FusionNet: An Unsupervised Convolutional Variational Network for Hyperspectral and Multispectral Image Fusion

Zhengjue Wang, Bo Chen, *Senior Member, IEEE*, Ruiying Lu, Hao Zhang, Hongwei Liu, *Member, IEEE*, and Pramod K. Varshney, *Life Fellow, IEEE*

***Abstract*— Due to hardware limitations of the imaging sensors, it is challenging to acquire images of high resolution in both spatial and spectral domains. Fusing a low-resolution hyperspectral image (LR-HSI) and a high-resolution multispectral image (HR-MSI) to obtain an HR-HSI in an unsupervised manner has drawn considerable attention. Though effective, most existing fusion methods are limited due to the use of linear parametric modeling for the spectral mixture process, and even the deep learning-based methods only focus on deterministic fully-connected networks without exploiting the spatial correlation and local spectral structures of the images. In this paper, we propose a novel variational probabilistic autoencoder framework implemented by convolutional neural networks, in order to fuse the spatial and spectral information contained in the LR-HSI and HR-MSI, called FusionNet. The FusionNet consists of a spectral generative network, a spatial-dependent prior network, and a spatial-spectral variational inference network, which are jointly optimized in an unsupervised manner, leading to an end-to-end fusion system. Further, for fast adaptation to different observation scenes, we give a meta-learning explanation to the fusion problem, and combine the FusionNet with meta-learning in a synergistic manner. Effectiveness and efficiency of the proposed method are evaluated based on several publicly available datasets, demonstrating that the proposed FusionNet outperforms the state-of-the-art fusion methods.**

***Index Terms*— Hyperspectral images, multispectral images, image fusion, probabilistic generative model, convolutional neural network, meta-learning.**

I. Introduction

THE rich spectral information available in hyperspectral images (HSIs) is considered to be very promising and valuable for applications in remote sensing, biomedicine, and various computer vision tasks [1]–[3]. However, due to hardware limitations of the imaging sensors, there is a tradeoff between the spatial and the spectral resolution in the images, resulting in degradation of the spatial resolution of the HSIs that provide fine spectral resolution. [4]. For current sensors, especially sensors used for remote sensing, it is quite challenging to acquire high-resolution (HR) HSIs [5], [6]. On the contrary, HR multispectral images (MSIs) with much fewer spectral bands can be easily obtained. Improving the resolution of a image by the aid of another kind of image captured from the same scene is commonly considered in image enhancement [7]. This has resulted in an increasing trend of fusing a low-resolution (LR) HSI and an HR-MSI to obtain an HR-HSI in an unsupervised manner, often referred to as hyperspectral and multispectral image fusion [5].

Recently, linear spectral mixture model based HSI and MSI fusion has drawn considerable attention, due to its sound physical description of the observed spectra. Popular approaches [8]–[20] retrieve the HR-HSIs through linear factorization with the help of different prior knowledge or constraints. Despite the high level of performance achieved, the above linear models are limited due to the insufficient ability of parametric modeling of the spectral mixture process which is actually nonlinear, and it is hard to be expressed by accurate physics-inspired modeling of the optical behavior [21].

Although the past few years have witnessed the tremendous success of deep learning in various applications, only a few works have been devoted to unsupervised deep HSI and MSI fusion [6]. In [6], Qu *et al*. present an unsupervised sparse Dirichlet-net (uSDN) containing two coupled autoencoders, showing state-of-the-art performance. However, the uSDN optimizes the two autoencoders separately, and may not make full use of interactions between the LR-HSI and HR-MSI during fusion. In addition, with fully-connected structures, it does not consider the spatial correlation and local spectral structures of the images. Therefore, constructing a deep model with more effective information extraction and more harmonious information fusion in an unsupervised manner is worthy of further study.

As we know, deep probabilistic generative models are proficient in representing the underlying data distribution and modeling prior knowledge naturally, which have shown

Manuscript received August 7, 2019; revised February 24, 2020; accepted June 6, 2020. Date of publication June 29, 2020; date of current version July 13, 2020. The work of Bo Chen was supported in part by the Program for Oversea Talent by the Chinese Central Government, in part by the 111 Project under Grant B18039, and in part by NSFC under Grant 61771361. The work of Hongwei Liu was supported by NSFC for Distinguished Young Scholars under Grant 61525105 and in part by the Shaanxi Innovation Team Project. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dong Tian. *(Corresponding author: Bo Chen.)*

Zhengjue Wang, Bo Chen, Ruiying Lu, Hao Zhang, and Hongwei Liu are with the National Laboratory of Radar Signal Processing, Collaborative Innovation Center of Information Sensing and Understanding, Xidian University, Xi'an 710071, China (e-mail: zhengjuewang@163.com; bchen@mail.xidian.edu.cn).

Pramod K. Varshney is with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: varshney@syr.edu).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TIP.2020.3004261

excellent unsupervised data expressive ability, such as the deep belief networks in [22], [23], the variational autoencoder [24], while convolutional neural networks (CNNs) are powerful in capturing the local structures of images. However, to the best of our knowledge, no effort has been devoted to solving the unsupervised fusion problem by exploiting their advantages.

In traditional unsupervised hyperspectral and multispectral image fusion methods [6], [8]–[20], different parameters are learned for different image pairs. In particular, the model needs to be retrained for every fusion task, which is time-consuming, since the remote sensors often monitor the Region of Interest (RoI) continuously. Although using the parameters, learned from a training dataset composed of lots of LR-HSI and HR-MSI image pairs, directly for a future fusion task is quite efficient, distribution mismatch between training and testing data often exists. Thus, constructing a model with the ability for fast adaptation to different fusion tasks is highly desirable for real applications. Meta-learning has shown promise in training a model for a variety of learning tasks and adapting quickly as more tasks become available. However, to the best of our knowledge, there is neither a meta-learning explanation for the fusion problem, nor an unsupervised fusion model combined with meta learning in a synergistic manner.

Contributions: In this paper, we start by modeling a novel variational probabilistic autoencoder framework for unsupervised HSI and MSI fusion. Under this framework, we use convolutional networks as nonlinear functions to express the distributions in the probabilistic model, presenting an unsupervised convolutional variational network, called FusionNet. The following key components constitute our fusion methodology:

(1) A shared deconvolutional decoder network is designed to describe the shared spectral characteristics of the observed LR-HSI and the target HR-HSI, while the shared latent representation represents the spatial correspondence between the observed HR-MSI and the target HR-HSI.

(2) We employ a spatial-dependent prior on the latent variables, which further enhances the ability of information fusion in the latent space.

(3) A convolutional encoder network containing feature extraction and feature embedding is designed to realize an end-to-end fusion system. Specifically, the local spectral information in the LR-HSI and the spatial-spectral information in the HR-MSI are extracted via a 1D and a 2D convolutional networks, respectively, and then these two kinds of information are fused to infer the latent representations of the target HR-HSI via convolutional embedding.

(4) With principled probabilistic formulation, the whole network is optimized by maximizing the evidence lower bound (ELBO) of the joint full likelihood of LR-HSI and HR-MSI, leading to an efficient inference scalable to large observation scenes.

(5) For fast adaptation to different tasks, we give the fusion problem a meta-learning explanation, and update the parameters of FusionNet via meta-learning.

The rest of this paper is organized as follows. Section II reviews some related works. The methodology of the proposed variational probabilistic autoencoder framework and FusionNet are introduced in Section III. Experimental results are presented in Section IV to demonstrate the effectiveness and efficiency of our method. Section V concludes this paper.

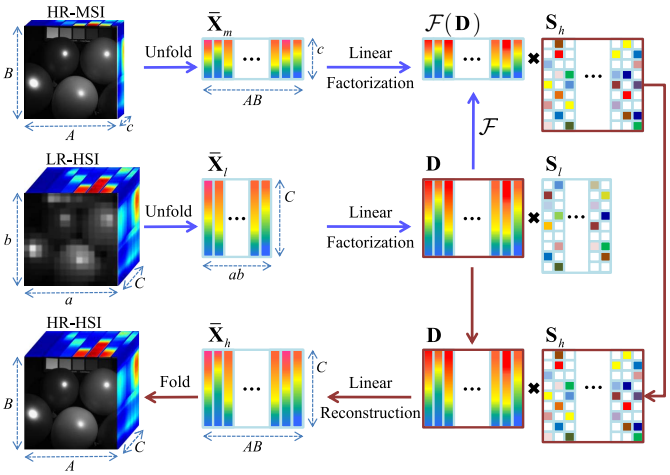


Fig. 1. Illustration of the matrix decomposition based fusion methods.

II. Related Work

Let $\mathbf{X}_l \in \mathbb{R}^{a\times b\times C}$ denote the acquired a-by-b LR-HSI with C channel bands, and we assume the availability of an HR-MSI $\mathbf{X}_m \in \mathbb{R}^{A\times B\times c}$ of the same scene, where $c \ll C$, $a \ll A$, $b \ll B$. The objective is to fuse the spectral and spatial information produced by $\mathbf{X}_l$ and $\mathbf{X}_m$ and recover an HR-HSI $\mathbf{X}_h \in \mathbb{R}^{A\times B\times C}$.

A. Linear Decomposition Based Fusion Methods

1) Matrix Decomposition: Recently, the strategy of associating the fusion task with linear spectral mixture models has drawn considerable attention, as shown in Fig. 1, where each pixel is represented as a linear combination of the reflectance from a small number of distinct materials [10]. Specifically, the unfolded matrix of the target HR-HSI, $\bar{\mathbf{X}}_h \in \mathbb{R}^{C\times AB}$, can be represented as the product of the spectral signatures $\mathbf{D}$ and the mixture proportions $\mathbf{S}_h$, *i.e.*,

$$\bar{\mathbf{X}}_h \approx \mathbf{D}\mathbf{S}_h. \tag{1}$$

Compared with HR-HSI, LR-HSI and HR-MSI have poor resolution in spatial and spectral dimensions, respectively. Let us denote the unfolded LR-HSI and HR-MSI as $\bar{\mathbf{X}}_l \in \mathbb{R}^{C\times ab}$ and $\bar{\mathbf{X}}_m \in \mathbb{R}^{c\times AB}$, respectively. Since LR-HSI and HR-HSI are observations w.r.t. the same scene and have the same spectral resolution, they are composed of the same set of spectral signatures $\mathbf{D}$, but $\bar{\mathbf{X}}_l$ has different mixture proportions $\mathbf{S}_l$, *i.e.*,

$$\bar{\mathbf{X}}_l \approx \mathbf{D}\mathbf{S}_l. \tag{2}$$

According to the available spectral response function (SRF) $\mathcal{F} : \mathbb{R}^C \to \mathbb{R}^c$ of the imaging sensor that describes the sensitivity to optical radiation of different wavelengths [5], the unfolded HR-MSI can be expressed as:

$$\bar{\mathbf{X}}_m \approx \mathcal{F}(\bar{\mathbf{X}}_h) \approx \mathcal{F}(\mathbf{D})\mathbf{S}_h. \tag{3}$$

To sum up, $\bar{\mathbf{X}}_h$ has shared spectral signatures with $\bar{\mathbf{X}}_l$, and shared proportions with $\bar{\mathbf{X}}_m$, as shown in Fig. 1. Thus, approaches based on matrix factorizations w.r.t. $\bar{\mathbf{X}}_l$ and $\bar{\mathbf{X}}_m$ have been actively investigated [8]–[20].

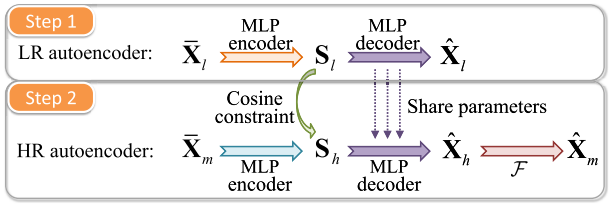


Fig. 2. Simplified structure of the uSDN in [6]. After the first-step learning, the decoder is fixed for the second step. In the HR autoencoder, only the encoder needs to be optimized by minimizing the reconstruction error of $\bar{\mathbf{X}}_m$ constrained by the cosine similarity of the latent representations.

2) Tensor Decomposition: Considering that unfolding the 3D images into matrices destroys the spatial characteristics of the images, a sparse tensor factorization based method is proposed in [25], in order to better exploit the inherent spatial and spectral characteristics.

Despite the improved performances achieved, all the above methods are limited by the linear mixture assumption. Because, in practice, light typically interacts multiple times with the materials making up the mixture [21], which leads to complex nonlinear interactions w.r.t. the individual abundances, and accurate physics-inspired parametric modeling of the optical behavior is a difficult task. On the contrary, deep learning techniques are proficient in representing the underlying data structure automatically, which has motivated the researchers to construct neural network based fusion methods.

B. Neural Network Based Fusion Methods

1) Coupled Autoencoders: Inspired by the matrix decomposition based methods, where $\bar{\mathbf{X}}_l$ and $\bar{\mathbf{X}}_h$ have shared global parameters, and $\bar{\mathbf{X}}_m$ and $\bar{\mathbf{X}}_h$ have shared local parameters, an unsupervised sparse Dirichlet-net (uSDN) is presented in [6]. As shown in Fig. 2, the fusion process is realized by assuming that $\bar{\mathbf{X}}_l$ and $\bar{\mathbf{X}}_h$ have a shared decoder, that $\bar{\mathbf{X}}_m$ and $\bar{\mathbf{X}}_h$ have shared latent representations, and that the latent representations of $\bar{\mathbf{X}}_l$ and $\bar{\mathbf{X}}_h$ are similar in cosine distance. However, the two autoencoders are separately optimized. They firstly train the LR autoencoder, and then they fix the decoder parameters and only update the encoder parameters of the HR autoencoder. To reduce the spectral distortion, a similarity constraint is imposed on the latent representation $\mathbf{S}_h$ every 10 iterations during the optimization of the HR autoencoder. The sequential optimization may not make full use of the interaction between the two sub-problems, since the feedback from the HR-MSI information cannot influence the learning at the first stage. Besides, the uSDN is constructed using multilayer perceptrons (MLPs) with the unfolded images as inputs, without considering the spatial information. In addition, the fully-connected operation treats each pixel, *i.e.*, a spectral vector, as a whole, incapable of exploiting the local spectral structures.

2) CNNs: Due to their ability to capture the local structures of the data, CNNs have shown great success in image processing, *e.g.*, single image super-resolution [26], [27], which inspires us to construct an unsupervised convolutional fusion model. Although Yang *et al.* [28] and Xie *et al.* [29] proposed CNN based models to obtain an HR-HSI, our work is done independently and parallelly and derived from different considerations and focus on different problem backgrounds.

Methods in [28] and [29] aim at solving the regression problem from the LR-HSI and HR-MSI inputs to the target HR-HSI in a supervised manner. Specifically, in [28], the objective function is to minimize the mean square error of the unfolded matrix of HR-HSI as:

$$\min ||\bar{\mathbf{X}}_h - \text{MLP}\left[\text{CNN}(\mathbf{X}_l) \oplus \text{CNN}(\mathbf{X}_m)\right]||_F^2, \tag{4}$$

where, $\oplus$ represents vector-concatenation. In other words, the information fusion in (4) is mainly realized by concatenating the features extracted by the CNNs. In [29], they believed that the target $\bar{\mathbf{X}}_h$ can be represented by the columns in $\bar{\mathbf{X}}_m$ and a to-be-estimated matrix $\mathbf{\Omega}$, *i.e.*,

$$\bar{\mathbf{X}}_h = \bar{\mathbf{X}}_m\mathbf{U} + \mathbf{\Omega}\mathbf{V} \tag{5}$$

with coefficient matrices $\mathbf{U}$ and $\mathbf{V}$. By solving this problem with iterative algorithm, they presented a deep network with $\mathbf{X}_l$ and $\mathbf{X}_m$ as inputs to approximate $\mathbf{\Omega}$ and finally output the reconstructed HR-HSI.

However, acquiring the supervised information, *i.e.*, HR-HSI, especially for remotely sensed HR-HSI, is challenging for the currently available sensors [5], [6]. This is because, for high spatial resolution, the sensor must have a small instantaneous field of view, which leads to the decrease of the signal to noise ratio (SNR) of the spectral images. Whereas, in order to improve the SNR, one has to widen the bandwidth allowing more light to enter into the sensor while acquiring the individual bands, which may degenerate the spectral resolution [4]. In this case, Xie *et al.* [29] tried to use Wald protocol [30], [31] to create the supervised training data. They downsample both of the HR-MSI and LR-HSI in spatial domain, so that the original LR-HSI is treated as the "target HR-HSI" corresponding to the downsampled images. However, it is often challenging as the LR-HSI has small spatial dimension or the scaling factor is large, since the downsampled images can not provide enough data volume for the training. Therefore, how to construct a CNN model to solve the unsupervised fusion task should be redesigned from the very beginning.

III. FusionNet

In this section, we start by formulating the joint probabilistic generative processes of the observed LR-HSI and HR-MSI, and the unknown HR-HSI. Specifically, the conditional likelihood of a pixel, expressed by a deep structure with a nonlinear latent variable, implies the data-driven nonlinear spectral mixture process and the spectral relationships between the pixels in these images, while the prior distributions describe the relationships between the LR image and the HR images in the latent space. Considering that the true posterior in this case is intractable, we leverage two recognition models to infer the latent variables, resulting in a novel variational probabilistic autoencoder framework for the unsupervised HSI and MSI fusion task. To further explore the spatial and spectral characteristics, we specify the functions in the framework via CNNs,

resulting in an unsupervised convolutional variational network, called FusionNet.

A. Spectral Generative Network

In a hyperspectral image, observed data corresponding to each pixel is a vector describing the observed spectrum. Denote the i-th pixel in $\mathbf{X}_l$ and the j-th pixel in $\mathbf{X}_h$ as $\boldsymbol{x}_i^{(l)} \in \mathbb{R}^c$, $\boldsymbol{x}_j^{(h)} \in \mathbb{R}^C$, respectively. As claimed in [10]–[12], although $\mathbf{X}_l$ and $\mathbf{X}_h$ have different spatial resolutions, they are observations with respect to the same scene and have the same spectral characteristics. Different from employing the linear factorization approach [10]–[12], [25] or a deterministic network [6], we use two probabilistic generative (or decoder) models with shared parameters $\boldsymbol{\theta}$ to model the distributions of the hyperspectral pixels $\{\boldsymbol{x}_i^{(l)}\}_{i=1}^{ab}$ and $\{\boldsymbol{x}_j^{(h)}\}_{j=1}^{AB}$, formally stated as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^{(l)}|\boldsymbol{z}_i^{(l)}) = \mathcal{N}(\boldsymbol{\mu}_p(\boldsymbol{z}_i^{(l)}), \boldsymbol{\sigma}_p^2(\boldsymbol{z}_i^{(l)})\mathbf{I}), \tag{6}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_j^{(h)}|\boldsymbol{z}_j^{(h)}) = \mathcal{N}(\boldsymbol{\mu}_p(\boldsymbol{z}_j^{(h)}), \boldsymbol{\sigma}_p^2(\boldsymbol{z}_j^{(h)})\mathbf{I}), \tag{7}$$

where, $\boldsymbol{z}_i^{(l)} \in \mathbb{R}^k$ and $\boldsymbol{z}_j^{(h)} \in \mathbb{R}^k$ are the latent variables, corresponding to the mixture proportions of the pixels $\boldsymbol{x}_i^{(l)}$ and $\boldsymbol{x}_j^{(h)}$, respectively, $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ denotes a Gaussian distribution with mean vector $\boldsymbol{\mu}$ and diagonal covariance matrix $\boldsymbol{\sigma}^2\mathbf{I}$. Both $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ are non-linear functions w.r.t. the latent variable with parameters $\boldsymbol{\theta}$, realized by neural networks, where the subscript p (q in (17) and (20), or the *prior* in (16)) is to highlight that the $\boldsymbol{\mu}(\cdot)$ and $\boldsymbol{\sigma}^2(\cdot)$ belong to the generative model (inference model or prior model).

As observed in Fig. 6 later, there are local structural similarities among different pixels, marked by the circles. To better explore the spectral characteristics, a 1D deconvolutional neural network (DCNN) is used to realize the non-linear functions $\boldsymbol{\mu}_p$ and $\boldsymbol{\sigma}_p^2$ with high expressive ability, *i.e.*,

$$\boldsymbol{h}_p(\cdot) = \text{1D-DCNN}_1(\cdot) \tag{8}$$

$$\boldsymbol{\mu}_p(\cdot) = \text{1D-DCNN}_2(\boldsymbol{h}_p(\cdot)), \tag{9}$$

$$\boldsymbol{\sigma}_p(\cdot) = \exp(\text{1D-DCNN}_3(\boldsymbol{h}_p(\cdot))), \tag{10}$$

where, $\boldsymbol{h}_p(\cdot)$ is a hidden layer, the deconvolutional operations take place along the spectral dimension, the subscript of 1D-DCNN is used to distinguish the structure with others. Thus, the parameters of the likelihoods in (6) and (7) are actually network parameters, namely $\boldsymbol{\theta}$ contains the DCNN parameters that need to be learned. Example architectures of these DCNNs can be seen in Table I.

According to the sensor-specific spectral response function $\mathcal{F} : \mathbb{R}^C \to \mathbb{R}^c$, the spectral degradation model is stated as:

$$\boldsymbol{x}_j^{(m)} = \mathcal{F}(\boldsymbol{x}_j^{(h)}) + \boldsymbol{n}_m. \tag{11}$$

By assuming $\boldsymbol{n}_m$ to be isotropic Gaussian noise, the multispectral pixels $\{\boldsymbol{x}_j^{(m)}\}_{j=1}^{AB}$ are distributed as:

$$p_{\boldsymbol{\eta},\alpha}(\boldsymbol{x}_j^{(m)}|\boldsymbol{x}_j^{(h)}) = \mathcal{N}(\mathcal{F}(\boldsymbol{x}_j^{(h)}), \alpha^2\mathbf{I}), \tag{12}$$

where, $\boldsymbol{\eta}$ contains the parameters of $\mathcal{F}$, which is known, while α is the unknown variance of the noise that needs inferring. According to (7) and (12), the hierarchical generative process from the abundance vector $\boldsymbol{z}_j^{(h)}$ to the observed multispectral pixel $\boldsymbol{x}_j^{(m)}$ is derived as follows:

$$p(\boldsymbol{x}_j^{(m)}|\boldsymbol{z}_j^{(h)}) = \int p(\boldsymbol{x}_j^{(m)}|\boldsymbol{x}_j^{(h)}) p(\boldsymbol{x}_j^{(h)}|\boldsymbol{z}_j^{(h)}) d\boldsymbol{x}_j^{(h)} \tag{13}$$

$$= \mathbb{E}_{\boldsymbol{x}_j^{(h)} \sim p(\boldsymbol{x}_j^{(h)}|\boldsymbol{z}_j^{(h)})}[p(\boldsymbol{x}_j^{(m)}|\boldsymbol{x}_j^{(h)})], \tag{14}$$

where the target pixel $\boldsymbol{x}_j^{(h)}$ is an intermediate latent variable. Therefore, the HR-HSI can be retrieved by (7) after inferring the latent variable $\boldsymbol{z}_j^{(h)}$, and optimizing the parameters $\boldsymbol{\theta}$ and α.

TABLE I
NETWORK ARCHITECTURE FOR THE AVIRIS DATASET

module	operation	kernel size	output size
$q_{\boldsymbol{\phi}_l}(\boldsymbol{z}_i^{(l)}\|\boldsymbol{x}_i^{(l)})$			
1D-CNN$_1$	conv.	(1*1*9)*16	16*16*188*16
	conv.	(1*1*5)*4	16*16*188*4
	conv.	(1*1*5)*1	16*16*188
Linear$_1$	fully conn.	188*30	16*16*30
Linear$_2$	fully conn.	188*30	16*16*30
$q_{\boldsymbol{\phi}_h}(\boldsymbol{z}_j^{(h)}\|\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)})$			
1D-CNN$_2$	conv.	(1*1*5)*1	16*16*188
2D-CNN	conv.	(5*5*6)*188	16*16*188
1D-CNN$_3$	conv.	(1*1*9)*16	16*16*188*16
	conv.	(1*1*5)*4	16*16*188*4
	conv.	(1*1*5)*1	16*16*188
Linear$_3$	fully conn.	188*30	16*16*30
Linear$_4$	fully conn.	188*30	16*16*30
$p_{\boldsymbol{\theta}}(\boldsymbol{x}_j^{(h)}\|\boldsymbol{z}_j^{(h)})$ or $p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^{(l)}\|\boldsymbol{z}_i^{(l)})$			
1D-DCNN$_1$	fully conn.	30*188	16*16*188
	dconv.	(1*1*5)*4	16*16*188*4
	dconv.	(1*1*5))*16	16*16*188*16
1D-DCNN$_2$	dconv.	(1*1*9)*1	16*16*188
1D-DCNN$_3$	dconv.	(1*1*9)*1	16*16*188
$p_{\boldsymbol{\omega}}(\boldsymbol{z}_j^{(h)}\|\overline{\boldsymbol{z}}_j^{(l)})$			
–	fully conn.	30*30	16*16*30

B. Spatial-Dependent Prior Network

A probabilistic model is able to model prior knowledge naturally, where a more accurate prior assumption is beneficial to model performance. In our model, there are two kinds of latent variables.

For the LR latent variable $\boldsymbol{z}_i^{(l)}$, a commonly used prior distribution

$$p(\boldsymbol{z}_i^{(l)}) = \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{15}$$

is employed. Whereas, for the HR latent variable $\boldsymbol{z}_j^{(h)}$, it may not be reasonable to follow the same prior, since such a choice does not make full use of the prior knowledge that $\mathbf{X}_l$ and $\mathbf{X}_h$ have a close spatial relationship. More importantly, according to Bayes rule, with $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as the prior, the posterior of the $\boldsymbol{z}_j^{(h)}$ would be $p_{\{\boldsymbol{\theta},\boldsymbol{\eta},\alpha\}}(\boldsymbol{z}_j^{(h)}|\boldsymbol{x}_j^{(m)})$, only dependent on $\boldsymbol{x}_j^{(m)}$, which weakens the effect of $\mathbf{X}_l$ in information fusion. Based on these considerations, a spatial-dependent prior network is proposed.

In the observation space, $\mathbf{X}_l$ is a spatial blurred counterpart w.r.t. $\mathbf{X}_h$. In the latent space, $\{\boldsymbol{z}_i^{(l)}\}_{i=1}^{ab}$ and $\{\boldsymbol{z}_j^{(h)}\}_{j=1}^{AB}$ form the 3D latent representations as $\mathbf{Z}_l \in \mathbb{R}^{a\times b\times k}$ and $\mathbf{Z}_h \in \mathbb{R}^{A\times B\times k}$, respectively. We first expand $\mathbf{Z}_l$ by duplicating its values at

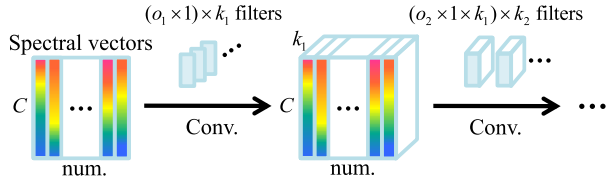
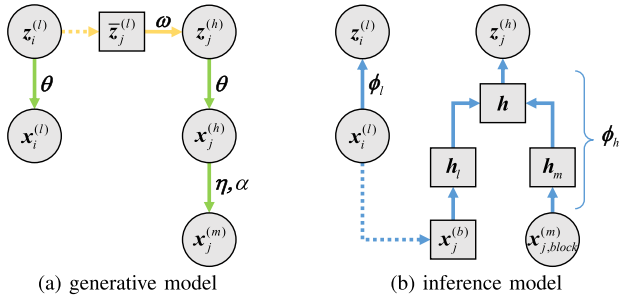


Fig. 3. Illustration of the convolutional operations along the spectral dimension.



(a) generative model (b) inference model

Fig. 4. Illustration of the variational probabilistic autoencoder framework of the FusionNet. Circles are stochastic variables and squares are deterministic variables. The green, yellow, and blue solid lines indicate the generative, prior, and recognition models, respectively. The yellow and blue dashed lines indicate duplicate and bicubic operations, respectively.

each pixel in its spatial neighborhood, having $\overline{\mathbf{Z}}_l \in \mathbb{R}^{A\times B\times k}$, as shown in Fig. 5. And then, a spatial-dependent prior is proposed to enhance the information interaction, formally stated as:

$$p_{\boldsymbol{\omega}}(\boldsymbol{z}_j^{(h)}|\overline{\boldsymbol{z}}_j^{(l)}) = \mathcal{N}(\boldsymbol{\mu}_{prior}(\overline{\boldsymbol{z}}_j^{(l)}), \mathbf{I}), \tag{16}$$

where, $\overline{\boldsymbol{z}}_j^{(l)}$ is the j-th pixel in $\overline{\mathbf{Z}}_l$, $\boldsymbol{\mu}_{prior}$ is realized by a fully-connected network parameterized by $\boldsymbol{\omega}$, with detailed architecture shown in Table I. As a result, the posterior distribution w.r.t. $\boldsymbol{z}_j^{(h)}$ is $p_{\{\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\eta},\alpha\}}(\boldsymbol{z}_j^{(h)}|\boldsymbol{x}_j^{(m)}, \overline{\boldsymbol{z}}_j^{(l)})$, related to both the LR-HSI and HR-MSI.

In contrast to [6] that adds constraints on the latent representations every few iterations, which may need exhaustive tuning, the proposed spatial-dependent prior is involved with every-step of joint optimization, which is more principled. More details will be discussed in Section III-D.

C. Spatial-Spectral Variational Inference Network

Due to the complicated forms of the likelihoods and priors, the true posteriors w.r.t. $\boldsymbol{z}_i^{(l)}$ and $\boldsymbol{z}_j^{(h)}$ are intractable. To realize efficient inference and learning with a intractable posterior, constructing a recognition (or encoder) model, *i.e.*, a variational distribution parameterized by neural networks, to approximate the true posterior distribution is commonly used [23], [24]. In our model, we employ two recognition models to infer the latent variables $\boldsymbol{z}_i^{(l)}$ and $\boldsymbol{z}_j^{(h)}$, for which details are introduced in the following.

According to (6), (15), and Bayes Rule, the true posterior distribution w.r.t. $\boldsymbol{z}_i^{(l)}$ is $p_{\boldsymbol{\theta}}(\boldsymbol{z}_i^{(l)}|\boldsymbol{x}_i^{(l)})$, related to $\boldsymbol{x}_i^{(l)}$. To approximate it, the variational distribution in the LR recognition model is defined as

$$q_{\boldsymbol{\phi}_l}(\boldsymbol{z}_i^{(l)}|\boldsymbol{x}_i^{(l)}) = \mathcal{N}(\boldsymbol{\mu}_{q,l}(\boldsymbol{x}_i^{(l)}), \boldsymbol{\sigma}_{q,l}^2(\boldsymbol{x}_i^{(l)})\mathbf{I}), \tag{17}$$

where, the subscript l (or h in (20)) is to highlight that the $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ belong to the LR (or HR) recognition model. In order to capture intra-correlation across different spectral bands, a 1D CNN is used to realize the nonlinear functions in (15), as:

$$\boldsymbol{\mu}_{q,l}(\cdot) = \text{Linear}_1(\text{1D-CNN}_1(\cdot)), \tag{18}$$

$$\boldsymbol{\sigma}_{q,l}(\cdot) = \exp(\text{Linear}_2(\text{1D-CNN}_1(\cdot))). \tag{19}$$

where, the "Linear(·)" represents a linear fully-connected layer, the convolutional operations take effect along the spectral dimension, as shown in Fig. 3. An example architecture can be seen in Table I.

According to (7), (12), (16), and Bayes Rule, the true posterior distribution w.r.t. $\boldsymbol{z}_j^{(h)}$ is $p_{\{\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{\eta},\alpha\}}(\boldsymbol{z}_j^{(h)}|\boldsymbol{x}_j^{(m)}, \overline{\boldsymbol{z}}_j^{(l)})$, related to both the LR-HSI and HR-MSI, as described in Sec. III-B. Therefore, the HR recognition model should be related to them in a similar manner. Inspired by single image super-resolution methods [26], [32] that preprocess the LR image with bicubic interpolation, we up-scale the LR-HSI $\mathbf{X}_l$ via bicubic interpolation to form the LR inputs of the HR recognition model, denoted as $\mathbf{X}_b \in \mathbb{R}^{A\times B\times C}$, whose j-th pixel is $\boldsymbol{x}_j^{(b)}$. The HR recognition model is stated as:

$$q_{\boldsymbol{\phi}_h}(\boldsymbol{z}_j^{(h)}|\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)}) = \mathcal{N}(\boldsymbol{\mu}_{q,h}(\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)}), \boldsymbol{\sigma}_{q,h}^2(\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)})\mathbf{I}), \tag{20}$$

where, $\boldsymbol{x}_{j,block}^{(m)}$ is the spatial neighboring block of $\boldsymbol{x}_j^{(m)}$. Eq. (20) indicates that the HR recognition model aims to fuse the spatial-spectral information in the HR-MSI and the spectral information in the LR-HSI. For this purpose, the nonlinear functions in (20) are implemented by a neural network containing three modules, *i.e.*, feature extraction, feature fusion, and feature embedding.

The spatial-spectral information in the HR-MSI and the spectral information in the LR-HSI are extracted via a 2D CNN along the spatial dimension and a 1D CNN along the spectral dimension, respectively, *i.e.*,

$$\boldsymbol{h}_l = \text{1D-CNN}_2(\mathbf{X}_b), \tag{21}$$

$$\boldsymbol{h}_m = \text{2D-CNN}(\mathbf{X}_m). \tag{22}$$

And then, the features $\boldsymbol{h}_l$ and $\boldsymbol{h}_m$ are fused by element summation as:

$$\boldsymbol{h} = \boldsymbol{h}_l + \boldsymbol{h}_m, \tag{23}$$

followed by feature embedding as:

$$\boldsymbol{\mu}_{q,h} = \text{Linear}_3(\text{1D-CNN}_3(\boldsymbol{h})), \tag{24}$$

$$\boldsymbol{\sigma}_{q,h} = \exp(\text{Linear}_4(\text{1D-CNN}_3(\boldsymbol{h}))). \tag{25}$$

Detailed network architecture of the HR recognition model can be seen in Table I.

D. Joint Optimization

To sum up, the proposed variational probabilistic autoencoder framework is illustrated in Fig. 4. Further, by concatenating the spectral generative network, spatial-dependent

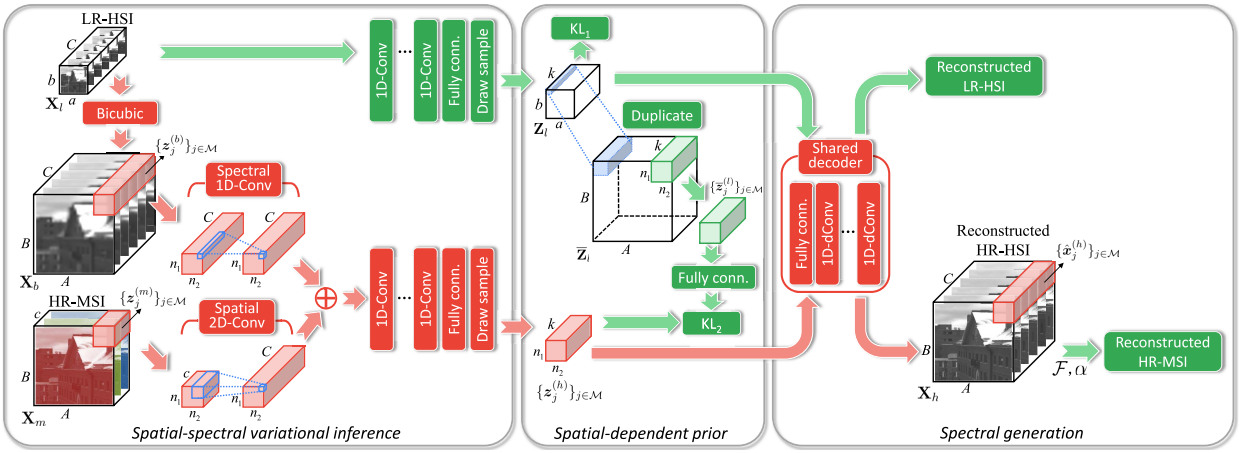


Fig. 5. Architecture of the proposed FusionNet. The spatial-spectral variational inference network aims to infer the latent representations, where the green path represents (17)-(19), the red path corresponds to (20)-(25), and the blue cubes indicate 1D- and 2D-convolution operations. The spatial-dependent prior network imposes constraints on the latent representations via the KL-divergence terms in (29) and (30), where the blue cubes indicate duplicating a pixel in $\mathbf{Z}_l$ as a block in $\overline{\mathbf{Z}}_l$. The spectral generation network aims to generate the hyperspectral pixels via (6)-(10), and then the multispectral pixels via (12). During training, both the green and red paths are active with mini-batch update. After the network is fully trained, only the red path is used to obtain the HR-HSI. To illustrate the spatial correspondence between the hyperspectral and multispectral mini-batches, we choose pixels with index $j \in \mathcal{M}$ as an example mini-batch.

prior network, and the spatial-spectral variational inference network together, an unsupervised convolutional variational network, FusionNet, is obtained, as shown in Fig. 5.

To simultaneously infer the latent representations $z_i^{(l)}$ and $z_j^{(h)}$, and learn the network parameters $\Theta = \{\boldsymbol{\theta}, \alpha, \boldsymbol{\phi}_l, \boldsymbol{\phi}_h, \boldsymbol{\omega}\}$, we can maximize the ELBO of the joint full likelihood of the LR-HSI and HR-MSI as

$$\mathcal{L}(\Theta; \mathbf{X}_l, \mathbf{X}_m) = \sum_{i=1}^{ab}(E_1 - KL_1) + \sum_{j=1}^{AB}(E_2 - KL_2) \quad (26)$$

where,

$$E_1 = \mathbb{E}_{q_{\boldsymbol{\phi}_l}(z_i^{(l)}|\boldsymbol{x}_i^{(l)})}\big[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i^{(l)}|z_i^{(l)})\big], \quad (27)$$

$$E_2 = \mathbb{E}_{q_{\boldsymbol{\phi}_h}(z_j^{(h)}|\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)})}\big[\log p_{\{\boldsymbol{\theta},\boldsymbol{\eta},\alpha\}}(\boldsymbol{x}_j^{(m)}|z_j^{(h)})\big], \quad (28)$$

$$KL_1 = \mathcal{KL}\big[q_{\boldsymbol{\phi}_l}(z_i^{(l)}|\boldsymbol{x}_i^{(l)})||p(z_i^{(l)})\big], \quad (29)$$

$$KL_2 = \mathbb{E}_{q_{\boldsymbol{\phi}_l}(z_i^{(l)}|\boldsymbol{x}_i^{(l)})}\mathcal{KL}\big[q_{\boldsymbol{\phi}_h}(z_j^{(h)}|\boldsymbol{x}_{j,block}^{(m)}, \boldsymbol{x}_j^{(b)})||p_{\boldsymbol{\omega}}(z_j^{(h)}|\overline{z}_j^{(l)})\big]. \quad (30)$$

The term $\mathcal{KL}[\cdot||\cdot]$ represents the Kullback-Leibler (KL) divergence. E_1 and E_2 respectively represent expected negative reconstruction errors of the pixels in LR-HSI and HR-MSI with the latent representations drawn from the variational distributions, describing the capability of reconstructing the data, while KL_1 and KL_2 act as regularizers. Detailed derivations of the ELBO can be seen in the supplementary file.

Compared with the two-step models inferring the LR and HR latent representations separately [6], [10], [12], [25], and with the alternating and iterative parameter inference [15]–[18], our method is able to take advantage of the joint optimization. Namely, during every iteration, all parameters and latent representations are updated under the synergistic effects of the LR and HR data representation and constraints, which results in better information interaction and fusion. In addition, it should be pointed out that the KL_2 in (30) measures the KL divergence between the HR variational distribution and the spatial-dependent prior given $z_i^{(l)}$ drawn from the LR inference model. In other words, the spatial-dependent prior further promotes the interactions between optimizing the LR and HR recognition models.

By applying the stochastic gradient variational Bayes method [24], the ELBO $\mathcal{L}$ can be efficiently optimized based on stochastic gradient methods such as Adam [33], where several patches randomly sampled from $\mathbf{X}_l$ and $\mathbf{X}_m$ are treated as a mini-batch. Therefore, our approach can deal satisfactorily with large scenes, *e.g.*, remote sensing imagery [16]. After the model is well trained, the HR recognition model and generative model lead to a fusion path to obtain the retrieved HR-HSI, as the red route shown in Fig. 5.

E. FusionNet With Fast Adaptation

As the remote sensors often perform continuous monitoring, a sequence of historical data recording similar scenes is available. These additional LR-HSI and HR-MSI image pairs can be used for learning purposes. Although supervised fusion methods, *e.g.*, [28], are good at training and testing machanisms, it is infeasible for them to accomplish such learning task, due to a lack of supervised information, *i.e.*, HR-HSI of the historical scene. On the contrary, the FusionNet does not require any supervised information. Thus, the available additional data can be treated as training data to learn the parameters of FusionNet in an unsupervised manner, while the incoming LR-HSI and HR-MSI image pair can be treated as testing data, directly using the well-learned parameters for fusion without any update. Although this strategy helps with efficient testing, one is not sure that the learned model matches the testing data, since the distribution of testing data often differs from that of the training data.

Motivated by [34], we give the fusion problem a meta-learning explanation to solve this problem. Suppose we have

Algorithm 1 FusionNet by Meta-Training

Require: β: step size hyperparameter;
Require: C: number of gradient descent updates of inner loop
Randomly initialize Θ
while not done **do**
  Sample a batch of N tasks $\mathcal{T} = \{\mathcal{T}_n\}_{n=1,\ldots,N}$
  for all $\mathcal{T}_n$ **do**
    Sample a mini-batch of data $\mathbf{D}_{\mathcal{T}_n,pre}$
    Initialize $\Theta'_n \leftarrow \Theta$
    for $c = 1$ **to** C **do**
      $\Theta'_n \leftarrow \Theta'_n - \beta\nabla_{\Theta'_n}[-\mathcal{L}(\Theta'_n; \mathbf{D}_{\mathcal{T}_n,pre})]$
    end for
    Sample another mini-batch of data $\mathbf{D}_{\mathcal{T}_n,obj}$
  end for
  Obtain the meta-objective: $\sum_n -\mathcal{L}(\Theta'_n; \mathbf{D}_{\mathcal{T}_n,obj})$
  Update the meta parameters Θ via Adam.
  $\Theta \leftarrow \Theta - AdamUpdate[\sum_n -\mathcal{L}(\Theta'_n; \mathbf{D}_{\mathcal{T}_n,obj})]$
end while

observed a variety of LR-HSI and HR-MSI image pairs, every image pair corresponding to a fusion task. Our goal of meta-training is to learn the parameters of FusionNet on these tasks, such that, during meta-testing, a few iterations on the new task can produce good results. In other words, we intend to learn good initial parameters (or meta parameters) such that the model has maximal performance on a new task after a few updates, *i.e.*, fast adaptation. To accomplish this, we adopt and embody the meta-learning method in [34] to train the FusionNet.

Formally, during the meta-training stage, consider a batch of N fusion tasks $\mathcal{T} = \{\mathcal{T}_n\}_{n=1,\ldots,N}$, with Θ as the meta-parameters of the FusionNet. As adapting to task $\mathcal{T}_n$, Θ is updated as Θ'_n. For example, using one-step stochastic gradient update, the updated parameters Θ'_n is computed via:

$$\Theta'_n = \Theta - \beta\nabla_{\Theta}[-\mathcal{L}(\Theta; \mathbf{D}_{\mathcal{T}_n,pre})], \tag{31}$$

where, β is the learning rate, $\mathbf{D}_{\mathcal{T}_n,pre}$ is a mini-batch containing patches in the LR-HSI and HR-MSI from task $\mathcal{T}_n$, $\mathcal{L}(\Theta; \mathbf{D}_{\mathcal{T}_n,pre})$ is the ELBO on the likelihood of $\mathbf{D}_{\mathcal{T}_n,pre}$. An extension of the multiple gradient update can be seen in Algorithm 1. We hope that the task-specific Θ'_n performs well on task $\mathcal{T}_n$. For this purpose, the meta-objective is stated as:

$$\min_{\Theta} \sum_n -\mathcal{L}(\Theta'_n; \mathbf{D}_{\mathcal{T}_n,obj}) \tag{32}$$

where, $\mathbf{D}_{\mathcal{T}_n,obj}$ is another mini-batch sampled from $\mathcal{T}_n$. The meta-objective aims to optimize the meta-parameters Θ such that a few task-specific updates will produce maximally effective behavior on that task. The overall algorithm is outlined in Algorithm 1.

After obtaining the optimal meta-parameters, we treat them as the initial parameters for the incoming fusion task, helping the FusionNet to perform well on that task with much fewer updates than before.

IV. Experiments

A. Comparison Approaches and Metrics

We compare our approach with the state-of-the-art unsupervised fusion methods.

- **SNNMF** [15]: A linear sparse non-negative matrix factorization method that encourages non-negativity on the spectral signatures and both non-negativity and sparsity on the mixture proportions.
- **GSOMP** [10]: A generalized simultaneous OMP based method that separately estimates the spectral signatures and mixture proportions via solving two matrix decomposition problems.
- **SSR** [19]: A subspace regularization method that optimizes a linear matrix factorization problem via an Alternating Direction Method of Multipliers (ADMM) approach.
- **BSR** [11]: A Bayesian sparse representation approach that is derived from the GSOMP, but employs Bayesian methods to realize the dictionary learning and sparse coding.
- **HBP-GP** [12]: A hierarchical beta process with Gaussian process prior model that represents the spectral smoothness and spatial consistency using Gaussian processes and a hierarchical beta-Bernoulli process, respectively.
- **NLSTF** [25]: A non-local sparse tensor factorization based approach that constructs a two-step model.
- **uSDN** [6]: An unsupervised sparse Dirichlet-net that separately optimize two autoencoders.

To evaluate the quality of the fusion images, three commonly used metrics are considered in our study, namely, the root mean square error (RMSE) measured on 8-bit images, the spectral angle mapper (SAM) [14], [18] given in degrees, and the structural similarity (SSIM) [35]. Let us denote the retrieved image as $\mathbf{Y}_h$, and its j-th pixel as $\boldsymbol{y}_j^{(h)}$. The definitions of these metrics are as follows:

RMSE The root mean square error measures the numerical similarity between the target HR-HSI $\mathbf{X}_h$ and the retrieved image $\mathbf{Y}_h$, where the scale is in the range of 8 bit, i.e., $0-255$.

$$\text{RMSE}(\mathbf{X}_h, \mathbf{Y}_h) = \sqrt{\frac{\|\mathbf{Y}_h - \mathbf{X}_h\|_F^2}{ABC}}, \tag{33}$$

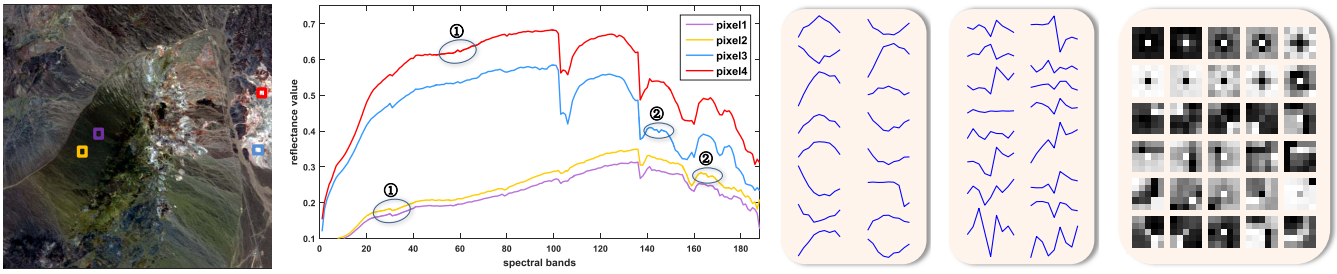
where, $\|\cdot\|_F$ is the Frobenius norm. The smaller the RMSE value, the better the fusion performance.

SAM The spectral angle mapper describes the angle between the target pixel $\boldsymbol{x}_j^{(h)}$ and the retrieved pixel $\boldsymbol{y}_j^{(h)}$, $j = 1, \ldots, AB$. The overall SAM, expressed in degrees, is obtained by averaging over the whole image as

$$\text{SAM} = \frac{1}{AB}\sum_{j=1}^{AB} \arccos \frac{\boldsymbol{y}_j^{(h)T}\boldsymbol{x}_j^{(h)}}{\|\boldsymbol{y}_j^{(h)}\|_2\|\boldsymbol{x}_j^{(h)}\|_2}, \tag{34}$$

where, $\|\cdot\|_2$ is the l_2-norm. The smaller the SAM value, the less the spectral distortion.

SSIM We firstly measure the structural similarity of every spectral band, and then the overall SSIM is obtained by



(a) (b) (c) (d) (e)

Fig. 6. Visualizations on the spectral structure and the learned filters. (a) shows the true color image of the AVIRIS Cuprite "sc04". Each curve in (b) represents the spectral vector corresponding to the pixel with the same color in (a). The first-layer convolutional filters and the last-layer deconvolutional filters are shown on (c) and (d), respectively. Some learned 2D filters are shown in (e).

averaging over all spectral bands as

$$\text{SSIM} = \frac{1}{C}\sum_{k=1}^{C}\frac{(2\mu_x^k\mu_y^k + d_1)(2\sigma_{xy}^k + d_2)}{\big((\mu_x^k)^2 + (\mu_y^k)^2 + d_1\big)\big((\sigma_x)^2 + (\sigma_y)^2 + d_2\big)}, \tag{35}$$

where, μ_x^k and μ_y^k are the means of the k-th spectral band in $\mathbf{X}_h$ and $\mathbf{Y}_h$, respectively, $(\sigma_x^k)^2$ and $(\sigma_y^k)^2$ are the variances of the k-th spectral band in $\mathbf{X}_h$ and $\mathbf{Y}_h$, respectively, σ_{xy}^k is the correlation coefficient between the k-th spectral images in $\mathbf{X}_h$ and $\mathbf{Y}_h$, d_1 and d_2 are constants. We follow the setting in [35]. The higher the SSIM score, the better the fusion quality.

TABLE II
QUANTITATIVE RESULTS ON THREE DATASETS

Dataset	CAVE			Harvard			AVIRIS	
Criterion	RMSE	SAM	SSIM	RMSE	SAM	SSIM	RMSE	SAM
SNNMF [15]	4.38	17.85	0.918	2.46	4.93	0.973	NA	NA
SSR [19]	4.71	22.00	0.945	3.08	5.59	0.820	NA	NA
GSOMP [10]	5.44	12.23	0.960	3.10	4.34	0.971	1.52	1.64
BSR [11]	5.19	12.93	0.955	2.64	4.48	0.974	1.48	1.61
HBP-GP [12]	NA	NA	NA	NA	NA	NA	1.18	1.41
NLSTF [25]	2.60	6.83	0.980	1.78	3.12	0.982	NA	NA
uSDN [6]	4.09	6.95	NA	1.78	4.05	NA	NA	NA
FusionNet	**1.88**	**6.79**	**0.983**	**1.36**	**2.94**	**0.984**	**1.07**	**1.35**

B. Datasets

Evaluations are conducted on commonly used two ground-based datasets, the CAVE [36] and the Harvard [37] datasets, and one remote sensing dataset, the AVIRIS Cuprite.[1]

The CAVE dataset consists of 32 hyperspectral images of everyday objects, *e.g.*, statue, food, captured at a wavelength interval of 10nm in the range 400−700nm and with the size of $512 \times 512 \times 31$. The Harvard dataset contains 50 real-world indoor and outdoor images with 31 spectral bands ranging from 420 nm to 720 nm at an interval of 10 nm. The spatial resolution of the Harvard images is 1392×1040 pixels, while only the top left 1024×1024 pixels are used for the convenience of the spatial down-sampling process. We consider the hyperspectral images from the two datasets as ground truth images to simulate LR-HSIs and HR-MSIs. Following [6], [10], [11], we down-sample a ground truth image by averaging over 32×32 disjoint spatial blocks to generate the LR-HSI. The HR-MSI is acquired by integrating a ground truth HSI over the spectral dimension using the SRF derived from the Nikon D700 camera.[2]

The AVIRIS Cuprite, a more challenging dataset, is remotely sensed by the NASA's Airborne Visible and Infrared Imaging Spectrometer (AVIRIS) over the Cuprite mining district in Nevada [38], containing four images with the size of $512 \times 512 \times 224$, and ranging from 370nm to 2500nm. Following [10], [11], 36 bands are removed considering water absorptions and low signal-to-noise in these bands, resulting images with 188 bands considered as the ground truth. The LR-HSIs are simulated as before with scaling factor 32, while the HR-MSIs are obtained using a binary matrix which directly select six bands of the ground truth, corresponding to the wavelengths 480nm, 560nm, 660nm, 830nm, 1650nm, and 2220nm, because these bands roughly correspond to the visible and mid-infrared channels of NASA-Landsat 7 satellite.

C. Settings

The parameters of FusionNet are initialized by random sampling from $\mathcal{N}(0, 0.01)$, and optimized via Adam [33] with default parameters and 500 epochs. For each mini-batch, we randomly crop a 16×16 MSI cube and use all the pixels in the LR-HSI, since the LR-HSI has only hundreds of pixels as the scaling factor is large. The network architecture for the AVIRIS dataset is shown in Table I, where the output size corresponds to a mini-batch, and the activations are tanh. The architecture for the CAVE and Harvard datasets is given in the supplementary file.

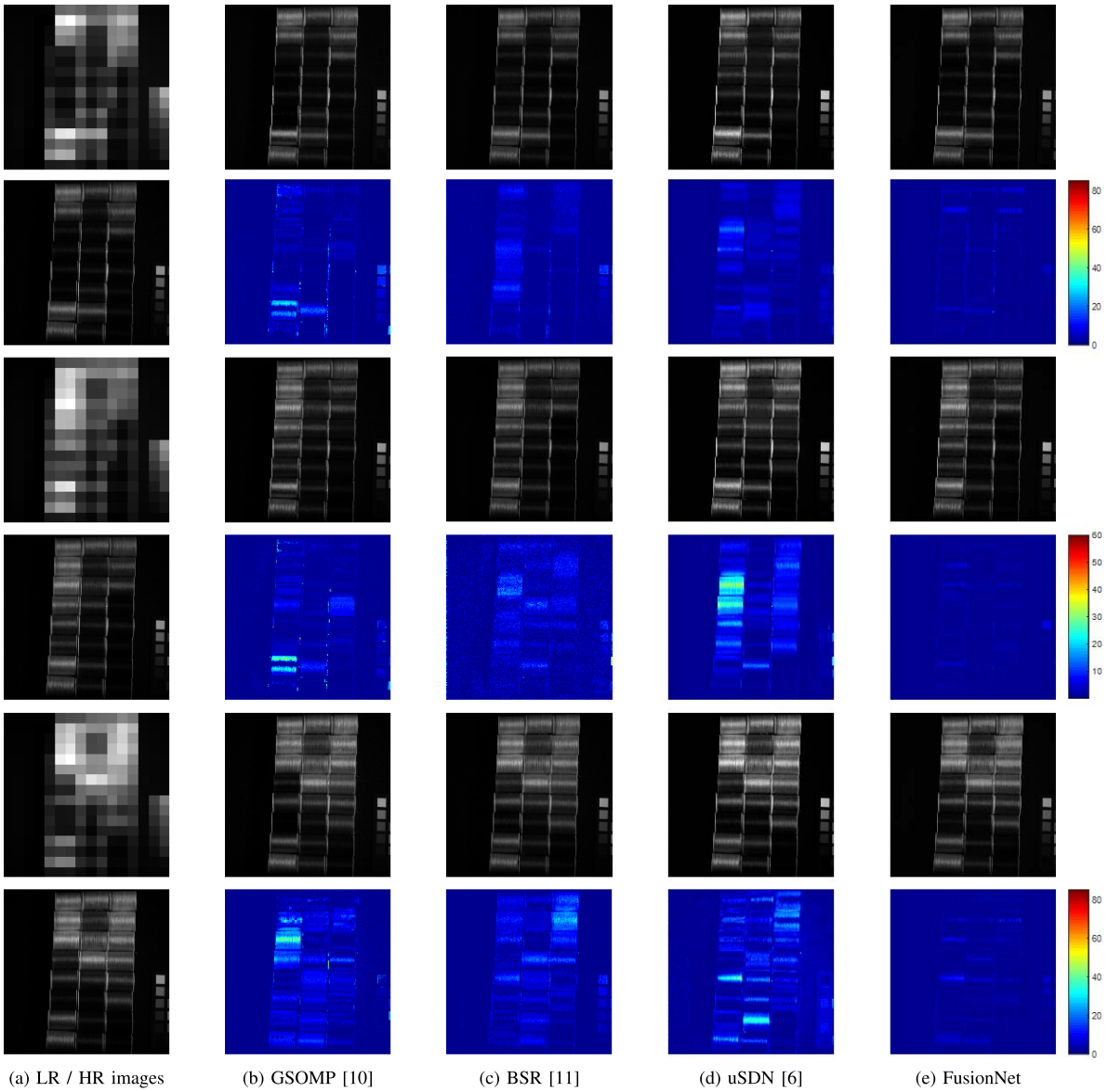
D. Quantitative and Qualitative Results

The average RMSE and SAM corresponding to the CAVE, Harvard, and AVIRIS datasets are list in Table II. The best results are marked in bold for clarity. The comparison results are obtained based on publicly available references [6], [10]–[12], [25]. Since the authors of [6], [15], [19], [25] did not perform experiments on the challenging AVIRIS dataset, and the authors of HBP-GP did not report the average results on the CAVE and Harvard dataset, and the code

[1] Available at http://aviris.jpl.nasa.gov/data/free_data.html
[2] Available at https://www.maxmax.com/spectral_response.htm

TABLE III
THE RMSE RESULTS ON SOME BENCHMARKED IMAGES

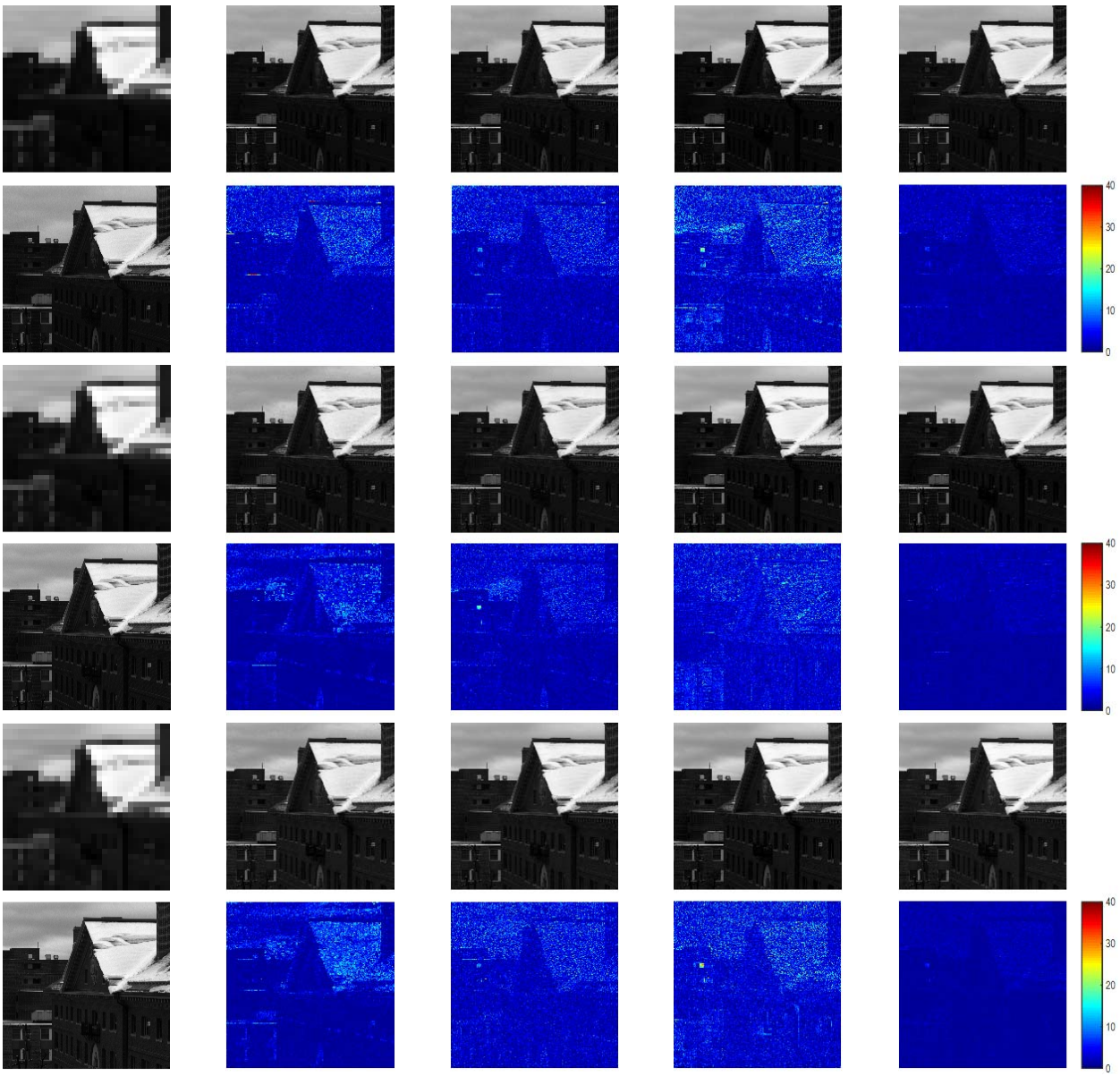
Dataset	CAVE					Harvard		AVIRIS	
Image	balloon	CD	cloth	photo	spool	img1	imgb5	sc02	sc04
BS [14]	14.2	15.3	17.6	11.3	15.2	10.9	14.7	NA	NA
SSR [19]	14.9	20.3	14.8	4.6	12.5	4.4	5.4	NA	NA
MF [8]	2.3	7.9	6.0	3.3	8.4	3.9	2.4	1.55	2.73
SSFM [9]	2.4	8.1	7.6	3.7	6.1	4.3	2.6	1.56	2.68
GSOMP [10]	2.3	7.5	4.0	2.2	5.0	1.2	0.9	1.54	1.53
BSR [11]	2.1	5.3	4.0	1.6	4.6	1.1	0.9	1.65	1.57
HBP-GP [12]	1.9	5.3	3.7	NA	NA	NA	0.8	1.32	1.36
NLSTF [25]	1.3	5.3	3.7	NA	NA	NA	NA	NA	NA
uSDN [6]	1.8	4.8	3.7	2.0	5.3	2.0	0.7	NA	NA
FusionNet w/o SD prior	1.10	4.22	2.89	1.44	2.55	1.22	0.60	1.22	1.25
FusionNet	**1.02**	**3.90**	**2.85**	**1.42**	**2.32**	**0.78**	**0.58**	**1.11**	**1.19**



(a) LR / HR images (b) GSOMP [10] (c) BSR [11] (d) uSDN [6] (e) FusionNet

Fig. 7. Fusion results of the image "spool" from the CAVE dataset with scaling factor 32 at 460nm, 540nm, and 620nm, corresponding to the 1-2 rows, 3-4 rows, and 5-6 rows, respectively. In column (b)-(e), the reconstructed images are shown in grayscale, while the absolute-error images are shown in color.

is not available, we indicate the non-availability by entering "NA" in Table II. Besides, considering the fact that all these authors did not provide the SSIM results on the AVIRIS dataset, for fairness, we do not report the corresponding



(a) LR / HR images (b) GSOMP [10] (c) BSR [11] (d) uSDN [6] (e) FusionNet

Fig. 8. Fusion results of the image "img1" from the Harvard dataset with scaling factor 32 at 460nm, 540nm, and 620nm, corresponding to the 1-2 rows, 3-4 rows, and 5-6 rows, respectively. In column (b)-(e), the reconstructed images are shown in grayscale, while the absolute-error images are shown in color.

comparisons in Table II. The average SSIM score of the FusionNet is 0.984 for the AVIRIS dataset.

Among the fusion methods considered for comparison, SNNMF, SSR, GSOMP, BSR, and HBP-GP are matrix factorization based methods. Compared with them, the tensor decomposition based method NLSTF further exploits the spatial information, while the autoencoder based method uSDN better explores the underlying nonlinear data structure. As expected, both NLSTF and uSDN show better RMSE and SAM performance, and NLSTF also show better SSIM performance. Information fusion in NLSTF and uSDN is mainly realized via sharing decoder parameters. In addition, NLSTF and uSDN are two-step models, leading to relatively poor interactions between the LR-HSI and HR-MSI during the fusion process. On the contrary, the proposed FusionNet concentrates on effective information extraction, *i.e.*, spatial correlation and local spectral structure, and harmonious information fusion via spectral generation, spatial-dependent prior, spatial-spectral variational inference, and joint optimization, leading to superior RMSE, SAM, and SSIM performance than the other methods. In addition, the FusionNet outperforms other methods when applied on the remotely sensed AVIRIS dataset, showing its potential in practical applications.

According to [6], [10]–[12], [25], some images from these datasets are often used as benchmarks. Following them, we also list the RMSE scores for each of these images, as shown in Table III. The outstanding RMSE scores further illustrate the effectiveness of the proposed FusionNet.

To better demonstrate the fusion results in spectral domain, Fig. 9 shows the RMSE curves as functions of the wavelengths

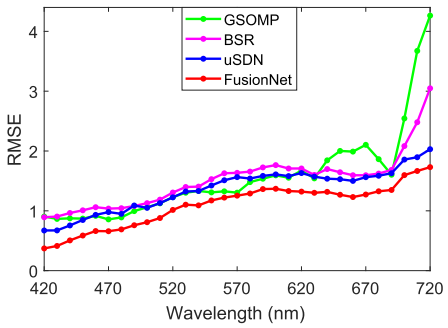
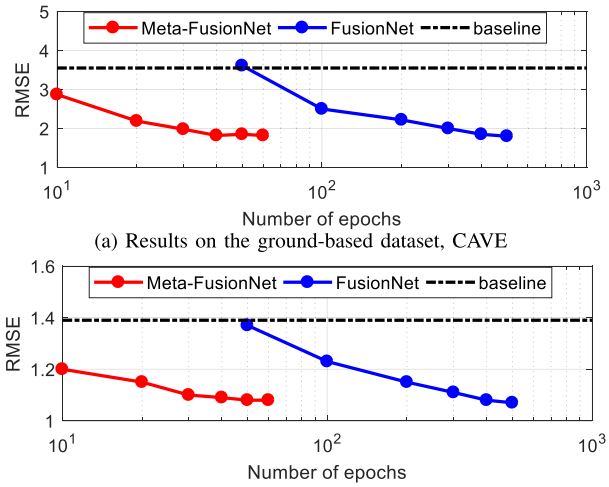


Fig. 9. Spectral retrieval results of the img "imgb5" from the Harvard dataset, presenting the RMSE curves as functions of the wavelengths of spectral bands.

of spectral bands on image "imgb5" from the Harvard dataset. It can be seen that the FusionNet outperforms in every spectral band.

For qualitative performance analysis, Figs. 7 and 8 show some reconstructed images and the corresponding absolute-error images. Clearly, the FusionNet is good at recovering the details of the target HR-HSI and has less artifacts. More visual results can be found in the supplementary file.



(a) Results on the ground-based dataset, CAVE

(b) Results on the remote sensing dataset, AVIRIS

Fig. 10. Convergence comparison on (a) CAVE and (b) AVIRIS. The red lines indicate the RMSE performance against meta-testing epochs. The blue lines indicate the RMSE performance against learning epochs. We adopt traditional learning strategy to train the FusionNet on training images, and to test on the other images, whose average RMSE is treated as the baseline.

E. Discussion

1) Spatial-Dependent Prior: For better illustration of the effectiveness and efficiency of the proposed spatial-dependent (SD) prior, we perform a comparison experiment where the SD prior is replaced by $\mathcal{N}(\mathbf{0}, \mathbf{I})$, termed as "FusionNet w/o SD prior". As is clear from the RMSE results listed in Table III, the proposed FusionNet and the FusionNet w/o SD prior consistently perform better than the others, illustrating the effectiveness of spectral generative network and spatial-spectral variational network. Besides, the SD prior further assists FusionNet to fuse spatial information, improving the performance.

2) Spectral and Spatial Correlations: According to the spectral visualization shown in Fig. 6, similar materials have similar reflectance spectra, (see pixels 1 and 2, pixels 3 and 4), while different materials show clear spectral differences (see pixels between 1, 2, and 3, 4). However, there exist local structural similarities among different pixels, as marked by the circles. The 1D convolutional and 1D deconvolutional filters are aimed at extracting and retrieving these local structures, respectively. As shown in Fig. 6, the learned 1D filters show various structures with high diversity, validating the effectiveness of employing convolutional networks to describe the spectral characteristics. Besides, the learned 2D filters also show diverse structures, such as arc, corner, edge, point, center-surrounding, assisting the FusionNet to capture the spatial information.

3) FusionNet With Meta-Learning: The performance of FusionNet with meta-learning is evaluated based on a ground based dataset and a remote sensing dataset, *i.e.*, CAVE and AVIRIS. For the AVIRIS, the meta-parameters are learned from three quarters of the dataset with 10 epochs, and are tested on the other image. For the CAVE dataset, we randomly choose half of the dataset, *i.e.*, 16 images, for meta-training tasks, and the others are used for meta-testing tasks. Since, compared with images from AVIRIS, the differences among images from CAVE are more significant, the meta-parameters for CAVE are learned with more epochs, *i.e.*, 25 epochs. During the meta-training stage, the step size hyperparameter β is set as 10^{-5}, while the number of gradient descent updates of inner loop C is set as 5. The network architectures are set as before. After obtaining the meta-parameters, we treat them as the initial-parameters for the meta-testing tasks, termed as "Meta-FusionNet".

Using the same data partition, we adopt a traditional learning strategy to train the FusionNet on the meta-training images, and to test on the other images, whose average RMSE is treated as the baseline. Namely, for the incoming images, the baseline model does not perform any optimization on these images. Due to the differences between training and testing images, the baseline model does not perform satisfactorily, as shown in Fig. 10. In contrast, the optimization of the original FusionNet is concentrated on the fusion task at hand, achieving better results along with the learning epochs. Further, the Meta-FusionNet utilizes the information in both the extra training tasks and the task at hand, achieving good results with faster convergence than the original FusionNet. In other words, the meta-training exploits training tasks to learn good initial parameters, while the meta-testing further specializes the model on a certain task. In order to better understand the meta-testing, we look deeper into the meta-filters and observe their evolutions during meta-testing. As shown in Fig. 11, each meta-filter is updated as 16 different filters to adapt 16 fusion tasks. Surprisingly, the updated 16 filters maintain the tendency of the corresponding meta-filter. For example, the fourth meta-filter looks like letter "V". After meta-testing, the corresponding 16 filters also behave as letter "V", higher or lower, or more straight or more slanting. Such phenomenon further demonstrates that the parameters learned via meta-training

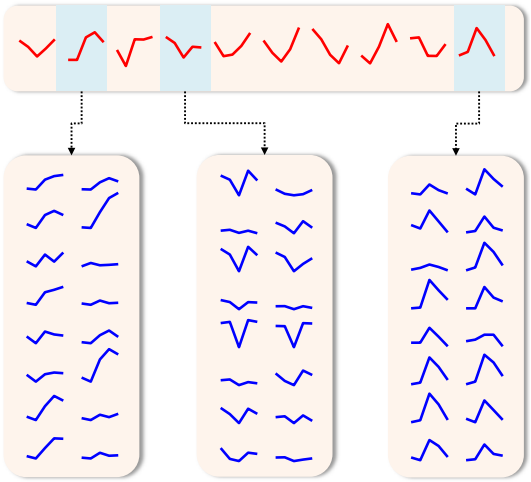


Fig. 11. The evolutions of the meta-filters on CAVE dataset. After meta-training, we obtain the meta-parameters, where the last-layer deconvolutional filters are shown in red color. After meta-testing, the meta-parameters are updated to adapt 16 fusion tasks, *i.e.*, each filter is updated as 16 ones, shown in blue color.

TABLE IV
COMPARISONS ON THE LEARNING TIME

Dataset	FusionNet	Meta-FusionNet	
		meta-training	meta-testing
CAVE	40min	14h	3min
AVIRIS	1.6h	1.75h	12min

have good generalization performance. The learning time of the proposed models are summarized in Table IV, which are evaluated on the Tensorflow platform [39] with one 1080Ti GPU. The results for the FusionNet and meta-testing correspond to one image, while the meta-training time corresponds to all training images. Compared with CAVE, images from the AVIRIS dataset have more spectral bands, which need more time for the FusionNet to train the model. With more training tasks and training epochs, the Meta-FusionNet spends more time to finish the meta-training for the CAVE than the AVIRIS. Fortunately, the meta-training process can be further accelerated with more GPUs. Thanks to the offline meta-training, the Meta-FusionNet increases the running speed as testing on a new fusion task.

V. CONCLUSION

We have proposed an unsupervised convolutional variational network, FusionNet, for the task of hyperspectral and multispectral image fusion. FusionNet consists of a spectral generative network, a spatial-dependent prior network, and a spatial-spectral variational inference network, which are jointly optimized in an unsupervised manner. To the best of our knowledge, this is the first effort solving such an unsupervised fusion task via convolutional network, and the first effort that applies meta-learning for fast adaptation to various fusion tasks. Experimental results on three benchmark datasets validate the effectiveness and efficiency of the proposed model quantitatively and qualitatively, demonstrating the superiority of the proposed approach over state-of-the-art. For future works, FusionNet can be extended by jointly considering information fusion and subsequential remote sensing tasks, *e.g.*, simultaneous image fusion and pixel classification.

REFERENCES

[1] G. Lu and B. Fei, "Medical hyperspectral imaging: A review," *J. Biomed. Opt.*, vol. 19, no. 1, Jan. 2014, Art. no. 10901.

[2] M. Govender, K. Chetty, and H. Bulcock, "A review of hyperspectral remote sensing and its application in vegetation and water resource studies," *Water SA*, vol. 33, no. 2, pp. 145–151, 2007.

[3] J. M. Bioucas-Dias *et al.*, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 2, pp. 354–379, Apr. 2012.

[4] R. C. Patel and M. V. Joshi, "Super-resolution of hyperspectral images using compressive sensing based approach," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. I–7, pp. 83–88, Jul. 2012.

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

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

[7] S. Lu, X. Ren, and F. Liu, "Depth enhancement via low-rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3390–3397.

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

[9] B. Huang, H. Song, H. Cui, J. Peng, and Z. Xu, "Spatial and spectral image fusion using sparse matrix factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1693–1704, Mar. 2014.

[10] R. Dian, S. Li, and L. Fang, "Non-local sparse representation for hyperspectral image super-resolution," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 63–78.

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

[12] N. Akhtar, F. Shafait, and A. Mian, "Hierarchical beta process with Gaussian process prior for hyperspectral image super resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 103–120.

[13] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of hyperspectral and multispectral images," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3176–3180.

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

[15] E. Wycoff, T.-H. Chan, K. Jia, W.-K. Ma, and Y. Ma, "A non-negative sparse promoting algorithm for high resolution hyperspectral imaging," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 1409–1413.

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

[17] Y. Zhou, L. Feng, C. Hou, and S.-Y. Kung, "Hyperspectral and multispectral image fusion based on local low rank and coupled spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5997–6009, Oct. 2017.

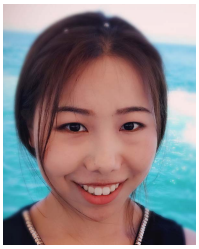
[18] W. Dong *et al.*, "Hyperspectral image super-resolution via non-negative structured sparse representation," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2337–2352, Mar. 2016.

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

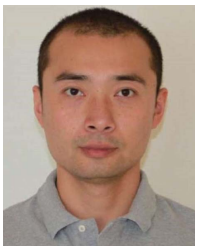
[20] Q. Wei, N. Dobigeon, and J.-Y. Tourneret, "Bayesian fusion of multispectral and hyperspectral images with unknown sensor spectral response," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 698–702.

[21] R. Heylen, M. Parente, and P. Gader, "A review of nonlinear hyperspectral unmixing methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 1844–1868, Jun. 2014.

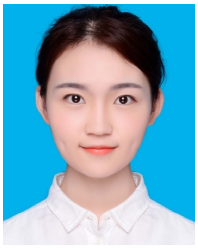
[22] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.
[23] H. Zhang, B. Chen, D. Guo, and M. Zhou, "WHAI: Weibull hybrid autoencoding inference for deep topic modeling," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.
[24] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2013, pp. 1–14.
[25] R. Dian, L. Fang, and S. Li, "Hyperspectral image super-resolution via non-local sparse tensor factorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3862–3871.
[26] C. Dong, C. L. Chen, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.
[27] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1637–1645.
[28] J. Yang, Y.-Q. Zhao, and J. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, p. 800, May 2018.
[29] Q. Xie, M. Zhou, Q. Zhao, D. Meng, W. Zuo, and Z. Xu, "Multispectral and hyperspectral image fusion by MS/HS fusion net," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–14.
[30] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.
[31] 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 *Proc. 18th Int. Conf. Geoinform.*, Jun. 2010, pp. 1–4.
[32] Z. Wang, B. Chen, H. Zhang, and H. Liu, "Variational probabilistic generative framework for single image super-resolution," *Signal Process.*, vol. 156, pp. 92–105, Mar. 2019.
[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. for Learn. Represent.*, 2015, pp. 1–13.
[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1–13.
[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[36] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, "Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.
[37] A. Chakrabarti and T. Zickler, "Statistics of real-world hyperspectral images," in *Proc. CVPR*, Jun. 2011, pp. 193–200.
[38] R. O. Green *et al.*, "Imaging spectroscopy and the airborne visible/infrared imaging spectrometer (AVIRIS)," *Remote Sens. Environ.*, vol. 65, no. 3, pp. 227–248, Sep. 1998.
[39] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.



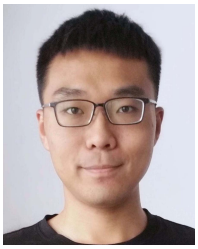
Zhengjue Wang received the B.S. and M.S. degrees in electronic engineering from Xidian University, Xi'an, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with Xidian University. Her research interests include probabilistic model and deep learning, and their applications in image super-resolution, hyperspectral image fusion, and natural language processing.



Bo Chen (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Xidian University, Xi'an, China, in 2003, 2006, and 2008, respectively, all in electronic engineering. He became a Postdoctoral Fellow, a Research Scientist, and a Senior Research Scientist with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA, from 2008 to 2012. From 2013, he has been a Professor with National Laboratory for Radar Signal Processing, Xidian University. His current research interests include statistical machine learning, statistical signal processing, and radar automatic target detection and recognition. He received the Honorable Mention for the 2010 National Excellent Doctoral Dissertation Award and is selected into Oversea Talent by the Chinese Central Government, in 2014.



Ruiying Lu received the B.S. degree in telecommunication engineering from Xidian University, Xi'an, China, in 2016, where she is currently pursuing the Ph.D. degree. Her research interests include deep learning for image processing and natural language processing.



Hao Zhang received the B.S. and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2012 and 2019, respectively. He is currently working as a Postdoctoral Researcher with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. His research interests include statistical machine learning and its combination with deep learning, and the natural language processing.

Hongwei Liu (Member, IEEE) received the M.S. and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 1995 and 1999, respectively. From 2001 to 2002, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. He is currently a Professor with the National Laboratory of Radar Signal Processing, Xidian University. His research interests include radar automatic target recognition, radar signal processing, and adaptive signal processing.

Pramod K. Varshney (Life Fellow, IEEE) received the B.S. degree (Hons.) in electrical engineering and computer science from the University of Illinois at Urbana–Champaign, in 1972, and the M.S. and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana–Champaign, in 1974 and 1976, respectively. Since 1976, he has been with Syracuse University, Syracuse, NY, where he is currently a Distinguished Professor in electrical engineering and computer science and the Director of the Center for Advanced Systems and Engineering (CASE). He is the author of *Distributed Detection and Data Fusion* (New York: Springer Verlag, 1997). His current research interests include distributed sensor networks and data fusion, detection and estimation theory, wireless communications, and image processing. He was a recipient of the 1981 ASEE Dow Outstanding Young Faculty Award. He was elected a fellow of the IEEE for his contributions in the area of distributed detection and data fusion, in 1997. In 2000, he received the Third Millennium Medal from the IEEE and Chancellor's Citation for exceptional academic achievement with Syracuse University. He was the President of the International Society of Information Fusion, during 2001. He was also a recipient of the IEEE 2012 Judith A. Resnik Award, the ECE Distinguished Alumni Award from UIUC, in 2015, and the ISIF's Yaakov Bar-Shalom Award for a Lifetime of Excellence in Information Fusion, in 2018. He is on the editorial board of the *Journal on Advances in Information Fusion*.