Multi-scale spatial fusion and regularization induced unsupervised auxiliary task CNN model for deep super-resolution of hyperspectral image.

HA, V.K., REN, J., WANG, Z., SUN, G., ZHAO, H. and MARSHALL, S.

2022

Multi-scale spatial fusion and regularization induced unsupervised auxiliary task CNN model for Deep Super-Resolution of Hyperspectral Image

Viet Khanh Ha, Jinchang Ren, *Senior Member*, *IEEE*, Zheng Wang, Genyun Sun, Huimin Zhao and Stephen Marshall

***Abstract*—Hyperspectral images (HSI) features rich spectral information in many narrow bands but at a cost of a relatively low spatial resolution. As such, various methods have been developed for enhancing the spatial resolution of the low-resolution HSI (Lr-HSI) by fusing it with high-resolution multispectral images (Hr-MSI). The difference in spectrum range and spatial dimensions between the Lr-HSI and Hr-MSI have been fundamental but challenging for multispectral/hyperspectral (MS/HS) fusion. In this paper, a multi-scale spatial fusion and regularization induced auxiliary task (MSAT) based CNN model is proposed for deep super-resolution of HSI, where a Lr-HSI is fused with a Hr-MSI to reconstruct a high-resolution HSI (Hr-HSI) counterpart. The multi-scale fusion is used to efficiently address the discrepancy in spatial resolutions between the two inputs. Based on the general assumption that the acquired Hr-MSI and the reconstructed Hr-HSI share similar underlying characteristics, the auxiliary task is proposed to learn a representation for improved generality of the model and reduced overfitting. Experimental results on five public datasets have validated the effectiveness of our approach in comparison with several state-of-the-art methods.**

***Index Terms*—Hyperspectral image (HSI); super-resolution (SR), multi-scale spatial fusion, auxiliary task, convolutional neural networks (CNN).**

I. Introduction

Hyperspectral images (HSI) consist of contiguous bands from across the electromagnetic spectrum, where each pixel in the image scene is composed of a spectral vector as its profile or signature. With the rich spectral characteristics, HSI has been successfully applied in a wide range of applications, such as precision agriculture [1], [2], [3], [4], target detection [5], image enhancement [6], [7], [8], land cover analysis [9], as well as measurement of chemical substances [10], and change detection [11]. In fact, there is always the inevitable trade-off between the spatial and spectral resolutions in captured the HSI, which means that images can not be acquired with both high spatial and high spectral resolutions at the same time.

Recent developments in image super-resolution (SR) have heightened the need for hyperspectral image super-resolution (HSI-SR). Image SR aims to reconstruct a high-resolution (HR) image from one or several low-resolution (LR) images. Due to the high spectral dimension of HSI, the reconstructed Hr-HSI only from Lr-HSIs usually contains spectral and/or spatial distortion. Given the auxiliary information such as the panchromatic (PAN) image, Red, Green, and Blue (RGB) image, or multispectral image (MSI), the fusion-based HSI-SR has received increasing attention recently. Originated from image pan-sharpening [12], [13], [14], HSI-SR is a combination of Lr-HSI and Hr-MSI to create a single Hr-HSI.

Recently, various techniques have been proposed for MS/HS fusion. Typically, the HSI SR approaches can be roughly categorized into three classes: i.e., dictionary-based sparse representation, maximum a-posteriori-based Bayesian, and deep learning. In sparse representation approaches, the source images are represented by a dictionary and the corresponding sparse coefficients, where the matrix factorization and the tensor factorization are most commonly used. The matrix factorization can help to decompose high dimensional data and fuse MS/HS data [15], [16]. Dong *et al.* proposed a non-negative structured sparse representation (NSSR) [17] method to jointly estimate the dictionary and the sparse coefficients based on the prior knowledge of the spatial-spectral sparsity in the source images. As the observed Lr-HSIs and Hr-MSIs can capture the same scene as the target Hr-HSIs, they are assuming to share the same underlying spectral materials or *endmembers*. Lanaras *et al.* proposed the coupled spectral unmixing (CSU) [18] method for the fusion problem, where the Lr-HSI and Hr-MSI are alternatively unmixed to estimate the spectral endmembers and abundances.

By extending the matrix factorization to higher-order tensors, tensor factorization can extract the underlying factors in high-order dataset [19], [20], [21]. Dian *et al.* proposed non-local sparse tensor factorization (NLSTF) [19] to reconstruct HSI-SR in a cube-by-cube manner, by assuming that each cube is formed by a core coefficient tensor and dictionaries of width mode, height mode and spectral mode. In [19], the non-local spatial self-similarity of Hr-MSI is exploited through a clustering method to constrain the spatial correlation in the

This work was supported in part by the Shanxi Hundred People Plan of China, in part by the ETF Scholarship from the Faculty of Engineering, University of Strathclyde, and in part by the Government Scholarship of Vietnam. *(Corresponding author: Jinchang Ren.)*

Viet Khanh Ha and S. Marshall are with Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow G1 1XQ, UK.

J. Ren and H. Zhao are with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China, J, Ren is also with the National Subsea Centre, Robert Gordon University, Aberdeen, U.K (email: jinchang.ren@ieee.org).

Z. Wang is with the School of Computer Software, Tianjin University, Tianjin 300354, China.

G. Sun is with the College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China.

Hr-HSI. In [22], they proposed a low tensor train rank representation (LTTR) method by considering the Hr-HSI as a four-dimension tensor with non-local LTTR prior from HR-HSI to regularize the fusion problem. Various tensor factorization based approaches have been used for the SR fusion problem, including non-local patch tensor sparse representation [23], and subspace-based low tensor multi-rank regularization [24], etc. The issue of the factorization based methods is that there is not a single unique decomposition thus it is difficult to determine the basic elements or factorization rank. Some prior information of the HR-HSI are introduced to regularize the SR problem in previously mentioned work, including priors of spectral unmixing [18], nonlocal spatial similarities [17], sparse priors [19], and nonlocal LTTR prior [22] et al.

As a different framework, Bayesian approaches typically estimate the posterior distribution with the maximum a posteriori (MAP) based on the prior knowledge and the observation model [25], [26], [27], [28], [29]. Since the HSI-SR problem is usually ill-posed, the Bayesian methods define an appropriate prior distribution for the scene of interest to regulate problem. Instead of incorporating simple Gaussian prior, sparse representation is used as sparsity promoted Gaussian to regulate the problem. Akhtar *et al.* [25] proposed a Bayesian dictionary learning and sparse coding algorithm for HSI-SR that has shown improved performance. In [26], [27], Wei *et al.* introduced subspace transformation and a regularization to cope with ill-posed inverse problem. Later, a Sylvester equation-based explicit solution was integrated into the Bayesian MS/HS fusion [28] to significantly decrease the computational complexity. In [29], a method called *Hysure* was proposed that use a form of vector total variation (VTV) [30] for the regularizer. The major drawback of the Bayesian methods is that prior assumptions of distribution, which typically derive from well-known distributions may not observe in real-world datasets.

Different from the conventional methods, deep learning-based methods impose fewer assumptions on the prior knowledge of to-be-estimated Hr-HSI and still achieve good result for MS/HS fusion. It has previously been used to solve the pan-the sharpening problems [12], [31], [32], and later developed for MS/HS fusion in non-blind fusion with both supervised [14], [33], [34], [35], and unsupervised [36], [37] ones or blind fusion [38]. Recent work has shown hybrid methods by combining the deep learning with sparse representation [39], [40]. The drawback of deep learning-based methods is a lack of specific designs for MS/HS fusion, which often use a generic CNNs-based framework being designed for other tasks or different types of images, thus not effective. There is little attention to the characteristics of HSI, for example, spectral low-rankness. Although the low-rank property is not focused in our work, this implies the possibility of introducing additional regularization to HSI reconstruction.

In our paper, we first propose a novel CNN architecture to fuse the Lr-HSI and the Hr-MSI in a progressive manner. To address the spatial difference between the Hr-MSI and the Lr-HSI, recent CNN-based approaches [41], [42] have upsampled images from the Lr-HSI to the image that has a size of Hr-HSI. This strategy would increase computational demand without compromising the SR performance. Second, the regularization methods in generic CNNs-based framework is insufficient for specific tasks or image types, additional constraints are therefore needed to regulate HSI-SR solution. The current deep learning-based methods have considered only Hr-HSI as the ground-truth for the supervised task while paying less attention on unsupervised features of the Hr-MSI. As the Hr-MSI and the Hr-HSI are both Hr images and capture the same scene, we exploit representation from Hr-MSI through unsupervised learning to improve generality of our MS/HS fusion network.

The major contributions of the proposed Multi-Scale spatial fusion and Auxiliary Task (MSAT) are two-fold:

1. A multi-scale spatial and spectral architecture is proposed, which can efficiently and effectively exploit the spatial and spectral features from both Hr-MSIs and Lr-HSIs.

2. An auxiliary unsupervised task is proposed, which acts as an additional form of regularization to further improve the generalization performance of the supervised task. This can not only significantly improve the performance of our proposed MSAT model but also that of other CNN models when tested on five publicly available datasets.

The remainder of this paper is organized as follows. In Section II, a review of the related MS/HS fusion methods is given. Section III formulates the problem of the MS/HS fusion and details our proposed MSAT model for MS/HS fusion. In Section IV, experimental results on five public HSI datasets and discussions are represented. Some concluding remarks and future directions are given in Section V.

II. Related Work

A. Joint learning via progressively downsampling and upsampling process

Jointly learning operation in a CNN-based model is to combine features using the summation or concatenation of the tensors, which normally requires tensors to have the same spatial dimension. Since the observed Hr-MSI and Lr-HSI have different spatial resolutions, two stages are employed for the fusion framework, as detailed below. In the first stage, the Hr-MSI is progressively downsampled into multi-scales and then fused with the LrHS images of the same spatial size. For the second phase, there are three commonly used upsampling techniques for image super-resolution, i.e., pre-upsampling, post-upsampling and progressive-upsampling. When the upsampling factor is large, the first two techniques increase either the parameters of the network or the difficulty of training. The progressive upsampling method, however, allows the training to gradually shift its attention from the large-scale structure of image to finer-scale details, instead of having to learn all scales simultaneously. Therefore, the architecture appears similar to the U-Net [43], which can not only significantly reduce the learning difficulty but also improve the performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3176969, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

B. Multi-Task learning

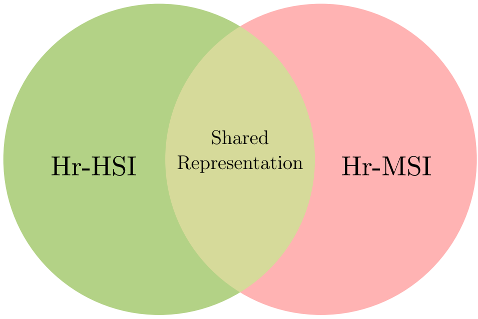


Fig. 1: Both the observed Hr-MSI and estimated Hr-HSI share some common spatial representation.

Multi-task learning has improved the generalization performance [44]. Apart from directly reconstructing Hr-HSI in a supervised manner, we introduce an unsupervised auxiliary task, aiming to reconstruct a Hr-MSI from the corrupted Hr-MSI. Intuitively, the observed Hr-MSI and estimated Hr-HSI should share similar spatial information, as shown in Fig. 1; Otherwise, the MS/HS fusion task becomes trivial.

This shared representation is essential for estimating both the Hr-MSI and Hr-HSI, where this feature is representative for the Hr-MSI data and also crucial for estimating the Hr-HSI. Directly estimating of the Hr-HSI from any given Lr-HSI and Hr-MSI is likely an under-constrained problem. This means solutions can be found to well fit the data but often fail to extract the underlying patterns in the data, result in poor generalization. Introducing an auxiliary task for reconstructing Hr-MSI will train the model to find the solution over a small area of the intersection of two tasks rather than on a broader area of a single task. Therefore, this can help the network to achieve faster and better convergence. Moreover, the auxiliary task acts as a regularizer by introducing a reductive bias, where the number of possible solutions can be reduced.

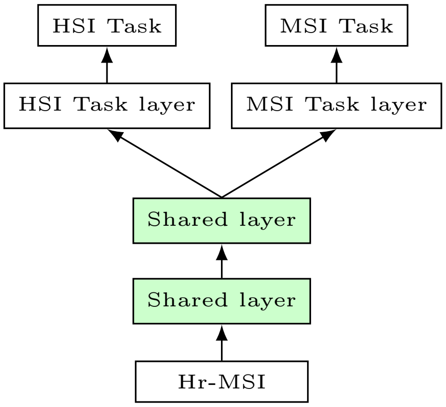


Fig. 2: Hard parameter sharing for multi-task learning in deep neural networks.

Hard parameter sharing is the most commonly used approach in multi-task learning with neural networks, as shown in Fig. 2. It is generally applied by sharing the hidden layers between all tasks while keeping several task-specific output layers. When training jointly, both Hr-MSI and Hr-HSI tasks can extract mutually important features to reduce the total error of reconstruction, i.e. enabling the shared layer to capture the common features of both. This is equivalent to the sparse representation-based method, e.g. the NLSTF [14] method, that uses the non-local self-similarity of Hr-MSI to impose the spatial constraints on the estimated Hr-HSI.

C. Denoising with the autoencoders

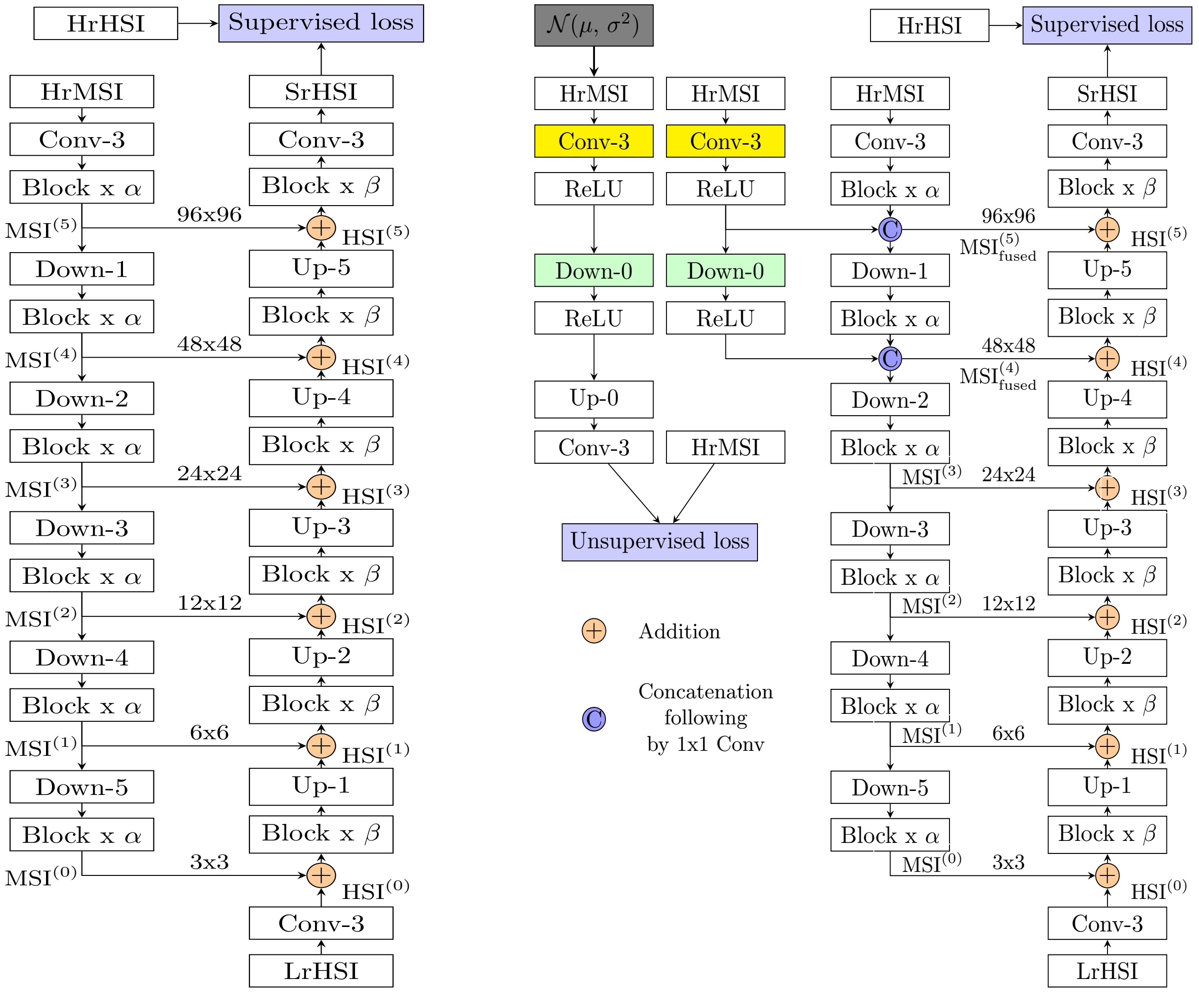
Given Hr-MSIs as the high-resolution images, an autoencoder can also be used as an auxiliary task for learning a compressed representation of Hr-MSIs, which is then used to impose regularization on the HSI SR. The convolutional autoencoder is an unsupervised learning method, which first learns the representations by performing the convolution and downsampling on the input. These representations are then decoded by up-sampling and convolutions to reconstruct the original image of input. The denoising autoencoders [45], is an extension to the classical autoencoder, which reconstructs the input from a corrupted version of it.

D. Fusion based HSI Super-Resolution

Borrowing spatial information from the high-resolution auxiliary image (e.g., RGB, PAN, MSI) is commonly used in the MS/HS fusion-based HSI-SR methods. The estimated Hr-HSI is assumed to share the spatial information with the auxiliary image and also similar spectral information with the Lr-HSI. The relationship between the Hr-HSI and Hr-MSI was analyzed in [46], where the camera spectral sensitivity that generates the Hr-MSI was exploited from the Hr-HSI before being applied to improve the Hr-HSI reconstruction. To exploit correlations in both the spectral and spatial domains, the sparse representation methods are used to estimate the key elements of Hr-HSI in both the source images. Though these approaches have achieved competitive performance, the handcraft prior between the input images and the target image is needed. Most recently, some deep learning-based methods [33], [34], [35], [36], [37], [38], [39], [40], have gradually become popular due to their superior performance and fewer assumptions in this context.

III. The proposed approach

For notational convenience, all Lr-HSI, Hr-MSI, and Hr-HSI are denoted as two-dimensional matrices. Let the matrix representing the Lr-HSI be $\mathbf{Z} \in R^{C\times hw}$ with C bands and spatial dimension *hw*, and let denote $\mathbf{Y} \in R^{c\times WH}$ the obtained Hr-MSI with c spectral bands and spatial dimension *WH*. The goal is to estimate the Hr-HSI, present as $\mathbf{X} \in R^{C\times WH}$, with both high spatial and spectral resolutions. In general, Hr-MSI has much higher spatial resolution than Lr-HSI ($HW \gg hw$), and Lr-HSI has a much higher spectral resolution than the Hr-MSI ($C \gg c$).



(a) Baseline architecture without auxiliary task.

(b) MSAT architecture.

Fig. 3: The architecture of the proposed MSAT. The same yellow or green colour boxes indicates shared variables between the supervised and unsupervised tasks.

The Lr-HSI can be regarded as a spatially down-sampled version of the Hr-HSI:

$$\mathbf{Z} = \mathbf{XBS} \tag{1}$$

where $\mathbf{B} \in R^{WH \times WH}$ represents a convolution between the point spread function (PSF) of the sensor and the Hr-HSI band, and $\mathbf{S} \in R^{WH \times wh}$ is a downsampling matrix.

Similarly, the Hr-MSI, e.g. a RGB/PAN image, can be taken as a spectrally downsampled version of the Hr-HSI:

$$\mathbf{Y} = \mathbf{RX} \tag{2}$$

where $\mathbf{R} \in R^{c \times C}$ is the corresponding camera spectral response function.

The problem of HSI-SR can be solved by learning the mapping between $\mathbf{X}$ and the coupled $\mathbf{Y}$, $\mathbf{Z}$ below in a fully convolutional fashion using the gradient descent. The proposed multi-task objective is represented as:

$$\underset{\theta,\psi}{\operatorname{argmin}} \parallel \mathbf{f}(\mathbf{X}|\boldsymbol{\theta},\mathbf{Y},\mathbf{Z}) - \mathbf{X} \parallel_2^2 + \gamma \parallel \mathbf{g}(\mathbf{Y}|\boldsymbol{\psi},\widetilde{\mathbf{Y}}) - \mathbf{Y} \parallel_2^2 + \eta R(\mathbf{X}) \tag{3}$$

where $\mathbf{f}(\mathbf{X}|\boldsymbol{\theta},\mathbf{Y},\mathbf{Z})$ and $\mathbf{g}(\mathbf{Y}|\boldsymbol{\psi},\widetilde{\mathbf{Y}})$ are the outputs of the proposed network; $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are trainable parameters of two sub-networks; γ and η denote two pre-defined trade-off parameters. During the multi-task learning, part of $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is shared, as illustrated in Fig. 2. The first and second terms are the pixel-wise L_2 distance between the network outputs and the corresponding ground-truth $\mathbf{X}$ and $\mathbf{Y}$, respectively. The final term refers to the L_2 regulation.

There are two major objectives for designing our fusion network. One is to reduce the spatial discrepancy between the two observed data. The other is to improve the generalization of representation by sharing the main supervised task with an unsupervised auxiliary task. These representations are not only

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3176969, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

useful to support the decision for the supervised task but also work as a regularizer for more effective HSI SR [47].

We detail our MS/HS fusion network in Fig. 3, in which Fig. 3a illustrates the baseline architecture and Fig. 3b a baseline architecture extended with the proposed auxiliary task. The construction of our model involves a top-down pathway, a bottom-up pathway, an auxiliary task, and some lateral connections, as introduced below.

Top-down pathway (MSI branch): In this pathway, the given training Hr-MSI is progressively down-sampled with a scaling factor of 2 into five hierarchical spatial levels, starting from an image sized of 96 × 96 × 3 to 3 × 3 × 31. Often, there are many layers that produce output maps of the same size, which are defined in the same network *stage* or *level*. Let $Y^{(s-1)}$ and $Y^{(s)}$ denote the input and output feature maps of the s-th *level* in the MSI branch, and the relation between $Y^{(s-1)}$ and $Y^{(s)}$ is formulated by:

$$Y^{(s-1)} = Resblock(Downsample(Y^{(s)})) \tag{4}$$

where $Resblock(\cdot)$ and $Downsample(\cdot)$ denote respectively the ResNet block and a downsample operation using a convolution layer with *strike* = 2. The highest *level* (s = 5) is the feature maps extracted from the observed Hr-MSI without downsampling.

An auxiliary task (Denoising branch): In the proposed model (Fig. 3b), a *light* denoising autoencoder (DAE) is introduced as an auxiliary task, which is trained to reconstruct the original observation $\mathbf{Y}$ from its corrupted version $\widetilde{\mathbf{Y}}$ by minimizing the error between the input $\mathbf{Y}$ and its reconstruction $\mathbf{g}(\mathbf{Y}|\psi, \widetilde{\mathbf{Y}})$ from the corrupted $\widetilde{\mathbf{Y}}$. With the presence of noise, the DAE is forced to learn the representation of the data, which later is able to reconstruct the original input. The corrupted $\widetilde{\mathbf{Y}} = \mathbf{Y} + \mathcal{N}(\mu, \sigma^2)$ is used to train the DAE with the clean version Y fed into the Encoder to extract the underlying representation for both tasks. Formally, the representation of Hr-MSI at multiple levels $\bar{Y}_s, \bar{Y}_{s-1}, \dots, \bar{Y}_0$ are extracted as follows:

$$\hat{Y}_s, \hat{Y}_{s-1}, \dots, \hat{Y}_0 = \textbf{Encoder}(\widetilde{Y}) \tag{5}$$

$$\dot{Y}_s, \dot{Y}_{s-1}, \dots, \dot{Y}_0 = \textbf{Decoder}(\hat{Y}_s, \hat{Y}_{s-1}, \dots, \hat{Y}_0) \tag{6}$$

$$\bar{Y}_s, \bar{Y}_{s-1}, \dots, \bar{Y}_0 = \textbf{Encoder}(Y) \tag{7}$$

We then use features $\bar{Y}_s, \bar{Y}_{s-1}, \dots, \bar{Y}_0$ that come after the ReLU activation for a supervised task, and our model is trained in an end-to-end manner.

Lateral connections between the main task and the auxiliary task: The DAE relies on a certain number of training (noisy) examples to learn the representations/patterns before transferring them to the main task. Our main task should therefore decide when to use such information and when to forget irrelevant ones. The simple mechanism is to use a 1×1 convolution layer. However, one problem with this is that the main task may neglect shared representations by setting the kernel parameters to zeros. The total loss is then minimized by decreasing each supervised and unsupervised loss separately. The representations learnt by the DAE thus are of no use to the main task. To avoid this unwanted effect, we introduce a compression mechanism by taking advantage of a 1×1 convolution layer. Concretely, the feature maps extracted from Hr-MSI in both the denoising task and the main task are concatenated and sent to a 1×1 convolution layer. This layer performs dimensionality reduction and forces the main task to utilize the information from the denoising task.

$$Y^{(l)}_{fused} = Conv_{1x1}(Concatenate(Y^{(l)}, \bar{Y}^{(l)})) \tag{8}$$

Bottom-up pathway (HSI branch): The bottom-up pathway hallucinates higher resolution features by up-sampling the spatial feature maps from lower levels of the Lr-HSI.

$$X^{(s-1)} = Upsample(Resblock(X^{(s-2)})) \tag{9}$$

where $Resblock(\cdot)$ denotes ResNet block and $Upsample(\cdot)$ is a upsample operation using a transposed convolution layer. The up-sampled map is then merged with the corresponding top-down map by element-wise addition.

$$\hat{X}^{(s-1)} = X^{(s-1)} + Y^{(s-1)}_{fused} \tag{10}$$

The top-down pathway is rich in spatial information while the bottom-up pathway contains a high level of spectral information. To build a deep network without changing the network topology, the parameters α and β control the depth of the network. Only one residual block (α = 1, β = 1) is used at a certain spatial levels unless stated otherwise. Our residual block is derived from the MobileNetV1 [48], in which the conventional 3 × 3 convolution is replaced by a 3 × 3 depth-wise separable convolution. The down-sampling and up-sampling blocks refer to one-step convolution with stride = 2 and a transpose convolution, respectively.

IV. Experiments

A. Experimental Database

For performance evaluation, we conduct experiments on five public benchmark datasets: CAVE [49], Harvard [50], ICVL [51], Chikusei [52], and a spaceborne images of *Roman Colosseum* acquired by World View-2. The CAVE dataset [49] comprises 32 indoor HSIs captured under controlled illumination. The images have 31 spectral bands with a spatial dimension of 512×512 pixels, and a spectral sampling gap of 10nm from 400nm to 700nm. The Harvard dataset [50] has 50 indoor and outdoor images, recorded under daylight illumination, where 27 images were under artificial or mixed illumination. With a spatial size of 1392×1040 pixels, each HSI has 31 spectral bands, with a 10-nm spectral sampling gap within [420, 720] nm. The ICVL dataset [51] contains 201 HSIs of real-world indoor and outdoor scenes, has 31 spectral bands each ranging from 400nm to 700nm at a 10nm increment. We use only the top left 1024×1024 pixels for convenience of the spatial down-sampling. The Chikusei scene [52] is an airborn HS image taken over Chikusei, Ibaraki, Japan. The image has a spatial dimension of 2517 × 2335 pixels, comprising 128 bands in the spectral range from 363 to 1018 nm. We select a 500 × 2210 pixel-size image from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3176969, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing

the top area of the original data for training. Besides, we extract 16 non-overlapped 448 × 448 images as the testing set. The sample images of the Roman Colosseum contain an Hr-MSI (RGB image) of size 1676 × 2632 × 3 and Lr-HSI image of size 419 × 658 × 8. We select 208 × 658 and 836 × 2632 pixels image from Lr-HSI and Hr-MSI for training and the remaining for testing data. Since the ground-truth is not available, we follow Wald's training strategy for simulated experiments. The original images are filtered by a 9 × 9 Gaussian smoothing kernel and downsampled by a factor of 4. The Lr-HSI, therefore, can be treated as the ground-truth Hr-MSI. The original HSIs from the four other databases are used as the ground-truth images. We downsample the Hr-HSIs by averaging the 32×32 disjoint spatial blocks to generate the Lr-HSIs. The Hr-MSI (RGB image) of the same scene are stimulated by down-sampling X with a spectral model using a spectral dowmsampling matrix derived from the response of a Nikon D700 camera. The CAVE, Harvard and ICVL datasets are split into a training set of 20 images, 30 images, and 75 images and a test set of 12 images, 20 images and 25 images, respectively.

To prepare the training samples, we extract 96×96 overlapped patches from the training images as reference Hr-HSI images. The Hr-HSI, Hr-MSI and Lr-HSI images are sized of 96×96×S, 96×96×3 and 3×3×S, respectively, where S refers to the number of spectral bands in each experimental datasets. We use a fixed weighting factor γ within [1e-3, 1e-2] to balance the supervised loss and the unsupervised loss. When γ is too small, i.e. 1e-4, the problem (3) reduce to solving one single-task learning problem. On the other hand, when γ is too large, i.e. 1e-1, the auxiliary task can prevent the primary task from reconstructing the details.

B. Quantitative Metrics

Four quantitative picture quality indices (PQI) are utilized for performance evaluation, which include the root-mean-square error (RMSE), structural similarity (SSIM) [53], spectral angle mapping (SAM) [54] and the relative dimensionless global error in synthesis (ERGAS) [55].

The RMSE between the reconstructed and the original HSIs is defined as the average RMSE of all bands, e.g.,

$$\mathbf{RMSE}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{S}\sum_{i=1}^{S} RMSE(X^i, \hat{X}^i) \tag{11}$$

where X^i and $\hat{X}^i$ denote the *i*th band images of the ground-truth $\mathbf{X} \in \mathbb{R}^{N_W \times N_H \times S}$ and the estimated Hr-HSI $\hat{\mathbf{X}} \in \mathbb{R}^{N_W \times N_H \times S}$, respectively, and $RMSE(X^i, \hat{X}^i) = \sqrt{\frac{\sum_{j=1}^{N} \| X_j^i - \hat{X}_j^i \|_2}{N}}$ where $N = N_H \times N_W$. The RMSE is commonly used to compare the difference between two images by computing the variation in pixel values. The reconstructed image is close to the reference image when the RMSE value is near zero.

The structure similarity index measure is defined as the average value of all bands, i.e.,

$$\mathbf{SSIM}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{S}\sum_{i=1}^{S} SSIM(X^i, \hat{X}^i) \tag{12}$$

where $SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$, μ and σ are the mean intensity and the standard deviation, C_1, C_2 are two constants. The SSIM is used to compare the local patterns of pixel intensities between the two compared images and its values is range between 0 to 1. The value 1 indicates the reference and reconstructed images are identical.

The spectral angle mapper (SAM) is defined as an angle between the estimated pixel $\hat{x}_j$ and the ground truth pixel x_j over the whole image:

$$\mathbf{SAM}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{N}\sum_{j=1}^{N} arcos \frac{\hat{x}_j^T x_j}{\| \hat{x}_j \|_2 \| x_j \|_2} \tag{13}$$

The SAM is performed on a pixel-by-pixel base. A value of SAM equal to zero indicates no spectral distortion.

Finally, the ERGAS is defined as:

$$\mathbf{ERGAS}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{100}{d}\sqrt{\frac{1}{S}\sum_{i=1}^{S} \frac{MSE(\hat{X}^i, X^i)}{\mu_{\hat{X}^i}^2}} \tag{14}$$

where $\mu_{\hat{X}^i}$is the mean of $\hat{X}^i$ and $MSE(\hat{X}^i, X^i)$ is the mean squared error between $\hat{X}^i$ and X^i, *d* is a spatial downsampling factor. The ERGAS is used to determine the image's quality in terms of the normalised average error of each band. Increased ERGAS indicates that the reconstructed image is distorted, whereas decreased ERGAS means that the reconstructed image is more similar to the reference image.

C. Training details

For hardware and software settings, all experiments are implemented on the TensorFlow with CUDA 9.0 and cuDNN back-ends with a GPU of NVIDIA GeForce GT 1030. We trained the model with 40,000 iterations with a batch size of 16. The ADAM optimization [56] algorithm was used with an initial learning rate of 0.00035, which reduces by 30% after every 10,000 iterations. Only the flipping was used to augment the data. Additional Gaussian noise added to the original inputs is zero-mean with a variance within [0.05, 0.2].

D. Experimental Results

We set up our MSAT with α = 0 and β = 1 for small training dataset of CAVE and Harvard, and α = 0 and β = 2 for ICVL. Since our method needs training, we compare the performance on the testing set instead of the full dataset. The comparion methods include: non-local sparse tensor factorization (NLSTF) [1] [19], non-negative structured sparse representation (NSSR) [2] [17], and low tensor-train rank

[1] https://github.com/renweidian/NLSTF

[2] https://see.xidian.edu.cn/faculty/wsdong/HSI_SR _Project.htm

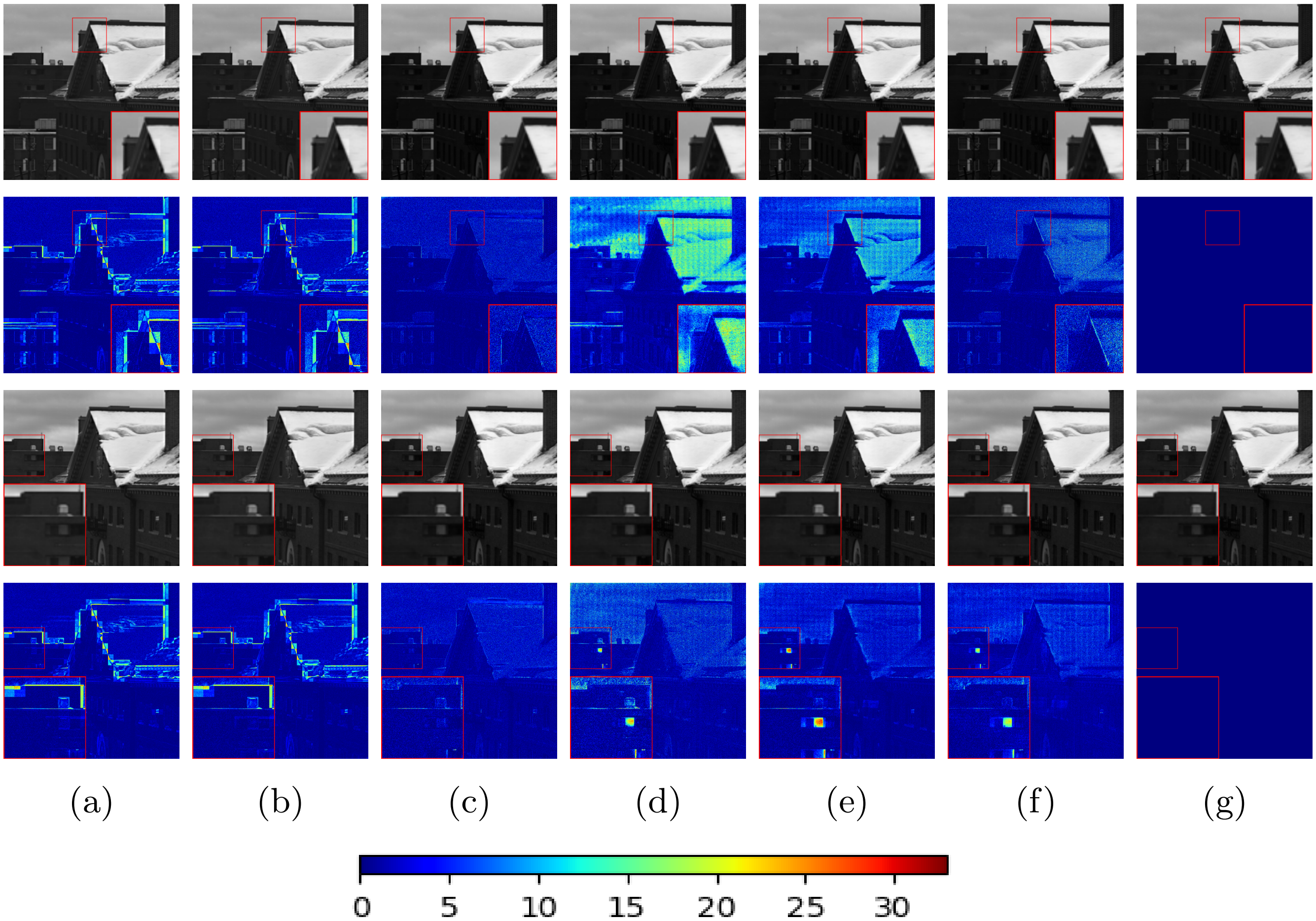


Fig. 4: First and second row: the reconstructed images and the corresponding error images of the compared methods for Harvard at 460nm band. Third row and Fourth row: reconstructed images and corresponding error images of the compared methods for Harvard at 620nm band. (a) the NLSTF method [19] (RMSE = 3.33, ERGAS = 0.19, SAM = 2.34, SSIM = 0.96). (b) the NSSR method [17] (RMSE = 3.36, ERGAS = 0.20, SAM = 2.51, SSIM = 0.96). (c) the LTTR method [22] (RMSE = 1.87, ERGAS = 0.161, SAM = 2.27, SSIM = 0.972). (d) the HSRnet method [41] (RMSE = 3.12, ERGAS = 0.193, SAM = 2.59, SSIM = 0.963). (e) the MoG-DCN method [42] (RMSE = 2.62, ERGAS = 0.189, SAM = 2.41, SSIM = 0.972). (f) Proposed MSAT (RMSE = 2.37, ERGAS = 0.173, SAM = 2.38, SSIM = 0.972). (g) Ground-truth.

representation (LTTR) [3] [22] methods, which represent the state-of-the-art sparse representation based approaches; the hyperspectral super resolution network (HSRnet) [4] [41] and the model-guided deep convolutional network (MoG-DCN) [5] [42] represent the state-of-the-art deep learning-based SR methods. Table I shows the average results of the compared methods on the CAVE testing set, where the best results are highlighted in bold for clarity. As seen, the proposed method achieves the better performance than all others in terms of ERGAS, SAM and SSIM, although the RMSE is not the least. With just a few samples used for training suggests that our model has the potential to further improve the RMSE scores when more training images are available.

TABLE I: Average quantitative results of the compared methods using 12 testing images on the CAVE dataset.

Method	RMSE↓	ERGAS ↓	SAM↓	SSIM↑
NLSTF [19]	3.136±1.236	0.461±0.301	6.567±2.413	0.976±0.012
NSSR [17]	2.772±1.294	0.417±0.314	5.697±2.019	0.980±0.011
LTTR [22]	**2.640**±1.590	0.377±0.259	6.238±2.248	0.982±0.010
HSRnet [41]	3.360±1.700	0.387±0.330	4.784±1.144	0.980±0.010
MoG-DCN [42]	3.330±1.676	0.370±0.277	4.573±0.986	0.984±0.007
MSAT	3.245±1.610	**0.361±0.254**	**4.245±0.929**	**0.985±0.005**

The quantitative averages on the Harvard database are compared in Table II. Although none of these methods can consistently outperform others, the LTTR [22] seems to perform better on the Harvard dataset. The proposed approach achieves the competitive results in terms of RMSE and SSIM, where the ERGAS and SAM are slightly worse than others. Fig. 4 shows a reconstructed image from the Harvard test dataset. As

[3] https://github.com/renweidian/LTTR

[4] https://github.com/liangjiandeng/HSRnet

[5] https://github.com/chengerr/Model-Guided-Deep-Hyperspectral-Image-Super-resolution

TABLE II: Average quantitative results of the compared methods over 20 testing images on the Harvard dataset.

Method	RMSE↓	ERGAS ↓	SAM↓	SSIM↑
NLSTF [19]	2.658±1.303	**0.314±0.206**	3.362±1.720	0.974±0.014
NSSR [17]	2.520±1.237	0.340±0.217	3.228±1.596	0.975±0.014
LTTR [22]	2.187±1.147	0.340±0.213	**3.093±1.292**	**0.979**±0.011
HSRnet [41]	2.622±1.333	0.361±0.263	3.461±1.167	0.973±0.016
MoG-DCN [42]	2.208±1.126	0.349±0.238	3.393±1.497	0.979±0.011
MSAT	**2.184±1.053**	0.353±0.228	3.392±1.490	**0.979±0.010**

TABLE III: Average results of the compared methods (25 testing images, 75 training images).

Method	RMSE↓	ERGAS ↓	SAM↓	SSIM↑
NLSTF [19]	1.725±0.632	0.122±0.047	1.062±0.373	0.991±0.003
NSSR [17]	1.735±0.603	0.128±0.047	1.048±0.352	0.991±0.003
LTTR [22]	1.125±0.389	0.079±0.040	0.997±0.316	0.994±0.001
HSRnet [41]	1.652±0.563	0.109±0.043	1.092±0.361	0.996±0.001
MoG-DCN [42]	1.236±0.382	0.079±0.042	1.032±0.340	**0.998**±0.002
MSAT	**1.034±0.322**	**0.065±0.035**	**0.990±0.306**	**0.998±0.0005**

the NLSTF [19] method is actually a variation of the NSSR [17] algorithm, visual inspection validates that the former closely resembled patterns in the latter. The reconstructed images from three deep learning-based methods also follow the closely mirrored patterns. Among them, the LTTR [22] and the proposed MSAT recover more spatial details of the HSI.

Obviously, deep learning-based methods require sufficient features by a grant from a larger amount of training data or properties of the datasets. As a result, small training dataset, as well as the high training/test split ratio from CAVE (20 images/12 images ≈ 62.5/37.5%) or Harvard (30 images/20 images ≈ 60/40%), will cause high variance in training of model or overfitting. Another issue is unrepresentative training dataset, which means that the data available during training is insufficient to capture the model, relative to the validation dataset. Without increasing the model complexity, we randomly choose 100 images from the ICVL dataset where 75 images are used for training and remaining 25 for testing. The performance of our method now consistently outperforms the compared methods significantly with a more considerable margin, as shown in Fig. 5 and Table III. As seen from Table III, the proposed MSAT method significantly outperforms the compared models of NLSTF [19], NSSR [17], LTTR [22], HSRnet [41], and MoG-DCN [42] in terms of all the four quantitative metrics. Furthermore, our model produced

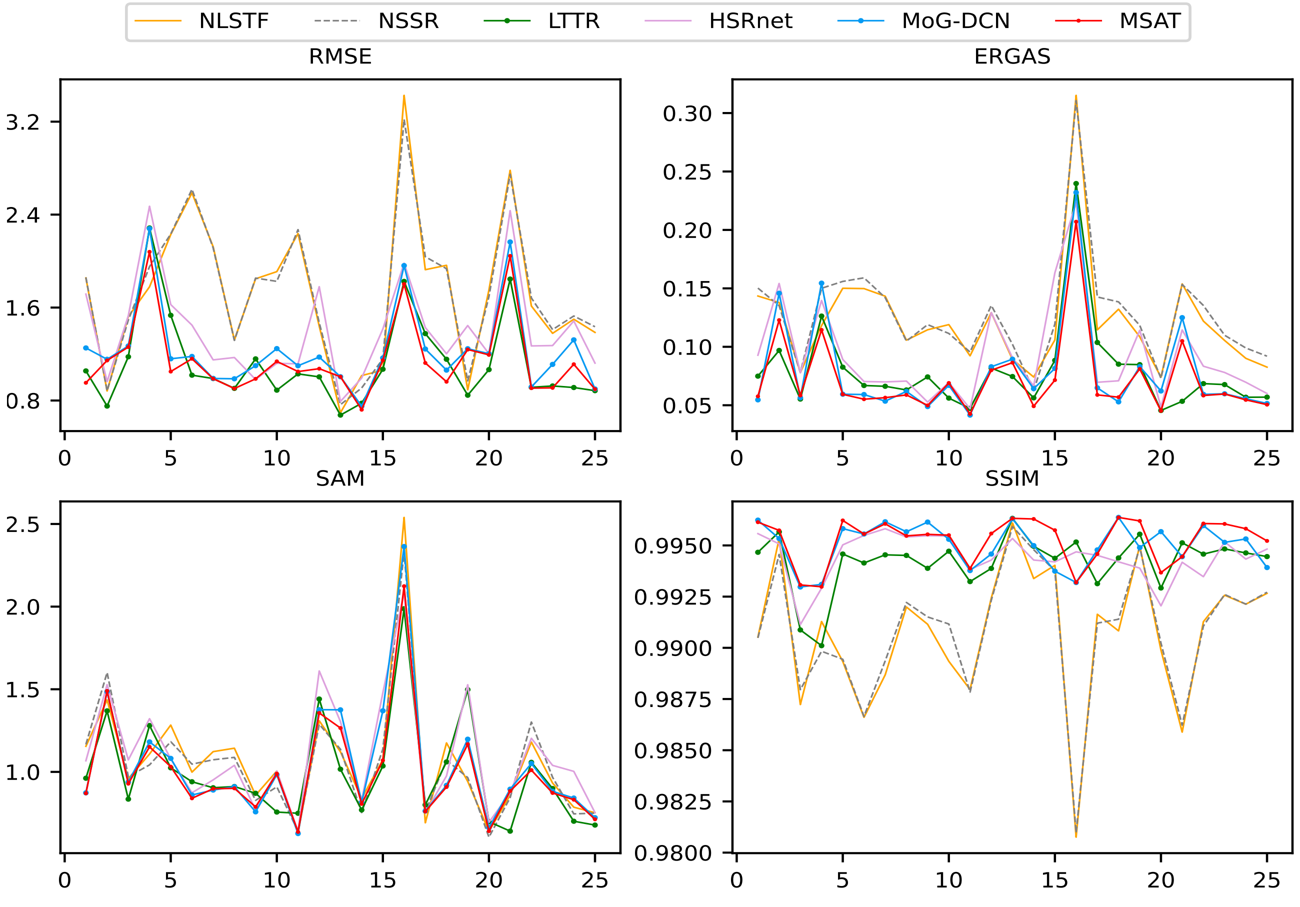


Fig. 5: The comparison of our proposed CNN-based method vs. five approaches on the testing set of 25 images from the ICVL dataset.

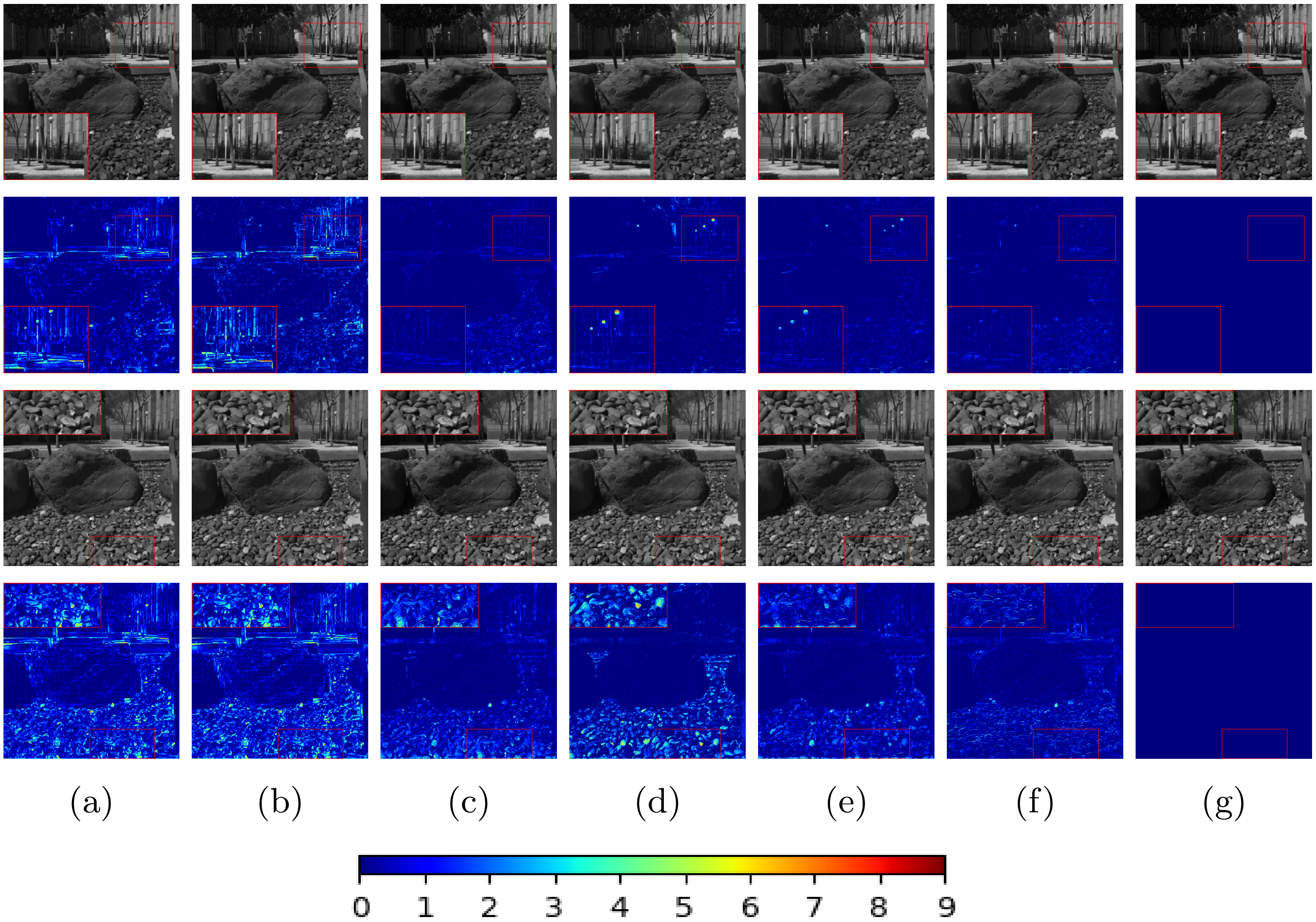


Fig. 6: The reconstructed images and corresponding error images of the compared methods for ICVL at 460nm band (first two rows) and at 620 nm (the last two rows). (a) the NLSTF method [19] (RMSE = 1.96, ERGAS = 0.13, SAM = 1.17, SSIM = 0.99). (b) the NSSR method [17] (RMSE = 1.93, ERGAS = 0.13, SAM = 1.07, SSIM = 0.99). (c) the LTTR method [22] (RMSE = 1.15, ERGAS = 0.085, SAM = 1.06, SSIM = 0.994). (d) the HSRnet method [41] (RMSE = 1.36, ERGAS = 0.091, SAM = 1.07, SSIM = 0.994). (e) the MoG-DCN method [42] (RMSE = 1.13, ERGAS = 0.067, SAM = 0.098, SSIM = 0.995). (f) Proposed MSAT (RMSE = 0.96, ERGAS = 0.05, SAM = 0.90, SSIM = 0.996). (g) Ground-truth.

consistently lower variance around the average score than all others. In Fig. 6 and Fig. 7, we also show the reconstructed images and the error images, where the test results are for an outdoor image *BGU_0403-1419-1* and an indoor image *objects _0924-1629* from the ICVL dataset. The NLSTF [19] and NSSR [17] again perform worse as shown in the changed brightness while the LTTR [22] and the proposed MSAT approaches perform better regarding the well preserved spatial and spectral structures. The HSRnet [41] and the MoG-DCN [42] are still unable to surpasses the LTTR [22] in ICVL dataset.

Table IV compares the quantitative average of all compared methods using 16 testing images on the Chikusei dataset. As the training and test samples are cropped from the same image, they have common features and do not suffer from overfitting and unrepresentative training dataset. Fig. 8 shows the composition of test samples with bands of 70, 100, and 36 as a false-color image with the error image given in all three channels. As seen, the three sparse representation-based approaches perform worse compare to deep learning-based methods. The proposed method significantly outperforms three sparse representation-based methods with a large margin while still perform better than the HSRnet [41] and the MoG-DCN [42]. The composition image obtain from the proposed method is closet to the ground-truth, while other methods show obvious unsatisfactory reconstruction.

TABLE IV: Average results of the compared methods over 16 testing samples in the Chikusei dataset.

Method	**RMSE↓**	**ERGAS ↓**	**SAM↓**	**SSIM↑**
NLSTF [19]	2.553±0.673	0.478±0.056	2.776±0.661	0.971±0.007
NSSR [17]	3.944±1.114	0.772±0.112	3.895±0.923	0.943±0.015
LTTR [22]	4.532±1.324	0.683±0.121	3.111±0.525	0.952±0.013
HSRnet [41]	2.324±0.435	0.827±0.163	2.938±0.486	0.970±0.005
MoG-DCN [42]	1.355±0.259	0.483±0.060	2.642±0.316	0.989±0.001
MSAT	**0.934±0.116**	**0.475±0.053**	**2.090±0.303**	**0.992±0.001**

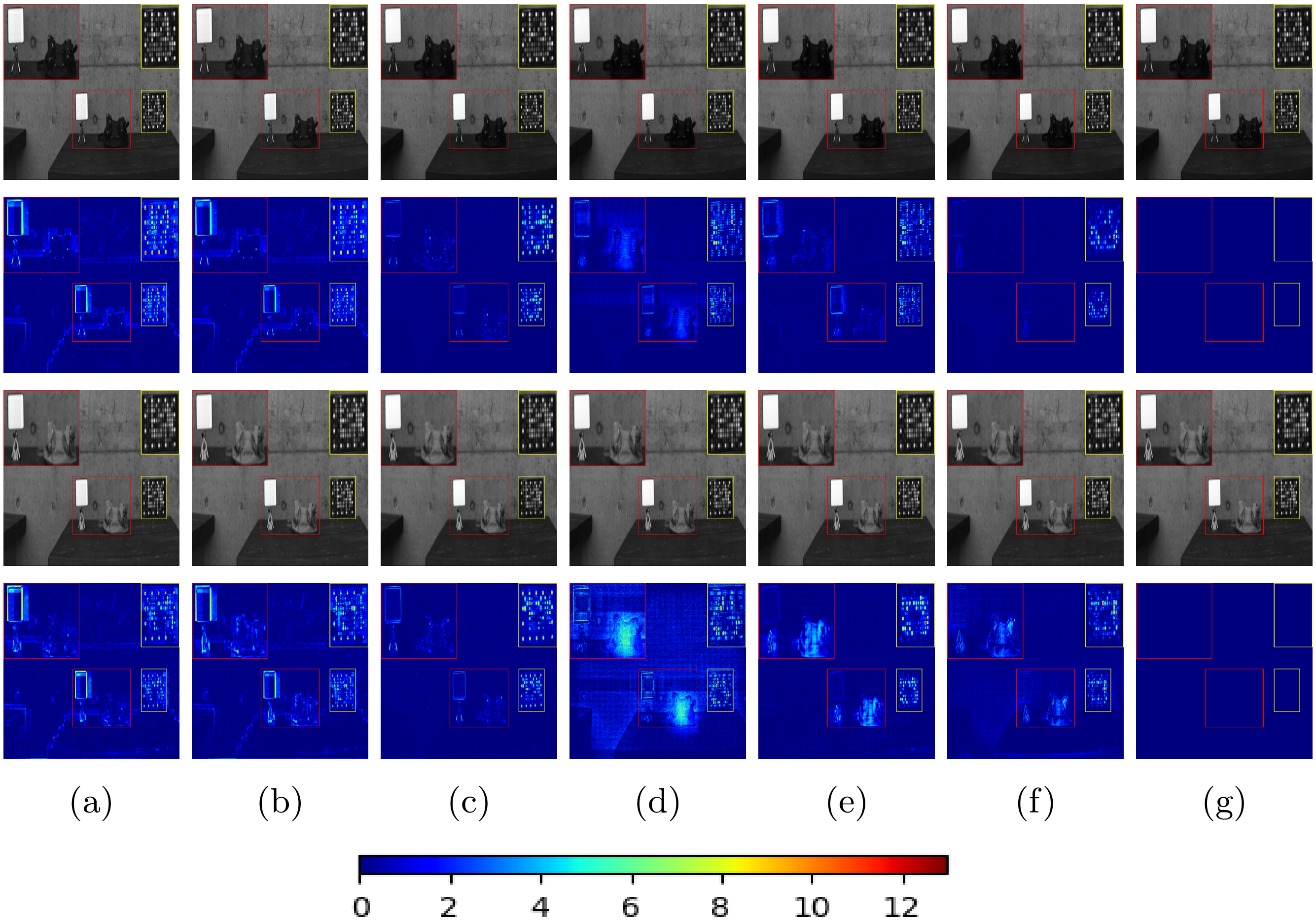


Fig. 7: The reconstructed images and corresponding error images of the compared methods for ICVL at 540nm band (first two rows) and at 620 nm (the last two rows). (a) the NLSTF method [19] (RMSE = 1.75, ERGAS = 0.07, SAM = 0.64, SSIM = 0.98). (b) the NSSR method [17] (RMSE = 1.69, ERGAS = 0.07, SAM = 0.60, SSIM = 0.99). (c) the LTTR method [22] (RMSE = 1.26, ERGAS = 0.548, SAM = 0.69, SSIM = 0.992). (d) the HSRnet method [41] (RMSE = 1.48, ERGAS = 0.067, SAM = 0.62, SSIM = 0.990). (e) the MoG-DCN method [42] (RMSE = 1.22, ERGAS = 0.534, SAM = 0.68, SSIM = 0.993). (f) Proposed MSAT (RMSE = 1.19, ERGAS = 0.04, SAM = 0.64, SSIM = 0.993). (g) Ground-truth.

The fusion result on real spaceborne HS dataset is shown in Fig. 9. As the ground-truth Hr-HSIs are unavailable, we follow the procedure of training and measure the performance by comparing the result image with upsampled image of Lr-HSI. As seen, the result image obtained from our proposed method is much closer to Lr-HSI and Hr-MSI. Furthermore, Fig. 10 compares the performance of three deep learning-based HS/MS fusion methods over the validation set. The HSRnet [41] performs worst among the three methods, while the MoG-DCN [42] cannot outperform our smaller-size baseline model. An introduced auxiliary task provides a consistent gain in generality and achieves the best performance.

E. The effectiveness of multi-scale image decomposition and auxiliary task

We performed an ablation study to verify the effect of Hr-MSI decomposition and the proposed auxiliary task used in training on the CAVE dataset, where the L2 regularization is turned off for a fair comparison in these evaluations. We denote *w/o 3×3* as the case without Hr-MSI decomposition to the spatial size of 3×3 whilst keeping other settings the same. We observed that the more scales the Hr-MSI is decomposed, the better performance it delivers. As shown in Fig. 11, the lowest reconstruction loss in both the training and validation is achieved when the Hr-MSI is decomposed into the maximum scales of five, of which the final scale has a spatial size equal to the that of the Lr-HSI. Reducing one level of decomposition may result in performance degradation. Each smaller scale of the image contains features to approximate the original image, and the early applying of the joint-training can help to refine information in a coarse-to-fine manner. Although the Lr-HSI does not decompose further from the size of 3×3, the results shown in Fig. 11 suggest that joint learning from smallest levels would reduce the reconstruction error. Finally, the combination of both five-level decompositions and an unsupervised loss induced by the auxiliary task has significantly outperformed all others after about only 10 epochs during

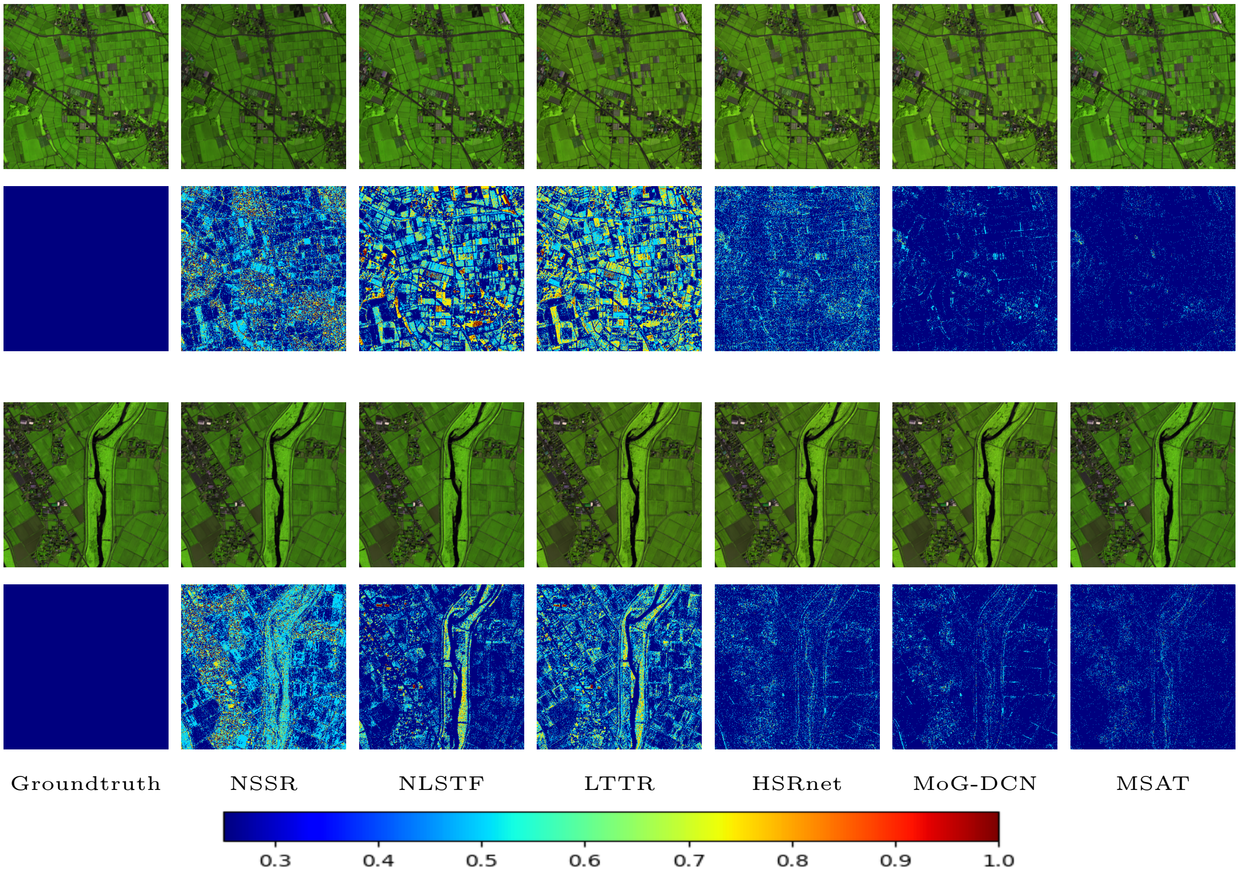


Fig. 8: The HSI-SR results on the Chikusei dataset of all competing methods. First and Fourth row: the false-color image with bands (70, 100, 36). Second and Fifth row: the corresponding error images compared to the ground-truth.

the training or about 5 epochs during the validation. The turbulences at 8 and 15 epochs indicate the outlier of the unsupervised features from the auxiliary task. Although they do not degrade the final performance, reducing noise level in the auxiliary task or global learning rate can avoid these spikes.

Table V shows the testing results with and without the auxiliary task on the ICVL dataset. As seen, the introduced auxiliary task does improve the overall performance in both shallow and deeper networks. The accuracy, however, does not improve further while increasing the number of the residual blocks. One possible reason here is that our lightweight model can sufficiently fit with the 75 training images, thus increasing the depth of the model can not produce further improvement.

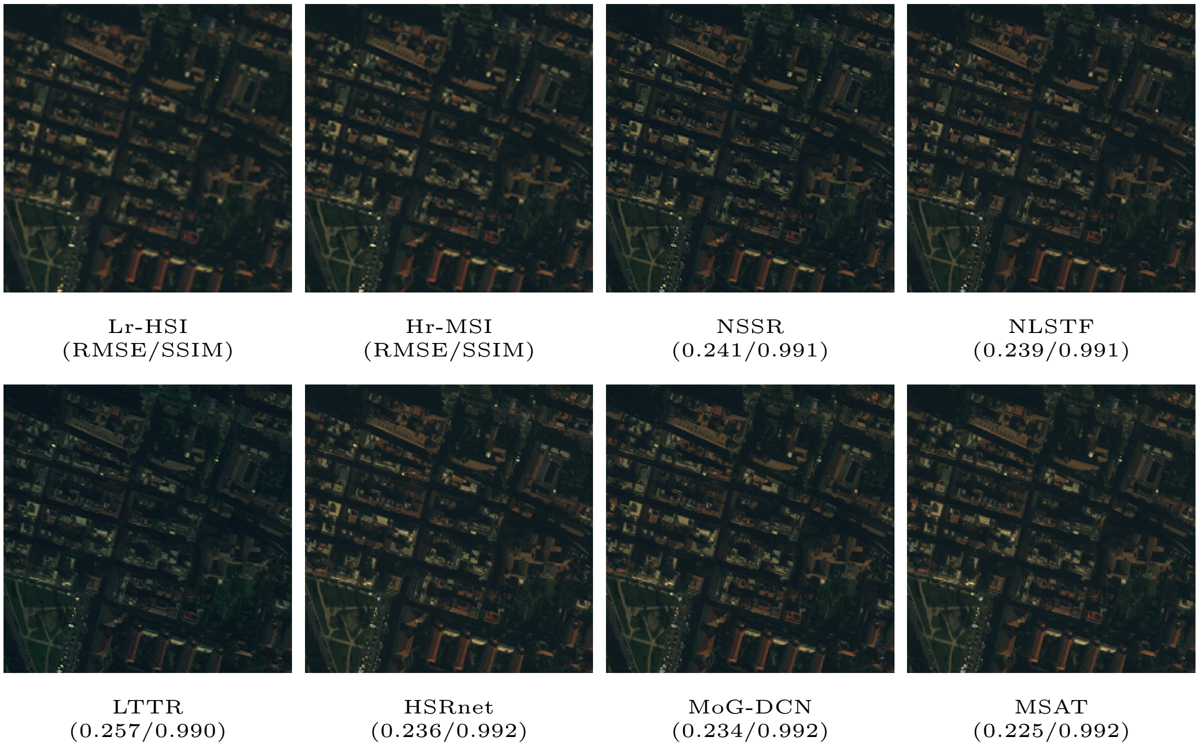
TABLE V: Average performance of the Baseline network (without the proposed auxiliary task) and MSAT (with the auxiliary task) over testing images of the ICVL dataset.

Method	RMSE↓	ERGAS ↓	SAM↓	SSIM↑
Baseline ($\beta = 1$)	1.368±0.450	0.086±0.043	1.043±0.327	0.994±0.0012
MSAT ($\beta = 1$)	**1.154±0.336**	**0.072±0.035**	**0.998±0.314**	**0.995±0.0010**
Baseline ($\beta = 2$)	1.258±0.328	0.079±0.038	1.041±0.347	**0.998±0.0005**
MSAT ($\beta = 2$)	**1.034±0.322**	**0.065±0.035**	**0.990±0.306**	**0.998±0.0005**

To further demonstrate the effectiveness of multi-scale reconstruction, we include comparisons with other CNN-based methods, such as SRCNN [57] and VDSR [58], where pre-upsampling is used. The SRCNN [57] model has only 3 simple convolutional layers while the VDSR [58] contains 20 convolutional layers. In addition, we re-conduct experiments with a more powerful architecture based on the ResNet, namely HSI-ResNet, with the same configurations as the ResNet including the number of blocks, optimization method of network training, epoch number, training and testing samples etc. The HSI-ResNet does not fuse Lr-HSI and Hr-MSI at multi-stages as we have done in the proposed MSAT model. As shown in Fig. 12, the Lr-HSI is spatially upsampled before concatenated with the Hr-MSI. The CNN network consists of five residual blocks, which has a similar depth as our model. Table VI illustrates that progressive fusion at multiple stages has obviously the advantage over a single-stage fusion.

F. Tuning the noise level in denoising autoencoder

We trained several denoising autoencoders with different noise levels to understand the qualitative effect of the noise through different datasets. The variation of RMSE, ERGAS, SAM, and SSIM value when varying the noise levels from



Lr-HSI (RMSE/SSIM) | Hr-MSI (RMSE/SSIM) | NSSR (0.241/0.991) | NLSTF (0.239/0.991)

LTTR (0.257/0.990) | HSRnet (0.236/0.992) | MoG-DCN (0.234/0.992) | MSAT (0.225/0.992)

Fig. 9: The Hr-MSI (RGB) and Lr-HSI images are of the left bottom area of *Roman Colosseum* acquired by World View-2. The composite image of the HS image with bands 5-3-2 as R-G-B is displayed.

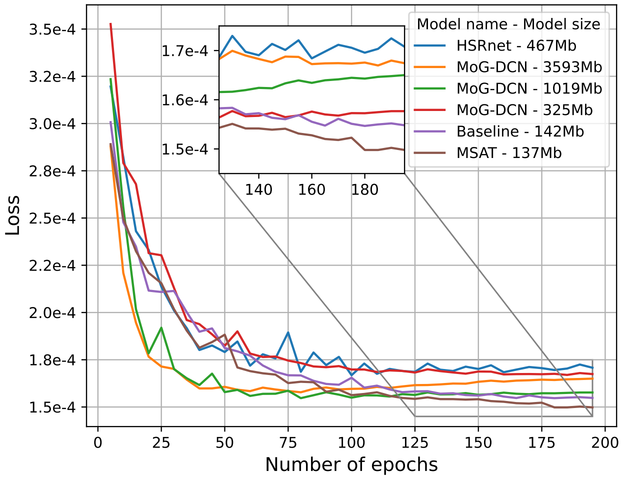


Fig. 10: Comparison of the proposed MSAT to two deep learning-based methods (HSRnet [41] and MoG-DCN [42]) over the validation set in the *Roman Colosseum* dataset.

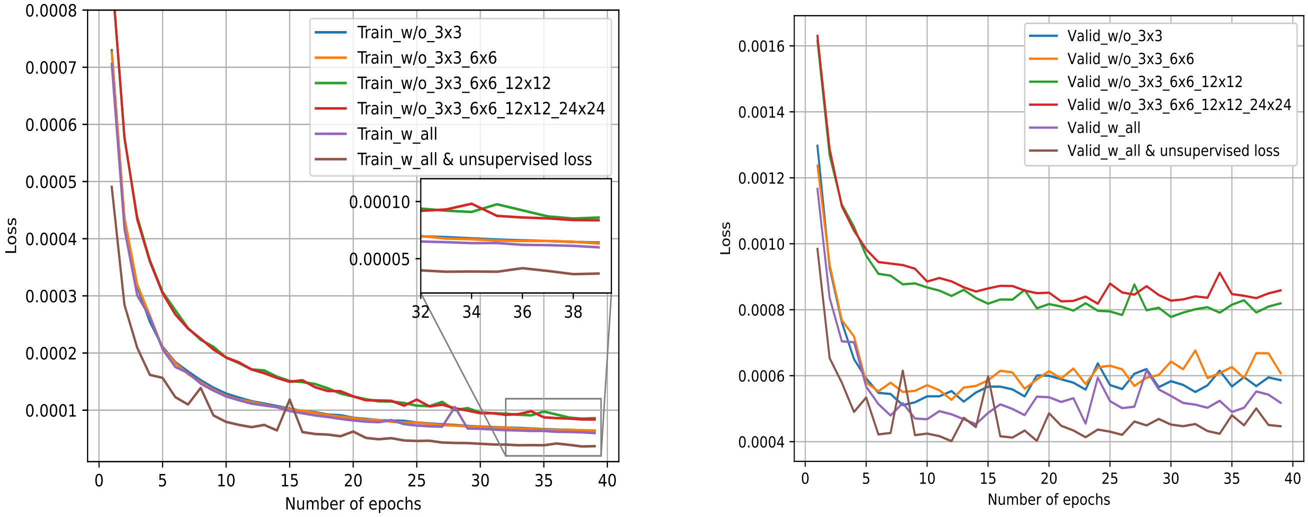
TABLE VI: Quantitative results on CAVE dataset. Our baseline indicate that our model do not include auxiliary task.

Method	**RMSE↓**	**ERGAS ↓**	**SAM↓**	**SSIM↑**
SRCNN [57]	4.320±2.216	0.543±0.407	6.165±1.466	0.961±0.018
VDSR [58]	4.135±2.151	0.492±0.372	5.983±1.360	0.970±0.014
HSI-ResNet	3.960±1.873	0.435±0.308	5.328±1.196	0.977±0.007
Our Baseline	**3.454**±**1.692**	**0.384**±**0.252**	**4.810**±**1.066**	**0.981**±**0.006**

0.0 to 0.3 for CAVE, Harvard, ICVL, Chikusei, and Roman Colosseum datasets are shown in Fig. 13. As can be seen from the figure, as the value of noise increases, the performance metric also begins to improve, then plateau, and then degrade for all datasets. The appropriate noise levels were discovered to be dependent on the quality of collected images as well as the number of training samples. Adding a large amount of noise to noisy images could degrade the performance. The images in the CAVE dataset, for example, are clean and contain fewer noises than those in the Harvard and the ICVL datasets. Therefore, applying a large noise level ($\sigma = 0.2$) leads to improve performance for the CAVE dataset, while increasing errors for the Harvard and the ICVL. As the training set for the Chikusei and Roman Colosseum datasets is limited, only the top part of an image is used, the smaller noise level of 0.05 is the most appropriate.

G. Robustness to noise

In practice, noise from various aspects can corrupt Lr-HSIs and Hr-MSIs even during image acquisition, transmission, and compression. To test the noise robustness of all compared methods, we add the Gaussian noise to the Lr-HSI and Hr-MSI inputs and then fuse them to produce a HR-HSI. The SNRs of the noisy Lr-HSI and Hr-MSI are set to 20dB and 25dB, respectively. The quality metric values in the noisy case are shown in Table VII and visually compared with those noise-free ones (as referred to Table I) in Fig. 14. As seen, the performance of NLSTF [19], NSSR [17], and



(a) The training loss

(b) The validation loss

Fig. 11: The training and validation loss of model with different level of decomposition and with/without unsupervised loss.

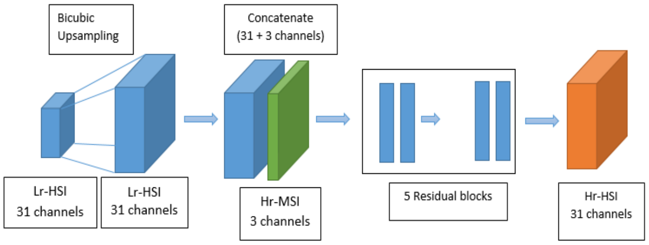


Fig. 12: HSI ResNet model.

LTTR [22] methods drops faster than three deep learning-based methods in all four metrics and degenerates sharply in the SAM measure. The RMSE of the LTTR [22] increases from 2.640 ± 1.590 to 4.064 ± 1.913 by $53.9\% \pm 20.3\%$ while our proposed approach is more robust, increasing only from 3.245 ± 1.610 to 4.282 ± 1.712 or by $31.9\% \pm 6.1\%$. The architecture of the MoG-DCN contains autoencoders that are robust to noise. The RMSE of the MoG-DCN [42] increases from 3.330 ± 1.676 to 4.390 ± 788 by $31.9\% \pm 6.6\%$.

TABLE VII: Quantitative results of a noisy case on the CAVE dataset.

Method	RMSE↓	ERGAS ↓	SAM↓	SSIM↑
NLSTF [19]	4.806 ±1.873	0.695 ±0.275	20.065±7.070	0.851±0.047
NSSR [17]	4.900±**1.642**	0.714±0.329	19.062±7.092	0.850±0.053
LTTR [22]	**4.064**±1.913	0.577±0.259	15.621±5.811	0.902±0.051
HSRnet [41]	4.582±1.845	0.577±0.350	10.078±3.929	0.894±0.058
MoG-DCN [42]	4.390±1.788	0.546±0.274	9.528±3.253	**0.902**±0.056
MSAT	4.282±1.712	**0.490±0.254**	**9.436±3.085**	**0.902±0.050**

H. Feature map

Differing from RGB images, HSIs have the characteristics of high spectral resolution across many narrow bands. Therefore, it is not straightforward to interpret the meaningful feature maps at lower layers, which typically display features in a spatial manner. To visualise the features learn from our CNN-based network, we select one testing image from the CAVE dataset, running on a forward path to show the learnt feature maps from the fifth (top) block in Fig. 15. It is worth noting that the transposed convolutions are used when up-sampling the input feature map at each stage. This is a well-known operation which may introduce severe checkerboard artifacts and tend to be most prominent with a higher up-sampling scale factor [59]. The checkerboard pattern can be observed in the feature maps of Fig. 11b and Fig. **??**, where they have shown that the feature maps extract from the model without the unsupervised loss will suffer more from horizontal and vertical stripes in the final prediction. By contrast, the feature maps from the model with the proposed additional unsupervised loss can successfully suppress such artifacts.

V. Conclusion

In this paper, we have presented an effective CNN-based method for fusing the observed Lr-HSIs and Hr-MSIs to reconstruct the HSI-SR. By decomposing the Hr-MSIs into multiple spatial scales, the discrepancy between the observed Lr-HSI and Hr-MSI is facilitated, which allows our model to be able to gradually learn the high-resolution features from Lr-HSIs and spatial-reduced feature from Hr-MSIs. In addition, we integrating with a proposed auxiliary task in the training procedure can help to improve the generalization capability of the CNN models. The proposed auxiliary task does not only regulate the model from overfitting, but also redirect the main

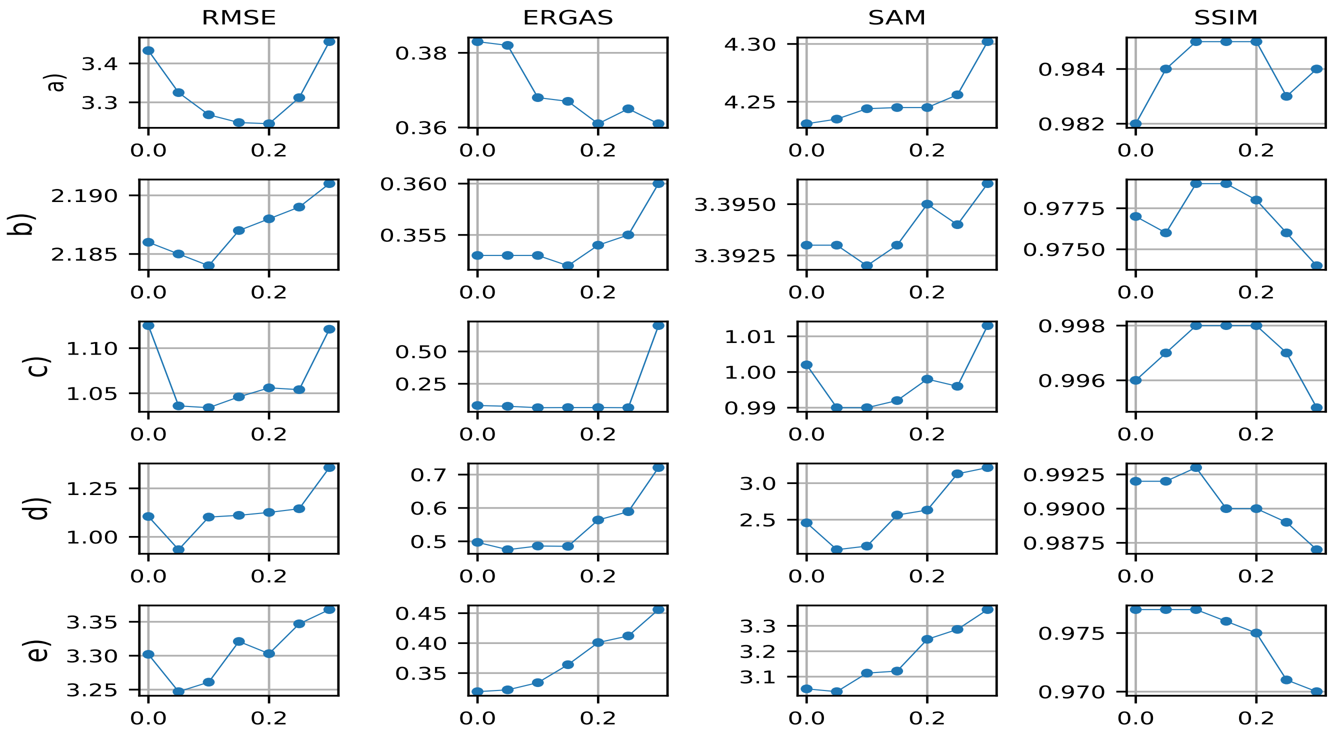


Fig. 13: The variation of RMSE, ERGAS, SAM, and SSIM with the noise levels σ in our denoising autoencoder for five datasets. (a) the CAVE. (b) the Harvard. (c) the ICVL. (d) the Chikusei. (e) the Roman Colosseum. We select σ = 0.2 for the CAVE dataset, σ = 0.1 for both Harvard and the ICVL datasets, σ = 0.05 for both Chikusei and the Roman Colosseum, respectively.

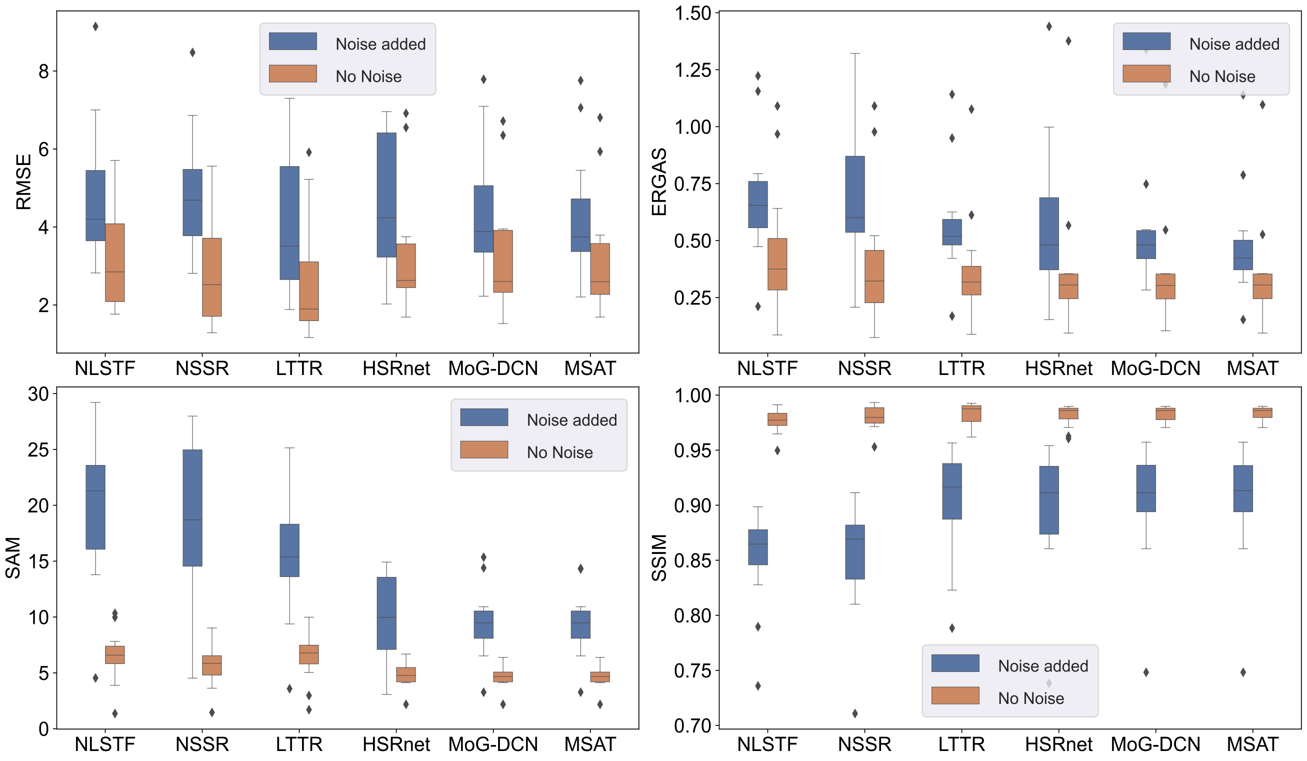
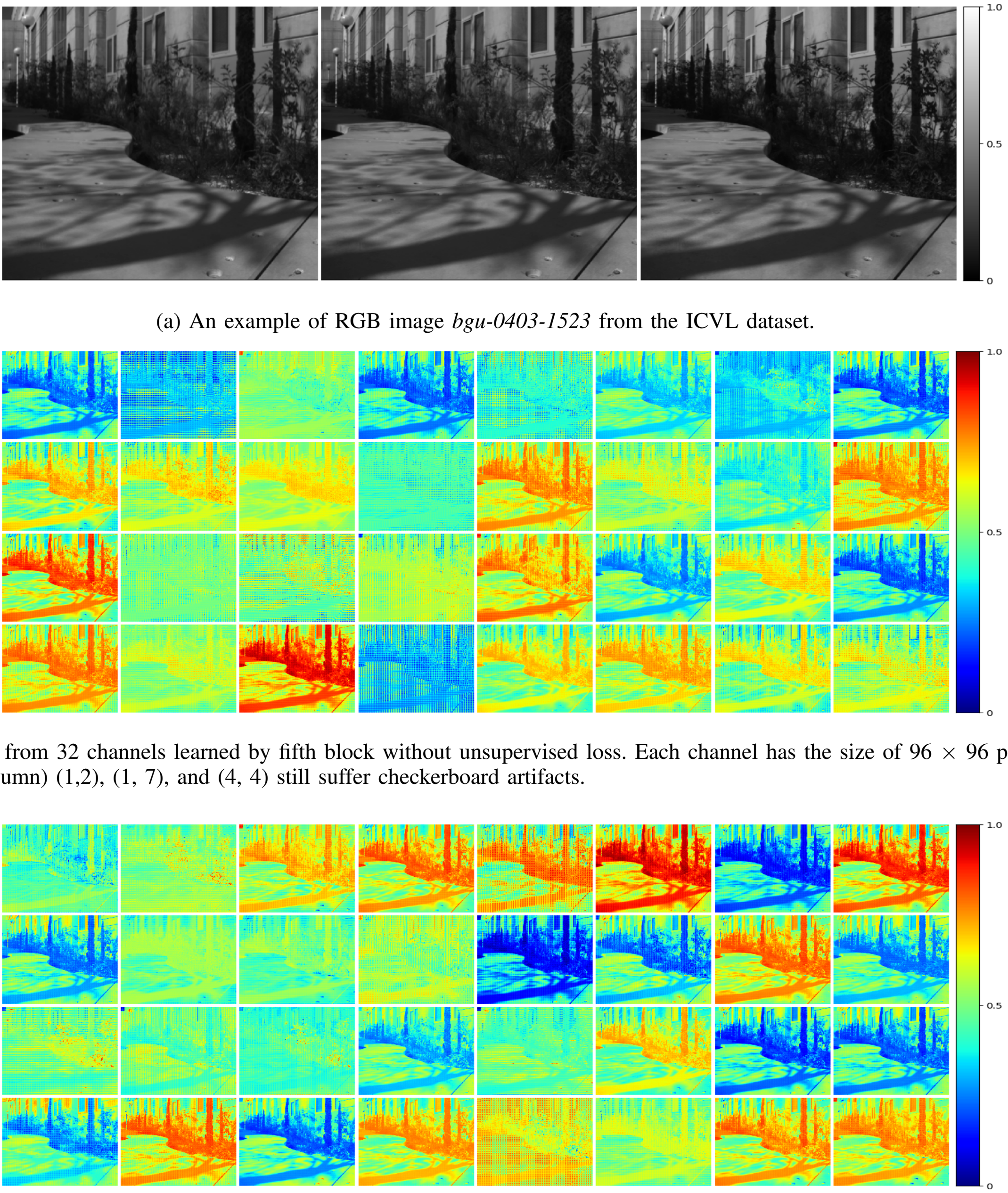


Fig. 14: Quantitative result of noisy cases on CAVE testing set.



(a) An example of RGB image *bgu-0403-1523* from the ICVL dataset.

(b) Feature maps from 32 channels learned by fifth block without unsupervised loss. Each channel has the size of 96 × 96 pixels. Feature maps at (row, column) (1,2), (1, 7), and (4, 4) still suffer checkerboard artifacts.

(c) Feature maps from 32 channels learned by fifth block with unsupervised loss. Each channel has the size of 96 × 96 pixels. Only feature map at (row, column) (4, 5) has a checkerboard artifact.

Fig. 15: Visualization of feature maps learned by the fifth block of our reconstruction network: (a) 3 channels of the observed RGB image; (b) Without using the proposed unsupervised auxiliary loss. (c) With the unsupervised auxiliary loss.

task to learn representations with sped-up and better model convergence. Furthermore, the auxiliary task deriving from the denoising autoencoder is more resistant to noise than all compared methods. Our testing results on five public datasets have demonstrated that the proposed method can provide improvements over the-state-of-the-art methods in term of both objective assessment and subjective visual quality. In future research, we plan to study an adaptive balance between the primary task and the proposed auxiliary task though training. In addition, a natural progression of this work is to investigate other auxiliary tasks for improving the performance of primary task.

Acknowledgment

The authors would greatly thank colleagues from the Image Processing Group in Strathclyde University for their valuable suggestions.

References

[1] Zhong, Z., Li, J., Clausi, D.A. and Wong, A., “Generative adversarial networks and conditional random fields for hyperspectral image classification,” *IEEE Trans. Cybern.*, vol. 50, pp. 3318-3329, 2019.

[2] H. Li, G. Xiao, T. Xia, Y. Y. Tang, and L. Li, “Hyperspectral image classification using functional data analysis,” *IEEE Trans. Cybern.*, vol. 44, no. 9, pp. 1544-1555, Sept, 2014.

[3] Y. Zhou, and Y. Wei, “Learning hierarchical spectral–spatial features for hyperspectral image classification,” *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1667-1678, Jul, 2016.

[4] J. Zabalza, J. Ren, Z. Wang, S. Marshall and J. Wang, “Singular spectrum analysis for effective feature extraction in hyperspectral imaging,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1886-1890, Nov, 2014.

[5] C. Zhao, X. Li, J. Ren and S. Marshall, “Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery”, *Int J Remote Sens.*, vol. 34 no. 24, pp. 8669-8684, Dec, 2013.

[6] J. Tschannerl, J. Ren, H. Zhao, F. Kao, S. Marshall and P. Yuen, “Hyperspectral image reconstruction using Multi-colour and Time-multiplexed LED illumination,” *Opt. Lasers. Eng*, vol. 121, pp. 352-357, Oct, 2019.

[7] Dian, R., Li, S., Fang, L., Lu, T. and Bioucas-Dias, J.M., “Nonlocal sparse tensor factorization for semiblind hyperspectral and multispectral image fusion,” *IEEE Trans. Cybern.*, vol. 50, no. 10, pp.4469-4480, Oct, 2019.

[8] Y. Chen, W. He, N. Yokoya, and T-Z. Huang, “Hyperspectral image restoration using weighted group sparsity-regularized low-rank tensor decomposition,” *IEEE Trans. Cybern.*, vol. 50, no. 8, pp.3556-3570, 2019.

[9] J. Zabalza, J. Ren, J. Zheng, H. Zhao, C. Qing, Z. Yang, P. Du and S. Marshall, “Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging”, *Neurocomputing*, vol. 185, pp. 1-10, 2016.

[10] J. Tschannerl, J. Ren, F. Jack, J. Krause, H. Zhao, W. Huang and S. Marshall, “Potential of UV and SWIR hyperspectral imaging for determination of levels of phenolic flavour compounds in peated barley malt,” *Food Chem.*, vol. 270, pp. 105-112, Jan, 2019.

[11] X. Lu, Y. Yuan, and X. Zheng, “Joint dictionary learning for multispectral change detection,” *IEEE Trans. Cybern.*, vol. 47, no. 4, pp. 884-897, Apr, 2017.

[12] Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X. and Paisley, J, ”PanNet: A deep network architecture for pan-sharpening,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5449-5457.

[13] Fu, X., Lin, Z., Huang, Y. and Ding, X., ”A variational pan-sharpening with local gradient constraints,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10265-10274.

[14] Fu, X., Wang, W., Huang, Y., Ding, X. and Paisley, J., ”Deep Multiscale Detail Networks for Multiband Spectral Image Sharpening,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2090-2104, 2020.

[15] R. Kawakami, J. Wright, Y. Tai, Y. Matsushita, M. Ben-Ezra, and K. Ikeuchi, “High-resolution hyperspectral imaging via matrix factorization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2329–2336.

[16] N. Akhtar, F. Shafait, and A. Mian, “Sparse spatio-spectral representation for hyperspectral image super-resolution,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 63–78.

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

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

[19] textcolorblueDian, R., Fang, L. and Li, S, “Hyperspectral image super-resolution via non-local sparse tensor factorization,” In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5344-5353.

[20] Zhang, K., Wang, M., Yang, S. and Jiao, L., “Spatial–spectral-graph-regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion,” *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 11, no. 4, pp.1030-1040, 2018.

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

[22] Dian, R., Li, S. and Fang, L., ”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, 2019.

[23] Xu, Y., Wu, Z., Chanussot, J. and Wei, Z., ”Nonlocal patch tensor sparse representation for hyperspectral image super-resolution,” *IEEE Trans. Image Process.*, vol. 28, no. 6, pp.3034-3047, 2019.

[24] Dian, R. and Li, S., ”Hyperspectral image super-resolution via subspace-based low tensor multi-rank regularization,”*IEEE Trans. Image Process.*, vol. 28, no. 10, pp.5135-5146, 2019.

[25] N. Akhtar, F. Shafait, and A. Mian, “Bayesian sparse representation for hyperspectral image super resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3631–3640.

[26] Q. Wei, J. M. B. Dias, N. Dobigeon, and J.-Y. Tourneret, “Hyperspectral and multispectral image fusion based on a sparse representation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul, 2015.

[27] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, “Bayesian fusion of multi-band images,” *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 6, pp. 1117–1127, Sep, 2015.

[28] Q. Wei, N. Dobigeon, and J.-Y. 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.

[29] M. Simoes, J. B. Dias, L. 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.

[30] X. Bresson and T.F Chan, “Fast dual minimization of the vectorial total variation norm and applications to color image processing,” *Inverse Probl Imaging.*, vol. 2, no. 4, p. 455-484, 2008.

[31] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, “A new pan-sharpening method with deep neural networks,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May, 2015.

[32] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, “Pansharpening by convolutional neural networks,” *Remote Sens.*, vol. 8, no. 7, pp. 594, 2016.

[33] Xie, Q., Zhou, M., Zhao, Q., Meng, D., Zuo, W. and Xu, Z., ”Multi-spectral and hyperspectral image fusion by MS/HS fusion net,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1585-1594.

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

[35] Chang, Y., Yan, L., Fang, H., Zhong, S. and Liao, W., ”HSI-DeNet: Hyperspectral image restoration via convolutional neural network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp.667-682, 2018.

[36] Zhang, L., Nie, J., Wei, W., Zhang, Y., Liao, S. and Shao, L., ”Unsupervised Adaptation Learning for Hyperspectral Imagery Super-Resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3073-3082.

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

[38] Wang, W., Zeng, W., Huang, Y., Ding, X. and Paisley, J., ”Deep Blind Hyperspectral Image Fusion,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4150-4159.

[39] Dian, R., Li, S. and Kang, X., ”Regularizing Hyperspectral and Multispectral Image Fusion by CNN Denoiser,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1124-1135, March, 2021.

[40] Dian, R., Li, S., Guo, A. and Fang, L., ”Deep hyperspectral image sharpening,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol 29, no. 11, pp. 5345-5355, Feb, 2018.

[41] Hu, J.F., Huang, T.Z., Deng, L.J., Jiang, T.X., Vivone, G. and Chanussot, J., ”Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1-15, 2021.

[42] Dong, W., Zhou, C., Wu, F., Wu, J., Shi, G. and Li, X., ”Model-guided deep hyperspectral image super-resolution,” *IEEE Trans. Image Process.*, vol 30, pp.5754-5768, 2021.

[43] Ronneberger, O., Fischer, P. and Brox, T., ”U-net: Convolutional networks for biomedical image segmentation,” in *Med. Image. Comput. Comput. Assist. Interv.*, 2015, pp. 234-241.

[44] Caruana, R., ”Multitask learning, ” *Mach. Learn.*, vol. 28, no. 1, pp.41-75, 1997.

[45] Vincent, P., Larochelle, H., Bengio, Y. and Manzagol, P.A., "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2018, pp. 1096-1103.

[46] Fu, Y., Zhang, T., Zheng, Y., Zhang, D. and Huang, H., "Joint camera spectral sensitivity selection and hyperspectral image recovery," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 788-804.

[47] Le, L., Patterson, A. and White, M., "Supervised autoencoders: Improving generalization performance with unsupervised regularizers," *Adv. Neural Inf. Process. Syst.*, pp.107-117, 2018.

[48] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H., "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[49] 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, 2010.

[50] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*", 2011, pp. 193–200.

[51] Arad, B. and Ben-Shahar, O., "Sparse recovery of hyperspectral signal from natural RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 19-34.

[52] Yokoya, N. and Iwasaki, A., Airborne hyperspectral data over Chikusei. Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, 2016.

[53] 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, 2004.

[54] R. H. Yuhas, J. W. Boardman, and A. F. Goetz., "Determination of semi-arid landscape endmembers and seasonal trends using convex geometry spectral unmixing techniques," 1993.

[55] L. Wald., "Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions," *Presses deslEcole MINES*, 2002.

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

[57] C. Dong, C. C. Loy, K. M. He, X. O. Tang, "Image superresolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295-307, 2016.

[58] J. Kim, J. Kwon Lee, K. Mu Lee., "Accurate image superresolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646-1654.

[59] Odena, A., Dumoulin, V. and Olah, C., "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p.e3, 2016.