	**0.9994**	**0.9048**	**0.3185**	**45.5478**	**0.0039**	**0.9994**

D. Comparison With State-of-the-Art Methods

The quality of fused images on the Houston dataset in the scale factor of 8 is evaluated visually in Fig. 3. The first row is the output of each method, the second row is the SAM error heatmap to show the spectral reconstruction quality in pixelwise domain, and the third row is the mean relative

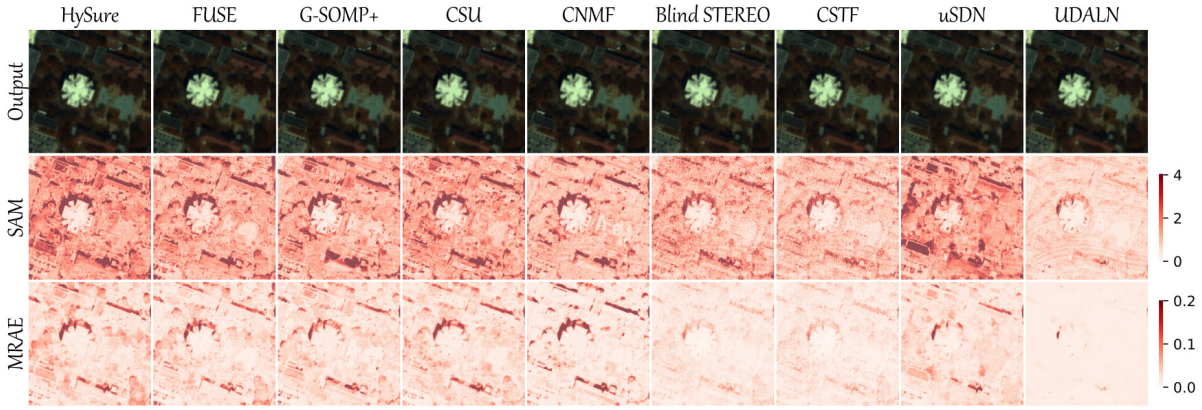

Fig. 3. Fusion results of nine methods on the Houston dataset in the scale factor of 8.

absolute error (MRAE) heatmap to show the magnitude of pixelwise error. Though it is hard to discern the differences among the outputs of all the methods, it can be distinguished from SAM and MRAE maps that our method provides better reconstruction results with a lighter color.

The quantitative results of all the methods on two datasets with two scale factors of 5 and 8 are also summarized in Tables II and III, respectively. Tensor-factorization-based STEREO and CSTF show good performance because of their exploitation of 3-D spatial and spectral structure information, while uSDN underperforms in spectral restoration due to the dependence on prior information of the degradation model. Thanks to the ability of degradation adaptive learning and the training strategy, our method shows competitive results and gives the best values in almost all the metrics, which verifies the faithful reconstruction ability of UDALN.

IV. Conclusion

In this letter, a new unsupervised DL-based HSI-MSI fusion method using a three-stage training strategy is proposed to obtain a desired HHSI without requiring the prior information of the degradation model and large training samples. By incorporating degradation adaptive learning into HSI-MSI fusion, spatial and spectral transformations are learned by three simple and effective modules. Experiments on two hyperspectral datasets indicate the reconstruction effectiveness of our method in both qualitative and quantitative perspectives.

References

[1] D. Hong *et al.*, "Interpretable hyperspectral artificial intelligence: When nonconvex modeling meets hyperspectral remote sensing," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 2, pp. 52–87, Jun. 2021.

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

[3] M. Selva, B. Aiazzi, F. Butera, L. Chiarantini, and S. Baronti, "Hyper-sharpening: 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.

[4] 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.

[5] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3373–3388, Jun. 2015.

[6] N. Yokoya, T. Yairi, 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.

[7] C. Lanaras, E. Baltsavias, and K. Schindler, "Hyperspectral super-resolution by coupled spectral unmixing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3586–3594.

[8] N. Akhtar, F. Shafait, and A. Mian, "Sparse spatio-spectral representation for hyperspectral image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2014, pp. 63–78.

[9] 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.

[10] C. I. Kanatsoulis, X. Fu, N. D. Sidiropoulos, and W.-K. Ma, "Hyperspectral super-resolution: A coupled tensor factorization approach," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6503–6517, Dec. 2018.

[11] J. Yang, Y.-Q. Zhao, and J. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, May 2018.

[12] 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.

[13] W. Wei, J. Nie, Y. Li, L. Zhang, and Y. Zhang, "Deep recursive network for hyperspectral image super-resolution," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 1233–1244, 2020.

[14] Q. Xie, M. Zhou, Q. Zhao, Z. Xu, and D. Meng, "MHF-Net: An interpretable deep network for multispectral and hyperspectral image fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1457–1473, Mar. 2022, doi: 10.1109/TPAMI.2020.3015691.

[15] J. Yang, L. Xiao, Y.-Q. Zhao, and J. C.-W. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022, doi: 10.1109/TGRS.2021.3080697.

[16] Y. Qu, H. Qi, and C. Kwan, "Unsupervised sparse Dirichlet-net for hyperspectral image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2511–2520.

[17] Y. Fu, T. Zhang, Y. Zheng, D. Zhang, and H. Huang, "Hyperspectral image super-resolution with optimized RGB guidance," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11661–11670.

[18] L. Zhang, J. Nie, W. Wei, Y. Zhang, S. Liao, and L. Shao, "Unsupervised adaptation learning for hyperspectral imagery super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3070–3079.

[19] K. Zheng *et al.*, "Coupled convolutional neural network with adaptive response function learning for unsupervised hyperspectral super resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2487–2502, Mar. 2021.

[20] J. Yao, D. Hong, J. Chanussot, D. Meng, X. X. Zhu, and Z. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Oct. 2020, pp. 208–224.

[21] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based multi-branch BP neural networks," *Remote Sens.*, vol. 11, no. 10, p. 1173, May 2019.

[22] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.