Decoupled-and-Coupled Networks: Self-Supervised Hyperspectral Image Super-Resolution With Subpixel Fusion

Danfeng Hong, Senior Member, IEEE, Jing Yao, Member, IEEE, Chenyu Li, Student Member, IEEE, Deyu Meng, Member, IEEE, Naoto Yokoya, Member, IEEE, and Jocelyn Chanussot, Fellow, IEEE

Abstract—Enormous efforts have been recently made to super-resolve hyperspectral (HS) images with the aid of high spatial resolution multispectral (MS) images. Most prior works usually perform the fusion task by means of multifarious pixel-level priors. Yet, the intrinsic effects of a large distribution gap between HS–MS data due to differences in the spatial and spectral resolution are less investigated. The gap might be caused by unknown sensor-specific properties or highly mixed spectral information within one pixel (due to low spatial resolution). To this end, we propose a subpixel-level HS super-resolution (HS-SR) framework by devising a novel decoupled-and-coupled network (DC-Net), to progressively fuse HS–MS information from the pixel level to subpixel level and from the image level to feature level. As the name suggests, DC-Net first decouples the input into common (or cross-sensor) and sensor-specific components to eliminate the gap between HS–MS images before further fusion and then thoroughly blends them by a model-guided coupled spectral unmixing (CSU) net. More significantly, we append a self-supervised learning module behind the CSU net by guaranteeing material consistency to enhance the detailed appearance of the restored HS product. Extensive experimental results show the superiority of our method both visually and quantitatively and achieve a significant improvement in comparison with the state of the art (SOTA).

Index Terms—Data fusion, deep learning (DL), hyperspectral (HS) image, self-supervised, spectral unmixing, super-resolution.

1. INTRODUCTION

RECENT hyperspectral (HS) images are characterized by abundant and detailed spectral information, which enables the recognition of materials at a more accurate level compared with RGB or multispectral (MS) images. In recent years, HS imaging has been garnering increasing attention in a wide range of applications related to computer vision, such as image classification [1], [2], [3], image super-resolution [4], [5], [6], object detection, and tracking [7], [8], [9], [10], to name a few.

HS images are capable of capturing more subtle discrepancies between different objects, which, to a great extent, is benefited from its high spectral resolution. As a trade-off, HS imaging systems are usually designed to acquire the data at a high spatial sampling distance (i.e., sacrificing spatial resolution), thereby limiting the range of potential applications in practice [11]. Fortunately, MS imaging systems, with a broader spectral bandwidth, can provide finer spatial information. Therefore, fusing the low-resolution HS (LRHS) and high-resolution MS (HRMS) image pair is an intuitive and feasible solution to generate the high-resolution HS (HRHS) product, also known as MS-guided HS super-resolution (HS-SR). Fig. 1(a) illustrates the observation models for HRMS and LRHS images degraded from the HRHS image.

Over the last decade, a variety of model-driven HS-SR algorithms have been successfully developed [12]. These approaches model the underlying relationships between HS–MS data by various handcrafted priors and the knowledge of relevant sensor characteristics, e.g., spectral response function (SRF) and point spread function (PSF) of different camera systems. Very recently, the great success of convolutional neural networks (CNNs) has also made significant progress on the HS-SR task [13], [14], [15], [16], [17]. Owing to the powerful representation ability, CNNs-based models can excavate the intrinsic properties that lie in the HS (or MS) image more effectively and blend HS and MS images in a more compact way. Nevertheless, those existing, either model-based or data-driven (e.g., CNNs), fusion approaches rarely consider underlying problems that impact the HS-SR performance to a great extent, e.g., different sensor-specific properties leading to the weak affinity between original HS and MS data, highly mixed pixels widely existed in both HS and MS images [see Fig. 1(a)]. Mixed pixels inevitably degrade the performance of HR-SR due to the inherent ambiguity in the types and

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
The rest of this article is organized as follows. Section II introduces the related works from the perspectives of traditional models and deep models in detail. We then elaborate on the proposed DC-Net and detail each newly developed module. Experiential results are conducted in comparison with currently existing state-of-the-art (SOTA) HS-SR approaches in Section IV. Finally, Section V draws a conclusion with a possible future outlook.

II. RELATED WORK

A. Traditional Models

HS pansharpening is a heuristic way to perform the HS–MS fusion [24], which has been widely applied in the HS image super-resolution task. Component substitution (CS) and multisresolution analysis (MRA) are the two main types of pansharpening techniques. The former aims to inject detailed information about MS images into the LrHS image, thereby generating the HrHS product. The latter is to pansharpen the HS image by linearly combining MS bands to synthesize an HrHS band using regression techniques.

One representative technique that fuses the HS–MS images is the subspace-based approach, which is usually designed to enhance the spatial resolution of HS images by means of matrix factorization [11], [25], [26], [27], [28] or Bayesian estimation [29], [30]. Kawakami et al. [11] first decomposed the LrHS image into a spectral dictionary and corresponding sparse representations via matrix factorization, and the HrHS image can be then formed by using shareable sparse coefficients estimated from the RGB observation with a sampled spectral dictionary. Akhtar et al. [26] jointly considered different physical properties of materials in the scene, e.g., sparsity, non-negativity, and spatial structure, to improve the quality of the HrHS image. Dong et al. [28] structured the sparse coding model with the application to HS-SR. Yokoya’s algorithm [25] is designed to perform the HS–MS fusion via an effective couple matrix factorization approach, which can be well explained from the spectral unmixing perspective. Similarly, Lanaras et al. [27] further extended [25] by jointly unmixing the input HS and MS data and fusing them in the latent subspace. In [29], the spectral dictionary is learned by using a parameter-free Bayesian model. Using the learned dictionary, the HrHS image can be reconstructed with potentially sparse assumptions. Furthermore, the same investigators used the Gaussian process as the priors on the basis of the
Bayesian model to super-resolve the HS image [30]. Beyond matrix factorization, Zhang et al. [5] and Dian et al. [31] have attempted to model the HS-SR issue in the tensor form to achieve better structural representations.

B. DL-Based Models

These traditional methods have shown competitive performance in the HS–MS fusion task, yet they, to a great extent, rely on cross-sensor calibration and strong hand-crafted priors. Inspired by the recent success of deep learning (DL) techniques, CNNs-based approaches have been garnering growing attention in the HS-SR task [13], [14], [15], [16], [17], [32]. Dian et al. [13] proposed a two-stage HS image pansharpening framework by using CNNs-based prior training to refine the initialized fusion results obtained by traditional optimization methods. Qu et al. [14] trained an end-to-end unsupervised Dirichlet network to restore the HrHS image. Xie et al. [15] sought to open the “black box” and presented an interpretable deep supervised model for the MS/HS fusion task. Yao et al. [16] attempted to explain an unsupervised deep super-resolution network from the spectral unmixing perspective by introducing cycle consistency and cross-attention mechanism. Zhu et al. [32] proposed a supervised deep residual network in a progressive way for the HS image super-resolution. Furthermore, the field of RS has witnessed numerous advancements in multimodal image fusion and analysis techniques [33], [34], [35], [36]. These sophisticated approaches have found widespread application in various practical scenarios, including pansharpening, classification, unmixing, geospatial object detection, and more.

III. DECOUPLED-AND-COUPLED NETWORKS

A. Overview

A large distribution gap between HS–MS images possibly caused by unknown sensor-specific information (e.g., HS–MS) and highly mixed spectral pixels has been shown to be a challenging and potential problem in the HS-SR task [12], [25], [27]. Focusing on it, we provide a point-to-point solution by designing a novel network architecture, i.e., DC-Net, which consists of two subnetworks. D-Net aims to decompose the original HS–MS data into the common content (or structure) and sensor-specific features and implement crossover recombination on the sensor-specific information to improve the compatibility between HS–MS images for better subsequent fusion. With the input of recombined HS–MS images, C-Net performs the subpixel-level fusion by decomposing the mixed pixels into fractional coefficients (termed abundances) and pure spectral signatures (termed endmembers), yielding an interpretable CSU-guided HS-SR network. It should be noted that DC-Net (i.e., D-Net and C-Net) can be regarded as a novel progressive fusion framework from pixel level to subpixel level and from the image level to feature level. Furthermore, a simple but effective self-supervised subnetwork (S-Net) in C-Net is devised to improve the quality of the HrHS product by correcting the material correspondence between HS–MS images. Fig. 2 illustrates the architecture overview of the proposed DC-Net.

B. Problem Formulation

Let \( X \in \mathbb{R}^{H \times W \times L} \), \( Y \in \mathbb{R}^{H \times W \times l} \), and \( Z \in \mathbb{R}^{H \times W \times L} \) be \( \text{LrHS} \) with \( hw \) pixels by \( L \) bands, \( \text{HrMS} \) with \( HW \) pixels by \( l \) bands, and \( \text{HrHS} \) images with \( HW \) pixels by \( L \) bands, respectively (\( h < H, w < W, l < L \)); then \( X \in \mathbb{R}^{L \times hw} \), \( Y \in \mathbb{R}^{l \times HW} \), and \( Z \in \mathbb{R}^{L \times HW} \) are defined as the unfolded matrices. The observation models for \( \text{LrHS} \) and \( \text{HrMS} \) images from the \( \text{HrHS} \) image can be written as follows:

\[
X = ZR + N_x \\
Y = HZ + N_y
\]

(1)

where \( H \in \mathbb{R}^{l \times L} \) and \( R \in \mathbb{R}^{H \times hw \times hw} \) denote the SRF and PSF to degrade the spectral and spatial resolutions of \( Z \), respectively, and \( N_x \) and \( N_y \) are the observed noises.

Ideally, the observations in (1) can be further unfolded by using the spectral mixing model as follows:

\[
X = SAR + N_x = S\tilde{A} + N_x \\
Y = HSA + N_y = S\tilde{A} + N_y
\]

(2)

where \( S \geq 0 \) and \( A \geq 0 \) denote a collection of pure spectral signatures (i.e., endmembers) and corresponding fractional coefficients (i.e., abundances) that meet the sum-to-one constraint (i.e., \( 1^T \tilde{A} = 1^T \)), respectively. The two constraints offer valuable physical insights regarding the endmembers and abundances, enhancing the interpretability of our observation model and ensuring its suitability for real-world scenarios. \( \tilde{S} \) and \( \tilde{A} \) can be explained as the degraded \( S \) and \( A \) in spectral and spatial domains with the use of SRF and PSF. Thus, estimating \( Z \) is equivalent to finding \( S \) and \( A \), i.e., \( Z = SA \), by solving problem (2).

Model (2) has been proven to be effective for the problem of mixed pixels in the HS–MS fusion [12], [13], [14], [16], [25], [27], [31]. However, it seldom considers the high coupling between pixels in the same data source. For this reason, we learn the mappings \( (f_s, f_y) \) for transforming \( X \) and \( Y \) to a latent image space (e.g., \( \tilde{X} \) and \( \tilde{Y} \)), where the pixel information can be easily separated between the same data source and better fused across different data sources. Therefore, the resulting model is

\[
\tilde{X} = f_s(X) = f_s(S\tilde{A} + N_x) \\
\tilde{Y} = f_y(Y) = f_y(S\tilde{A} + N_y)
\]

(3)

In the following, we will build the DC-Net architecture step-by-step to perform the HS–MS fusion in model (3).

C. Decoupled Network (D-Net)

Owing to the powerful image-to-image transformation ability of GANs [37], [38], [39], we develop the D-Net to decouple the input images and recombine them (see Fig. 2), making it possible for the recombined images to be fused more sufficiently. D-Net embeds the HS–MS images onto a common content space \( \mathcal{C} \) and sensor-specific spaces, e.g., \( \mathcal{C}_X \) and \( \mathcal{C}_Y \). This process can be performed by cross-sensor (or domain-common) encoders \( \{E_X^C, E_Y^C\} \), sensor-specific encoders \( \{E_X^S, E_Y^S\} \), generators \( \{G_X, G_Y\} \), and
D. Coupled Network (C-Net)

Linking to the D-Net, C-Net unmixes the recombined HS and MS images into endmembers (S and Ș) and abundances (A and (IntPtr 2. Illustrative overview of the proposed DC-Net. It consists of two subnetworks: D-Net and C-Net. With the adversarial loss $L_{adv}$ between the common contents ($\tilde{X}$ and Ș) of X and Ȳ, the reconstruction loss $L_{rec}$ and the self-supervised module (i.e., S-Net), we can learn an end-to-end HS-SR network more effectively without any prior training and better enhance detailed appearances.

Fig. 2. Illustrative overview of the proposed DC-Net. It consists of two subnetworks: D-Net and C-Net. With the adversarial loss $L_{adv}$ between the common contents ($\tilde{X}$ and Ș) of X and Ȳ, the reconstruction loss $L_{rec}$ and the self-supervised module (i.e., S-Net), we can learn an end-to-end HS-SR network more effectively without any prior training and better enhance detailed appearances.

$W$ is the weights of C-Net, where $W_{f,v}$ and $W_{g,v}$ are specified to be a $1 \times 1$ linear convolutional kernel, which can be well explained as endmembers, i.e., S and Ș. Note that we adopt the clamp function to meet the abundance non-negative constraint (ANC), while the abundance sum-to-one constraint (ASC), i.e., $1^T A = 1^T$, can be guaranteed in the form of regularization. Using (5), the to-be-estimated HRHS image Ș can be obtained by $\hat{Z} = SA = f_v(g_v(\hat{Y}; W_{g,v}); W_{f,v})$.

*Cycle Consistency:* Coupled factors, i.e., SRF (H) and PSF (R), play a dominant role in the performance improvement of C-Net. Fig. 3 shows an SRF example with respect to RGB bands. Unlike previous methods that assume the two functions to be known, we attempt to automatically learn them by designing a cycle consistency mechanism in networks, enabling it to be more applicable in reality. For simplicity, the cycle can be written as $(X, Y) \xrightarrow{D-Net} (\tilde{X}, \tilde{Y}) \xrightarrow{C-Net} \hat{Z} \xrightarrow{(H,R)} (X, Y)$. More specifically, given the spectrum of the $i$th pixel in Ș (denoted as $z_i$) and the corresponding spectrum of the $i$th pixel and the $j$th channel in Y (denoted as $y_{i,j}$), the degradation process can be then expressed as follows:

$$y_{i,j} = \int_{\phi} h_j(\eta) z_i(\eta) d\eta$$

(6)

where $\phi$ denotes the support set with respect to the wavelength $\eta$, $h_j(\eta) d\eta$ is the continuous representation of $H$, and $N_H$ is a constant for scaled normalization. In our case, $H$ can be approximated using a set of $1 \times 1 \times L$ convolutional kernels, e.g., $w_j$, in a discretized way, i.e.,

$$y_{i,j} = z_j \otimes w_j = \sum_{\phi} \frac{w_j(\eta)}{N_w} z_i(\eta).$$

(7)

Similarly, the spatial downsampling process performed by the PSF (R) can also be simulated by means of convolutional operators with the same scaling kernel and stride sizes.

As a result, the cycle consistency for both spatial and spectral degradation can be well organized in the C-Net as follows:

$$X = \hat{Z} R, \quad Y = H \hat{Z}$$

(8)
which is a good fit for the observation models in (1) and helps us form a closed cycle chain.

E. Self-Supervised Subnetwork (S-Net)

In C-Net, the spectral unmixing process for HS and MS images is relatively independent. This might lead to chaotic semantic correspondences on abundance maps (or endmembers) of HS–MS images, e.g., between \( A \) and \( \hat{A} \) or \( S \) and \( S \). To this end, we design an effective self-supervised learning module (S-Net) in C-Net to obtain more plausible abundance maps and endmembers, further yielding a more accurate HS reconstruction result. Fig. 4 illustrates the S-Net.

S-Net aims at correcting the correspondence of abundance maps of HS–MS images by learning high-level semantic alignment or consistency (in both order and category of materials), enabling a one-to-one match between \( A \) and \( \hat{A} \). We call it as material or semantic consistency. The abundance maps of the same materials should be attracted to each other (as positive samples); otherwise, they are repelled (as negative samples). Considering the computational cost (since hundreds of abundance maps and spectral bundles usually need to be considered), we employ the same grouping strategy on \( A \) and \( \hat{A} \) to make a group-to-group match. \( A \) and \( \hat{A} \) can be grouped as follows:

\[
A = [A_1, \ldots, A_k, \ldots, A_m] \\
\hat{A} = [\hat{A}_1, \ldots, \hat{A}_k, \ldots, \hat{A}_m]
\]

(9)

where \( \Phi_{A,p} \) and \( \Phi_{A,q} \) denote the grouping sets with respect to \( A \) and \( \hat{A} \) for \( \forall p, q \in \{1, \ldots, m\} \), respectively. We experimentally set the grouping parameter \( m \) to be 8. Furthermore, the representation vectors of \( \Phi_{A,p} \) and \( \Phi_{A,q} \), denoted as \( z_a \), can be obtained via a two-stream encoder or CNN, e.g., \( f_{\Phi_a} \) and \( f_{\Phi_q} \), which both consist of two blocks, i.e., [conv-relu-pool-fc]. The main difference lies in the pooling size and stride for \( f_{\Phi_q} \), larger than those for \( f_{\Phi_a} \), to guarantee the same size for final output features. We then have \( z_{A,p} = f_{\Phi_a}(\Phi_{A,p}) \) and \( z_{A,q} = f_{\Phi_q}(\Phi_{A,q}) \). In our S-Net, only when \( p = q \), \( z_{A,q} \) is the positive sample of \( z_{A,p} \); otherwise, \( (p \neq q) \), and they are the negative sample pair.

F. Network Training

1) Loss Functions: As shown in Fig. 2, there are several important loss functions that need to be carefully considered.

In D-Net, the adversarial loss between the common content of \( X \) and \( Y \) can be expressed as follows:

\[
\mathcal{L}_{\text{adv}} = \mathbb{E}_x \left[ \frac{1}{2} \log D_\mathcal{C}(X) + \frac{1}{2} \log (1 - D_\mathcal{C}(X)) \right] + \mathbb{E}_y \left[ \frac{1}{2} \log D_\mathcal{C}(Y) + \frac{1}{2} \log (1 - D_\mathcal{C}(Y)) \right]
\]

(10)

where \( X = E_\mathcal{C}(X) \) and \( Y = E_\mathcal{C}(Y) \).

The reconstruction loss stemmed from coupled convolutional autoencoders and the cycle consistency can be computed by means of (5) and (8)

\[
\mathcal{L}_{\text{rec}} = \|\hat{X} - X\|_1 + \|\hat{Y} - Y\|_1 + \|\hat{X} - Z_{i,R}\|_1 + \|\hat{Y} - Z_{i,H}\|_1 + \|\hat{X} - X\|_1 + \|\hat{Y} - Y\|_1
\]

(11)

where the \( \ell_1 \)-norm measure is applied to enhance the detailed perception in image reconstruction [40], and \( Z_{i,R} \) and \( Z_{i,H} \) denote \( \hat{Z}R \) and \( \hat{Z}H \), respectively.

Moreover, the ASC is satisfied using the following loss:

\[
\mathcal{L}_{\text{ASC}} = \|1^T - 1^T A\|_1 + \|1 - 1^T \hat{A}\|_1
\]

(12)

In S-Net, the InfoNCE loss [41] is optimized to measure the distances between positive and negative samples

\[
\mathcal{L}_{\text{self}} = -\mathbb{E}_\Phi \left[ \log \frac{\exp(f_{\Phi}(\Phi_{A,p})^T f_{\Phi}(\Phi_{A,q}))}{\sum_{q=1}^m \exp(f_{\Phi}(\Phi_{A,p})^T f_{\Phi}(\Phi_{A,q}))} \right]
\]

(13)

where \( \Phi_{A,p} \) and \( \Phi_{A,q}^T \) denote the anchor sample and the positive sample, respectively.

In summary, our DC-Net is trained by minimizing the following overall loss parameterized by the set \( \{\alpha, \beta, \gamma\} \):

\[
\mathcal{L} = \mathcal{L}_{\text{rec}} + \alpha \mathcal{L}_{\text{adv}} + \beta \mathcal{L}_{\text{ASC}} + \gamma \mathcal{L}_{\text{self}}
\]

(14)

The penalty parameters are established through empirical and experimental fine-tuning. In network training, different
Fig. 5. Visualization of super-resolved HS images obtained by all compared methods (CAVE: chart and stuffed toy), where an ROI zoomed in four times (bottom right) and the difference image with GT (bottom left) are highlighted.

Fig. 6. Visualization of super-resolved HS images obtained by all compared methods (CAVE: stuffed toys), where an ROI zoomed in four times (bottom right) and the difference image with GT (bottom left) are highlighted.

| Method   | Ref.       | PSNR | SAM | ERGAS | SSIM | UQI   |
|----------|------------|------|-----|-------|------|-------|
| BSR      | CVPR’15 [29]| 38.13| 4.54| 0.60  | 0.948| 0.963 |
| CSU      | ICCV’15 [27]| 36.96| 5.20| 0.73  | 0.898| 0.937 |
| NSSR     | TIP’16 [28]| 42.30| 3.12| 0.48  | 0.970| 0.984 |
| NLSTF    | CVPR’17 [31]| 37.04| 6.11| 0.70  | 0.901| 0.925 |
| uSDN     | CVPR’18 [41]| 40.02| 2.96| 0.41  | 0.964| 0.981 |
| MHF-Net  | CVPR’19 [42]| 42.43| 3.01| 0.44  | 0.977| 0.982 |
| GDD      | ECCV’20 [43]| 41.58| 3.04| 0.43  | 0.974| 0.984 |
| CUCaNet  | ECCV’20 [44]| 41.79| 2.84| 0.40  | 0.975| 0.988 |
| DC-Net   | –          | 42.08| 2.78| 0.38  | 0.977| 0.989 |
| DC-Net-S | –          | 42.56| 2.55| 0.34  | 0.979| 0.992 |

Table II

Quantitative Performance Comparison Between the Same SOTA Super-Resolution Methods in Terms of Five Different PQIs on the Chikusei Dataset

regularizations assume distinct roles, each making a substantial contribution to the overall process. Although their overall importance is generally on par, a more precise quantitative ranking of their significance is available in the following ablation analysis.

2) Implementation Details: Our networks are implemented on the PyTorch platform, and the Adam [42] optimizer is used to train the networks for 10,000 epochs with a batch size of 1. The base learning rate starts with 0.005, which can be gradually updated by a linear decay schedule [43]. We initialize all convolutional layers in our networks with Kaiming initialization [44]. Furthermore, the hyperparameters are experimentally determined using a grid search, and an early stopping strategy is considered in the network training when the validation loss fails to decrease.

IV. EXPERIMENTAL RESULTS

In this section, we first review the datasets and experimental settings. Next, we verify our networks with extensive ablation studies. Finally, the performance evaluation on simulated and real datasets is conducted to show the superiority of DC-Net in comparison with SOTA methods.

A. Dataset and Setup

Two public HS–MS datasets are used to evaluate the performance of DC-Net, i.e., CAVE dataset (indoor) [45] and Chikusei dataset (remote sensing) [12]. The CAVE dataset consists of 32 indoor HrHS images with the size of $512 \times 512 \times 31$, covering the spectral wavelength...
from 400 to 700 nm. The Chikusei dataset acquired by the VNIR-C sensor comprises of $2517 \times 2335$ pixels and 128 spectral bands. We randomly crop the scene into 12 nonoverlapped subimages with the spatial size of $576 \times 448$. Moreover, we divide the two datasets into training, validation, and testing sets in the proportions of 8:8:16 and 4:4:4, respectively, where training and validation sets are used to determine optimal hyperparameters.

For the two datasets, the HrMS images are generated by degrading the HrHS images in the spectral domain with SRFs, while the LrHS images can also be simulated by spatially downsampling the HrHS images with the Wald et al.’s [46] protocol of a 32 sampling ratio. In addition, the Nikon D700 sensor and Landsat-8 MS sensor [47] are selected as the SRFs for the CAVE and Chikusei datasets, respectively.

Evaluation Metrics: We evaluate the fusion performance quantitatively in terms of five widely used picture quality indices (PQIs): peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM) [48], erreur relative globale adimensionnelle de synthèse (ERGAS) [49], structure similarity (SSIM) [50], and universal image quality index (UIQI) [51].

B. Comparison With SOTAs

We evaluate the performance of DC-Net both quantitatively and qualitatively in comparison with SOTA models in the HS-SR task, including traditional SOTA methods: BSR [29], CSU [27], NSSR [28], and NLSTF [31], and DL-based SOTA methods: uSDN [14], MHF-Net [15], GDD [17], and CUCaNet [16], as well as the DC-Net without S-Net module. We maintain the same experimental configurations for all compared methods as in their original literature as much as possible.

1) Experiments on CAVE Data: Table I lists the average results over 16 testing images on the CAVE dataset to make a quantitative comparison between all compared SOTA methods and the DC-Net and DC-Net-S in terms of five PQIs.

Overall, BSR, as a pioneer work in HS-SR, reconstructs the high-quality HS image from the Bayesian statistic point of view and achieves competitive performance. Despite relatively lower PSNR and SSIM (see BSR), CSU can obtain better PQIs in SAM, ERGAS, and UIQI, owing in large part to the use of the CSU mechanism. Beyond the CSU, NLSTF fuses HS and MS images in a coupled tensor form with nonlocal spatial information preservation, yielding better performance in all indices except the UIQI. By virtue of non-negativity and structured modeling, NSSR outperforms other traditional SOTA methods in HrHS image reconstruction. As an early stage DL-based method, the fusion performance of uSDN is moderately inferior to that of other DL-based competitors but is clearly higher than that of those non-DL algorithms. It is more noteworthy that the results obtained by unsupervised GDD and CUCaNet are basically comparable to those of the supervised MHF-Net (even better in certain indices). DC-Net behaves superiorly compared with the existing SOTA DL approaches, implying that the decoupled-and-coupled strategy creates the "one plus one greater than two" effect. Furthermore, S-Net module can learn high-level semantic consistency.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 8. Visual evaluation on CAVE and Chikusei datasets, respectively. The GT, the residual maps between GT maps, and super-solved HS results obtained by compared methods (e.g., NSSR, uSDN, MHF-Net, and CUCaNet) and our proposed DC-Net and DC-Net-S. The RMSE values of HS images for these methods along the spectra (covering the wavelength) are also given corresponding to the six studied scenes.

from abundance maps and embed interpretable and physically meaningful information into the DC-Net, enabling the best performance (i.e., DC-Net-S).

Figs. 5 and 6 visualize the super-resolved HS image in the 11th band (≈500 nm) for chart and stuffed toy. Furthermore, two regions of interest (ROIs) are selected to highlight the visually detailed differences of all compared methods. By and large, DL-based models outperform evidently compared with traditional methods, particularly in the coarser-grained structure of objects. The results of the proposed methods are superior to other unsupervised ones (e.g., uSDN, GDD, and CUCaNet), while DC-Net-S performs better than DC-Net in terms of detailed appearances. Despite similar visual results (see MHF-Net), our networks, either DC-Net or DC-Net-S, are capable of better recovering textural details and color cues that approximate ground truth (GT).

2) Experiments on Chikusei Dataset: Quantitative performance comparison on the Chikusei dataset is given in Table II, where there is a basically identical trend with that on the CAVE dataset (see Table I). Remote sensing images (due to the lower spatial resolution) are not as complicated as indoor images; all these methods, therefore, tend to super-resolve HS images with higher quality on the Chikusei dataset (see CAVE). Significantly, the performance of our DC-Net-S is still superior to other competitors with respect to all PQIs, demonstrating its superiority in the HS-SR task. However, the most noteworthy point lies in that MHF-Net obtains competitive results (fully comparable to DC-Net without S-Net module), since its accuracy relies greatly on supervised training on plenty of sample pairs. Fig. 7 gives a scene example and shows the composite HrHS images obtained by all competing methods with an ROI zoomed in four times for better observing visual differences.
Fig. 9. Visualization of super-resolved HrHS fused by WV-3 HrMS (image courtesy Maxar) and Hyperion LrHS using our method (real data of Cuprite: three subregions).

3) Visual Quality Evaluation: To evaluate the visual quality of HS super-solved results more finely, we select several representative images from CAVE (four scenes) and Chikusei (two scenes) datasets, respectively, and show the corresponding residual maps between GT maps and super-solved HS results obtained by different compared methods, e.g., NSSR, uSDN, MHF-Net, and CUCaNet, and our proposed DC-Net and DC-Net-S, in Fig. 8. Furthermore, the RMSE values of HS images in each band are shown along spectra corresponding to the aforementioned six scenes. Intuitively, our methods perform better than other competitors from the perspective of residual maps, which demonstrates that our methods can provide finer spatial details. More importantly, the stability and generalization ability of our DC-Nets are higher than compared networks in different scenes, showing the effectiveness to a great extent. The bandwise RMES results obtained by our models are obviously similar and closer to the GT in the six investigated scenes. This, to some extent, indicates that the sequentially spectral information can be well recovered and retained by our approaches.

C. Evaluation on Real Data

To effectively assess the performance of our DC-Net, we curated a dataset comprising real, registered MS and HS scenes acquired from the WorldView-3 and Hyperion satellite missions, respectively. We focused on three specific subregions to facilitate a visual quality comparison of the composite HrHS products, as depicted in Fig. 9. Each subregion consists of HrMS and LrHS images, with the dimensions of $640 \times 640$ pixels and eight bands, maintaining a ground sampling distance (GSD) of 7.5 m, and $160 \times 160$ pixels with 167 spectral channels at a GSD of 30 m, respectively. Unfortunately, a real HrHS image is not available for reference.

In Fig. 9, it is evident that both DC-Net and DC-Net-S successfully restore HrHS results. While both methods produce clear results, DC-Net-S exhibits a higher degree of visual fidelity. In comparison with DC-Net, DC-Net-S achieves structural and textural appearances that closely match the HrMS image and, in some aspects, even surpass it. Furthermore, its color and brightness more closely align with those of the LrHS image. This highlights the superiority of the proposed methods in the HS-SR task. However, it is important to acknowledge that some spectral degradation is visually apparent. Therefore, there is room for improvement in terms of reconstruction performance.

D. Model Analysis

1) Ablation Studies: We investigate the performance gain by stepwise adding different modules (or subnetworks, i.e., D-Net and S-Net) in networks. We also study the importance of two physical constraints (ANC and ASC) related to the spectral unmixing model and their effects on the quality of HS-SR. To verify the effectiveness of the ablation analysis, we report the average results in terms of PSNR and SAM indices over eight images on the validation set of the CAVE dataset by maximizing the stepwise performance using optimal parameters, as listed in Table III.

In detail, C-Net without ANC and ASC yields relatively poor performance, which, to some extent, indicates the importance of the two critical physical constraints in the spectral unmixing-inspired HS-SR network. By turning on the ANC, the fusion results of C-Net increase dramatically (beyond four PSNR values and around halved SAM value). It is better still that, the joint use of ANC and ASC can further bring the performance improvement in both PSNR and SAM compared with ANC used alone in C-Net. As expected, DC-Net after plugging the D-Net performs observably better than C-Net at an increase of about 1.5 PSNR value and a decrease of just over two SAM values. More remarkably, the self-supervised module (S-Net) is capable of breaking through the existing “bottleneck” in DC-Net by enforcing the semantic consistency on abundance maps ($\mathbf{A}$ and $\tilde{\mathbf{A}}$).

2) Computational Analysis: We list the size of network parameters (i.e., DC-Net-S) and the inference time per image on a PC with one NVIDIA GeForce GTX 1080Ti GPU in Table IV. Our network parameters only have 3.43 MB, and
since our method is unsupervised, the inference time is also desirable and acceptable in such a PC setting.

V. CONCLUSION

In this article, we propose a novel subpixel-level HS-SR framework, i.e., DC-Net, by fully considering the effects of affinity of the same data sources and exclusivity across HS–MS data and mixed pixels within one pixel. Inspired by spectral unmixing, DC-Net is capable of utilizing the intrinsic properties of HS–MS images effectively for the fusion task, which is more applicable to various real cases. We further optimize the network performance by designing an effective self-supervised module. Extensive experiments conducted on simulated and real data show the superiority of DC-Net in HS-SR over current SOTA fusion methods.

We found, however, that there exists spectral degradation in the HS-SR products of the real data. Consequently, in our future endeavors, we will prioritize the development of novel modules or the adoption of advanced network architectures aimed at preserving spectral sequentiality to the greatest extent possible while enhancing resolution.

REFERENCES

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

[2] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, “Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework.” IEEE Trans. Geosci. Remote Sens., vol. 61, 2023, Art. no. 5514415.

[3] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, “Multimodal fusion transformer for remote sensing image classification,” IEEE Trans. Geosci. Remote Sens., vol. 61, 2023, Art. no. 5515620.

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

[5] L. Zhang, W. Wei, C. Bai, Y. Gao, and Y. Zhang, “Exploiting clustering manifold structure for hyperspectral imagery super-resolution,” IEEE Trans. Image Process., vol. 27, no. 12, pp. 5969–5982, Dec. 2018.

[6] W. He, N. Yokoya, and X. Yuan, “Fast hyperspectral image recovery of dual-camera compressive hyperspectral imaging via non-iterative subspace-based fusion,” IEEE Trans. Image Process., vol. 30, pp. 7170–7183, 2021.

[7] N. M. Nasrabadi, “Hyperspectral target detection: An overview of current and future challenges,” IEEE Signal Process. Mag., vol. 31, no. 1, pp. 34–44, Jan. 2014.

[8] G. Tochon, J. Chanussot, M. Dalla Mura, and A. L. Bertozzi, “Object tracking by hierarchical decomposition of hyperspectral video sequences: Application to chemical gas plume tracking,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 8, pp. 4567–4585, Aug. 2017.

[9] X. Wu, D. Hong, and J. Chanussot, “UIU-net: U-net in U-net for infrared small object detection,” IEEE Trans. Image Process., vol. 32, pp. 364–376, 2023.

[10] C. Li, B. Zhang, D. Hong, J. Yao, and J. Chanussot, “LRR-net: An interpretable deep unfolding network for hyperspectral anomaly detection,” IEEE Trans. Geosci. Remote Sens., vol. 61, 2023, Art. no. 5513412.

[11] R. Kawakami, Y. Matsushita, J. Wright, M. Ben-Ezra, Y.-W. Tai, and K. Ikeuchi, “High-resolution hyperspectral imaging via matrix factorization,” in Proc. CVPR, Jun. 2011, pp. 2329–2336.

[12] N. Yokoya, C. Grohfeldt, and J. Chanussot, “Hyperspectral and multispectral data fusion: A comparative review of the recent literature,” IEEE Geosci. Remote Sens. Mag. Replaces Newsletter, vol. 5, no. 2, pp. 29–56, Jun. 2017.

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

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

[15] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, “Multispectral and hyperspectral image fusion by MS/HS fusion net,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1583–1594.

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

[17] T. Uezato, D. Hong, N. Yokoya, and W. He, “Guided deep decoder: Unsupervised image pair fusion,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 87–102.

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

[19] P. V. Arun, K. M. Buddhiraju, and A. Porwal, “CNN based sub-pixel mapping for hyperspectral images,” Neurocomputing, vol. 311, pp. 51–64, Oct. 2018.

[20] J. Yao, D. Hong, L. Xu, D. Meng, J. Chanussot, and Z. Xu, “Sparsity-enhanced convolutional decomposition: A novel tensor-based paradigm for blind hyperspectral unmixing,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5505014.

[21] Y. Fang, Y. Wang, L. Xu, Y. Chen, A. Wong, and D. A. Clausi, “Unsupervised Bayesian subpixel mapping autoencoder network for hyperspectral images,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2023, Art. no. 5513112.

[22] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014, pp. 2672–2680.

[23] N. Keshavarz and M. Alam, “Spectral unmixing,” IEEE Signal Process. Mag., vol. 19, no. 1, pp. 44–57, Jan. 2002.

[24] L. Loncan et al., “Hyperspectral pansharpening: A review,” IEEE Geosci. Remote Sens. Mag. Replaces Newsletter, vol. 3, no. 3, pp. 27–46, Sep. 2015.

[25] 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.

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

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

[28] W. Dong et al., “Hyperspectral Image super-resolution via non-negative structured sparse representation,” IEEE Trans. Image Process., vol. 25, no. 5, pp. 2337–2352, May 2016.

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

[30] N. Akhtar, F. Shafait, and A. Mian, “Hierarchical beta process with Gaussian process prior for hyperspectral image super-resolution,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 103–120.

[31] R. Dian, L. Fang, and S. Li, “Hyperspectral image super-resolution via non-local sparse tensor factorization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5344–5353.

[32] Z. Zhu, J. Hou, J. Chen, H. Zeng, and J. Zhou, “Hyperspectral image super-resolution via deep progressive zero-centric residual learning,” IEEE Trans. Image Process., vol. 30, pp. 1423–1438, 2021.

[33] D. Hong et al., “Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks,” 2023, arXiv:2309.16499.

[34] R. Hang, Q. Liu, and Z. Li, “Spectral super-resolution network guided by intrinsic properties of hyperspectral imagery,” IEEE Trans. Image Process., vol. 30, pp. 7256–7265, 2021.
Dr. Hong is an Editorial Board Member of Remote Sensing and an Editorial Advisory Board Member of the ISPRS Journal of Photogrammetry and Remote Sensing. He was a recipient of the Best Reviewer Award of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS) in 2021 and 2022 and the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) in 2022, the Jose Bioucas Dias Award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing, Evolution in Remote Sensing (WHISPERS) in 2021, the Remote Sensing Young Investigator Award in 2022, the IEEE Geoscience and Remote Sensing Society (GRSS) Early Career Award in 2022, and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters) in 2022. He is an Associate Editor of IEEE TGRS.

Jing Yao (Member, IEEE) received the Ph.D. degree in mathematics from Xi’an Jiaotong University, Xi’an, China, in 2021. From 2019 to 2020, he was a Visiting Student at Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, and the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. He put his recent research interests on hyperspectral and multimodal remote sensing image analysis, mainly including optimization and deep learning-based methods for image processing and interpretation tasks. Dr. Yao was a recipient of the Jose Bioucas Dias Award for recognizing the outstanding paper at the Workshop on Hyperspectral Imaging and Signal Processing, Evolution in Remote Sensing (WHISPERS) in 2021. He serves as a Guest Editor for Remote Sensing.

Chenyu Li (Student Member, IEEE) received the B.S. and M.S. degrees from the School of Transportation, Southeast University (SEU), Nanjing, China, in 2018 and 2021, respectively, where she is currently pursuing the Ph.D. degree with the Department of Mathematics. She is a co-supervised Ph.D. Student at the Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China. Her research interests include hyperspectral remote sensing, interpretable artificial intelligence, and target detection.

Danfeng Hong (Senior Member, IEEE) received the M.Sc. degree (summa cum laude) in computer vision from the College of Information Engineering, Qingdao University, Qingdao, China, in 2015, and the Dr.-Ing. degree (summa cum laude) from the Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), Munich, Germany, in 2019. From 2019 to 2021, he was a Research Scientist and led the Spectral Vision Working Group, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He has been an Adjunct Scientist at the GIPSA-Lab, Grenoble Institute of Technology (Grenoble INP), CNRS, Université Grenoble Alpes, Grenoble, France. Since 2022, he has been a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China. His research interests include artificial intelligence, remote sensing big data analysis, multimodal interpretation, and their applications in Earth vision.

Deyu Meng (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 2001, 2004, and 2008, respectively, all in applied mathematics. From 2012 to 2014, he took his two-year sabbatical leave with Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Mathematics and Statistics, Xi’an Jiaotong University. His research interests include model-based deep learning, variational networks, and meta-learning.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees in aerospace engineering from the Department of Aeronautics and Astronautics, The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively.

He was an Assistant Professor at The University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow with the German Aerospace Center (DLR), Oberpfaffenhofen, Germany, and the Technical University of Munich (TUM), Munich, Germany. Since 2018, he has been with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Team. Since 2020, he has been with The University of Tokyo, where he is currently an Associate Professor. His research is focused on the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management and environmental assessment.

Dr. Yokoya won the first place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was a Highly Cited Researcher (Clarivate Analytics) in 2022. He was the Chair from 2019 to 2021 and the Co-Chair of the IEEE GRSS IADF TC from 2017 to 2019. He was an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) from 2018 to 2021. He has been an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2021.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

From 2015 to 2017, he was a Visiting Professor at the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. Since 1999, he has been with the Grenoble INP, where he is currently a Professor of signal and image processing. He has been a Visiting Scholar at Stanford University, Stanford, CA, USA; KTH Royal Institute of Technology, Stockholm, Sweden; and National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor at the University of Iceland, Reykjavik, Iceland. He holds the AXA Chair in remote sensing and is an Adjunct Professor at the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008 and the Institut Universitaire de France from 2012 to 2017. He was the Founding President of the IEEE Geoscience and Remote Sensing French Chapter from 2007 to 2010, which received the 2010 IEEE Geoscience and Remote Sensing Society (GRSS) Chapter Excellence Award. He has received multiple outstanding paper awards. He is a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair from 2009 to 2011 and the Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He was the Vice President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia, from 2017 to 2019. He was an Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015 and an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING. In 2014, he served as a Guest Editor for IEEE Signal Processing Magazine. He is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the PROCEEDINGS OF THE IEEE.