Spatial and Spectral Extraction Network With Adaptive Feature Fusion for Pansharpening

Kai Zhang, *Member, IEEE,* Anfei Wang, Feng Zhang, Wenxiu Diao, Jiande Sun, Lorenzo Bruzzone, *Fellow,* IEEE

***Abstract*—Pansharpening methods based on deep neural networks (DNNs) have been attracting great attention due to their powerful representation capabilities. In this paper, to combine the feature maps from different sub-networks efficiently, we propose a novel pansharpening method based on a spatial and spectral extraction network (SSE-Net). Differently from the other methods based on DNNs that directly concatenate the features from different sub-networks, we design adaptive feature fusion modules (AFFMs) to merge these features according to their information content. First, the spatial and spectral features are extracted by the sub-networks from low spatial resolution multispectral (LR MS) and panchromatic (PAN) images. Then, by fusing the features at different levels, the desired high spatial resolution multispectral (HR MS) images are generated by the fusion network consisting of AFFMs. In the fusion network, the features from different sub-networks are integrated adaptively and the redundancy among them is reduced. Moreover, spectral ratio loss and gradient loss are defined to ensure the effective learning of spatial and spectral features. The spectral ratio loss captures the nonlinear relationships among the bands in the MS image to reduce the spectral distortions in the fusion result. Extensive experiments were conducted on QuickBird and GeoEye-1 satellite datasets. Visual and numerical results demonstrate that the proposed method produces better fusion results when compared with literature techniques. The source code is available at https://github.com/RSMagneto/SSE-Net.**

***Index Terms*—Pansharpening, spatial extraction network, spectral extraction network, adaptive feature fusion, spectral ratio loss, remote sensing.**

I. Introduction

WITH the development of imaging technology, more and more multispectral (MS) and panchromatic (PAN) images have been collected and widely used in various applications, such as land-cover classification [1] and change detection [2]. However, very high spatial resolution MS (HR MS) images are not available due to sensor physical limitations [3]. These images can be achieved for some applications by using pansharpening techniques [4]-[5]. Most remote sensing satellites, such as QuickBird and GeoEye-1, acquired low spatial resolution (LR) MS images and PAN images simultaneously. In these cases, pansharpening has been used to produce HR MS images, which can efficiently integrate the spatial and spectral information present in LR MS and PAN images, respectively. In other words, pansharpening aims to enhance the spatial resolution of the MS image by leveraging on the PAN image.

Over the past decades, a large number of pansharpening methods have been explored and presented in the literature. They can be categorized into four groups: 1) component substitution (CS) based methods; 2) multiresolution analysis (MRA) based methods; 3) spatial and spectral degradation model (SSDM) based methods, and 4) deep learning (DL) based methods.

In the first kind of methods, a specific transformation is usually applied to the up-sampled LR MS image to separate the spatial and spectral components. Then, the PAN image is considered to replace the spatial component of the LR MS image. Finally, the HR MS image is obtained through the corresponding inverse transformation of the spectral components and the new spatial component. Different projections can be considered in these methods, such as intensity-hue-saturation (IHS) transformation [6], principal component analysis [7], and Gram-Schmidt (GS) transformation [8]. For example, adaptive IHS [9] was proposed to extract the spectral features according to the edge information in the LR MS image. Nonlinear IHS [10] was utilized to estimate a more reasonable intensity component by local and global synthesis approaches. CS-based methods are widely used due to their simple principles and fast implementation. However, the spectral distortions in the fusion results of these methods cannot be ignored. Recently, the band-dependent spatial detail (BDSD) [11] model and its variants [12]-[13] were explored because they can better preserve the spectral information in the fused image.

In MRA-based methods, it is assumed that the spatial resolution of the LR MS image can be enhanced by injecting the spatial details existing in the PAN image. Many MRA techniques are employed to infer the spatial details from the

This work was supported in part by the Natural Science Foundation of China under Grant 61901246, in part by the China Postdoctoral Science Foundation Grant under Grant 2019TQ0190 and 2019M662432, in part by the China Scholarship Council under Grant 202008370035, in part by the Scientific Research Leader Studio of Ji'nan under Grant 2021GXRC081, the Ningxia Natural Science Foundation under Grant 2021AAC03045, and in part by the Joint Project for Smart Computing of Shandong Natural Science Foundation under Grant ZR2020LZH015. *(Corresponding author: Kai Zhang.)*

Kai Zhang is with the School of Information Science and Engineering, Shandong Normal University, Ji'nan 250358, China, and also with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: zhangkainuc@163.com).

A. Wang, F. Zhang, W. Diao, and J. Sun are with the School of Information Science and Engineering, Shandong Normal University, Ji'nan 250358, China (e-mail: Wanganfei1009@163.com, fengzhangpl@163.com, diaowx0920@163.com, jiandesun@hotmail.com).

L. Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy (e-mail: lorenzo.bruzzone@unitn.it).

PAN image, such as contourlet [14] and curvelet [15]. For instance, the coupled multiresolution decomposition [16] was developed through the combination of modulation transfer functions (MTFs) and wavelet decomposition. Non-subsampled contourlet transform (NSCT) [17] was introduced to efficiently capture the directional details of the PAN image. Inspired by the MRA framework, support value transformation (SVT) [18] was proposed for the high-frequency analysis in LR MS and PAN images. Then, support tensor transformation (STT) [19] was further presented to depict the relationships among the bands of the LR MS image. Although these methods behave well in spectral preservation, the fused images contain spatial distortions caused by the wrong estimation of gain coefficients.

In the third category, the LR MS image is viewed as the spatial degradation result of the HR MS image by down-sampling and blurring. The PAN image can be modeled as the linear combination result of all bands in the HR MS image. Then, the spatial and spectral relationships are formulated as image restoration models, which are solved through optimization algorithms [20]. Due to the ill-posedness of the observation models, different priors are employed to regularize the solution space of the fusion result. As the prevalent regularization, sparse prior was introduced into the pansharpening task [21]. Ghahremania *et al*. [22] constructed a more comprehensive dictionary to ensure better reconstruction results. Moreover, a multi-scale dictionary was trained in [23] to provide a more accurate representation. Inspired by the non-negativity of pixel values, coupled sparse nonnegative matrix factorization was presented [24], in which sparse prior was considered to produce better fusion results. To characterize the structural sparsity, the low-rank property was also exploited for pansharpening [25]. Besides, Palsson *et al*. [26] adopted the total variation (TV) to regularize the spatial and spectral degradation models. In the fused image, spatial details are enhanced by TV. For these methods, the computational time cannot be overlooked due to the number of iterations in the optimization algorithms. Furthermore, it is difficult to describe the degradation models accurately because the relationships between the source images and HR MS images are complex.

Recently, pansharpening methods based on deep neural networks (DNNs) have been developed and achieved good fusion results thanks to their capability of nonlinear approximation [27]. For most of them, the up-sampled LR MS image is directly concatenated with the PAN image, and then the concatenated images are regarded as the inputs of DNNs. For example, Masi *et al*. [28] fed the concatenated images into a convolution neural network (CNN) with three layers, called PNN, to achieve end-to-end training. Subsequently, Scarpa *et al*. [29] further promoted PNN by exploring different architectures. In [30], a novel work was proposed to plug well-trained CNN into an iterative optimization problem. Through the combination of CNN and the degradation model, the generalization of the proposed method was improved greatly. Dian *et al*. [31] further incorporated the learned deep priors with the Sylvester equation derived from the sharpening task and effectively obtain good fusion performance. Ma *et al*. [32] utilized the generative adversarial network (GAN) to obtain the fused image from the concatenation of LR MS and PAN images. Diao *et al*. [33] proposed a multiscale GAN to enrich the spatial details in the LR MS image progressively. Compared with the combination in the original domain, PanNet [34] considered the high-frequency information of LR MS and PAN images as the input of DNNs. Then, a multiscale dilated network [35] was constructed to boost the performance of PanNet. Besides, Jiang *et al*. [36] merged the gradient maps of LR MS and PAN images to enhance the spatial details in the fusion result. Moreover, some methods introduced the concatenation operation into feature levels instead of the original image level. For instance, a two-stream fusion network (TFNet) [37] was defined for pansharpening, in which the feature maps from different sub-networks were all bundled together as the input of the reconstruction network. Inspired by the colorization framework [38], Ozcelik *et al*. [39] injected the spectral information from the LR MS image into the network of spatial details. Moreover, to reduce the complementary information in the feature domain, some methods designed different fusion rules to integrate the feature maps. For example, the bidirectional pyramid network [40] adopted the additive block to merge the feature maps from different networks. Zhang *et al*. [41] used a spatial attention module to strengthen the details in feature maps, which included average- and max-pooling paths.

Although good performance is achieved by the methods based on DNNs, three issues need to be considered. First, the spatial and spectral information in LR MS and PAN images is not efficiently extracted by the sub-networks. In previous methods, the feature maps from different sub-networks are directly put together and fed into the following networks. This is difficult to distinguish whether the spatial and spectral information is learned adequately from the source images in the end-to-end training. Second, the feature maps from the sub-networks cannot be adaptively integrated, which results in spatial distortions in the fusion result. Compared with direct concatenation in the feature domains, the feature maps should be further analyzed to reduce the redundancy among them. Some methods select “choose-max” or average rules to fuse the feature maps, but the hand-crafted strategies ignore the information content in the source images. Third, in DNNs-based methods, the mean squared/absolute error between the fused image and the reference image is generally adopted as loss functions, which cannot capture the nonlinear relationships among the bands of the MS image.

To address the problems mentioned above, we present a new pansharpening method based on a spatial and spectral extraction network (SSE-Net), which can adaptively fuse the feature maps from different sub-networks. The proposed method assumes that the observed LR MS image can be translated into its corresponding PAN image and vice versa because both of them are acquired on the same scene. Thus, a spatial extraction network is constructed, whose input and output are LR MS and PAN images, respectively. By reconstructing the PAN image from its corresponding LR MS image, the spatial information in the spatial extraction network is learned while suppressing the spectral information. Similarly, we send the PAN image into the spectral extraction network to generate the LR MS image. The spectral information in the spectral extraction network is enriched gradually and the spatial information is eliminated. Then, the spatial and spectral information extracted by the two sub-networks flows into the fusion network, where adaptive feature fusion modules (AFFMs) are designed to integrate the feature maps. According to the content of the feature maps, weight maps are generated adaptively in AFFMs to fuse the spatial and spectral features from the sub-networks. Through AFFMs, the redundancy among the feature maps from different sub-networks is reduced efficiently. Moreover, to learn subtle spatial structures, the spatial extraction network combines the reconstruction loss with the gradient constraint of the PAN image. For the spectral extraction network, an interdependency loss is derived to further depict the nonlinear relationship among the bands of the MS image, by which the spectral prior can be sufficiently captured. Finally, taking into account the loss between the fused image and the reference image, the proposed SSE-Net can be trained for spatial and spectral information preservation. The experimental results demonstrate that the proposed method achieves competitive results when compared with the state-of-the-art methods. Compared with methods as in [42], the method in this paper introduces image cross-domain reconstruction into the sub-networks to more efficiently extract features from the source images. Then, the features from sub-networks are adaptively integrated. To the best of our knowledge, this is the first time that image cross-domain reconstruction is applied to the fusion of PAN and LR MS images. Accordingly, the main contributions of this paper are:

1) We define the spatial and spectral extraction sub-networks, which can simultaneously achieve the feature extraction and the generation of the corresponding LR MS and PAN images, through the formulation of the image cross-domain reconstruction. Thus, the spatial and spectral information can be extracted explicitly and effectively.

2) To adaptively fuse the spatial and spectral information of LR MS and PAN images in the feature space, we design a novel fusion module, AFFM, to infer the combination weights of spatial and spectral features from the source image content. Compared to the "choose-max" and average rules, the adaptive weights can better reduce the redundancy among features and restrain the spatial and spectral distortions in the fused image.

3) To model the nonlinear relationships among the bands of the MS image, we design an interdependency loss to constrain the pixel value ratio between the paired bands of the MS image. The spectral information in the MS image can be captured more accurately due to the nonlinearity of the loss.

The remainder of the paper is organized as follows. Section II introduces the proposed SSE-Net in terms of spatial and spectral extraction networks, fusion network, and loss functions. Section III presents the experimental results obtained on several datasets derived from different satellites. Ablation studies and parameter analysis are also given. Finally, the conclusion is given in Section IV.

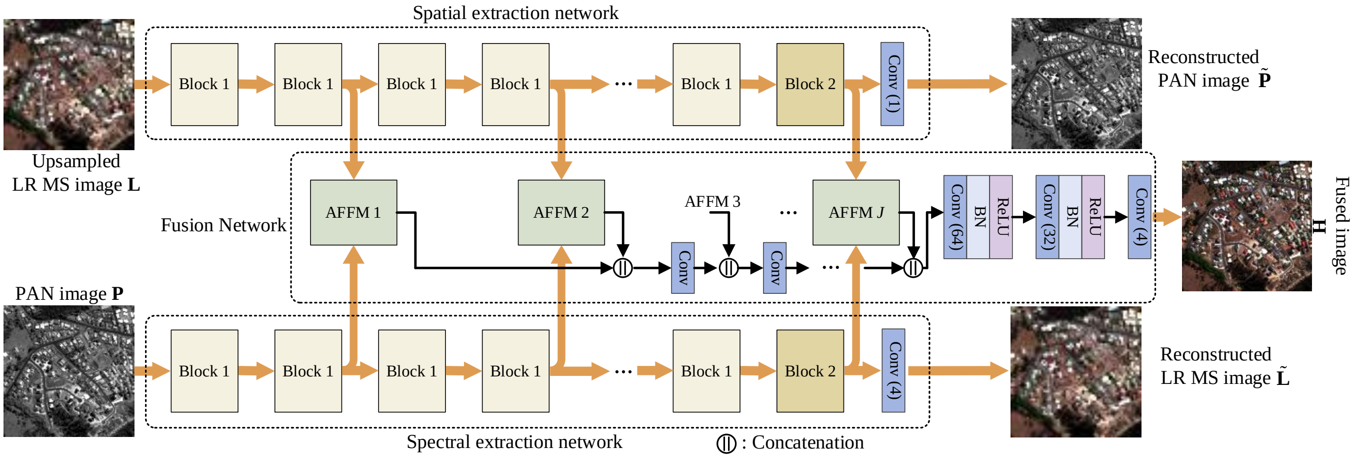


Fig. 1. Overall framework of the proposed SSE-Net for pansharpening. (i) denotes the number of filters in the convolution layer.

II. Proposed Method

In this section, we present the architecture of the proposed SSE-Net first. The spatial and spectral extraction networks and the fusion network are introduced in detail. In addition, loss functions are defined carefully to measure the reconstruction errors of SSE-Net efficiently.

A. Overall Framework

Fig. 1 presents the architecture of the proposed SSE-Net, which consists of three parts: the spatial extraction network, the spectral extraction network, and the fusion network. The first two networks are responsible for the learning of spatial details and spectral information, respectively. By synthesizing the reconstructed PAN image in a cross-domain manner, the spatial extraction network can retain spatial details gradually with the suppression of spectral information in the up-sampled LR MS image $\mathbf{L}\in\mathbb{R}^{M\times N\times B}$, where M, N and B are the spatial and spectral dimensions of the image, respectively. Similarly, the spectral extraction network is introduced to generate the reconstructed LR MS image $\tilde{\mathbf{L}}\in\mathbb{R}^{M\times N\times B}$ from the PAN image $\mathbf{P}\in\mathbb{R}^{M\times N}$, in which the spectral information can be described well by eliminating the spatial information. The auxiliary reconstruction tasks of LR MS and PAN images further facilitate the learning of spatial and spectral features of sub-networks. In Fig. 1, Blocks 1 and 2 in the two networks share the same structures but the numbers of filters are different, as described in Section II.B. Then, the fusion network is defined to integrate the spatial and spectral feature maps from the two networks. We utilize AFFMs to make full use of the complementarity and reduce the redundancy among these feature maps. The weight maps for each feature map can be adaptively inferred by the designed AFFM. Finally, the fused image $\mathbf{H} \in \mathbb{R}^{M\times N\times B}$ is obtained by the fusion network from the recombined feature maps.

B. Spatial and Spectral Extraction Networks

The architecture of the spatial extraction network is illustrated in Fig. 2. Its input and output are the up-sampled LR MS and PAN images, respectively. From Fig. 2, we can see that the network is composed of two kinds of cascaded blocks, which both include a convolution layer, batch normalization (BN), and rectified linear unit (ReLU). For Block 1, the convolution layer consists of 32 filters with the size of 3×3 (See in Fig.1). In the spatial extraction network, blocks that are identical to Block 1 are cascaded for the feature extraction. Then, they are followed by Block 2. Block 2 also contains a convolution layer, BN, and ReLU, but it adopts 64 filters to obtain more feature maps with a filter size of 3×3. The last layer only contains one filter to reconstruct the PAN image with a single band. According to the network in Fig. 2, the spatial information is obtained through the reconstruction of the PAN image.

As illustrated in Fig. 1, the spectral extraction network shares the same architecture as the spatial extraction network except for the last layer. The number of filters in the last layer is decided by the number of bands in the MS image because the spectral extraction network aims to approximate the LR MS channels whose number depends on the considered sensors (e.g., 4 or 8 bands). Then, we can model the spectral information in the MS image by the spectral extraction network. The spatial and spectral features are extracted more effectively by the two sub-networks owing to the LR MS and PAN reconstruction constraints upon them.

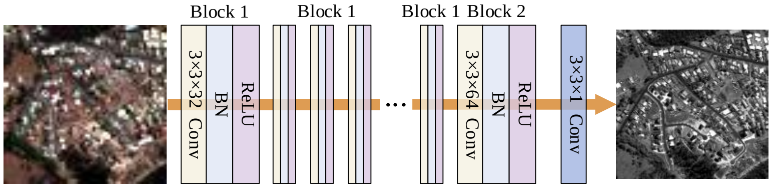


Fig. 2. Architecture of the spatial extraction network.

C. Fusion Network

In DNN-based pansharpening methods consisting of different sub-networks, feature maps are generally concatenated and fed into the following networks. Although all information in feature maps from different sub-networks is considered, the redundancy among them is not eliminated, which may lead to spatial or spectral distortions in the fusion result. Therefore, it is necessary to integrate these feature maps by reducing redundancy and promoting complementarity. Some literature methods [43] employ different fusion rules, such as “choose-max” and average rules, to integrate the spatial and spectral information in the feature domain. However, the combined weights of the spatial and spectral information in these cases are hand-crafted and neglect the content of the images to be fused.

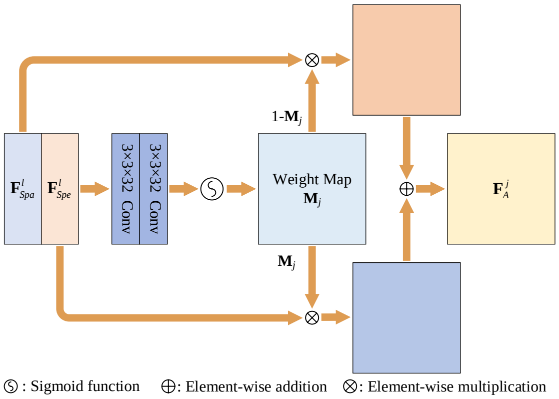


Fig. 3. Adaptive feature fusion module (AFFM).

Therefore, taking into account the redundant and complementary information, we design the AFFM to adaptively infer weight maps from the content of these feature maps. Fig. 1 shows that the fusion network is defined by a series of AFFMs. Fig. 3 shows the architecture of an AFFM, where $\mathbf{F}^j_A$ represents the output of the *j*th AFFM. $\mathbf{F}^l_{Spa}$ and $\mathbf{F}^l_{Spe}$ are the outputs of the *l*th layers in the spatial and spectral extraction networks, respectively. $\mathbf{M}_j$ is the weight map deduced from the concatenation of $\mathbf{F}^l_{Spa}$ and $\mathbf{F}^l_{Spe}$ after two convolution layers. Then, the context information of the source images is embedded into the fusion of spatial and spectral features by the weight map according to the following fusion operation:

$$\mathbf{F}^j_A = \mathbf{M}_j \otimes \mathbf{F}^l_{Spe} + \left(1-\mathbf{M}_j\right)\otimes \mathbf{F}^l_{Spa} \qquad (1)$$

Through $\mathbf{M}_j$, the integration of feature maps is achieved adaptively pixel-by-pixel. Then, the spatial and spectral distortions in the fusion result are reduced by the introduction of AFFMs. The number of channels in the weight map matches those of features from the two networks. The intermediate 1 to *J*-1 AFFMs have the same configuration as shown in Fig. 3. In the *J*th AFFM, each convolution layer in this module involves 64 filters.

When AFFMs achieve the integration of spatial and spectral feature maps, their outputs are employed to reconstruct the fusion result. The architecture used to integrate the outputs of AFFMs is shown in Fig. 4. We can see that the feature maps from the *j*th AFFM are combined with the output of the previous convolution layer. Then, the concatenated feature maps are fed into the following convolution layer. The same operation is implemented to generate the output of the *j*+1th AFFM. Each convolution layer in Fig. 4 contains 32 filters with the size of 3×3. Finally, concatenated features are obtained from the last AFFM and the last convolution layer, which are employed to generate the fused image.

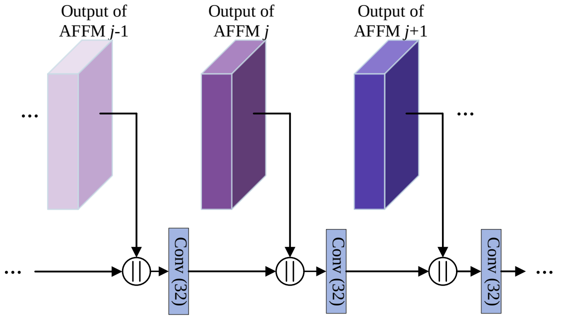


Fig. 4. Architecture of the proposed fusion network.

D. Loss Function

In the SSE-Net, the fusion and the reconstruction of LR MS and PAN images are simultaneously achieved. The fusion network is responsible for the generation of the fused image. The two sub-networks learn the spatial and spectral information by cross-domain reconstruction. So, different losses are introduced to efficiently capture the spatial and spectral information for fusion and balance the performance of each network. The following losses are exploited to train the proposed SSE-Net.

1) *Spatial reconstruction loss*

The spatial extraction network aims to synthesize the PAN image from the LR MS image. To enhance the spatial details in the PAN image, the gradients of the reconstructed PAN image $\tilde{\mathbf{P}}$ are also enforced to coincide with those of the original PAN image $\mathbf{P}$. So, combining the approximation error, the spatial reconstruction loss is summarized as:

$$L_{Spa} = \left\|\mathbf{P}-\tilde{\mathbf{P}}\right\|_1 + \omega_g \left\|\nabla\mathbf{P}-\nabla\tilde{\mathbf{P}}\right\|_1 \tag{2}$$

where ∇ denotes the gradient operator and ω_g is the weight parameter. Here, we use the mean absolute error (MAE) to measure the reconstruction loss. The PAN image can be reconstructed according to the minimization of (2),. Meanwhile, the spatial information is extracted by the spatial extraction network, which is then further used in the fusion network.

2) *Spectral reconstruction loss*

The spectral extraction network aims to synthesize the LR MS image from the PAN image $\mathbf{P}$. So, the output of the network should be as close to the original LR MS image $\mathbf{L}$ as possible. This is defined as:

$$\min\left\|\mathbf{L}-\tilde{\mathbf{L}}\right\|_1 \tag{3}$$

where $\tilde{\mathbf{L}}$ stands for the synthesized LR MS image by the spectral extraction network. Although this constraint is widely used, only a linear relationship is considered in (3), which cannot ensure sufficient reconstruction of the spectral information. Beyond the linear pixel-wise approximation, the nonlinear relationship among the bands of $\mathbf{L}$ should also be inherited into $\tilde{\mathbf{L}}$. Here, we make use of the ratio of pixel values between paired bands in the MS image to describe the relationship, which can be modeled as:

$$\min\sum_{\forall(p,q)\in\phi}\left\|\frac{\mathbf{L}_p}{\mathbf{L}_q}-\frac{\tilde{\mathbf{L}}_p}{\tilde{\mathbf{L}}_q}\right\|_1 \tag{4}$$

where ϕ represents the set of all possible combinations of spectral bands in the MS image. $\mathbf{L}_p$ and $\tilde{\mathbf{L}}_p$ represent the *p*th band in $\mathbf{L}$ and $\tilde{\mathbf{L}}$, respectively. $\div$ is the element-wise division. (4) is introduced into the spectral reconstruction loss so that spectral relationships can be extracted effectively in the spectral network. Accordingly, the spectral reconstruction loss is finally written as:

$$L_{Spe} = \left\|\mathbf{L}-\tilde{\mathbf{L}}\right\|_1 + \omega_r \sum_{\forall(p,q)\in\phi}\left\|\frac{\mathbf{L}_p}{\mathbf{L}_q}-\frac{\tilde{\mathbf{L}}_p}{\tilde{\mathbf{L}}_q}\right\|_1 \tag{5}$$

where ω_r is a regularization parameter to balance the two terms in (5).

3) *Fusion loss*

When the spatial and spectral features are extracted by the two sub-networks, the fusion network merges them to generate the fused image. The fusion loss is written as:

$$L_f = \left\|\mathbf{H}-\mathbf{R}\right\|_1 \tag{6}$$

where the fused image is denoted by $\mathbf{H}$ and $\mathbf{R}$ is the reference image. In this way, the desired fusion result is generated by the spatial and spectral information from different networks.

4) *Total loss*

Combining all losses, the total loss can be expressed as:

$$L = L_{Spa} + L_{Spe} + L_f \tag{7}$$

With the constraints of these losses, the sub-networks in the proposed SSE-Net are trained simultaneously. This allows us to achieve better performance in terms of spatial and spectral information preservation.

III. Experimental Results and Analysis

In this section, the performance of the proposed SSE-Net is assessed on two datasets derived from GeoEye-1 and QuickBird satellites. Some state-of-art methods are employed for comparison, including AWLP [44], Indusion [47], MTF-GLP [45], LRP [25], PNN [28], PanNet [34], MDSCC-GAN [39], and PSGAN [46]. In addition, we also present the fusion results obtained by the proposed SSE-Net with different architectures for a more comprehensive analysis.

A. Experimental Settings

1) *Datasets*: To fully compare the fusion performance of all methods, reduced-scale and full-scale experiments are conducted on two datasets derived from GeoEye-1 and QuickBird satellites. The dataset from the GeoEye-1 satellite was obtained from the urban area of Hobart, Australia on February 24, 2009. The resolutions of LR MS and PAN images in the dataset are 2.0m and 0.5m, respectively. The dataset from the QuickBird satellite includes LR MS and PAN images

acquired in Xi'an, China on September 30, 2008. Their spatial resolutions are 2.8m and 0.7m. For the experiments on reduced-scale datasets, the fusion results are directly compared with the original MS images, which are regarded as reference images. In reduced-scale datasets, original MS and PAN images are degraded in the spatial domain to produce the LR MS and PAN images to be fused. According to the spatial degradation model, the original images are smoothed by a Gaussian filter and then down-sampled by a factor of 4. Then, the original MS image is viewed as the reference image.

2) *Training and Test Details*: We generate the reduced-scale image pairs for the supervised training. Then, 83000 reduced-scale image pairs from the GeoEye-1 satellite are synthesized by blurring and downsampling according to Wald's protocol [5]. In the same way, 45000 image pairs from the QuickBird satellite are constructed for training. The original MS image is treated as the reference image. In the training data, the sizes of LR MS and PAN images are 16×16×4 and 64×64, respectively. The partition for the training and validation datasets is 95% and 5%. The training of SSE-Net and other DNN-based pansharpening methods is conducted by PyTorch on an NVIDIA 2080Ti GPU.

All methods are tested on 64×64×4 LR MS and 256×256 PAN images. So, the size of the fused image will be 256×256×4. For the test on the reduced-scale QuickBird dataset, 50 LR MS and PAN image pairs are used and the average indexes are listed for comparison. 50 pairs of LR MS and PAN images are prepared for the test of the full-scale QuickBird dataset. Similarly, we also utilize 50 reduced-scale and 50 full-scale image pairs to test the performance of the proposed method on the GeoEye-1 dataset.

For the proposed method, the batch size is 4 and all parameters in convolution layers are initialized by a Gaussian function with zero mean and standard deviation equal to 0.02. The learning rate is set as 0.001. The training is completed after 300 epochs. The regularization parameters ω_g and ω_r are all set as 0.1. Section III. I presents the influences of the regularization parameters on the fusion result in detail.

B. Evaluation Indexes

Some reference-based evaluation indexes are computed for comparison, including *Erreur Relative Globale Adimensionnelle de Synthèse* (ERGAS) [48], universal image quality index (UIQI) [49], spectral angle mapper (SAM) [50], root-mean-squared error (RMSE), and Q4 [51]. For RMSE, ERGAS, and SAM, the best values are 0. Smaller values mean better fusion results. UIQI and Q4 vary from 0 to 1, whose best values are 1. In the full-scale case, the fusion results are evaluated by no-reference indexes, such as D_λ, D_S, and QNR [52]. The spectral and spatial quality is measured by D_λ and D_S, respectively. For them, the values closer to 0 indicate better fusion results. QNR reflects the overall performance of the fused image. The eight indexes are introduced in detail below.

1) *Q4*: The Q4 is designed by modeling image distortions as a combination of three factors. It is defined as:

$$\text{Q4}=\frac{4\left|\sigma_{z_1z_2}\right|\times\left|\mu_{z_1}\right|\times\left|\mu_{z_2}\right|}{(\sigma_{z_1}^2+\sigma_{z_2}^2)(\mu_{z_1}^2+\mu_{z_2}^2)} \tag{8}$$

where z_1 and z_2 are two quaternions composed of the spectral vectors of the MS images, i.e. $z=a+\mathbf{i}b+\mathbf{j}c+\mathbf{k}d$. $\sigma_{z_1z_2}$ denotes the covariance between z_1 and z_2, μ_{z_1} and μ_{z_2} are the means of z_1 and z_2, respectively. $\sigma_{z_1}^2$ and $\sigma_{z_2}^2$ are the variances of z_1 and z_2. The highest value of Q4 is 1, and the lowest value is 0.

2) *SAM*: The SAM is used to calculate the similarity between two spectral vectors $\mathbf{v}$ and $\hat{\mathbf{v}}$. It is defined as follows:

$$\text{SAM=arccos}\left(\frac{\langle\mathbf{v},\hat{\mathbf{v}}\rangle}{\|\mathbf{v}\|_2\cdot\|\hat{\mathbf{v}}\|_2}\right) \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product and $\|\cdot\|_2$ denotes the vector L_2-norm. The optimal value of the SAM is 0.

3) *UIQI*: The UIQI is an index for estimating the global spectral quality of the fused image. UIQI consists of 3 factors: correlation loss, brightness distortion, and contrast distortion. It is formulated as:

$$\text{UIQI=}\frac{\sigma_{\mathbf{HR}}}{\sigma_{\mathbf{H}}\cdot\sigma_{\mathbf{R}}}\frac{2\mu_{\mathbf{H}}\mu_{\mathbf{R}}}{\mu_{\mathbf{H}}^2+\mu_{\mathbf{R}}^2}\frac{2\sigma_{\mathbf{H}}\sigma_{\mathbf{R}}}{\sigma_{\mathbf{H}}^2+\sigma_{\mathbf{R}}^2} \tag{10}$$

where $\sigma_{\mathbf{HR}}$ is the sample covariance of $\mathbf{H}$ and $\mathbf{R}$. $\mu_{\mathbf{H}}$ and $\mu_{\mathbf{R}}$ are the sample means of $\mathbf{H}$ and $\mathbf{R}$, respectively. The UIQI varies in the range [0,1], and its optimal value is 1.

4) *ERGAS*: Another commonly used global quality index is the ERGAS, defined as:

$$\text{ERGAS}=100\frac{p}{l}\sqrt{\frac{1}{B}\sum_{i=1}^{B}\left(\frac{\text{RMSE}(i)}{\text{M}(i)}\right)^2} \tag{11}$$

$$\text{RMSE}(i)=\sqrt{E[(\mathbf{H}_i-\mathbf{R}_i)^2]} \tag{12}$$

where $E(\cdot)$ is the average operation, RMSE(i) is the root mean square error between the ith band $\mathbf{H}_i$ and $\mathbf{R}_i$ in the fused image and the reference image. p and l are the spatial resolutions of PAN and MS images. $\text{M}(i)$ is the average value of the ith band of the MS image. Its optimal value is 0.

5) D_λ: D_λ is a spectral distortion index, derived from the spectral correlation between the fused MS image and the LR MS image. It is written as:

$$D_\lambda=\sqrt{\frac{1}{B(B-1)}\sum_{i=1}^{B}\sum_{j=i}^{B}\left|\text{UIQI}(\mathbf{H}_i,\mathbf{H}_j)-\text{UIQI}(\mathbf{L}_i,\mathbf{L}_j)\right|} \tag{13}$$

where $\mathbf{L}_i$ represents the ith band of the LR MS image. UIQI is exploited to measure the dissimilarities between couples of bands.

6) D_S: D_S is a spatial distortion index which is calculated as:

$$D_S=\sqrt{\frac{1}{B}\sum_{i=1}^{B}\left|\text{UIQI}(\mathbf{H}_i,\mathbf{P})-\text{UIQI}(\mathbf{L}_i,\bar{\mathbf{P}})\right|} \tag{14}$$

where $\bar{\mathbf{P}}$ is the degraded version of $\mathbf{P}$. The value of D_S is within [0,1], the lower the better.

7) *QNR*: The QNR is a jointly spectral and spatial quality index. It is the product of the spatial and spectral distortion indexes, which reflects the overall quality of the fused image. It is defined as:

$$\mathrm{QNR} = (1 - D_\lambda)^\alpha (1 - D_S)^\beta \qquad (15)$$

QNR is weighted by α and β. The highest value is 1, which is obtained when the spectral and spatial distortions are both 0.

C. Experiments on Reduced-Scale Datasets

In this section, we present experiments on reduced-scale datasets, where the reference image is utilized for the evaluation of the fused image. Fig. 5 shows the fusion results of all methods on the QuickBird dataset. In addition, some regions are chosen and magnified for further visual analysis. The selected regions are placed in the bottom right corner of the fused images. Absolute error maps between the reference image and the fused images are also displayed in the second and fourth rows of Fig. 5 to compare the reconstruction performance of different methods. Fig. 5(f) shows that the LRP fusion result suffers from significant spectral distortions. Compared with the reference image in Fig. 5(c), the spatial details are enhanced excessively in Indusion and MTF-GLP results (Figs. 5(e) and 5(g)), especially in the vegetation regions. The PanNet result in Fig. 5(i) shows some spatial differences compared with Fig. 5(c). Besides, The MDSCC-GAN fusion result in Fig. 5(j) is corrupted by some spectral distortions. By analyzing the zoomed region, we can find that some spatial artifacts are introduced into the result of PSGAN. The result of the proposed method in Fig. 5(l) shows a better quality and is more consistent with the reference image. From the absolute error maps of different methods, we also can see that the reconstruction errors of Indusion are more obvious than those of other methods. For DNN-based methods, PNN and MDSCC-GAN produce larger reconstruction errors in the regions containing trees and buildings. In general, the proposed SSE-Net has a better reconstruction performance when compared with other methods.

Table I lists the average values of different indexes computed on the fusion results of 50 LR MS and PAN image pairs. The best values in Table I are labeled in bold. One can see that the best Q4, SAM, and UIQI values are obtained by the proposed method. For ERGAS, SSE-Net also provides the best value, followed by PNN. So, the proposed method has a better overall performance.

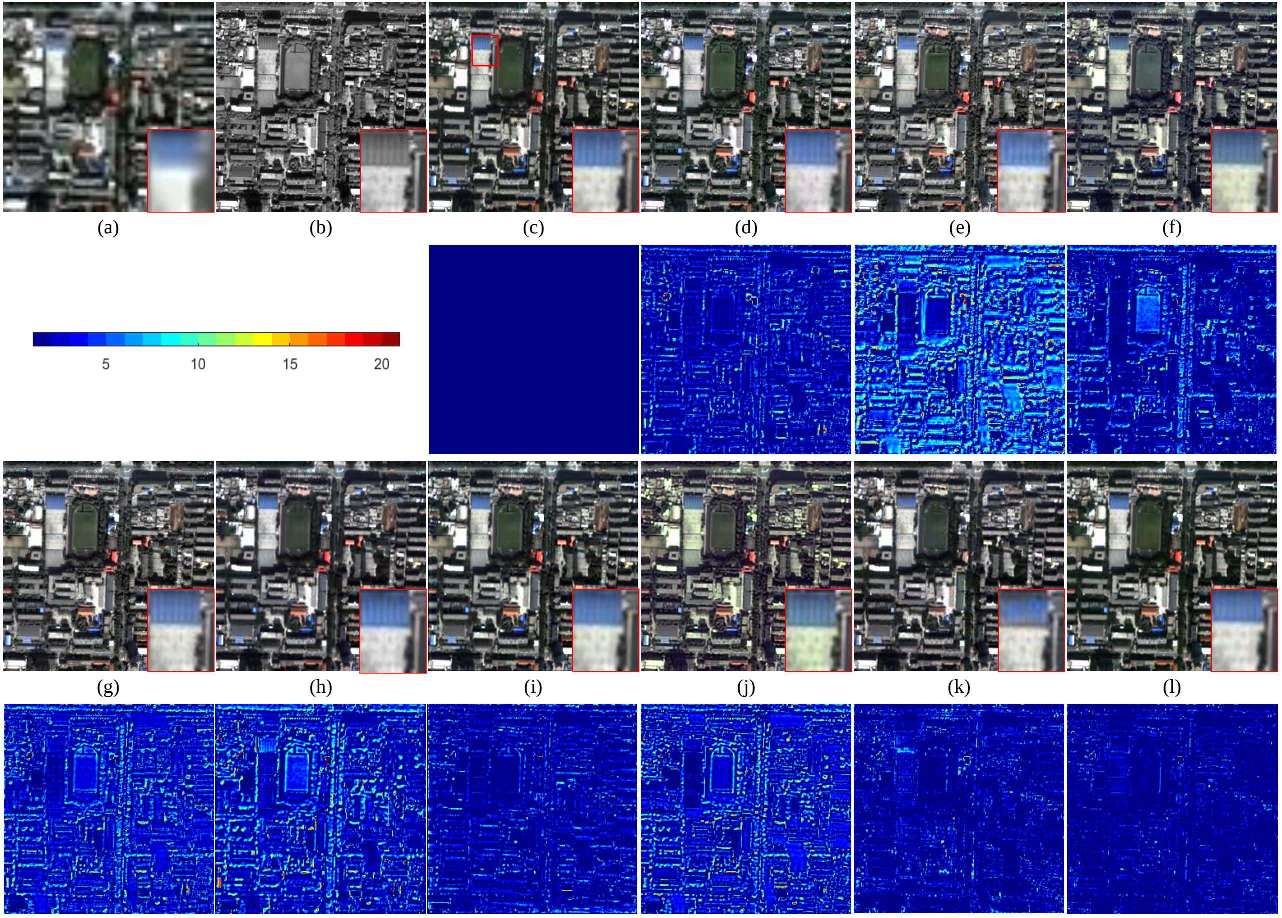


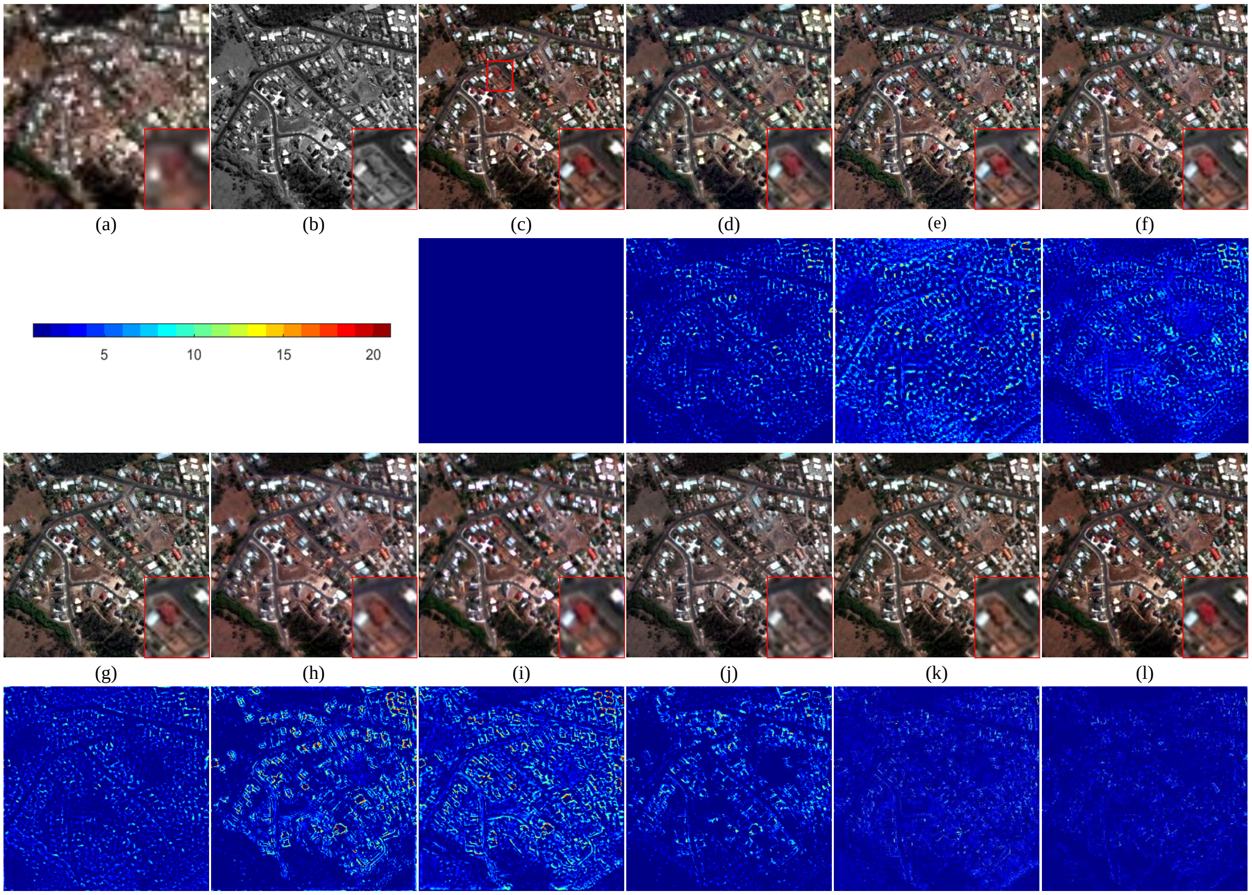
Fig. 5. Qualitative comparison of the fused images and the absolute error maps from different methods on the QuickBird dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) AWLP; (e) Indusion; (f) LRP; (g) MTF-GLP; (h) PNN; (i) PanNet; (j) MDSCC-GAN; (k) PSGAN; (l) Proposed SSE-Net.

Fig. 6 displays the fusion results of all methods on the GeoEye-1 dataset. In the figure, a region of interest in the fused image is selected and enlarged for a better visual comparison. The absolute error maps of all methods are also shown in Fig. 6. We can see that the result of Indusion in Fig. 6(e) behaves well in terms of spatial information enhancement, but introduces slight spectral distortions. The results of LRP and MTF-GLP in Figs. 6(f) and 6(g) have a similar performance in terms of spectral information. However, severe spectral distortions exist in the fused image of MDSCC-GAN. Compared with the reference image in Fig. 6(c), the results of PanNet and the proposed method show consistent spectral features. However, some blurring effects arise in the result of PanNet in Fig. 6(i). From the absolute error maps in Fig. 6, we can find that larger reconstruction errors are visible in the contours of buildings. For instance, obvious differences can be found in the error map of LRP. A similar performance also appears in the error maps of PNN and PanNet. The reconstruction errors of the proposed SSE-Net are closer to 0 than other methods.

Table II presents the average index values obtained by all methods on 50 LR MS and PAN reduced-scale image pairs from the GeoEye-1 satellite. The best values in Table II are labeled in bold. One can see that the proposed method has the best performance in terms of all indexes.

TABLE I
QUANTITATIVE EVALUATIONS ON 50 LR MS AND PAN IMAGE PAIRS FROM THE REDUCED-SCALE QUCKBIRD DATASET.

Evaluation index	**AWLP**	**Indusion**	**LRP**	**MTF-GLP**	**PNN**	**PanNet**	**MDSCC -GAN**	**PSGAN**	**Proposed SSE-Net**
Q4	0.9097	0.8639	0.9020	0.9103	0.9168	0.9133	0.8479	0.9237	**0.9449**
RMSE	14.7356	18.5029	21.3623	14.6803	12.8623	16.6743	22.1885	12.4505	**10.8455**
SAM	2.3071	2.7018	3.2570	2.2883	2.0292	2.2463	3.3305	1.9755	**1.5698**
UIQI	0.9224	0.8826	0.8416	0.9250	0.9474	0.8969	0.8799	0.9530	**0.9619**
ERGAS	0.8693	1.0981	1.1543	0.8630	0.7592	0.9192	1.2788	0.7232	**0.6282**



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

(g) (h) (i) (j) (k) (l)

Fig. 6. Qualitative comparison of the fused images and the absolute error maps from different methods on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) AWLP; (e) Indusion; (f) LRP; (g) MTF-GLP; (h) PNN; (i) PanNet; (j) MDSCC-GAN; (k) PSGAN; (l) Proposed SSE-Net.

TABLE II
QUANTITATIVE EVALUATIONS ON 50 LR MS AND PAN IMAGE PAIRS FROM THE REDUCED-SCALE GEOEYE-1 DATASET.

Evaluation index	**AWLP**	**Indusion**	**LRP**	**MTP-GLP**	**PNN**	**PanNet**	**MDSCC -GAN**	**PSGAN**	**Proposed SSE-Net**
Q4	0.7925	0.7571	0.7379	0.8009	0.7909	0.7648	0.7969	0.8059	**0.8148**
RMSE	26.0745	30.2596	50.4391	25.1462	27.6703	29.4105	26.3221	26.5761	**20.5044**
SAM	5.2666	5.4003	5.9174	5.1168	5.0444	4.9059	5.1800	4.4623	**3.6426**
UIQI	0.9432	0.9197	0.8948	0.9467	0.9306	0.9224	0.9354	0.9551	**0.9708**
ERGAS	1.6496	1.9299	3.2942	1.6049	1.7594	1.8598	1.7084	1.4468	**1.3709**

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3187025

D. *Experiments on Full-Scale Datasets*

The experiments are also conducted on full-scale datasets. Fig. 7 shows the fused images of all compared methods and a region of interest is magnified for more intuitive analysis. The zoomed areas are put on the bottom right corner of each fused image. From Fig. 7, we can see that the spectral information of the vegetation areas is distorted for most of the fused images. For instance, the color of the tree areas of the LRP fusion result in Fig. 7(e) becomes gray. The MTF-GLP fusion result in Fig. 7(f) has a similar performance to that in the result of PanNet in Fig. 7(h). Some spectral effects also can be found in the fused image of Fig. 7(j) from PSGAN, in which the color information is over-enhanced. We can find that the color of the result of the proposed method in Fig. 7(k) looks more natural. As for spatial information, all fused images contain clear textures or edges when compared with the LR MS image in Fig. 7(a). The results of PanNet and the proposed method are similar in terms of spatial details.

Table III reports the quantitive indexes computed on the fusion results from 50 LR MS and PAN image pairs. We label the best values in bold. The proposed SSE-Net obtains the best D_S and QNR values. The PSGAN achieves the best D_λ .

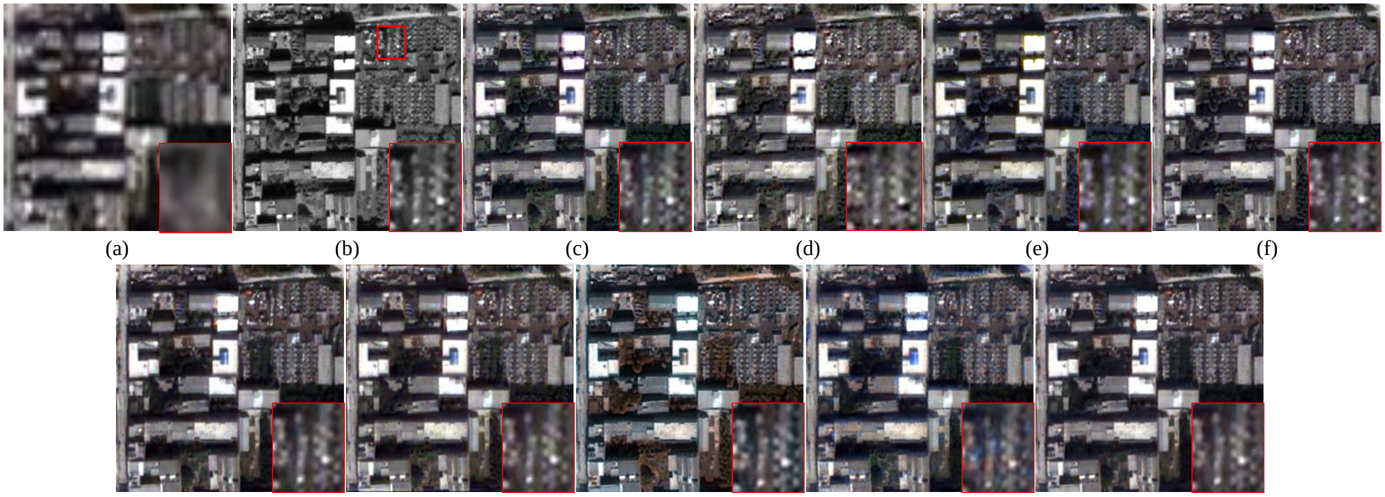

(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k)

Fig. 7: Qualitative comparison of the fused images from different methods on the QuickBird dataset. (a) LR MS image; (b) PAN image; (c) AWLP; (d) Indusion; (e) LRP; (f) MTF-GLP; (g) PNN; (h) PanNet; (i) MDSCC-GAN; (j) PSGAN; (k) Proposed SSE-Net.

TABLE III
QUANTITATIVE EVALUATIONS ON 50 LR MS AND PAN IMAGE PAIRS FROM THE FULL-SCALE QUICKBIRD DATASET.

Evaluation index	AWLP	Indusion	LRP	MTP-GLP	PNN	PanNet	MDSCC -GAN	PSGAN	Proposed SSE-Net
D_λ	0.0789	0.0493	0.2127	0.0656	0.0637	0.0566	0.1410	**0.0474**	0.0482
D_S	0.0609	0.0387	0.2101	0.0658	0.0396	0.0414	0.0803	0.0381	**0.0368**
QNR	0.8663	0.9139	0.6260	0.8748	0.9000	0.9053	0.7932	0.9163	**0.9173**

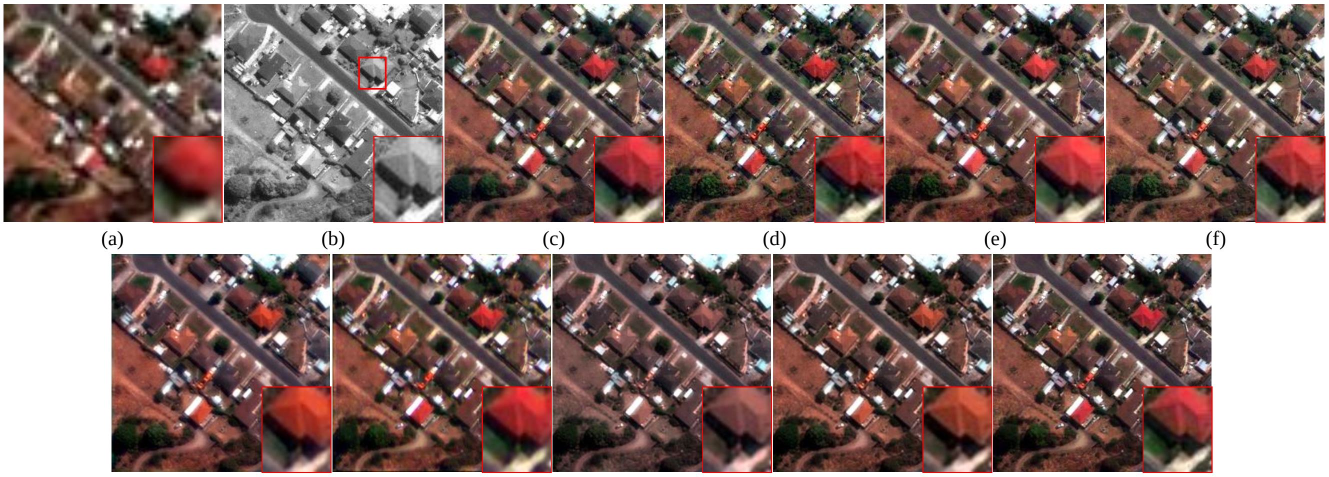

(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k)

Fig. 8. Qualitative comparison of the fused images from different methods on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) AWLP; (d) Indusion; (e) LRP; (f) MTF-GLP; (g) PNN; (h) PanNet; (i) MDSCC-GAN; (j) PSGAN; (k) Proposed SSE-Net.

The results on the GeoEye-1 dataset are illustrated in Fig. 8, where LR MS and PAN images are shown in Figs. 8(a) and 8(b). From Fig. 8, one can see that the results in Figs. 8(e) and 8(i) from LRP and MDSCC-GAN are affected by obvious spectral distortions. The other fusion results have a similar spectral appearance. The fused images of LRP and MTF-GLP in Fig.

8(e) and (f) present clear spatial textures. When compared with other fusion results, some blurring effects can be seen in the result of AWLP in Fig. 8(c). For the result of the proposed method, the spatial and spectral information is enhanced well.

Table IV lists the average indexes of all methods on 50 LR MS and PAN image pairs from the GeoEye-1 dataset. The proposed method has better performance than other methods in terms of D_λ. The D_S of the proposed SSE-Net is close to that of Indusion. Moreover, SSE-Net provides the best QNR, which assesses the overall performance of the fusion result.

TABLE IV
QUANTITATIVE EVALUATIONS ON 50 LR MS AND PAN IMAGE PAIRS FROM THE FULL-SCALE GEOEYE-1 DATASET.

Evaluation index	AWLP	Indusion	LRP	MTP-GLP	PNN	PanNet	MDSCC -GAN	PSGAN	Proposed SSE-Net
D_λ	0.1174	0.0868	0.0719	0.1238	0.0654	0.0602	0.0934	0.0784	**0.0590**
D_S	0.0576	**0.0404**	0.0992	0.0720	0.0567	0.1095	0.0613	0.0457	0.0460
QNR	0.8323	0.8764	0.8362	0.8140	0.8816	0.8369	0.8517	0.8802	**0.8977**

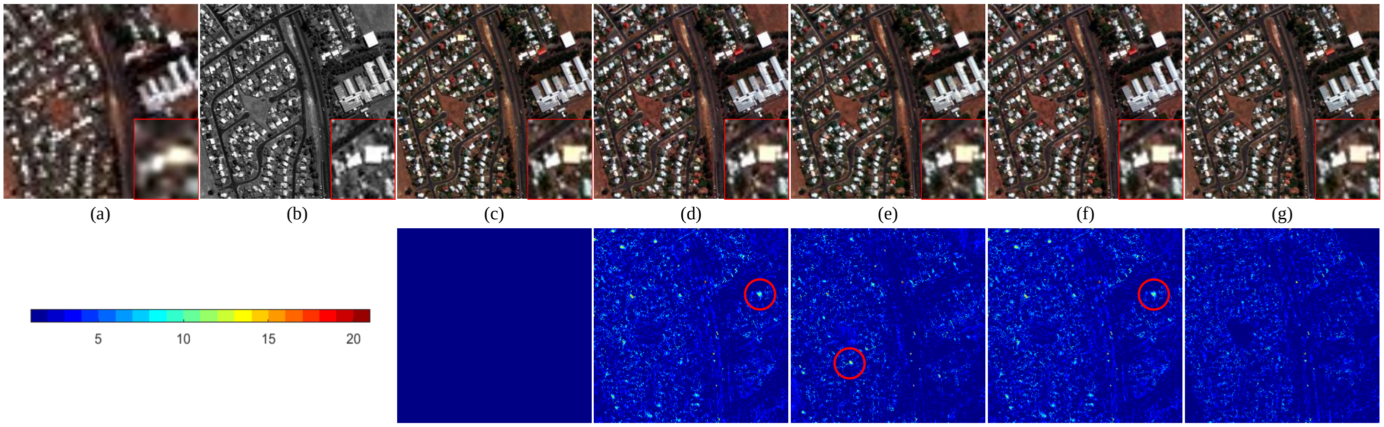


Fig. 9. Qualitative comparison of the fused images and the absolute error maps with different feature combinations on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) Concatenation; (e) Average; (f) Choose-Max; (g) Proposed SSE-Net.

E. Analysis of Feature Fusion Strategies

TABLE V
QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 9 (GEOEYE-1 DATASET).

Evaluation index	Concatenation	Average	Choose-Max	AFFM
Q4	0.8192	0.8179	0.8241	**0.8257**
SAM	**3.5994**	4.8750	3.6242	3.6546
UIQI	0.9767	0.9709	0.9767	**0.9770**
ERGAS	1.2590	1.5753	1.2203	**1.2068**

In this section, the effectiveness of the designed AFFM is analyzed. For a direct comparison, we replace AFFMs in the proposed framework with concatenation, average, and “choose-max”. Then, these networks are trained and tested on the GeoEye-1 dataset. Fig. 9 demonstrates the fusion results of LR MS and PAN images and their corresponding absolute error maps. Some areas with larger reconstruction errors are labeled by red circles. We can see that different combinations of feature maps have different influences on the fusion results. Some spectral distortions can be seen in the results obtained by concatenation and average (Figs. 9(d) and 9(e)). The fused image generated by the proposed technique in Fig. 9(g) is more similar to the reference image in Fig. 9(c). From the absolute error maps, one can see that the errors of AFFM are smaller than those of other strategies, especially in the circled areas. Thus, the absolute error maps demonstrate that the reconstruction performance is better when AFFM is introduced into the proposed SSE-Net. From the numerical results in Table V, we can also observe that the best Q4, UIQI, and ERGAS values are obtained when the proposed framework is equipped with AFFMs. Although the best SAM value is obtained by the concatenation strategy, the proposed SSE-Net provides a better overall performance, which confirms the effectiveness of AFFMs.

F. Ablation Study

In this section, we investigate the influences of spectral and spatial reconstruction loss functions on the fused images. Fig. 10 shows the fusion results when we remove the spectral or spatial reconstruction loss in (7). Moreover, the proposed network is also trained by only minimizing the fusion loss in (6). In this way, we can also investigate the effectiveness of the reconstruction tasks. Although the visual performance of Figs. 10(d)-10(g) is close, the error maps in the third and fifth rows of Fig. 10 demonstrate that the reconstruction precision is better when spectral and spatial reconstruction loss functions are both introduced. In addition, the evaluation results of the fused images are provided in Table VI. The best indexes are produced by the complete SSE-Net, which shows the effectiveness of the spectral and spatial reconstruction loss in (7).

Moreover, when the complete SSE-Net is tested, the reconstructed LR MS and PAN images are generated by the spectral extraction network and the spatial extraction network, respectively. We display them in Figs. 10(h) and 10(i). From Fig. 10, we can observe that there are large differences when the reconstructed LR MS and PAN images are compared to the images in Figs. 10(a) and 10(b). The same performance is also

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2022.3187025

found in their corresponding error maps. Although the sub-networks cannot reconstruct the LR MS and PAN images accurately, the introduction of the reconstruction tasks significantly contributes to the quality improvement of the fused image as shown in Table VI

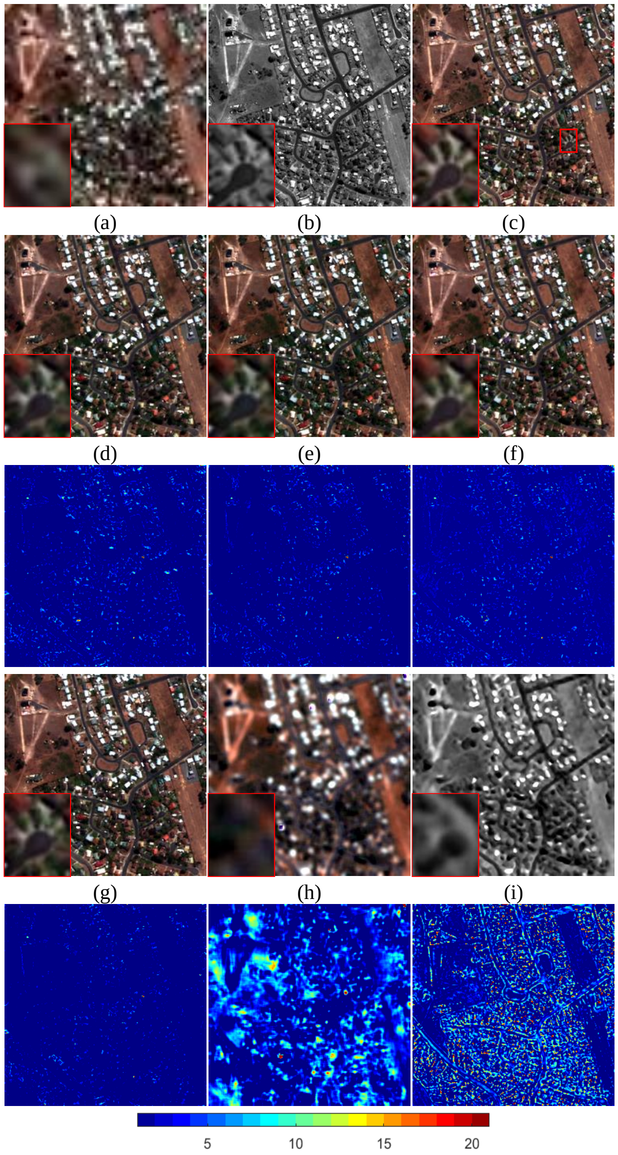


Fig. 10. Qualitative comparison of the fused images and the absolute error maps on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) w/o spectral reconstruction loss; (e) w/o spatial reconstruction loss; (f) Only fusion loss; (g) Complete SSE-Net; (h) Reconstructed LR MS image; (i) Reconstructed PAN image.

TABLE VI
QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 10 (GEOEYE-1 DATASET).

Evaluation index	w/o spectral reconstruction loss	w/o spatial reconstruction loss	only fusion loss	Complete SSE-Net
Q4	0.7962	0.7897	**0.7995**	0.7968
RMSE	28.8167	26.4876	24.9777	**24.0809**
SAM	4.2633	4.2866	4.1134	**3.9003**
UIQI	0.9494	0.9574	0.9633	**0.9654**
ERGAS	1.7230	1.5749	1.4940	**1.4253**

G. *Analysis of the Network Architecture*

AFFM can efficiently integrate the spatial and spectral feature maps from different networks. When more AFFMs are introduced into the framework, the depth of the spatial and spectral extraction networks increases. Then, the feature maps from different layers have different effects on the fused image. Here, we analyze the influences of the number of AFFMs on the fusion result.

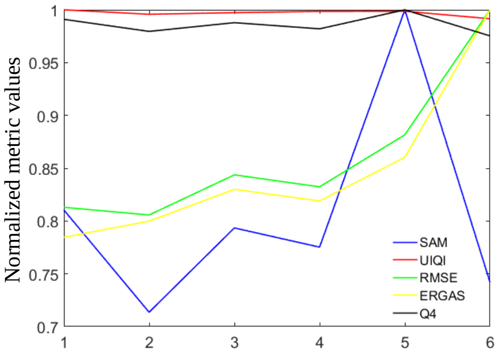


Fig. 11. Normalized index values provided by the proposed SSE-Net versus the number of AFFMs.

The experiments are conducted on the images in Figs. 5(a) and 5(b). The normalized index values are plotted in Fig. 11. From the figure, one can observe that RMSE and ERGAS become larger by increasing the number of AFFMs. SAM varies dramatically and the best SAM is obtained when two AFFMs are considered in the framework. On the contrary, there are only slight variations of UIQI and Q4. Considering the overall performance and representation capability of the network, we utilize four AFFMs to fuse the feature maps.

H. *Analysis of AFFM*

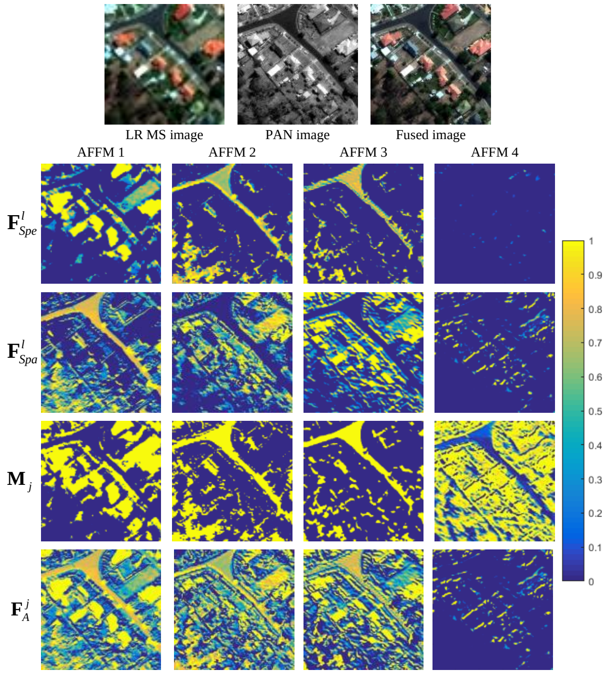
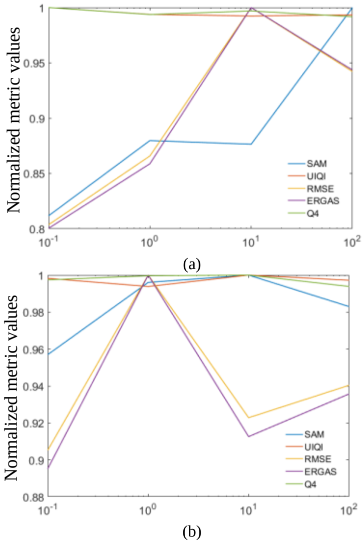


Fig. 12. Visualization of feature maps in AFFMs.

For a more comprehensive analysis, Fig.12 shows some feature maps from AFFMs. LR MS, PAN, and fused images are given in the first row in Fig.12, where source images are from the full-scale dataset of the GeoEye-1 satellite. According to the analysis in Section IV.F, we adopt four AFFMs in the SSE-Net. Some feature maps derived from the four AFFMs are selected and shown column by column. Different rows in Fig. 12 illustrate the inputs and outputs of AFFMs, which correspond to the variables in (1). By analyzing the figure, we can see that the feature maps from the spectral extraction network contain some low-frequency information. These feature maps pay more attention to the smooth areas of the source images. On the contrary, feature maps of the spatial extraction network contain more spatial details. The weight maps are inferred adaptively from the feature maps of spatial and spectral extraction networks. Thus, these feature maps can be efficiently fused. By analyzing the feature maps in the last row of Fig. 12, one can observe that they contain more information than those in $\mathbf{F}^{l}_{Spe}$ or $\mathbf{F}^{l}_{Spa}$.

I. Analysis of Parameters

During the training of SSE-Net, two parameters, ω_g and ω_r , have important effects on the fusion results. They are analyzed in this section. ω_g is responsible for the importance of the gradient information. The nonlinear spectral relationships are controlled by ω_r . Fig. 13 shows the behaviors of all indexes with different settings. In Fig. 13(a), ω_g increases from 0.1 to 100. One can see that SAM increases when ω_g becomes larger. RMSE and ERGAS first increase and then decrease. Q4 and UIQI are only slightly influenced by ω_g . The best values for these indexes are achieved when ω_g is equal to 0.1. Fig. 13(b) shows the results of all indexes when ω_r increases from 0.1 to 100. It can be seen that all values vary with the variations of ω_r . Through an analysis similar to that of ω_g , we set ω_r as 0.1 in the proposed SSE-Net.



(a)

(b)

Fig. 13. Normalized index values of SSE-Net versus different parameters. (a) ω_g ; (b) ω_r .

J. Running Time

In this section, we compare the computational time of all methods in terms of training and test (See Table VII). The traditional methods, including AWLP, Indusion, LRP, and MTP-GLP, are executed by MATLAB R2017a on the computer with Intel® Core™ i7-6700 processor, 3.4GHz, 16G memory. DNN-based methods are trained and tested on a server with an Intel® Core™ i7-9700 processor, 3.0GHz, an NVIDIA 2080Ti GPU, and 11G memory. From Table VII, we can see that the MDSCC-GAN requires more time training the network. The PNN takes about 14 hours for training due to fewer parameters. For the rest of the DNN-based methods, PSGAN has the best performance. Due to the introduction of AFFMs, the proposed method takes more time than PSGAN. Moreover, it has a competitive performance compared to MDSCC-GAN in terms of test time.

TABLE VII
TIME COMPARISON OF ALL METHODS.

Evaluation Index	AWLP	Indusion	LRP	MTP-GLP	PNN	PanNet	MDSCC -GAN	PSGAN	Proposed SSE-Net
Training time (h)	—	—	—	—	14.04	21.09	63.39	31.56	36
Test time (s)	0.42	0.32	12.51	0.54	0.02	9.68	0.13	0.008	0.18

IV. CONCLUSION

In this paper, a novel pansharpening method based on DNN is proposed, which is called SSE-Net. The proposed method extracts the spatial and spectral features from LR MS and PAN images by two sub-networks, which share the same architectures. To accurately fuse the features from different sub-networks, AFFMs are utilized to facilitate the integration of spatial and spectral information. In AFFMs, the weight maps are adaptively inferred from the content of the feature maps from sub-networks. With the introduction of AFFMs, the fusion network can achieve the reconstruction of the fused images efficiently. Moreover, the fusion network integrates the outputs of AFFMs to further reduce redundancy and promote complementarity. To reduce the blurring effects in the fused results, we use MAE loss to train the network. In addition, the spectral ratio loss is imposed on the proposed SSE-Net to characterize the interdependency of different bands in the MS image. Experimental results on different satellite datasets show the effectiveness of the proposed method. Compared with other literature methods, the proposed method not only enhances spatial details in the fused image but also better preserves spectral information. For future work, we plan to further investigate the influence of the introduction of auxiliary reconstruction tasks on the fusion result. Moreover, more efficient networks and loss functions will be explored to produce more accurate reconstructed results and fused images simultaneously.

REFERENCES

[1] S. Saha, L. Mou, X. Zhu, F. Bovolo, L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 607-611, 2021.

[2] G. Weikmann, C. Paris, L.Bruzzone, "TimeSen2Crop: a million labeled samples dataset of sentinel 2 image time series for crop type classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4699-4708, 2021.

[3] K. Zhang, M. Wang, S. Yang, "Multispectral and hyperspectral image fusion based on group spectral embedding and low-rank factorization," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 3, pp. 1363-1371, Mar. 2017.

[4] G. Vivone, L. Alparone, J. Chanussot, *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565-2586, May 2015.

[5] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301-1312, May 2008.

[6] T. Tu, P. Huang, C. Hung, C. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 53, no. 5, pp. 2565-2586, May 2004.

[7] H. R. Shahdoosti and H. Ghassemian, "Combining the spectral PCA and spatial PCA fusion methods by an optimal filter," *Inf. Fusion*, vol. 27, pp. 150-160, Jan. 2016.

[8] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6011875, Jan 2000.

[9] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, "An adaptive IHS pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746-750, Apr. 2010.

[10] M. Ghahremani and H. Ghassemian, "Nonlinear IHS: A promising method for pan-sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 1606-1610, Nov. 2016.

[11] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228-236, Jan. 2008.

[12] A. Garzelli, "Pansharpening of multispectral images based on nonlocal parameter optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096-2107, Apr. 2015.

[13] M. Imani, "Band dependent spatial details injection based on collaborative representation for pansharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4994-5004, Dec. 2018.

[14] K. Upla, M. Joshi, P. Gajjar "An edge preserving multiresolution fusion: Use of contourlet transform and MRF prior," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3210-3220, Jun. 2015.

[15] S. Devulapalli and R. Krishnan, "Synthesized pansharpening using curvelet transform and adaptive neuro-fuzzy inference system," *J. Appli. Remote Sensing*, vol. 13, no. 3, Sep. 2019.

[16] A. Kallel, L. Condat, J. M. Bioucas-Dias, J. Chanussot, and J. Xia, "Pansharpening: MTF-Adjusted pansharpening approach based on coupled multiresolution decompositions," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3124-3145, Jun. 2015.

[17] H. Li, F. Liu, S. Yang, K. Zhang, X. Su, L. Jiao, "Refined pan-sharpening with NSCT and hierarchical sparse autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5715-5725, Dec. 2016.

[18] S. Zheng, W. Shi, J. Liu, and J. Tian, "Remote sensing image fusion using multiscale mapped LS-SVM," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1313-1322, May 2008.

[19] Y. Xing, M. Wang, S. Yang, K. Zhang, "Pansharpening with multiscale geometric support tensor machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2503-2517, May 2018.

[20] C. Chen, Y. Li, W. Liu, J. Huang, "SIRF: Simultaneous satellite image registration and fusion in a unified framework," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4213-4224, Nov. 2015.

[21] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738-746, Feb. 2011.

[22] M. Ghahremania, Y. Liu, P. Yuen, A. Behera, "Remote sensing image fusion via compressive sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 34-48, Jun. 2019.

[23] R. Gogineni and A. Chaturvedi, "Sparsity inspired pan-sharpening technique using multi-scale learned dictionary," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 360-372, Dec. 2018.

[24] K. Zhang, M. Wang, S. Yang, Y. Xing, and R. Qu, "Fusion of panchromatic and multispectral images via coupled sparse non-negative matrix factorization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 12, pp. 5740-5747, Dec. 2016.

[25] S. Yang, K. Zhang and M. Wang, "Learning low-rank decomposition for pan-sharpening with spatial-spectral offsets," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3647-3657, Aug. 2018.

[26] F. Palsson, J. Sveinsson and M. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318-322, Jan. 2014.

[27] S. Hao, W. Wang, Y. Ye, L. Bruzzone, "A deep network architecture for super-resolution aided hyperspectral image classification with class-wise loss," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4650-4663, Aug. 2018.

[28] G. Masi, D. Cozzolino, L. Verdoliva and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, pp. 594, Jul. 2016.

[29] G. Scarpa, S. Vitale, D. Cozzolino, "Target-adaptive CNN based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 5443-5457, Aug. 2018.

[30] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by CNN denoiser," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 3, pp. 1124-1135, Mar. 2021.

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

[32] 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," *Inf. Fus.*, vol. 62, pp. 110-120, Oct. 2020.

[33] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, *et al.*, "ZeRGAN: zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Geosci. Remote Sens.*, early access, 2022, doi: 10.1109/TNNLS.2021.3137373.

[34] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE ICCV*, Oct. 2017, pp. 5449-5457.

[35] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2090-2104, May 2021.

[36] M. Jiang, H. Shen, J. Li, Q. Yuan, L. Zhang, "A differential information residual convolutional neural network for pansharpening," *ISPRS J. Photogramm. Remote Sens.*, vol. 163, pp. 257-271, May 2020.

[37] X. Liu, Q. Liu, Y. Wang, "Remote sensing image fusion based on two-stream fusion network,". *Inf. Fus.*, vol. 55, pp. 1-15, Mar. 2020.

[38] S. Wan, Y. Xia, L. Qi, Y. Yang, M. Atiquzzaman, "Automated colorization of a grayscale image with seed points propagation," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1756-1768, Jul. 2020.

[39] F. Ozcelik, U. Alganci, E. Sertel, G. Unal, "Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3468-3501, Apr. 2021.

[40] Y. Zhang, C. Liu, M. Sun and Yangjun Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549-5563, Aug. 2019.

[41] H. Zhang and J. Ma, "GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening," *ISPRS J. Photogramm. Remote Sens.*, vol. 172, pp. 223-239, Feb. 2021.

[42] Z. Jin, Y. Zhuo, T. Zhang, X. Jin, S. Jing, L. Deng, "Remote sensing pansharpening by full-depth feature fusion," *Remote Sens.*, vol. 16, pp. 466, Jan. 2022.

[43] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100-112, Jan. 2017.

[44] X. Otazu, M. González-Audícana, O.. Fors and J. Núñez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376-2385, Oct. 2005.

[45] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and Pan imagery," *Photogramm. Eng. RemoteSens.*, vol. 72, no. 5, pp. 591-596, May 2006.

[46] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A Generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227-10242, Dec. 2021.

[47] M. M. Khan, J. Chanussot, L. Condat, A. Montavert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 98-102, Jan. 2008.

[48] L. Wald, Weinberger, "Quality of high resolution synthesized images: Is there a simple criterion?," *Proc. 3rd Conf. Fusion Earth Data*, pp. 99-105, 2000.

[49] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81-84, Mar. 2002.

[50] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147-149.

[51] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313-317, Oct. 2004.

[52] L. Alparone et al., "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193-200, Feb. 2008.