Hyperspectral Pansharpening Based on Improved Deep Image Prior and Residual Reconstruction

Wele Gedara Chaminda Bandara, *Student Member, IEEE,* Jeya Maria Jose Valanarasu, *Student Member, IEEE,* and Vishal M. Patel, *Senior Member, IEEE*

Abstract*—Hyperspectral pansharpening aims to synthesize a low-resolution hyperspectral image (LR-HSI) with a registered panchromatic image (PAN) to generate an enhanced HSI with high spectral and spatial resolution. Recently proposed HS pansharpening methods have obtained remarkable results using deep convolutional networks (ConvNets), which typically consist of three steps: (1) up-sampling the LR-HSI, (2) predicting the residual image via a ConvNet, and (3) obtaining the final fused HSI by adding the outputs from first and second steps. Recent methods have leveraged Deep Image Prior (DIP) to up-sample the LR-HSI due to its excellent ability to preserve both spatial and spectral information, without learning from large data sets. However, we observed that the quality of up-sampled HSIs can be further improved by introducing an additional spatial-domain constraint to the conventional spectral-domain energy function. We define our spatial-domain constraint as the L_1 distance between the predicted PAN image and the actual PAN image. To estimate the PAN image of the up-sampled HSI, we also propose a learnable spectral response function (SRF). Moreover, we noticed that the residual image between the up-sampled HSI and the reference HSI mainly consists of edge information and very fine structures. In order to accurately estimate fine information, we propose a novel over-complete network, called HyperKite, which focuses on learning high-level features by constraining the receptive from increasing in the deep layers. We perform experiments on three semi-synthetic and one real HSI datasets to demonstrate the superiority of our DIP-HyperKite over the state-of-the-art pansharpening methods. The deployment codes, pre-trained models, and final fusion outputs of our DIP-HyperKite and the methods used for the comparisons will be publicly made available at https://github.com/wgcban/DIP-HyperKite.git.

***Index Terms*—Hyperspectral pansharpening, Hyperspectral image fusion, Deep Image Prior, Spatial and Spectral constraints, Over-complete representations.**

I. Introduction

HYPERSPECTRAL images (HSIs) with a large number of spectral bands have gained immense attention in the field of remote sensing due to its applications in broad research areas such as classification [1], unmixing [2], anomaly detection [3], change detection [4], etc. However, due to the limited incident energy available when capturing an image, hyperspectral imaging systems face trade-offs between spectral resolution, spatial resolution, and signal-to-noise ratio (SNR) [5]. For this reason, hyperspectral imaging systems can provide images with high spectral resolution but with low spatial resolution. In contrast, multispectral imaging systems can provide data with high spatial resolution but with fewer spectral bands (e.g., panchromatic images or multispectral images (MSIs) with three or four spectral bands). Low spatial resolution in HSIs leads to relatively poor performance in some practical remote sensing applications, such as road topology extraction [6], and spectral unmixing [7]. Therefore, full-resolution HSIs with high spatial and spectral resolution are desired. One way to obtain such ideal HSIs is to fuse high spectral resolution HSIs with high spatial resolution PAN/MSIs. This fusion process is called HS pansharpening in the remote sensing literature, which is indeed a form of super-resolution [8].

Traditional pansharpening methods can be mainly divided into five classes [5], [9]: (1) Component Substitution (CS), (2) Multi-Resolution Analysis (MRA), (3) Bayesian, (4) matrix factorization, and (5) variational. Component substitution methods rely on substituting the spatial component of the HSI with the MSI/PAN image. The family of CS contains algorithms such as Gram–Schmidt adaptive (GSA) [10], [11], principal component analysis (PCA) [12]–[14], and intensity-hue-saturation (IHS) [15]. Even though the CS methods usually generate pansharpened HSIs with accurate spatial information, sometimes they suffer from critical spectral distortions. The MRA approaches are based on injecting the spatial details obtained through the multi-scale decomposition of the MSI/PAN image into the HSI. In order to extract the spatial details from the PAN image, several algorithms have been proposed in the literature, such as decimated wavelet transform (DWT) [16], undecimated wavelet transform (UDWT) [17], smoothing filter-based intensity modulation (SFIM) [18], modulation transfer function with generalized Laplacian pyramid (MTF-GLP) [19], and MTF-GLP with high-pass modulation (MTF-GLP-HPM) [20]. In contrast to the CS methods, the MRA family performs better in spectral preservation, but is more sensitive to registration errors which may cause critical distortions in the spatial domain. Due to these inherent advantages and disadvantages of CS and MRA approaches, there have been works which attempted to combine both CS and MRA methods. One of the representatives of hybrid CS and MRA algorithm is guided filter PCA (GFPCA) [21]. The Bayesian-based methods also provide a convenient way to regularize the fusion methods by modeling the posterior distribution of the target HSI provided that the LR-HSI and MSI/PAN image. Examples of the algorithms based on the Bayesian inference framework include convex regularization under a Bayesian framework (abbreviated as Hysure) [22], naive Bayesian

This research was supported by NSF CAREER award 2045489.

Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu and Vishal M. Patel are with Whiting School of Engineering, The Johns Hopkins University, 3400 North Charles Street, Baltimore, MD 21218-2608 USA (email: wbandar1@jhu.edu; jvalana1@jhu.edu; vpatel36@jhu.edu).

Gaussian prior (abbreviated as BF) [23], and sparsity promoted Gaussian prior (abbreviated as BFS) [24]. Finally, the coupled non-negative matrix factorization (abbreviated as CNMF) is one of the examples for matrix factorization-based methods, which regularizes the fusion problem by using the priors of spectral unmixing [25]. More recently, the variational methods have gained a significant attention in HS pansharpening [26]–[28]. These methods perform pansharpening by modeling the relationship between PAN, LR-HSI, and HR-HSI images into an objective function based on some prior knowledge or certain assumptions. However, the fusion performance of traditional pansharpening approaches is generally limited due to their inadequate representation ability [29]. In addition, the algorithms mentioned above may result in severe quality degradation when the assumptions do not align with a particular dataset. Furthermore, most traditional pansharpening approaches typically reach the optimal solution through an iterative process, which is time-consuming and inefficient.

Recently, deep learning (DL) models based on convolutional neural networks (ConvNets) have also been introduced for the HS pansharpening problem due to ConvNets' excellent ability to learn high-level features automatically [8], [29]. ConvNet-based HS pansharpening methods generally consist of three steps,

1) *Up-sampling step*: Up-sampling the LR-HSI to the spatial resolution of the PAN image,
2) *Residual reconstruction step*: Concatenating the up-sampled HSI and PAN image along the spectral dimension and passing it through a residual learning network to learn the residual image,
3) *Final fusion step*: Obtaining the final fused HSI by adding the up-sampled HSI and the residual image.

There have been many methods proposed to up-sample LR-HSI to the spatial resolution of PAN. In the earliest studies, nearest-neighbor and bicubic interpolation were the famous methods to perform up-sampling. However, the methods mentioned above conduct upsampling on each band of the LR-HSI successively, thus ignoring the high spectral correlation of HSIs which may lead to spectral distortions [30], [31]. In order to minimize the spectral distortion, data-driven up-sampling techniques (i.e., deep super-resolution networks) have also been utilized in HS pansharpening. The LapSRN [32] network is an example of such a data-driven super-resolution method, which progressively super-resolves a LR image in a coarse-to-fine manner in a Laplacian pyramid framework. However, the LapSRN method requires a large number of images for training which is impractical in the HS domain due to the limited number of datasets available to the public. A remedy to the problem mentioned above was proposed by Ulyanov *et al.* [31] where they proposed a deep learning-based super-resolution framework called deep image prior (DIP). The proposed method uses a randomly initialized ConvNet to upsample an image, using its structure as an image prior, similar to bicubic upsampling. However, this method does not require any training but produces much cleaner results with sharper edges. Motivated by the super-resolution performance of DIP in the RGB domain, researchers have applied DIP to the HS pansharpening problem [30], [33] and achieved impressive results. However, we observed that the energy function defined in HS DIP up-sampling directly applies the energy function formulated for the RGB DIP process, where they only impose spectral-domain constraint by computing the L_1 distance between the down-sampled version of the target up-sampled HSI and the LR-HSI. However, the existing HS DIP methods do not impose any spatial-domain constraint by utilizing the available PAN image. We address this issue by introducing an additional spatial-domain constraint to the HS DIP process as our first contribution.

For residual reconstruction, various ConvNet architectures have been proposed in the literature to accurately predict the residual component between the up-sampled HSI and the reference HSI with less spectral and spatial distortion. Among those, Giuseppe *et al.* [34] was the first to introduce simple three-layer ConvNet architecture for the residual learning. Further, Lin *et al.* [35] improved the spatial and spectral prediction capability of Giuseppe's work (abbreviated as HyperPNN) by introducing spectral and spatial prediction modules. To further enhance the representational power of ConvNets, attention mechanisms [36] have also been introduced. Among those, Zheng *et al.* [30] proposed a spatial and spectral attention mechanism (abbreviated as DHP-DARN) for the residual learning in which they cascade several channel-spatial-attention residual blocks to adaptively learn more informative channel-wise and spatial-domain features simultaneously. More recently, Xu *et al.* [37] proposed a design (abbreviated as SDPNet) based on two encoder-decoder networks to extract deep-level features from two types of source images with densely connected blocks to strengthen feature propagation. However, we experimentally observed that most of the existing residual learning methods fail when predicting the high-frequency information, such as edges and delicate structures in the residual image. The main reason for this observation is due to the fact that the increasing receptive field of the network in the deep layers. Motivated by this observation, we introduce an over-complete network, called HyperKite, for residual reconstruction task as our second contribution, which constrains the receptive field from increasing in deep layers thus extracting more high-frequency information.

The main contributions of this paper are summarized as follows:

1) A novel spatial constraint is introduced for the DIP up-sampling process. To the best of our knowledge, this is the first study that integrates both spatial and spectral-domain constraints to the DIP up-sampling. The proposed spatial constraint significantly improves the spatial and spectral performance measures of the up-sampled HSIs.
2) An over-complete network, called HyperKite is proposed for the residual reconstruction, which is highly capable of extracting high-frequency information of the residual image by appropriately constraining the receptive field of the network.
3) We conduct extensive experiments to clearly demonstrate the improvements brought in from our contribu-

tions to the HS pansharpening. We compared the fusion performance of DIP-HyperKite with both conventional and deep learning-based approaches. The deployment codes, pre-trained models, and final fusion results of our DIP-HyperKite as well as the comparison methods in the results and discussion will be publicly made available at https://github.com/wgcban/DIP-HyperKite.git.

The rest of this paper is organized as follows. Section II provides some basics of DIP and over-complete representations. In Section III, the proposed DIP-HyperKite is described in detail. Section IV describes the datasets and performance metrics that we used in the experiments. In Section V, the experimental results on different datasets are presented. Finally, the conclusions are drawn in Section VI.

II. Related Work

A. DIP for HSI up-sampling

Generally, ConvNets have an excellent ability to learn realistic image priors from a large amount of visual data, placing them in leading positions on the benchmarks of various image processing tasks [38], [39]. Contrary to the general opinion on deep networks that they require large data to capture image priors, DIP [31] has shown that a randomly initialized network can capture low-level image statistics before any training. Concretely, in HS pansharpening, DIP can generate the up-sampled HSI $\mathbf{x}_{\text{dip}}$ of the LR-HSI $\mathbf{y}$ with spatial up-sampling factor β by taking a fixed randomly initialized vector $\mathbf{z}$ as the input, and utilizing the deep network as a parametric function $\mathbf{x}_{\text{dip}} = f_\theta(\mathbf{z})$. Next, the network is optimized over its parameters θ to obtain the up-sampled HSI $\mathbf{x}_{\text{dip}}$ as follows:

$$\mathbf{x}_{\text{dip}} = \min_{\mathbf{x}_{\text{dip}}} Q(\mathbf{x}_{\text{dip}};\mathbf{y}) + R(\mathbf{x}_{\text{dip}}), \tag{1}$$

where $Q(\mathbf{x}_{\text{dip}};\mathbf{y})$ is an energy function that controls the fidelity toward the LR-HSI $\mathbf{y}$, and $R(\mathbf{x}_{\text{dip}})$ is a regularization function based on prior knowledge. In [31], it has been shown that the regularization term $R(\mathbf{x}_{\text{dip}})$ can be implicitly substituted by the deep network. Therefore, the minimization problem in (1) has simplified to optimizing the network over its parameters θ as follows:

$$\theta^* = \arg\min_{\theta} Q(\mathbf{x}_{\text{dip}};\mathbf{y}) \text{ s.t. } \mathbf{x}_{\text{dip}} = f_\theta(\mathbf{z}), \tag{2}$$

where θ^*denotes the optimal set of parameters of the network. Furthermore, the most straightforward and commonly utilized energy function in HS pansharpening is that the L_1 distance [35] between the down-sampled version of the up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the LR-HSI $\mathbf{y}$ as follows:

$$Q(\mathbf{x}_{\text{dip}};\mathbf{y}) = \left\| d(\mathbf{x}_{\text{dip}}) - \mathbf{y} \right\|_1 \text{ s.t. } \mathbf{x}_{\text{dip}} = f_\theta(\mathbf{z}), \tag{3}$$

where $d(\cdot)$ denotes the down-sampling operator by a factor of β.

B. Over-complete ConvNets

Most of the current architectures in deep learning are "encoder-decoder" [40]–[42] based. Here, the encoder translates the high-dimensional input to a low-dimensional latent space while the decoder learns to take the latent low-dimensional representation back to a high-dimensional output. These type of architectures learn low-level features at their initial layers and high-level features at their deeper layers. These are termed under-complete networks as the input is taken to a lower spatial dimension in the latent space.

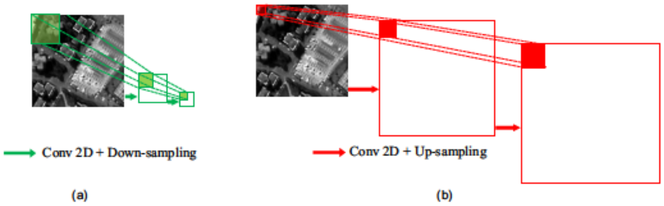

Fig. 1. (a) Effect of under-complete ConvNet on receptive field where the deeper layers focus on a larger region of the input thus extracting high-level/low-frequency information. (b) Effect of over-complete ConvNet on receptive field where the deeper layers focus on a much smaller region in the input thus extracting low-level/high-frequency information.

In signal processing, over-complete dictionaries are widely used for their highly robust characteristic [43]. The number of basis functions here are more than the number of input signal samples which enables a higher flexibility for capturing structure in data. In [44], over-complete auto-encoders were found to be better feature extractors for denoising when compared to under-complete auto-encoders. In an over-complete network [45], the encoder takes the input data to a higher spatial dimension unlike a traditional encoder. This is achieved by using an upsampling layer after every convolutional layer in the encoder. Using upsampling layers in the encoder causes the receptive field to be constrained in the deep layers. This causes the deep layers in the network to learn more fine-context high-frequency information when compared to under-complete networks. Increase in receptive field for an over-complete network can be generalized in an i^{th} layer as follows:

$$RF(\text{w.r.t } I) = \left(\frac{1}{2}\right)^{2(i-1)} \times k \times k, \tag{4}$$

where the initial receptive field of the conv filter is assumed to be $k \times k$ on the image I. This phenomenon has been visualized in Fig 1. As shown in Figure 1 (b), by employing an upsampling layer after every convolutional layer in the encoder, the over-complete network restricts the receptive field size to a smaller region which forces the network to learn very fine edges as it tries to focus heavily on smaller regions. This is completely different from the conventional under-complete architectures where they perform downsampling after each convolution block which makes the network to focus on a much larger region in the input as shown in Figure 1 (a).

Over-complete networks in deep learning is a new topic and was initially proposed for medical image segmentation of small anatomy [45]. It has since been successfully extended to solve fine-context requiring tasks like fine edge segmentation of 3D volumes [46], deep subspace clustering [47], MRI reconstruction [48], adversarial defense against videos [49] and image restoration problems like single image de-raining [50].

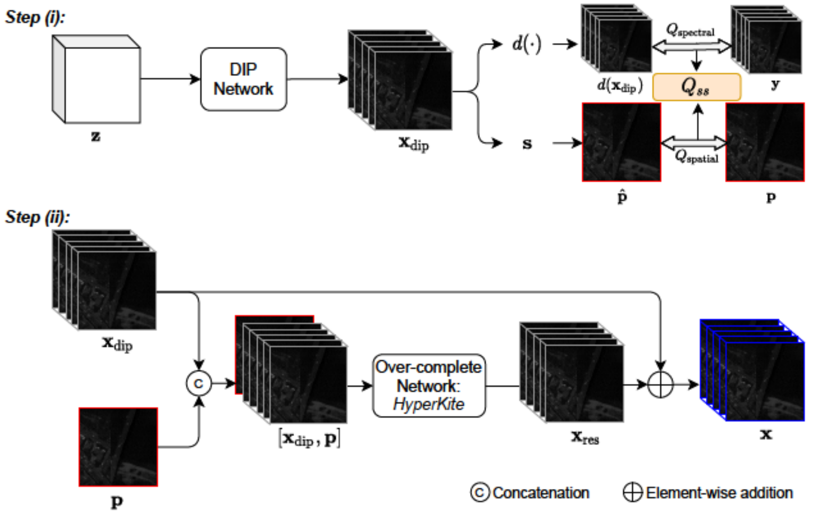

Fig. 2. The overall flowchart of our proposed DIP-HyperKite for HS pansharpening. In the first step, we up-sample the LR-HSI $\mathbf{y}$ via DIP process to obtain the up-sampled HSI $\mathbf{x}_{\text{dip}}$. The DIP process takes a fixed noise tensor $\mathbf{z}$ as input for a given LR-HSI $\mathbf{y}$, and produces the up-sampled HSI $\mathbf{x}_{\text{dip}}$ by optimizing the proposed spatial+spectral energy function Q_{ss} over the DIP network parameters θ. In the second step, we take the up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the PAN image $\mathbf{p}$ as inputs to predict the residual component $\mathbf{x}_{\text{res}}$ using our proposed over-complete network - HyperKite. Finally, the predicted residual image $\mathbf{x}_{\text{res}}$ is added to the up-sampled HSI $\mathbf{x}_{\text{dip}}$ to obtain the pansharpen HSI $\mathbf{x}$.

III. Methodology

The overall flowchart of the proposed DIP-HyperKite for HS pansharpening is shown in Figure 2. As can be seen from Figure 2 the proposed method consists of two main steps. In the first step, the LR-HSI $\mathbf{y} \in \mathbb{R}^{l \times w \times h}$ with $w \times h$ pixels and l spectral bands is up-sampled to the spatial resolution of the PAN image $\mathbf{p} \in \mathbb{R}^{1 \times \beta w \times \beta h}$, where β denotes the ratio between spatial resolution of $\mathbf{p}$ and $\mathbf{y}$. We denote the output from the DIP process as $\mathbf{x}_{\text{dip}} \in \mathbb{R}^{l \times \beta w \times \beta h}$. In the second step, we train an over-complete deep network which takes up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the corresponding PAN images $\mathbf{p}$ as inputs to predict the residual component $\mathbf{x}_{\text{res}}$ between the up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the reference HSI $\mathbf{x}_{\text{ref}}$.

A. Up-sampling via DIP

As shown in Figure 2, the low resolution HSI $\mathbf{y}$ is up-sampled to the spatial resolution of the PAN image $\mathbf{p}$ using the DIP. This recently introduced DIP method is different from the other existing up-sampling techniques such as bicubic interpolation, and LapSRN [51]. The main advantage of DIP over these conventional methods is that it does not require a large dataset for training. In other words, for each LR image $\mathbf{y}$, the DIP network takes a fixed random tensor $\mathbf{z}$ as an input and optimize the network parameters θ by minimizing the loss function Q which is defined in terms of the output up-sampled image $\mathbf{x}_{\text{dip}}$ and available LR-HSI $\mathbf{y}$ as given in (3). In contrast, the LapSRN network utilized in [52] is highly relied upon the RGB image datasets and the knowledge adaptation techniques. Furthermore, the bicubic and LapSRN methods up-sample each band in the HSI separately; thus ignoring the high spatial correlation between the spectral bands, which results in the loss of spatial details. Although the DIP method is capable of producing high-quality upsampling images compared to the other existing methods, it only utilizes the information from the LR-HSI $\mathbf{y}$, thus only imposing constraint on the spectral domain. However, we observed that the quality of the sampled HSIs can be further improved by incorporating an additional spatial constraint in the loss function using the available PAN image $\mathbf{p}$. In the next section we explain our novel spatial+spectral loss function.

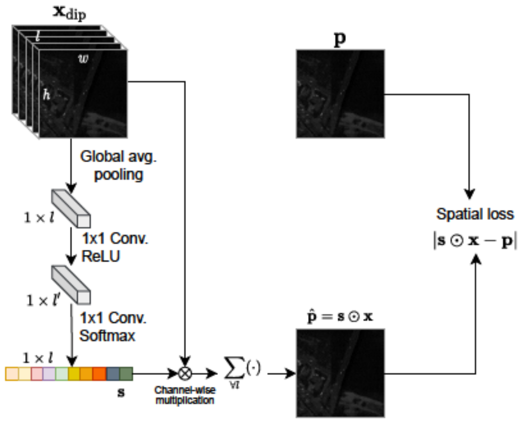
$\mathbf{x}_{\text{dip}}$
l
w
h
$\mathbf{p}$
Global avg. pooling
$1 \times l$
1x1 Conv. ReLU
$1 \times l'$
1x1 Conv. Softmax
$1 \times l$
$\mathbf{s}$
Channel-wise multiplication
$\sum_{\forall l}(\cdot)$
$\hat{\mathbf{p}} = \mathbf{s} \odot \mathbf{x}$
Spatial loss
$|\mathbf{s} \odot \mathbf{x} - \mathbf{p}|$

Fig. 3. The proposed learnable spectral response function $\mathbf{s}$, and the computational procedure of evaluating the spatial loss term Q_{spatial}. We take the up-sampled HSI $\mathbf{x}_{\text{dip}}$ as the input, and feed it in to a Global Average Pooling (GAP) layer, which yielding a vector with a single entry for each spectral band. Then we pass it through a gating mechanism by forming a bottleneck with two fully-connected (FC) layers (1×1 convolutions) around the non-linearity to learn the spectral response of each band. Next, we apply a Softmax activation function to obtain *normalized* spectral response $\mathbf{s}$, and then take the channel-wise multiplication followed by channel averaging to obtain the estimated PAN image $\hat{\mathbf{p}}$. Finally, we compute the the L_1 distance between the estimated PAN image $\hat{\mathbf{p}}$ and the reference PAN image $\mathbf{p}$ to obtain the spatial loss Q_{spatial}.

1) Proposed spatial+spectral energy function for HS DIP: As we discussed in Section II-A, the energy function given in (3) enforces a constraint only in spectral domain by defining the L_1 distance between up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the LR HSI

$\mathbf{y}$. Instead, we propose a loss function (denoted by Q_{ss}) for HS DIP, which enforces the constraints in both spatial and spectral domains as follows:

$$Q_{ss} = \underbrace{\left\| d(\mathbf{x}_{\text{dip}}) - \mathbf{y} \right\|_1}_{\text{spectral energy}} + \lambda \underbrace{\left\| \left(\sum_{i \in \forall l} \mathbf{s}[i] \odot \mathbf{x}_{\text{dip}}[i] \right) - \mathbf{p} \right\|_1}_{\text{spatial energy}}, \quad (5)$$

where $\mathbf{s} \in \mathbb{R}^{1\times l}$ denotes the spectral response function, $\mathbf{s}[i]$ (scalar) is the spectral response of i-th band, $\mathbf{x}[i] \in \mathbb{R}^{h\times w}$ is the i-th band image of the up-sampled HSI $\mathbf{x}_{\text{dip}}$, $\odot$ is the element-wise multiplication, and λ is a regularization constant. The first term in (5) enforces the spectral constraint on $\mathbf{x}_{\text{dip}}$ as in (3), and the additional second term enforces the constraint in spatial domain on $\mathbf{x}_{\text{dip}}$ by utilizing the available PAN image $\mathbf{p}$.

In the simplest case, the spectral response function can be approximated as the average across all spectral bands (i.e. $\mathbf{s}[i] = 1/l; \forall i \in [1, l]$) [53], [54]. In this scenario, the spatial loss term in (5) enforces that the average across all the spectral bands in up-sampled HSI $\mathbf{x}_{\text{dip}}$ to be close as possible to the PAN image $\mathbf{p}$, thus assuming a flat (i.e. uniform) spectral response. However, in general, this assumption is not valid as spectral response varies with wavelength coverage and different spectral bands describe the same semantic information across a wide spectral range with varying quality (i.e. PSNR) [55].

A recent attempt [55] estimates the spectral response function $\mathbf{s}$ by utilizing the larger eigenvalue of the structure tensor (ST) matrix (originally proposed in Harris corner detection algorithm [56]). However, this method cannot be directly utilized in an end-to-end deep learning network due to the difficulties encountered while performing back-propagation. In addition, it is highly computationally complex as it requires to compute derivatives of each band image along both x- and y-directions at each iteration of learning as part of constructing the structure tensor matrix. Instead, we propose a computationally lightweight and learnable spectral response function which can be easily integrated into the spatial loss term in (5) and can be simultaneously learned with DIP.

In this part we describe our novel way of estimating the spectral response function which is computationally lightweight, differentiable, and can be easily integrated into the existing DIP learning process. The overall computational procedure of estimating the spectral response function and thereby evaluating the spatial energy that we introduced for the DIP process in (5) is graphically depicted in Figure 3. First, we assume that the spectral response is proportional to the ratio of information in each spectral band. The next problem arises with this assumption is how do we quantify the information embedded in each spectral band. Motivated by recently proposed Squeeze-and-Excitation networks, we utilize global average pooling to quantify the global information present in each band. Formally, a statistic $\mathbf{q} \in \mathbb{R}^{1\times l}$ which quantifies the informative features in each spectral band is generated by shrinking the up-sampled HSI $\mathbf{x}_{\text{dip}}$ through its spatial dimensions $h \times w$ such that the i-th element in $\mathbf{q}$ is calculated as:

$$\mathbf{q}(i) = \frac{1}{h \times w} \sum_{\tilde{h}=1}^{h} \sum_{\tilde{w}=1}^{w} \mathbf{x}_i(\tilde{w}, \tilde{h}). \quad (6)$$

Next, we use a simple gating mechanism to capture the dependencies among spectral bands using the band-wise descriptor $\mathbf{q}$ that we obtained in the previous step. The gating mechanism consists of two Fully-Connected (FC) layers that give the network more flexibility to automatically learn the best spectral response function during the training process, resulting in better performance than direct normalization that does not involve any parameters to learn. We parameterize the spectral response function $\mathbf{s}$ by forming a bottleneck with two FC layers around the non-linearity as follows:

$$\mathbf{s} = \sigma(\mathbf{w}_2\, \delta(\mathbf{w}_1 \mathbf{q})), \quad (7)$$

where σ is the Sigmoid activation function, δ is the ReLU non-linearity, and $\mathbf{w}_1, \mathbf{w}_2$ are the learnable weight matrices. Here, we use Sigmoid activation to guarantee that the spectral responses of all the bands sump up to one.

2) DIP network: Figure 4 illustrates a U-Net like deep network that we used for the DIP method. The DIP network includes five down-sampling blocks $\mathbf{d}[i]$, five upsampling blocks $\mathbf{u}[i]$, and five skip-connection blocks $\mathbf{sk}[i]$ $(i = 1, 2, .., 5)$. We use stride convolutions as the down-sampling operator, bilinear up-sampling as the upsampling operator, and Lanczos2 as non-linearity. We initialize the input noise vector with uniform noise between 0 and 0.1. The Table I tabulates the values of all the hyperparameters of DIP network.

TABLE I
HYPERPARAMETER VALUES OF THE DIP NETWORK.

Hyperparameter	Value
z	$\mathbb{R}^{32\times\beta h\times\beta w} \sim U(0, 0.1)$
$n_d = n_u$	$[128, 128, 128, 128, 128]$
$k_d = k_u$	$[3, 3, 3, 3, 3]$
n_s	$[4, 4, 4, 4, 4]$
k_s	$[1, 1, 1, 1, 1]$
Optimizer	Adam
Number of iterations	1300
Learning rate	0.001
Weight decay	0.0001
Momentum	0.9
Batch size	4
LeakyReLU slope	0.2

B. Residual learning via over-complete HyperKite

Our motivation to design an over-complete network for the residual learning task emerged after observing the residual images between DIP up-sampled image $\mathbf{x}_{\text{dip}}$ and reference HSI $\mathbf{x}_{\text{ref}}$ as visualized in Figure 5. As we can see from Figure 5, the residual images correspond to different wavelength band mainly consists of boundary information like edges and other high-frequency components. In order to accurately capture this fine information, we design an over-complete HyperKite for the residual learning as shown in Figure 6.

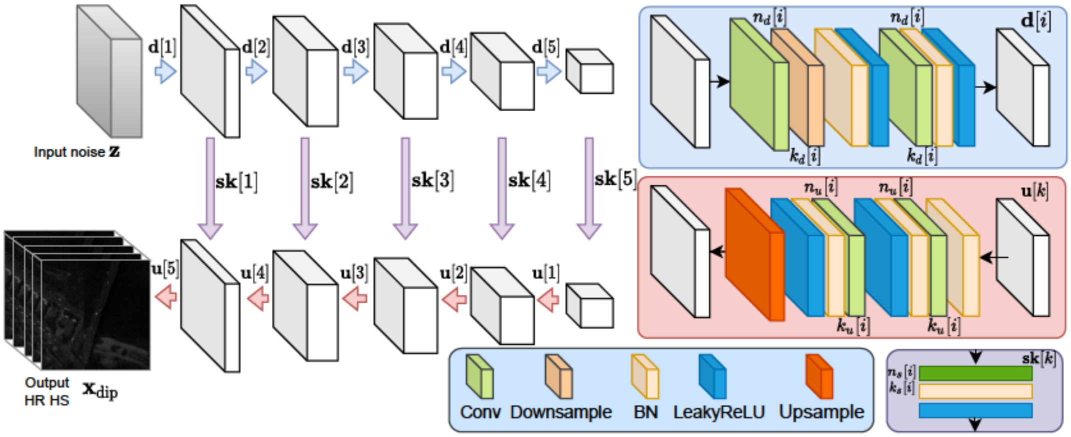

Fig. 4. The DIP network utilized for the up-sampling process. The DIP network is a U-Net like network which consists of five down-sampling blocks $\mathbf{d}[i]$, five upsampling blocks $\mathbf{u}[i]$, and five skip-connection blocks $\mathbf{sk}[i]$ $(i = 1, 2, .., 5)$. The values of all the hyperparameters of DIP network is summarized in Table I.

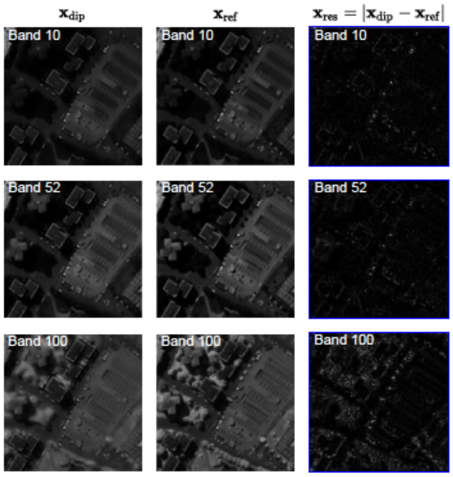
$\mathbf{x}_{\text{dip}}$
$\mathbf{x}_{\text{ref}}$
$\mathbf{x}_{\text{res}} = |\mathbf{x}_{\text{dip}} - \mathbf{x}_{\text{ref}}|$
Band 10
Band 10
Band 10
Band 52
Band 52
Band 52
Band 100
Band 100
Band 100

Fig. 5. We observed that the residual component $\mathbf{x}_{\text{res}}$ (see third column) between up-sampled HSI $\mathbf{x}_{\text{dip}}$ (see first column) and the reference HSI $\mathbf{x}_{\text{ref}}$ (see second column) mainly consists of boundary information and very fine structures. To support this observation we show the residual component $\mathbf{x}_{\text{res}}$ for three different wavelength bands (i.e. band 10, band 52, and band 100) in the Pavia Center data set which will be introduced in Section IV-A. This observation motivated us to use an over-complete network for the residual learning task, which is highly capable of learning low-level features such as fine edges and structures by transforming the input image into a higher dimension. We recommend that readers zoom in on this image to get a close-up view.

TABLE II
Hyperparameter values of HyperKite.

Hyperparameter	Value
n	$[32, 64, 128, 128, 64, 32, l]$
k	$[3, 3, 3, 3, 3, 3, 3]$
Optimizer	Adam
Num_it	2500
Learning rate	0.001
Weight decay	0.0001
Momentum	0.9
Batch size	4
LeakyReLU slope	0.2

The proposed HyperKite consists of an Initial Feature Extraction Network (IFEN), a High-dimensional Feature Mapping Network (HDFMN), and a Final Residual Reconstruction Network (FRRN). The input to the HyperKite $\mathbf{x}_{\text{in}}$ is obtained by concatenating the up-sampled HSI $\mathbf{x}_{\text{dip}}$ and the PAN image $\mathbf{p}$ along the spectral dimension (denoted as $[\mathbf{x}_{\text{dip}}, \mathbf{p}]$). The HyperKite starts with the IFEN layer, where one 3×3 convolutional layer is applied followed by Batch Normalization (BN) and LeakyReLU non-linearity to extract initial feature representation as:

$$\mathbf{F}_{\mathcal{D}_1} = f_{\text{IFEN}}(\mathbf{x}_{\text{in}}), \tag{8}$$

where $f_{\text{IFEN}}(\cdot)$ denotes the 3×3 convolution followed by LeakyReLU and batch normalization, $\mathbf{F}_{\mathcal{D}_1}$ denotes the extracted features transformed from $\mathbf{x}_{\text{in}}$ in $\mathcal{D}_1 \in \mathbb{R}^{n[0] \times \beta w \times \beta h}$ dimensional pixel-space, and $n[0]$ is the number of filters in the convolutional layer. Figure 7 (a) shows six example feature maps of $\mathbf{F}_{\mathcal{D}_1}$ for the 20-th patch of the Pavia Center dataset that we will introduce in Section IV. As we can see from the figure, the initial feature extraction network $f_{\text{IFEN}}(\cdot)$ extract low-level feature of the input $\mathbf{x}_{\text{in}}$. In order to capture high-level features that required for the residual learning, we successively transform the output of IFEN into three higher-dimensional pixel-spaces by utilizing the "bilinear" up-sampling denoted as $\mathcal{D}_2 \in \mathbb{R}^{n[2] \times 2\beta w \times 2\beta h}$, $\mathcal{D}_4 \in \mathbb{R}^{n[3] \times 4\beta w \times 4\beta h}$, and $\mathcal{D}_8 \in \mathbb{R}^{n[4] \times 8\beta w \times 8\beta h}$. Then we perform 3×3 convolution followed by BN and LeakyReLU to extract meaningful high-level features at each higher-dimensional space as:

$$\mathbf{F}_{\mathcal{D}_2} = f_{\mathcal{D}_2}(\uparrow \mathbf{F}_{\mathcal{D}_1}), \tag{9}$$

$$\mathbf{F}_{\mathcal{D}_4} = f_{\mathcal{D}_4}(\uparrow \mathbf{F}_{\mathcal{D}_2}), \tag{10}$$

$$\mathbf{F}_{\mathcal{D}_8} = f_{\mathcal{D}_8}(\uparrow \mathbf{F}_{\mathcal{D}_4}), \tag{11}$$

where $\uparrow$ denotes the "bilinear" interpolation by a factor of 2, $f_{\mathcal{D}_\text{d}}(\cdot)$: d $\in \{2, 4, 8\}$ denotes the 3×3 convolution layer followed by BN and LeakyReLU at the d-th higher-dimensional feature space. Next, we successively transform the extracted high-level features to the original dimensional space

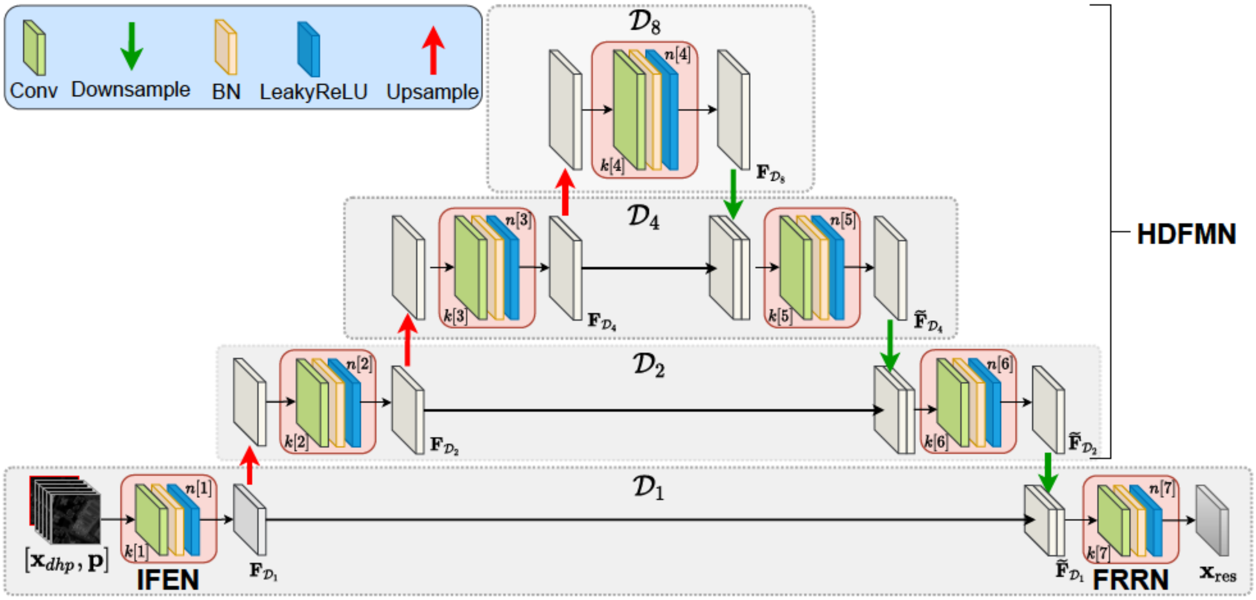
Conv
Downsample
BN
LeakyReLU
Upsample
$\mathcal{D}_8$
$n[4]$
$k[4]$
$\mathbf{F}_{\mathcal{D}_8}$
$\mathcal{D}_4$
$n[3]$
$k[3]$
$\mathbf{F}_{\mathcal{D}_4}$
$n[5]$
$k[5]$
$\widetilde{\mathbf{F}}_{\mathcal{D}_4}$
$\mathcal{D}_2$
$n[2]$
$k[2]$
$\mathbf{F}_{\mathcal{D}_2}$
$n[6]$
$k[6]$
$\widetilde{\mathbf{F}}_{\mathcal{D}_2}$
HDFMN
$\mathcal{D}_1$
$[\mathbf{x}_{dhp}, \mathbf{p}]$
$n[1]$
$k[1]$
IFEN
$\mathbf{F}_{\mathcal{D}_1}$
$\widetilde{\mathbf{F}}_{\mathcal{D}_1}$
$n[7]$
$k[7]$
FRRN
$\mathbf{x}_{\text{res}}$

Fig. 6. The proposed HyperKite architecture for the residual prediction task. We denote the kernel size and the number of filters associated with each convolution block (shown in red color box) as $k[\cdot]$ and $n[\cdot]$, respectively. The values of all hyperparameters for HyperKite is summarized in Table II.

$\mathcal{D}_1$ by employing "bilinear" downsampling and skip connections. Formally, we can define the operations of HDFMN as:

$$\widetilde{\mathbf{F}}_{\mathcal{D}_4} = \widetilde{f}_{\mathcal{D}_4}(\downarrow \mathbf{F}_8 \oplus \mathbf{F}_4), \tag{12}$$

$$\widetilde{\mathbf{F}}_{\mathcal{D}_2} = \widetilde{f}_{\mathcal{D}_2}(\downarrow \widetilde{\mathbf{F}}_{\mathcal{D}_4} \oplus \mathbf{F}_2), \tag{13}$$

$$\widetilde{\mathbf{F}}_{\mathcal{D}_1} = \downarrow \widetilde{\mathbf{F}}_{\mathcal{D}_2}, \tag{14}$$

where $\downarrow$ denotes the "bilinear" downsampling by a factor of 2, $\oplus$ denotes the feature concatenation operator, $f_{\mathcal{D}_\text{d}}(\cdot) : \text{d} \in \{0, 2, 4\}$ denotes the 3×3 convolution followed by BN and LeakyReLU at the d-th dimensional feature space, and $\widetilde{\mathbf{F}}_{\mathcal{D}_\text{d}}$ is the most relevant high-level features obtained at $\mathcal{D}_\text{d} \in \{0, 2, 4\}$ space. After flowing through all the downsampling layers (decoder blocks), a 3×3 convolutional layer is employed to recover the spectral dimension, and reconstruct the residual image $\mathbf{x}_\text{res}$ as:

$$\mathbf{x}_\text{res} = f_\text{FRNN}(\widetilde{\mathbf{F}}_{\mathcal{D}_1} \oplus \mathbf{F}_{\mathcal{D}_1}), \tag{15}$$

where f_FRNN denotes the 3×3 convolutional layer followed by BN and LeakyReLU employed at FRNN.

After carrying out DIP up-sampling and residual prediction of our DIP-HyperKite, the DIP up-sampled HSIs $\mathbf{x}_\text{dip}$ and $\mathbf{x}_\text{res}$ are created. Finally, we can obtain the fused HSI $\mathbf{x}$ by using $\mathbf{x}_\text{dip}$ and $\mathbf{x}_\text{res}$ as:

$$\mathbf{x} = \mathbf{x}_\text{res} + \mathbf{x}_\text{dip}. \tag{16}$$

To this end, we utilize L_1 loss to optimize HyperKite, which has been demonstrated as a superior choice for remote sensing image SR [5], [35] and also experimentally verified to be effective for improving the fusion accuracy. For the training set $\{\mathbf{x}_\text{in}^k, \mathbf{x}_\text{ref}^k\}^N$, where $\mathbf{x}_\text{in}^k$ is the k-th input, $\mathbf{x}_\text{ref}^k$ is the k-th reference HSI, and N is the total number of training HSIs in the training set. The L_1 loss function utilized for HyperKite training can be defined as follows:

$$L(\Theta) = \frac{1}{N} \sum_{k=1}^{N} \left\| \left(\mathbf{x}_\text{dip}^k + f_\text{HyperKite}(\mathbf{x}_\text{in}^k) \right) - \mathbf{x}_\text{ref}^k \right\|_1 . \tag{17}$$

Moreover, all the parameter details of our proposed HyperKite are summarized in Table II. We train our network in Pytorch framework using an NVIDIA Quadro 8000 GPU. We use Adam optimizer with a learning rate of 0.001, weight decay of 0.0001 and momentum 0.9 to train HyperKite. We use a batch size of 4 and train the network for 2500 epcochs.

IV. Experimental Settings

A. Datasets

To evaluate the performance of our proposed DIP-HyperKite for HS pansharpening, we conduct a series of experiments on three semi-synthetic and one real HSI datasets, which are described in detail below.

1) Pavia Center dataset: The Pavia Center scene was captured by the ROSIS camera [57]. The original HSI consists of 115 spectral bands spanning from 430 to 960 nm. The spatial size of the original image is 1096×1096 pixels, where a single pixel is equivalent to geometric resolution of 1.3×1.3 m^2. The thirteen noisy spectral bands in the original HSI were discarded, thus resulting in a HSI with 102 spectral bands spanning from 430 to 860 nm. In addition, a rectangular area of size 1096×381 pixels with no information at the center of the original HSI was also discarded, and the resulting "two-part" image with size of $1096 \times 715 \times 102$ was used for the experiments. Following the same experimental procedure outlined in [30], we also used only the top-left corner of the HSI with size of $960 \times 640 \times 102$, and partitioned it into 24 cubic patches of size $160 \times 160 \times 102$ with no overlap, which constituted the reference images ($\mathbf{x}_\text{ref}$) of Pavia Center

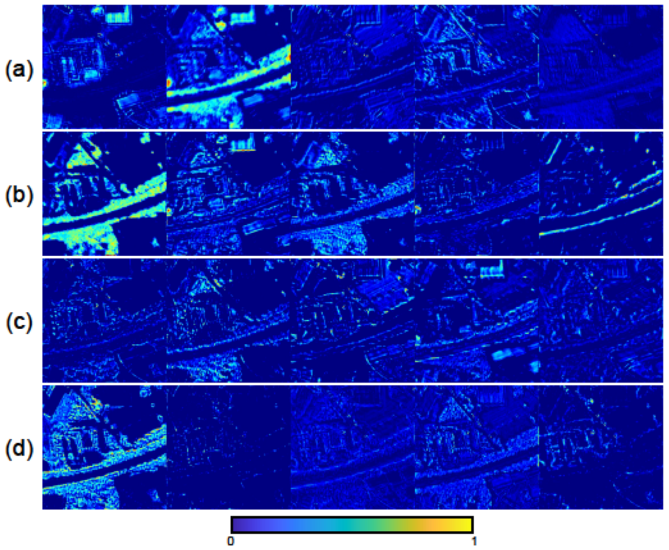

Fig. 7. Visualization of filter responses of HyperKite. (a) Feature maps from the first layer of encoder. (b) Feature maps from the second layer of encoder. (c) Feature maps from the third layer of encoder. (d) Feature maps from the third layer of encoder. By restricting the receptive field, HyperKite is able to focus on edges and smaller regions. Zoom in recommended.

data set. In order to generate PAN images ($\mathbf{p}$) and LR-HSIs ($\mathbf{y}$) corresponding to each HR-HSI, we utilize Wald's protocol [58]. Following the Wald's protocol, we generate PAN images ($\mathbf{p}$) of size 160×160 by averaging first 61 spectral bands of HR reference HSI. In order to generate LR-HSIs of size $40 \times 40 \times 102$, we spatially blurred the HR reference HSI with an 8×8 Gaussian filter, and then downsampled the result. The scaling factor (β) was set to 4 for the Pavia Center dataset. We randomly select 17 cubic patches for the training, and the rest of the seven patches forms the testing set of the Pavia Center dataset.

2) Botswana dataset: The Botswana scene was acquired by the Hyperion sensor on the NASA's Earth Observing 1 (EO-1) satellite. The original Botswana HSI consists of 242 spectral bands spanning from 400 to 2500 nm with spectral resolution of 10 nm. The spatial size of the original Botswana image is 1496×256 pixels. We remove the uncalibrated and noisy spectral bands in the original image, thus resulting in a HSI with 145 spectral bands. Following [30], we also use only the top-left corner of the HSI with size of $1200 \times 240 \times 145$, and partitioned it into 20 cubic patches of size 120×120 with no overlap, which constitute the reference images $\mathbf{x}_{\text{ref}}$ of the Botswana dataset. Next, we generate PAN images $\mathbf{p}$ of size 120×120 by averaging first 31 spectral bands of HR-HSI. We utilized same procedure mentioned for Pavia Center dataset to generate LR-HSIs $\mathbf{y}$ except we keep the down-sampling factor β as 3. We randomly select 14 cubic patches for training, and the rest of the patches are utilized for testing.

3) Chikusei dataset [59]: The Chikusei scene was captured by the Headwall Hyperspec-VNIR-C imaging sensor over the agricultural and urban areas in Chikusei, Japan. The original Chikusei HSI consists of 128 spectral bands spanning from 363 to 1018 nm. The spatial size of the Chikusei HSI is 2517×2335 pixels, where a single pixel is equivalent to geometric resolution of 2.5×2.5 m^2. We used top-left corner of the HSI with size of $2304 \times 2304 \times 128$, and partitioned it into 81 cubic patches of size $256 \times 256 \times 128$ with no overlap, which constituted the reference images $\mathbf{x}_{\text{ref}}$ of Chikusei dataset. Next, we generate PAN images and LR-HSIs following the same procedure mentioned for Pavia Center dataset. We randomly select 61 cubic patches for training, and the rest of the patches are utilized for testing.

4) Los Angeles Dataset: The Los Angeles dataset is a real HSI dataset that was acquired over a port in the city of Los Angeles. This dataset consists of LR-HSI and the HR-PAN image which were captured by the Hyperion sensor on the EO-1 satellite and the advanced land imager (ALI), respectively. The original LR-HSI consists of 242 spectral bands with a spatial resolution of 30 m. The spatial resolution of PAN image is 10 m. During the pre-processing, the uncalibrated and noisy spectral bands of LR-HSI were removed, thus resulting in a LR-HSI with 145 spectral bands for experimentation. Therefore, the well-trained model on the Botswana data set can be generalized to the Los Angeles data set. The size of the LR-HSI is $120 \times 120 \times 145$, and the size of the experimental PAN image is 360×360.

Note: The standard deviation (σ) of the Gaussian filter that we use to generate LR-HSIs is calculated as [30], [60]:

$$\sigma = \sqrt{\frac{1}{2 \times 2.7725887/\beta^2}} = 0.4247\beta. \tag{18}$$

B. Performance measures

In order to evaluate the quality of the proposed pansharpening method, we use different image quality measures. Following [30], we use reference-based metrics such as Cross-Correlation (CC), Spectral Angle Mapping (SAM), Root Mean Square Error (RMSE), Errur Relative Globale Adimensionnelle Desynthese (ERGAS), and Peak Signal to Noise Ratio (PSNR) to evaluate pansharpening performance on semi-synthetic datasets where the reference HSI is available. For the real HSI dataset where the reference image is not available to evaluate the above performance measures, we adopt no-reference based performance metrics such as spectral distortion (D_λ), spatial distortion (D_S), and Quality with No-Reference (QNR). These measures have been widely used in the HSI processing community and are appropriate for evaluating fusion in spectral and spatial resolutions.

1) CC: The CC metric characterizes the geometric distortion, and is defined as:

$$\text{CC}(\mathbf{x}, \mathbf{x}_{\text{ref}}) = \frac{1}{l} \sum_{i=1}^{l} \text{CCS}(\mathbf{x}^i, \mathbf{x}^i_{\text{ref}}), \tag{19}$$

where CCS denotes the cross-correlation for a single-band image as follows:

$$\text{CCS}(\mathbf{A}, \mathbf{B}) = \frac{\sum_{j=1}^{n} (\mathbf{A}_j - \mu_A)(\mathbf{B}_j - \mu_B)}{\sqrt{\sum_{j=1}^{n} (\mathbf{A}_j - \mu_A)^2 \sum_{j=1}^{n} (\mathbf{B}_j - \mu_B)^2}}, \tag{20}$$

where n is the total number of pixels in the image, and $\mu_A = \frac{1}{n} \sum_{j=1}^{n} \mathbf{A}_j$ is the sample mean of $\mathbf{A}$. The ideal value of CC is 1.0, which indicates that the two HSIs are highly correlated.

2) SAM: SAM is a spectral measure which is defined

$$\mathrm{SAM}(\mathbf{x}, \mathbf{x}_{\text{ref}}) = \frac{1}{n} \sum_{j=1}^{n} \mathrm{SAM}(\mathbf{x}_j, \mathbf{x}_{\text{ref}\,j}),$$

where given the vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^l$,

$$\mathrm{SAM}(\mathbf{a}, \mathbf{b}) = \arccos\left(\frac{< \mathbf{a}, \mathbf{b} >}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}\right), \tag{22}$$

where $< \mathbf{a}, \mathbf{b} >$ denotes the inner product between $\mathbf{a}$ and $\mathbf{b}$, and $\|\cdot\|_2$ is the L_2 norm. The SAM is a measure of spectral shape preservation. The SAM values reported in experiments are in degrees and thus belongs to $(-90, 90]$. optimal value of SAM is 0.0. The values of SAM reported in our experiments have obtained by averaging the values for all image pixels.

3) RSNR/ RMSE: The reconstruction SNR (RSNR) or root mean square error (RMSE) is related to the difference between the reference and fuse images, which is defined as follows:

$$\mathrm{RMSE}(\mathbf{x}, \mathbf{x}_{\text{ref}}) = \frac{1}{n \times l} \|\mathbf{x} - \mathbf{x}_{\text{ref}}\|_F^2, \tag{23}$$

$$\mathrm{RSNR}(\mathbf{x}, \mathbf{x}_{\text{ref}}) = 10 \log_{10}\left(\frac{\|\mathbf{x}_{\text{ref}}\|_F^2}{\|\mathbf{x} - \mathbf{x}_{\text{ref}}\|_F^2}\right). \tag{24}$$

4) ERGAS: Relative dimensionless global error in synthesis (ERGAS) calculates the amount of spectral distortion in the image. The ERGAS measure is defined as:

$$\mathrm{ERGAS} = 100 \frac{1}{d^2} \sqrt{\frac{1}{l} \sum_{i=1}^{l} \left(\frac{\mathrm{RMSE}(\mathbf{x}^i, \mathbf{x}_{\text{ref}}^i)}{\mu(\mathbf{x}_{\text{ref}}^i)}\right)}, \tag{25}$$

where d is the ratio between the linear resolution of the PAN image and the HSIs. defined as:

$$d = \frac{\text{PAN linear spatial resolution}}{\text{HS linear spatial resolution}}, \tag{26}$$

where $\mathrm{RMSE}(\mathbf{x}^i, \mathbf{x}_{\text{ref}}^i) = \frac{\|\mathbf{x}^i - \mathbf{x}_{\text{ref}}^i\|_F}{\sqrt{n}}$, and $\mu(\mathbf{x}_{\text{ref}}^i)$ is the sample mean of the i-th band of $\mathbf{x}_{\text{ref}}$. The ideal value of ERGAS is 0.

5) PSNR: PSNR also assess the fusion quality of each spectral bands, and the average PSNR is calculated as:

$$\mathrm{PSNR} = \frac{1}{l} \sum_{i=1}^{l} \left[10 \log_{10}\left(\frac{\max\left(\mathbf{x}_{\text{ref}}^i\right)}{\mathrm{RMSE}(\mathbf{x}^i, \mathbf{x}_{\text{ref}}^i)}\right)^2\right], \tag{27}$$

where $\max\left(\mathbf{x}_{\text{ref}}^i\right)$ is the maximum pixel value in the i-th band of $\mathbf{x}_{\text{ref}}$. A larger value of PSNR indicates a higher reconstruction quality in spatial information of the fusion result.

6) QNR: To quantitatively evaluate the pansharpening performance on real data sets which do not have reference HSI, we adopt the Quality with No-Reference (QNR) metric [30]. The QNR metric can be defined as:

$$QNR = (1 - D_\lambda)^\eta (1 - D_S)^\rho, \tag{28}$$

where D_λ defined the amount of spectral distortion, D_S quantifies the amount of spatial distortion, and η and ρ are two coefficients (usually set to 1). The amount of spectral distortion D_λ is calculated as:

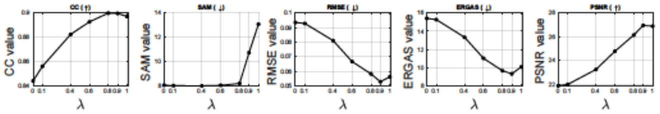

Fig. 8. The variation of CC, SAM, RMSE, ERGAS, and PSNR with the regularization constant λ in our spectral+spectral energy function Q_{ss} for Pavia Center dataset. We select $\lambda = 0.8$ as the optimal value of regularization constant for the Pavia Center dataset by considering all the performance metrics.

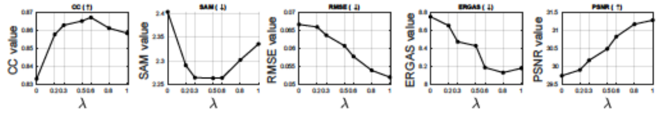

Fig. 9. The variation of CC, SAM, RMSE, ERGAS, and PSNR with the regularization constant λ in our spectral+spectral energy function Q_{ss} for Botswana dataset. We select $\lambda = 0.8$ as the optimal value of regularization constant for the Botswana dataset by considering all the performance metrics.

$$D_\lambda = \sqrt[\epsilon]{\frac{1}{l(l-1)} \sum_{q_1=1}^{l} \sum_{q_2=1, q_2 \neq q_1}^{l} \left\|Q_{\mathbf{y}}^{q_1,q_2} - Q_{\mathbf{x}}^{q_1,q_2}\right\|^\epsilon}, \tag{29}$$

where parameter ϵ is usually set to 1. $Q_{\mathbf{y}}^{q_1,q_2} = Q(\mathbf{y}^{q_1}, \mathbf{y}^{q_2})$ and $Q_{\mathbf{x}}^{q_1,q_2} = Q(\mathbf{x}^{q_1}, \mathbf{x}^{q_2})$ are denote the Q-*index* [61] which calculates the dissimilarities between couples of spectral bands for LR-HSI $\mathbf{y}$ and pansharpen HSI $\mathbf{x}$. The spatial distortion D_S is calculates as:

$$D_S = \sqrt[\delta]{\frac{1}{l} \sum_{q=1}^{b} \left\|Q(\mathbf{x}^q, \mathbf{p}) - Q(\mathbf{y}^q, \mathbf{p}_{\text{lr}})\right\|^\delta}, \tag{30}$$

where δ is typically set to 1, and $\mathbf{p}_{\text{lr}}$ denotes the simulated LR-PAN image with the same size of LR-HSI. The ideal values of D_λ, D_S, and QNR are 0, 0, and 1, respectively.

V. Results and Discussion

This section presents the results of our proposed DIP-HyperKite for HS pansharpening, and compares it with the state-of-the-art methods on the Pavia Center, Botswana, and Chikusei datasets. For better clarity, we divide this section into two parts. In the first part (section V-A), we highlight the contribution from our proposed spatial+spectral energy function for the DIP up-sampling process and compare it with available state-of-the-art up-sampling techniques such as nearest-neighbor, bicubic, LapSRN, and DIP with only spectral loss. In the second part (section V-B), we present the final fusion results that we obtain from our proposed HyperKite network and compare it with classical and deep-learning-based pansharpening approaches.

A. Effect of the proposed spatial+spectral energy function for the DIP up-sampling process

As we discussed in Section III-A, the recently proposed pansharpening methods such as DHP-DARN [30] and DHP

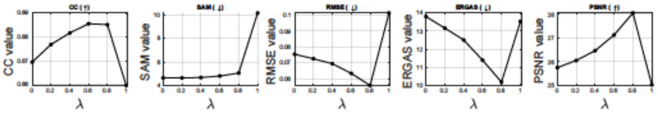

Fig. 10. The variation of CC, SAM, RMSE, ERGAS, and PSNR with the regularization constant λ in our spectral+spectral energy function Q_{ss} for Chikusei dataset. We select $\lambda = 0.8$ as the optimal value of regularization constant for the Chikusei dataset by considering all the performance metrics.

TABLE III
AVERAGE QUANTITATIVE RESULTS FOR DIFFERENT UP-SAMPLING TECHNIQUES ON THE PAVIA CENTER DATASET.

Method	CC	SAM	RMSE $\times 10^{-1}$	RSNR	ERGAS	PSNR
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
Nearest-neighbor	0.809	7.70	1.22	9.63	19.97	19.65
Bicubic	0.840	7.45	1.13	11.26	18.48	20.36
LapSRN [32]	0.843	**7.37**	1.12	11.56	18.16	20.49
DIP+spectral [30]	0.844	8.04	0.94	14.91	15.42	21.89
DIP+Q_{ss} (ours)	**0.900**	8.18	**0.58**	**24.36**	**9.66**	**26.15**
	(+6.6%)	-	(-37.8%)	(+63.4%)	(-37.3%)	(+19.5%)

TABLE V
AVERAGE QUANTITATIVE RESULTS FOR DIFFERENT UP-SAMPLING TECHNIQUES ON THE CHIKUSEI DATASET.

Method	CC	SAM	RMSE $\times 10^{-2}$	RSNR	ERGAS	PSNR
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
Nearest-neighbor	0.861	4.05	9.99	18.26	17.03	23.73
Bicubic	0.884	3.86	9.31	20.07	15.75	24.52
LapSRN [32]	0.885	**3.75**	8.53	21.37	14.33	25.06
DIP+spectral [30]	0.869	4.64	7.54	24.16	13.80	25.75
DIP+Q_{ss}(ours)	**0.885**	5.05	**5.56**	**29.69**	**10.18**	**28.06**
	(+0.1%)	-	(-26.3%)	(+22.9%)	(-26.2%)	(+8.9%)

[33] utilized the DIP process to up-sample the LR-HSI instead of using the nearest-neighbor, bicubic, or LapSRN techniques due to its excellent performance. However, we have observed that the quality of up-sampled HSI can be further improved by carefully redesigning the loss function used in the DIP optimization. Instead of only utilizing spectral constraint in the DIP loss function, we derived a novel loss function with spectral and spatial constraints. This section demonstrates the performance improvement brought by our proposed spatial+spectral loss function to the DIP up-sampling process. We compare DIP with the proposed spatial+spectral loss against the DIP with spectral loss only. Furthermore, to make the analysis more comprehensive, we also added a conventional up-sampling techniques used in the HS pansharpening domain, such as nearest-neighbor, and bicubic. Further, motivated by the experimental discussion in [30], we also added the results from LapSRN [32], which is trained on a large amount of RGB images.

1) Tuning the hyperparameter λ in our spatial+spectral energy function: We start our discussion with the effect of the regularization constant λ in our proposed spatial+spectral loss function as defined in (5). The variation of CC, SAM, RMSE, ERGAS and PSNR values when varying the regularization parameter λ from 0.0 to 1.0 for the Pavia Center, Botswana and Chikusei datasets are shown in Figure 8, Figure 9, and Figure 10, respectively. As can be seen from these figures, as the value of the regularization constant λ increases, the performance metrics also begin to improve, then hit a saturation point, and then degrade, for all three data sets. Therefore, we carefully select the regularization constant λ for each dataset by considering all the performance metrics. For example, consider the variation of the performance metrics with the regularization constant λ for the Pavia Center dataset which is shown in Figure 8. As we can see, when the value of the regularization constant increases from 0.0 to 0.8, we can see that CC, RMSE, ERGAS, and PSNR start to improve, and when λ increases beyond 0.8 the performance metrics start to degrade. Therefore, we set $\lambda = 0.8$ as the optimal value of the regularization constant of our proposed spatial+spectral energy term for the Pavia Center dataset. The variation in performance metrics with the regularization parameter λ for the Botswana and Chikusei datasets are also shown in Figure 9 and Figure 10, respectively. Following the same analysis we described for the Pavia Center dataset, we select $\lambda = 0.8$ as the optimal value of the regularization constant for the Botswana and Chikusei datasets. Note that the performance improvement bringing from our proposed spatial+spectral loss function for the DIP upsampling process. Under the optimal regularization constant ($\lambda = 0.8$), our spatial+spectral energy function improves the quality of up-sampled HSIs over the spectral loss (equivalent to $\lambda = 0$ point in Figure 8, 9, and 10) in-terms of CC, RMSE, ERGAS, and PSNR metrics by 6.64%, 37.8%, 63.4%, 37.3%, and 19.5%, respectively for the Pavia Center dataset. For the Botswana dataset, our proposed loss function improves CC, SAM, RMSE, ERGAS, and PSNR metrics over the DIP with spectral loss by 3.3%, 4.2%, 19.1%, 14.7%, 7.0%, and 5.2%, respectively. Similarly for the Chikusei dataset, our method improves CC, RMSE, ERGAS, and PSNR metrics compared to DIP with spectral loss by 1.8%, 26.3%, 22.9%, 26.2%, and 8.9%, respectively.

TABLE IV
AVERAGE QUANTITATIVE RESULTS FOR DIFFERENT UP-SAMPLING TECHNIQUES ON THE BOTSWANA DATASET.

Method	CC	SAM	RMSE $\times 10^{-2}$	RSNR	ERGAS	PSNR
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
Nearest-neighbor	0.854	2.52	7.87	29.03	9.08	28.87
Bicubic	0.852	2.42	7.67	29.60	8.77	29.17
LapSRN [32]	0.858	2.47	6.27	34.01	8.27	29.01
DIP+spectral [30]	0.833	2.40	6.66	32.91	8.75	29.75
DIP+Q_{ss} (ours)	**0.861**	**2.30**	**5.39**	**37.80**	**8.13**	**31.28**
	(+0.4%)	(-4.2%)	(-14.0%)	(+11.2%)	(-1.7%)	(+5.2%)

Discussion on the regularization constant λ: Let us first consider the case where the regularization constant λ is set to zero. This is equivalent to the case where we only have the spectral constraint. In this case, the DIP network minimizes the distance between the down-sampled version of the up-sampled HSI and the LR-HSI. Since the down-sampling operator acts as a low-pass filter in the frequency domain, what DIP network actually minimizes is that the distance between the low-pass version of the up-sampled HSI and the LR-HSI. Because of this reason, the up-sampled HSI from the DIP network trained only with spectral constraint lacks

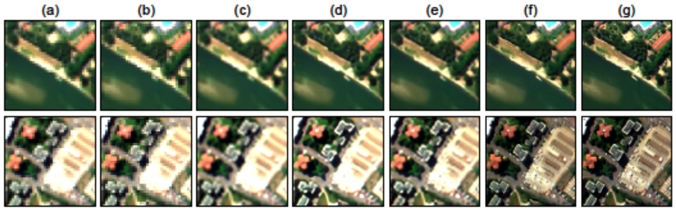

Fig. 11. Up-sampled images of 1-st patch (in 1-st row) and 11-th patch (in 2-nd row) of Pavia Center dataset. (a) LR-HSI. (b) Nearest-neighbor. (c) Bicubic. (d) LapSRN [32]. (e) DIP with only spectral energy [30]. (f) DIP with our spatial+spectral energy ($Q_{ss}; \lambda = 0.8$). (g) Reference. The RGB image is generated by utilizing the 10-th, 30-th, and 60-th bands of the HSI for blue, green and red bands, respectively.

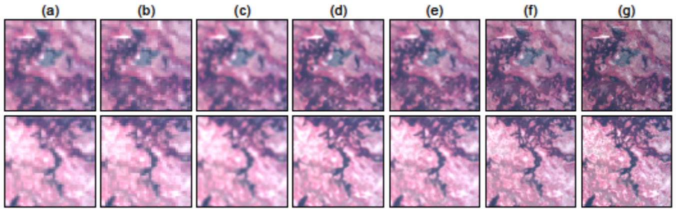

Fig. 12. Up-sampled images of 12-th (in first row) and 14-th (in second row) patch of Botswana dataset. (a) LR-HSI. (b) Nearest-neighbor. (c) Bicubic. (d) LapSRN [32]. (e) DIP with only spectral energy [30]. (f) DIP with our spatial+spectral energy ($Q_{ss}; \lambda = 0.8$). (g) Reference. The RGB image is generated by utilizing the 10-th, 35-th, and 61-th bands of the HSI for blue, green and red bands, respectively.

the high frequency components such as edge information and fine structures. Now let us consider the case where we have both spatial and spectral constraint in the DIP loss function. As we described in Section III-A, we combined the spatial and spectral constraints via regularization parameter λ. The value of λ controls the fidelity of the predicted PAN image towards the actual PAN image. Since the predicted PAN image and the up-sampled HSI are coupled via spectral response function, to make the predicted PAN image close as possible to the actual PAN image, the DIP network tries to predict some of the high-frequency components such as edges and fine structures in the PAN image, while maintaining the low-pass version of the up-sampled HSI close to the LR-HSI. Therefore, the regularization constant what actually controls is the amount of high-frequency components fused from PAN image to the up-sampled HSI. This explain the observation that we made from Figure 8, Figure 9, and Figure 10, where when the value of the regularization parameter increases the DIP network embed some of the high-frequency information to the up-sampled HSI, which ultimately helps to improve the quality of the up-sampled image. However, when the value of the regularization constant is large, the spatial loss term starts to dominate the loss function, and resulting in drop of spectral-domain performance metrics such as SAM and ERGAS. Therefore, we can achieve high-quality up-sampled HSIs by appropriately controlling the regularization parameter in spatial+spectral energy function.

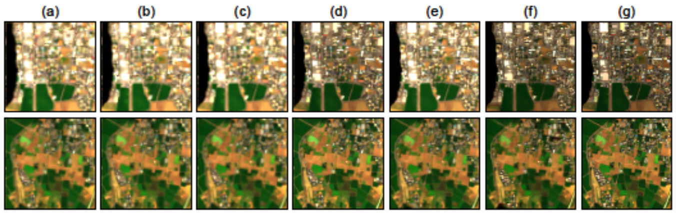

Fig. 13. Up-sampled images of 37-th (in first row) and 50-th (in second row) patch of Chikusei dataset. (a) LR-HSI. (b) Nearest-neighbor. (c) Bicubic. (d) LapSRN [32]. (e) DIP with only spectral energy [30]. (f) DIP with our spatial+spectral energy (Q_{ss} ($\lambda = 0.8$). (g) Reference. The RGB image is generated by utilizing the 12-th, 20-th, and 29-th bands for blue, green and red bands, respectively.

2) Comparison of DIP with the proposed spatial+spectral loss with state-of-the-art up-sampling techniques: In the previous section, we determined the optimal value of the regularization constant λ for our proposed spatial+spectral loss function for the three datasets. In this section, we compare DIP with our spatial+spectral loss against DIP with only spectral loss, and other commonly used up-sampling techniques such as nearest neighbor, bicubic, and LapSRN, both qualitatively and quantitatively.

Table III summarizes the quantitative results of nearest-neighbor, bicubic, LapSRN, and DIP up-sampling methods for the Pavia Center dataset. For this dataset, our proposed DIP method improves the quality of up-sampled images in terms of CC, SAM, RSNR, ERGAS, and PSNR performance measures by 6.6%, 37.3%, 63.4% 37.3%, and 19.5%, respectively. We have also noticed that this improvement is accompanied by a drop in the SAM index which is around 1.8% compared to the DIP with spectral loss. This fall in the SAM index is not that significant compared to the improvements we have achieved in terms of all other performance measures. Further, we can cross-verify these quantitative results with qualitative results that we have shown in Figure 11 for the Pavia Center dataset. We can see that the DIP up-sampled images with our proposed spatial+spectral constraint looks much more closer to the reference image, and have predicted very fine structures and edges compared to other upsampling methods.

We also summarize the quantitative results for different up-sampling methods for the Botswana dataset in Table IV. As we can see, DIP with the proposed spatial+spectral loss improves the quality of up-sampled images in terms of all the performance metrics by a significant margin: CC value increased by 0.4%, SAM value reduced by 4.3%, RMSE value reduced by 14.0%, RSNR value improved by 11.2%, ERGAS value reduced by 1.7%, and PSNR value value increased by 5.2%. Also, we can verify these quantitative results with the qualitative results shown in Figure 12 for the Botswana dataset. Similar to the qualitative results that we have observed for the Pavia Center dataset, we can see the the up-sampled images using DIP with our proposed spatial+spectral loss is much more closer to the reference HSI.

Finally, we summarize the quantitative results for different upsampling methods for the Chikusei dataset in Table V. In this case also, the performance of DIP up-sampled images with our proposed spatial+spectral loss outperforms five out of six performance measures that we considered for the analysis. As we can see from the Table V, our DIP method has increased the value of CC by 0.1%, has decreased the value of RMSE by 26.3%, has increased the RSNR by 22.9%, has

TABLE VI
THE AVERAGE QUANTITATIVE RESULTS ON THE PAVIA CENTER DATASET.

Method	CC	SAM	RMSE	RSNR	ERGAS	PSNR
			$\times10^{-2}$			
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
PCA [13]	0.845	8.92	3.45	34.32	6.64	31.26
GFPCA [21]	0.902	8.31	3.98	29.34	7.44	29.09
BF [23]	0.918	9.60	3.44	31.99	6.63	30.22
BFS [24]	0.925	8.10	3.05	34.37	6.00	31.09
SFIM [18]	0.946	6.76	2.55	37.47	5.43	32.61
GS [11]	0.961	6.62	2.55	38.08	4.95	32.93
GSA [11]	0.950	7.15	2.34	39.60	4.70	33.52
MTF-GLP-HPM [20]	0.955	6.81	2.25	40.70	4.77	33.97
CNMF [25]	0.960	6.64	2.20	40.79	4.39	34.14
MTF-GLP [19]	0.956	6.55	2.20	40.70	4.45	34.12
HySure [22]	0.966	6.13	1.80	44.60	3.77	35.91
HyperPNN [35]	0.967	6.09	1.67	48.62	3.82	36.70
DHP-DARN [30]	0.969	6.43	1.56	49.17	3.95	37.30
DIP-HyperKite (ours)	**0.980**	**5.61**	**1.29**	**51.72**	**2.85**	**38.65**

TABLE VII
THE AVERAGE QUANTITATIVE RESULTS ON THE BOTSWANA DATASET.

Method	CC	SAM	RMSE	RSNR	ERGAS	PSNR
			$\times10^{-2}$			
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
PCA [13]	0.946	2.22	1.74	57.10	2.89	28.17
GFPCA [21]	0.925	2.48	1.97	53.81	3.18	26.75
BF [23]	0.919	2.41	1.86	55.43	3.37	26.88
BFS [24]	0.918	2.39	1.85	55.52	3.38	26.91
SFIM [18]	0.890	3.31	2.56	48.30	2.98	27.27
GS [11]	0.949	2.17	1.68	57.55	2.74	28.32
GSA [11]	0.964	1.86	1.28	63.02	2.16	30.78
MGH [20]	0.962	1.90	1.33	62.23	2.15	30.47
CNMF [25]	0.951	2.28	1.38	60.90	2.48	29.63
MG [19]	0.963	1.88	1.32	62.23	2.16	30.45
HySure [22]	0.963	1.93	1.19	63.80	2.12	30.97
HyperPNN [35]	0.957	1.92	1.06	66.22	2.40	29.00
DHP-DARN [30]	0.954	1.91	1.05	66.22	2.35	29.98
DIP-HyperKite (ours)	**0.974**	**1.68**	**0.96**	**67.98**	**1.89**	**32.12**

decreased the ERGAS by 26.2%, and has increased the PSNR by 8.9% over the state-of-the-art results. Similar to the Pavia Center dataset, in this dataset also we have observed that the drop in SAM index; however this is negligible compared to the performance gained in terms of the other quantitative measures. Furthermore, following the similar trend with other datasets, we have included the qualitative results in Figure 13 for the Chikusei dataset. From the qualitative results also we can see that DIP with our spatial+spectral constraint is able to predict very fine structures and edges more accurately than the other methods.

In summary, we have shown that the DIP method with our proposed spatial+spectral constraints outperforms the state-of-the-art up-sampling methods with a significant margin in all the datasets that we have considered in this study. In the next section, we present final fusion results and compare them with state-of-the-art pansharpening algorithms, qualitatively and quantitatively.

B. Final fusion results on semi-synthetic HS datasets:

In this section we compare final fusion results from our DIP-HyperKite with the state-of-the-art pansharpening approaches such as PCA [13], GFPCA [21], BF [23], BFS [24], SFIM [18], GS [11], GSA [11], MTF-GLP-HPM [20], CNMF [25], MTF-GLP [19], HySure [22], HyperPNN [35], and DHP-DARN [30] for Pavia Center, Botswana, and Chikusei datasets.

TABLE VIII
THE AVERAGE QUANTITATIVE RESULTS ON THE CHIKUSEI DATASET.

Method	CC	SAM	RMSE	RSNR	ERGAS	PSNR
			$\times10^{-2}$			
	(↑)	(↓)	(↓)	(↑)	(↓)	(↑)
GFPCA [21]	0.883	4.76	1.98	34.22	7.00	37.05
BF [23]	0.903	5.15	1.94	34.40	6.62	37.89
BFS [24]	0.917	4.69	1.72	36.84	6.39	37.99
SFIM [18]	0.928	3.79	1.43	40.51	6.43	39.55
GS [11]	0.733	5.64	2.96	26.37	8.17	35.13
GSA [11]	0.943	3.52	1.42	40.73	4.30	41.38
MTF-GLP-HPM [20]	0.929	3.82	1.45	39.38	6.40	39.85
CNMF [25]	0.900	4.72	1.91	36.73	5.75	39.65
MTF-GLP [19]	0.938	3.81	1.52	39.38	4.41	41.05
HySure [22]	0.960	2.98	1.13	45.24	3.69	43.14
HyperPNN [35]	0.946	3.97	1.11	46.55	4.77	41.57
DHP-DARN [30]	0.953	3.60	1.05	46.66	4.44	42.24
DIP-HyperKite (ours)	**0.974**	**2.85**	**1.03**	**46.97**	**3.62**	**43.53**

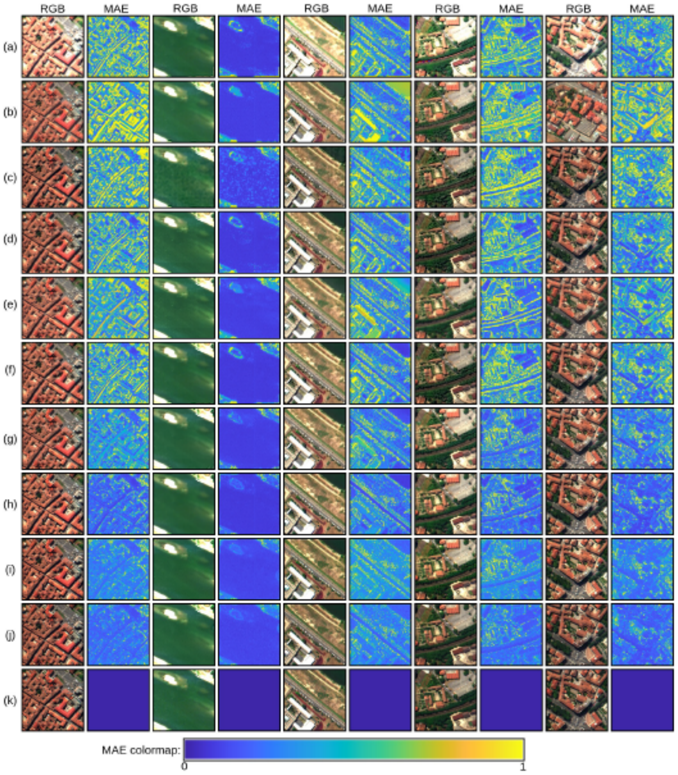

Fig. 14. Visual results generated by different pansharpening algorithms for the first (in first and second column), third (in third and fourth column), 12-th (in fifth and sixth column), 20-th (in seventh and eight column), 21-st (in nine and tenth column) patches of the Pavia Center dataset. (a) SFIM [18]. (b) GS [11]. (c) GSA [11]. (d) MTF-GLP-HPM [20]. (e) CNMF [25]. (f) MTF-GLP [19]. (g) HySure [22]. (h) HyperPNN [35]. (i) DHP-DARN [30]. (j) DIP-HyperKite (ours). (k) Reference.

a) Final Fusion results on the Pavia Center dataset: The average quantitative results for different pansharpening approaches on the testing set of the Pavia Center dataset are shown in Table VI. As can be seen from Table VI, our proposed HyperKite achieves the highest CC value compared to all the other pansharpening approaches that we have considered in this study. A higher CC value indicates that the fused HSI is closer to the actual HSI with less geometric

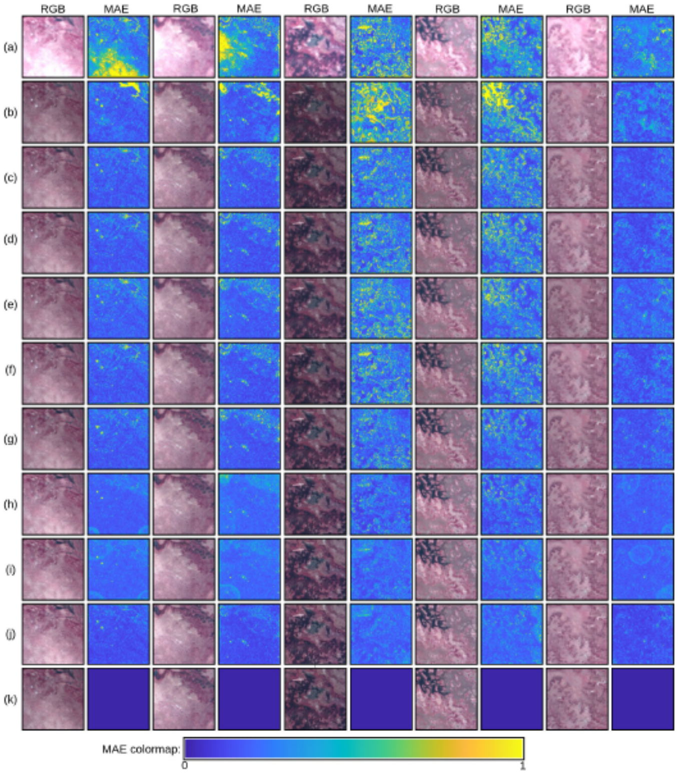

Fig. 15. Visual results generated by different pansharpening algorithms for the first (in first and second column), fourth (in third and fourth column), 12-th (in fifth and sixth column), 16-th (in seventh and eight column), 19-st (in nine and tenth column) patches of the Botswana dataset. (a) SFIM [18]. (b) GS [11]. (c) GSA [11]. (d) MTF-GLP-HPM [20]. (e) CNMF [25]. (f) MTF-GLP [19]. (g) HySure [22]. (h) HyperPNN [35]. (i) DHP-DARN [30]. (j) DIP-HyperKite (ours). (k) Reference.

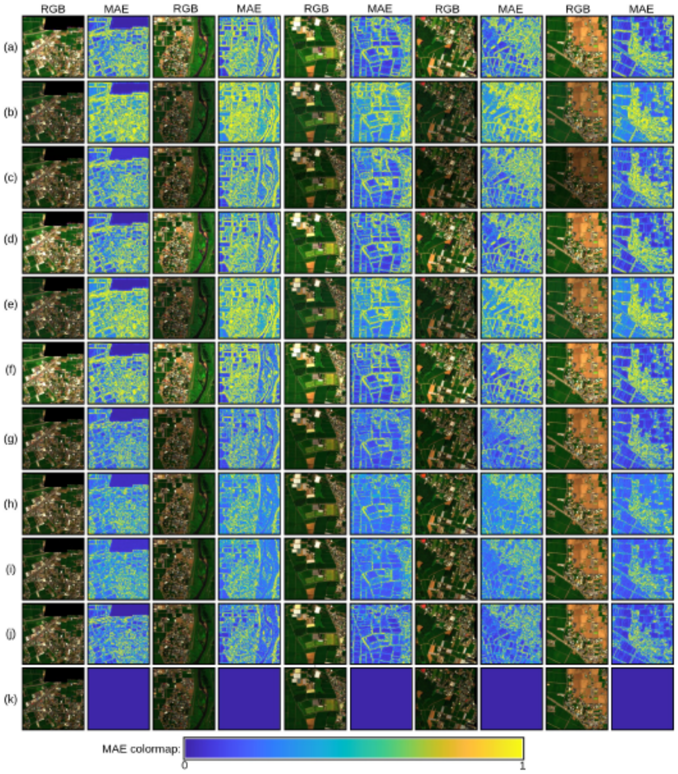

Fig. 16. Visual results generated by different pansharpening algorithms for the fifth (in first and second column), 13-th (in third and fourth column), 16-th (in fifth and sixth column), 27-th (in seventh and eight column), 32-nd (in nine and tenth column) patches of the Chikusei dataset. (a) SFIM [18]. (b) BF [23]. (c) GSA [11]. (d) MTF-GLP-HPM [20]. (e) BFS [24]. (f) MTF-GLP [19]. (g) HySure [22]. (h) HyperPNN [35]. (i) DHP-DARN [30]. (j) DIP-HyperKite (ours). (k) Reference.

distortion. Furthermore, our proposed DIP-HyperKite achieved the smallest values for SAM, RMSE, and ERGAS performance measures, indicating the best fusion performance over the other pansharpening approaches. Especially the smallest SAM and ERGAS indicate that our DIP-HyperKite can fuse HSIs with less spectral distortion than the state-of-the-art methods. In addition, our DIP-HyperKite improved the PSNR metric by 3.6% over the state-of-the-art value. To further verify the fusion quality of our proposed DIP-HyperKite, we present qualitative results in Figure 14 for the Pavia Center dataset. To better highlight the fusion quality between different pansharpening approaches, we have shown the Mean Absolute Error (MAE) plots along with the RGB composite image for each fused HSI. According to the figure, the MAE maps corresponding to our DIP-HyperKite are much purple than the other pansharpening approaches, indicating minor fusion error. This is mainly because of the ability of our HyperKite network to predict very fine structures and edges by constraining the receptive field of the deep network.

b) Final Fusion results on the Botswana dataset: The Table VII summarizes the average quantitative results of different fusion methods on the Botswana dataset. Similar to the Pavia Center dataset, we can see that our DIP-HyperKite outperforms all the other HS pansharpening approaches by a considerable margin. Concretely, our DIP-HyperKite has improved the CC by 1.03%, and PSNR by 3.71%. In addition, our method has reduced the SAM by 9.68%, RMSE by 8.57%, and ERGAS by 10.85%. Furthermore, we have shown qualitative results related to different pansharpening approaches on Botswana dataset in Figure 15. By observing the RGB images and MAE plots in Figure 15, we can see that the fusion results related to our method are much closer to reference image than the other pansharpening approaches.

c) Final Fusion results on the Chikusei dataset: In this section, we compare the qualitative and quantitative results on the Chikusei dataset. The average quantitative results of different pansharpening approaches on the Chikusei dataset is listed in Table VIII. Similar to the results we have observed for the other two datasets, for this dataset also our proposed DIP-HyperKite outperforms all the pansharpening approaches that we considered for the analysis. Our pansharpening method improves the CC, SAM, RMSE, RSNR, ERGAS, and PSNR performance measures over the state-of-the-art results by 1.45%, 19.0%, 6.67%, 0.67%, 18.5%, and 0.90%, respectively. To further highlight the fusion quality of our method we present the qualitative results of selected panshaprpening approaches for the Chikusei dataset is shown in Figure 16. By observing the RGB composite image and the MAE maps we can clearly see that the fusion quality of the proposed DIP-HyperKite is higher than the other pansharpening approaches.

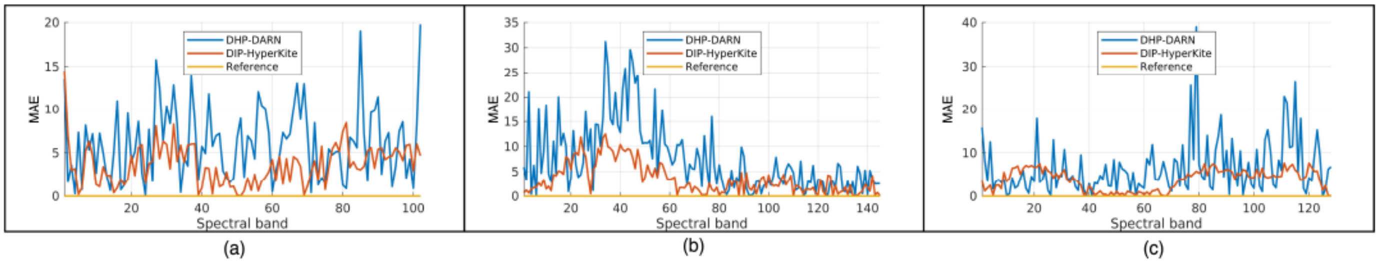

Fig. 17. The variation of MAE with spectral band for the 3-rd, 8-th, and 32-nd patch of (a). Pavia Center, (b). Botswana, and (c).Chikusei dataset, respectively.

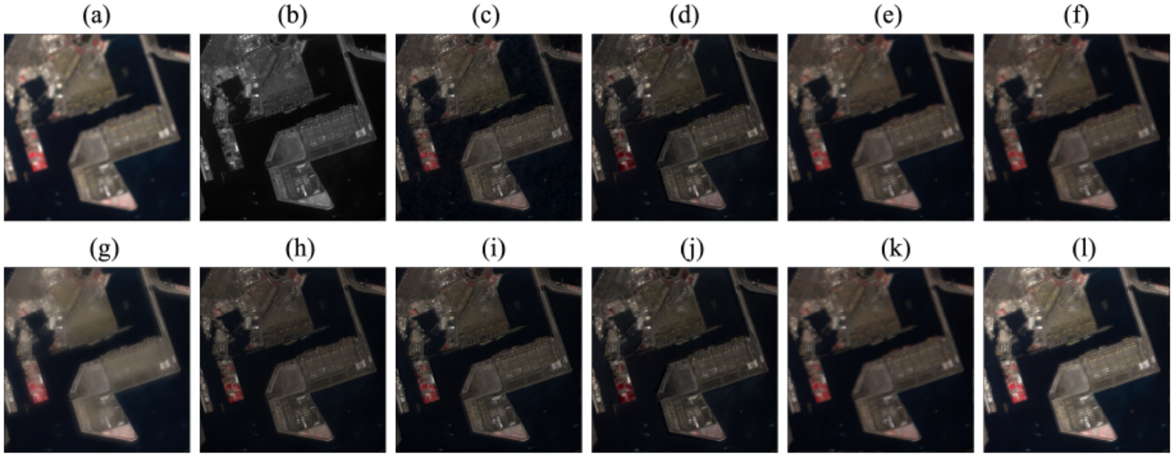

Fig. 18. Visual results obtained by different pansharpening methods for the Los Angeles dataset. (a) Real LR-HSI. (b) Real PAN image. (c) GSA [10]. (d) PCA [13]. (e) MG [19]. (f) MGH [20]. (g) GFPCA [21]. (h) CNMF [25]. (i) HySure [22]. (j) HyperPNN [35]. (k) DHP-DARN [30]. (l) DIP-HyperKite (ours). Note that the LR-HSI is zoomed in for visualization.

Variation of MAE with each spectral band: We visualize the variation of MAE with the spectral band for a randomly selected test image from the testing set of Pavia Center, Botswana, and Chikusei datasets in Figure 17 to further highlight the difference between the proposed DIP-HyperKite and state-of-the-art pansharpening method - DHP-DARN [30]. As shown in Figure 17, the proposed DIP-HyperKite results in overall low MAE across the spectral bands compared to the DHP-DARN. This further demonstrates the low spectral and spatial distortions introduced by our proposed DIP-HyperKite.

C. Final Fusion Results on a Real Hyperspectral Dataset

This section demonstrates the generalization capability of our proposed HyperKite on a real HSI dataset: the Los Angeles dataset. Since the Los Angeles dataset does not have reference HSI to compute the reference-based quantitative measures, we utilize three no-reference-based metrics following the previous works [30], namely D_λ, D_S and QNR as described in Section IV-B.

To obtain the pansharpen results on the Los Angeles dataset, we utilize the DIP-HyperKite trained on the Botswana dataset. The no-reference quantitative results and qualitative results for the Los Angeles dataset are presented in Table IX and Figure 18, respectively. The two Bayesian approaches that we previously used for comparisons on simulated HSI datasets are excluded because they require a corresponding blur matrix as an input to the algorithm, which is not available for the full-scale validation of the Los Angeles dataset.

TABLE IX
QUANTITATIVE RESULTS (NO-REFERENCE METRICS) FOR DIFFERENT PANSHARPENING METHODS ON THE LOSS ANGELES DATASET.

Metric	GSA	PCA	MG	MGH	GFPCA	CNMF	HySure	HyperPNN	DHP-DARN	DIP-HyperKite
$D_\lambda(\rightarrow 0)$	0.0750	0.0873	0.0478	0.0481	0.0697	0.0930	0.0529	0.0580	0.0459	**0.0431**
$D_S(\rightarrow 0)$	0.0942	0.1326	0.0432	0.0700	0.1891	0.1692	0.0485	0.0391	0.0372	**0.0347**
$QNR(\rightarrow 1)$	0.8378	0.7916	0.9111	0.8853	0.7544	0.7535	0.9011	0.9052	0.9186	**0.9240**

TABLE X
AVERAGE INFERENCE TIME PER HSI FOR DIFFERENT PANSHARPENING ALGORITHMS ON TESTING SET OF BOTSWANA DATASET.

Method	PCA	GFPCA	BF	BFS	SFIM	GS	GSA	MGH	CNMF	MG	HySure	Hyper-PNN	DHP-DARN	DIP	Hyper-Kite	DIP-HyperKite
Time(ms)	216.8	671.9	90.2	15536.1	222.6	191.6	227.0	347.2	437.5	341.6	3541.5	0.8	58305.2	58302.9	1.5	58304.4

According to the no-reference-based quantitative results shown in Table IX, we can see that our proposed DIP-HyperKite achieves outstanding results compared to existing pansharpening approaches. These quantitative results can be further validated with the qualitative results depicted in Figure 18. All these qualitative and visual results demonstrate the high potential and generalization capability of our proposed DIP-HyperKite for HS pansharpening.

D. Inference Time

Table X presents the average inference time per HSI on the testing set of Botswana dataset for different pansharpening methods. As we discussed previously, the proposed DIP-HyperKite consists of two steps where we first up-sample LR-HSI via DIP process, and then predict the residual image via HyperKite. Therefore, in Table X, we presents the average inference time for each step separately (i.e., DIP and HyperKite). As can be seen from the Table X, we observe relatively high inference time for our DIP-HyperKite and DHP-DARN [30] methods because they both utilize DIP for up-sampling which needs to be optimized for each LR-HSI separately during the testing. In average, the DIP up-sampling process takes about 58 s (58,000 ms) to up-sample a LR-HSI with spatial size of 40×40 by a scaling factor of 3 as shown in Table X for the Botswana dataset.

VI. Conclusion

In this paper, we have presented a novel approach for HS pansharpening, which mainly consists of three steps: (1) Up-sampling the LR-HSI via DIP, (2) Predicting the residual image via over-complete HyperKite, and (3) Obtaining the final fused HSI by summation. The previously proposed DIP methods for HS up-sampling only impose a constraint in the spectral-domain by utilizing LR-HSI. To better preserve both spatial and spectral information, we first exploited an additional spatial constraint to DIP by utilizing the available PAN image, thereby introduced both spatial and spectral constraints to the DIP. The comprehensive experiments conducted on three HS datasets showed that our proposed spatial+spectral loss function significantly improved the quality of up-sampled HSIs in CC, RMSE, RSNR, SAM, ERGAS, and PSNR performance measures. Next, in the residual prediction task, we have shown that the residual component between up-sampled HSI and the reference HSI primarily consists of edge information and very fine structures. Motivated by this observation, we proposed a novel over-complete deep-learning network for the residual prediction task. In contrast to the conventional under-complete representations, we have shown that our over-complete network is competent to focus on high-level features such as edges and fine structures by constraining the receptive field of the network. Finally, the fused HSI is obtained by adding the residual HSI and the up-sampled HSI. The comprehensive experiments conducted on three semi-synthetic and one real HS datasets demonstrated the superiority of our DIP-HyperKite over the other state-of-the-art results in terms of qualitative and quantitative evaluations.

References

[1] P. Ghamisi, J. Plaza, Y. Chen, J. Li, and A. J. Plaza, "Advanced spectral classifiers for hyperspectral images: A review," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 1, pp. 8–32, 2017.

[2] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 1844–1868, 2014.

[3] S. Matteoli, M. Diani, and G. Corsini, "A tutorial overview of anomaly detection in hyperspectral images," *IEEE Aerospace and Electronic Systems Magazine*, vol. 25, no. 7, pp. 5–28, 2010.

[4] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140–158, 2019.

[5] L. Loncan, L. B. De Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes *et al.*, "Hyperspectral pansharpening: A review," *IEEE Geoscience and remote sensing magazine*, vol. 3, no. 3, pp. 27–46, 2015.

[6] A. Mohammadzadeh, A. Tavakoli, and M. J. Valadan Zoej, "Road extraction based on fuzzy logic and mathematical morphology from pan-sharpened ikonos images," *The photogrammetric record*, vol. 21, no. 113, pp. 44–60, 2006.

[7] G. Licciardi, A. Villa, M. M. Khan, and J. Chanussot, "Image fusion and spectral unmixing of hyperspectral images for spatial improvement of classification maps," in *2012 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2012, pp. 7290–7293.

[8] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Information Fusion*, 2021.

[9] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geoscience and Remote Sensing Magazine*, 2020.

[10] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Jan. 4 2000, uS Patent 6,011,875.

[11] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms + pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, 2007.

[12] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive pca approach and contourlets," *IEEE transactions on geoscience and remote sensing*, vol. 46, no. 5, pp. 1323–1335, 2008.

[13] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens*, vol. 55, no. 1, pp. 339–348, 1989.

[14] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 5, pp. 2565–2586, 2014.

[15] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, "A new look at ihs-like image fusion methods," *Information fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[16] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," in *Fundamental Papers in Wavelet Theory*. Princeton University Press, 2009, pp. 494–513.

[17] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and statistics*. Springer, 1995, pp. 281–299.

[18] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.

[19] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Transactions on geoscience and remote sensing*, vol. 40, no. 10, pp. 2300–2312, 2002.

[20] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.

[21] W. Liao, F. Van Coillie, S. Gautama, A. Pizurica, and W. Philips, "Fusion of thermal infrared hyperspectral and vis rgb data using guided filter and supervised fusion graph."

[22] M. Simoes, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2014.

[23] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multiband images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1117–1127, 2015.

[24] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J.-Y. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.

[25] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 528–537, 2011.

[26] X. Tian, Y. Chen, C. Yang, and J. Ma, "Variational pansharpening by exploiting cartoon-texture similarities," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[27] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 265–10 274.

[28] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for p+ xs image fusion," *International Journal of Computer Vision*, vol. 69, no. 1, pp. 43–58, 2006.

[29] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[30] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[31] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[32] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[33] O. Sidorov and J. Y. Hardeberg, "Deep hyperspectral prior: Single-image denoising, inpainting, super-resolution," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3844–3851.

[34] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, no. 7, p. 594, 2016.

[35] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and B. Li, "Hyperpnn: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3092–3100, 2019.

[36] J. Li, R. Cui, B. Li, R. Song, Y. Li, Y. Dai, and Q. Du, "Hyperspectral image super-resolution by band attention through adversarial learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 6, pp. 4304–4318, 2020.

[37] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, and X. Guo, "Sdpnet: A deep network for pan-sharpening with enhanced information representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4120–4134, 2020.

[38] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2392–2399.

[39] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[43] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural computation*, vol. 12, no. 2, pp. 337–365, 2000.

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[45] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "Kiu-net: Towards accurate segmentation of biomedical images using over-complete representations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 363–373.

[46] ——, "Kiu-net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation," *arXiv preprint arXiv:2010.01663*, 2020.

[47] J. M. J. Valanarasu and V. M. Patel, "Overcomplete deep subspace clustering networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 746–755.

[48] P. Guo, J. M. J. Valanarasu, P. Wang, J. Zhou, S. Jiang, and V. M. Patel, "Over-and-under complete convolutional rnn for mri reconstruction," 2021.

[49] S.-Y. Lo, J. M. J. Valanarasu, and V. M. Patel, "Overcomplete representations against adversarial videos," *arXiv preprint arXiv:2012.04262*, 2020.

[50] R. Yasarla, J. M. J. Valanarasu, and V. M. Patel, "Exploring overcomplete representations for single image deraining using cnns," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[51] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 624–632.

[52] K. Li, W. Xie, Q. Du, and Y. Li, "Ddlps: Detail-based deep laplacian pansharpening for hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 8011–8025, 2019.

[53] M. Simões, J. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3373–3388, 2015.

[54] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Fast fusion of multi-band images based on solving a sylvester equation," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4109–4121, 2015.

[55] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, "Hyperspectral pansharpening with deep priors," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1529–1543, 2020.

[56] C. G. Harris, M. Stephens *et al.*, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[57] S. Holzwarth, A. Muller, M. Habermeyer, R. Richter, A. Hausold, S. Thiemann, and P. Strobl, "Hysens-dais 7915/rosis imaging spectrometers at dlr," in *Proceedings of the 3rd EARSeL workshop on imaging spectroscopy*, 2003, pp. 3–14.

[58] Y. Zeng, W. Huang, M. Liu, H. Zhang, and B. Zou, "Fusion of satellite images in urban area: Assessing the quality of resulting images," in *2010 18th International Conference on Geoinformatics*. IEEE, 2010, pp. 1–4.

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

[60] Q. Wei, "Bayesian fusion of multi-band images: A powerful tool for super-resolution," Ph.D. dissertation, Institut national polytechnique de Toulouse (INPT), 2015.

[61] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

Wele Gedara Chaminda Bandara (Student Member, IEEE) is a Ph.D. student in the Department of Electrical and Computer Engineering (ECE) at the Johns Hopkins University, USA. Before joining the Johns Hopkins, he worked as a graduate research student in the Department of Electrical and Electronic Engineering at the University of Peradeniya, Sri Lanka, for an NSF-funded project from 2019 to 2020. He graduated from the University of Peradeniya with first-class honors in Electrical and Electronic Engineering in 2019. His research interests include computer vision and image processing with applications in remote sensing, and hyperspectral image processing.

Jeya Maria Jose Valanarasu (Student Member, IEEE) is Ph.D. student in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University, USA. Prior to joining Hopkins, he graduated from NIT Trichy, India in 2019 with a Bachelor's degree in Instrumentation and Control Engineering. He also spent some time working in the Biomedical Engineering Department at National University of Singapore (NUS) as a visiting research intern. His research interests include image/3D segmentation, image enhancement, and image-to-image translation for computer vision and medical imaging tasks.

Vishal M. Patel [SM'15] is an Associate Professor in the Department of Electrical and Computer Engineering (ECE) at Johns Hopkins University. Prior to joining Hopkins, he was an A. Walter Tyson Assistant Professor in the Department of ECE at Rutgers University and a member of the research faculty at the University of Maryland Institute for Advanced Computer Studies (UMIACS). He completed his Ph.D. in Electrical Engineering from the University of Maryland, College Park, MD, in 2010. He has received a number of awards including the 2021 IEEE Signal Processing Society (SPS) Pierre-Simon Laplace Early Career Technical Achievement Award, the 2021 NSF CAREER Award, the 2021 IAPR Young Biometrics Investigator Award (YBIA), the 2016 ONR Young Investigator Award, the 2016 Jimmy Lin Award for Invention, A. Walter Tyson Assistant Professorship Award, Best Paper Awards at IEEE AVSS 2017 & 2019, IEEE BTAS 2015, IAPR ICB 2018, IEEE ICIP 2021, and two Best Student Paper Awards at IAPR ICPR 2018. He is an Associate Editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence, Pattern Recognition Journal, and serves on the Machine Learning for Signal Processing (MLSP) Committee of the IEEE Signal Processing Society. He serves as the vice president of conferences for the IEEE Biometrics Council.