A Unified Two-Stage Spatial and Spectral Network With Few-Shot Learning for Pan-Sharpening

Zhi Sheng, Feng Zhang, Jiande Sun, Yanyan Tan, Kai Zhang, *Member, IEEE*, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*—Recently, pan-sharpening methods based on deep learning (DL) have achieved state-of-the-art results. However, current existing DL-based pan-sharpening methods need to be trained repetitively for different satellite sensors to obtain satisfactory fusion performance and therefore require a large number of training images for each satellite. To deal with these issues, in this paper we propose a unified two-stage spatial and spectral network (UTSN) for pan-sharpening. A branch of networks is constructed for each different satellite, in which the spatial enhancement network (SEN) is shared to improve the spatial details in the fused images from different satellites. A spectral adjustment network (SAN) is employed to capture the spectral characteristics of the specific satellite. Through SAN, the spectral information in the intermediate image from SEN is refined to produce the final fusion results. Such a framework can integrate the datasets from different satellites together for sufficient training of SEN. The proposed method is able to achieve promising pan-sharpening results also for a new satellite with limited training images by only learning a new SAN on the few-shot datasets due to the simple but efficient structure of SAN. The experimental results show that the proposed method can produce state-of-the-art fusion results in both the standard and few-shot cases. The source code is publicly available at https://github.com/RSMagneto/UTSN.**

***Index Terms*—Pan-sharpening, spatial enhancement network, spectral adjustment network, few-shot learning, remote sensing.**

I. Introduction

Nowadays, an increasing number of satellites have been launched into space. These satellites carry different kinds of imaging sensors for data collection of the observed scenes. The obtained remote sensing images can contain various information in different observation manners. For example, a multispectral (MS) image can provide abundant spectral information coded in several bands. Sharp spatial details can be seen in a panchromatic (PAN) image. However, in general, the spatial resolution of a MS image is lower than that of a single band PAN image. This is caused by the essential tradeoff between spatial and spectral resolutions. Therefore, it is very difficult to obtain a high spatial and spectral resolution MS (HR MS) image due to physical limitations. To address the issue, image fusion techniques, also called pan-sharpening, have been developed in the literature to fuse the PAN and low spatial resolution MS (LR MS) images for the generation of the HR MS image. These techniques achieve an HR MS image with spatial and spectral resolutions of the PAN and LR MS images, respectively. Through pan-sharpening, more comprehensive scene information is described in the generated HR MS images, which will facilitate the subsequent tasks, such as change detection [1] and land cover classification [2].

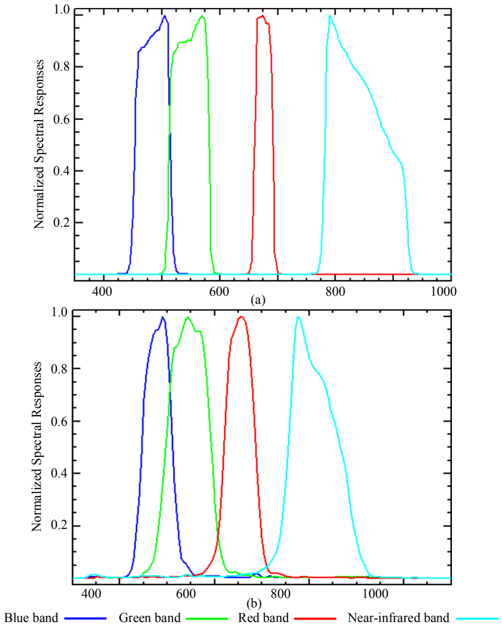


Fig. 1. Normalized spectral responses of different satellites: (a) GeoEye-1; (b) QuickBird.

Recently, deep learning (DL) has achieved considerable performance in the pan-sharpening field. For example, Ozcelik *et al.* [3] used the colorization scheme to fuse LR MS and PAN images. Xu *et al.* [4] designed an unfolding network to improve the spatial information in LR MS images. As more DL-based methods are proposed, the fusion performance of LR MS and

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

Z. Sheng, F. Zhang, J. Sun, Y. Tan, K. Zhang are with the School of Information Science and Engineering, Shandong Normal University, Ji'nan 250358, China (e-mail: sheng888zhi@qq.com, fengzhangpl@163. com, jiandesun@hotmail.com, yytan928@sdnu.edu.cn, zhangkainuc@163.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 images is further improved. However, it is necessary to further consider the issues related to the training and testing on multiple datasets from different satellites. For convenience, some pan-sharpening methods based on DL integrate the datasets from multiple satellites into a mixed dataset for training. Then, the LR MS and the PAN images to be fused from a specific satellite sensor are directly fed into the network trained on the mixed dataset for the fusion. Although many image pairs are included in the mixed dataset for training, the fusion result of a specific satellite sensor is influenced by the mapping information in the network learned from the data of the other satellite datasets. Fig. 1 displays an example of the spectral responses of different satellites. DL-based methods aim at learning a mapping function between the HR MS image and the pair of LR MS and PAN images. Due to the differences in the spectral response shown in Fig. 1, mappings for different satellites are distinct. Thus, when training on the dataset made up of the datasets from different satellites, the network will be affected by different spectral responses.

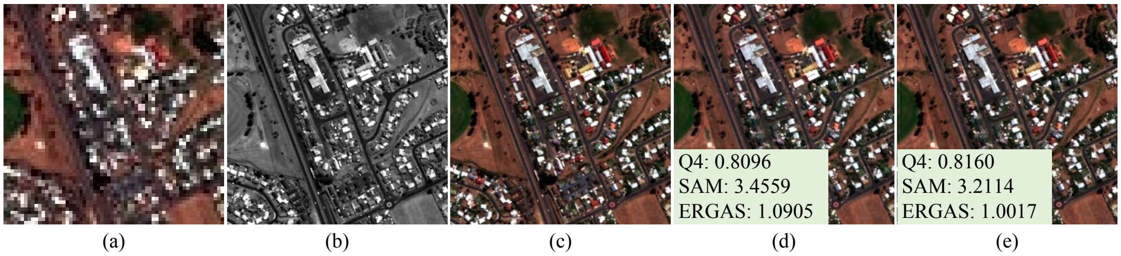


Fig. 2. Fusion results of PNN on different training datasets. (a) LR MS image; (b) PAN image; (c) Reference image; (d) Fused image on the mixed dataset; (e) Fused image on the dataset only from the GeoEye-1satellite.

Moreover, for further analysis, Fig. 2 compares the fusion results of PNN [21] trained from different datasets. PNN is first learned on a mixed dataset, which is made up of 250 pairs of 64×64 LR MS and 256×256 PAN images from the GeoEye-1 satellite and 250 LR MS and PAN image pairs with the same dimensions from the QuickBird satellite. Then, LR MS and PAN images in Figs. 2(a) and 2(b) are fed into the learned model to produce the fused image shown in Fig. 2(d). Then, another model of PNN is trained on the dataset only containing 500 LR MS and PAN image pairs from the GeoEye-1 satellite. The trained model is employed to fuse the images in Figs. 2(a) and 2(b). The fusion result is displayed in Fig. 2(e). Evaluation indexes are reported in Fig. 2 for a direct comparison. The indexes include global quality measurement (Q4) [5], spectral angle mapper (SAM) [6], and *Erreur Relative Globale Adimensionnelle de Synthèse* (ERGAS) [7]. Through comparison of the images in Figs. 2(d) and 2(e), one can see that the fusion result on the mixed dataset has inferior performance, which is caused by domain shift among different satellite datasets. Although the fused images are better if the test images and the training images are from the same satellite, this requires training the specific network for different satellites repeatedly. Repetitive training means more computational resources and limited operational flexibility.

In addition, existing DL-based pan-sharpening methods cannot deal with the few-shot cases where for some satellites there are not enough MS and PAN images to construct the training datasets. For example, the training image pairs are often insufficient for newly launched satellites. If DL-based methods are trained on limited data, it is difficult to produce satisfactory fusion results. Although the model trained on the datasets from other satellites can be employed to deal with this case, the fusion result is influenced by distinct spectral responses, as analyzed in Fig. 1.

To cope with the issues mentioned above, we propose a unified two-stage spatial and spectral network (UTSN), which improves the fusion performance and utilizes the datasets from different satellites for training simultaneously. In the proposed method, UTSN first utilizes the same spatial enhancement network (SEN) to enrich the spatial details in the fused image. Then a satellite-specific spectral adjustment network (SAN) is adopted to further improve the spectral information of the intermediate image from SEN. By SAN, the spectral information of the intermediate image is adjusted to adapt to the spectral response of the specific satellite. In UTSN, SEN is trained collaboratively by using the mixed dataset, which is made up of all datasets from different satellites. Then, the mixed dataset is divided into satellite-specific datasets for the training of the corresponding SANs. Through the training with the collaboration and division of datasets, a sufficiently effective SEN can be learned on the mixed dataset and UTSN can avoid the repetitive training of the whole model on different satellites.

Furthermore, for the data from a newly launched satellite, only the corresponding SAN needs to be trained. The SEN learned from the mixed dataset can be directly employed and cascaded with the SAN of the new satellite to produce the fusion result. Note that, most of the parameters of UTSN are included in SEN, which is learned on the mixed dataset. In SAN, only a channel attention block is used, which includes a small number of parameters. So, SAN can be trained sufficiently when the data from the new satellite is insufficient or few-shot. In this way, the proposed UTSN has a good generalization ability to the newly launched satellites. Experimental results on datasets acquired by 4 satellites show the effectiveness of UTSN in standard and few-shot pan-sharpening cases. To the best of our knowledge, this is the first time that few-shot learning is considered for pan-sharpening. The contributions of the proposed UTSN can be summarized as follows:

1) A unified framework is constructed to leverage all data from different satellites together for the learning of the network. In the framework, SEN is trained on the mixed dataset and is

shared among different satellites. We establish SEN to inject spatial details from the PAN image into the LR MS image level by level. To avoid the excessive injection of spatial information, the differences between the feature maps from different layers are added into LR MS images to preserve the subtle details in the fused image. Satellite-specific SAN then is used to adjust the spectral information in the result of SEN.

2) The proposed UTSN can be easily extended to the fusion of LR MS and PAN from a new unseen satellite when the training data of this satellite is insufficient or few-shot. Specifically, the new satellite shares the same SEN with other satellites for the enhancement of spatial details. Then, a satellite-specific SAN is trained for this new satellite on its few-shot dataset owing to the lightweight model and simple structure of SAN.

The remainder of the paper is organized as follows. In Section II, we present the proposed UTSN in terms of SEN, SAN, and loss functions. Section III demonstrates quantitative and qualitative results on different datasets. Finally, the conclusion is detailed in Section IV.

II. Related Work

Due to the decent performance, pan-sharpening has attracted a great deal of attention and many pan-sharpening methods have been proposed [8]-[9], which can be mainly classified into four groups: 1) component substitution (CS) based methods, 2) multiresolution analysis (MRA) based methods, 3) degradation model (DM) based methods, and 4) DL-based methods. They are introduced successively in Sections II.A-II.D. Besides, Section II.E gives some methods related to transfer learning to present its development in remote sensing fields.

A. *CS-Based Methods*

For CS-based pan-sharpening methods, many transformation methods are considered to divide the interpolated LR MS image into spatial and spectral components. Then, the spatial component is substituted by the histogram-matched PAN image. For example, intensity-hue-saturation (IHS) [10] was usually used for the extraction of the intensity component. In principal component analysis (PCA) [11], the spatial component of the LR MS image was modeled by the first principal component. It is obvious that the quality of the fusion result will be better if the spatial component of the LR MS image is highly correlated with the corresponding PAN image. However, it is difficult for these transformations to separate the spatial and spectral information in the LR MS image accurately. A nonlinear version of IHS [12] was extended to synthesize the intensity component in local and global ways, which can reduce the spectral distortions in the fused image. Shahdoosti *et al.* [13] applied PCA to spatial and spectral domains for better information preservation. CS-based methods are simple and easy to carry out. But their fusion results often suffer from spectral distortions.

B. *MRA-Based Methods*

For the second kind of method, many MRA tools are utilized to extract the spatial details from the PAN image, because these details are missing in the LR MS image. For instance, Shah *et al.* [14] used contourlets to capture the discontinuities of the spatial structures in the PAN image. Xing *et al.* [15] constructed a series of multiscale geometric support tensor filters, which focused on the analysis of the directional information in spatial details. Besides, injection gains also have important influences on the fusion result. So, Restaino *et al.* [16] estimated more accurate injection gains to alleviate the spectral distortions in the fused image. In [17], robust regression was combined with the generalized Laplacian pyramid (GLP), and then different injection coefficients were computed for each cluster. Compared to the methods based on CS, MRA-based methods have a good ability to preserve spectral information, but some dissimilar artifacts may appear in the fusion result owing to the mismatched filters in spatial details.

C. *DM-Based Methods*

The DM-based methods assume that the spatial and spectral degradation results of the HR MS image are LR MS and PAN images, respectively. Therefore, the fusion result can be obtained by solving the degradation models [18]. Inspired by the image restoration task, a wide range of priors are imposed to regularize the solution space of the degradation models. Li *et al.* [19] first embedded the degradation models into the compressed sensing framework, where the sparsity was used decently to obtain better solutions. In [20], the relationships between LR MS and HR MS images were rewritten as the formulation of robust PCA [21], and the correlation among the bands of the MS image was characterized by low-rank properties. Subsequently, Zhang *et al.* [22] took the low-rank and sparse priors into account together to produce the fused image. In addition, other priors, such as non-negativity [23] and consistency priors [24], are also utilized in the third kind of method. Although these methods can enhance the spatial and spectral information in the fused images well, their computational complexity is non-negligible, which results from the iterative optimization algorithms.

D. *DL-Based Methods*

During the past several years, DL-based pan-sharpening methods have been greatly boosted due to the powerful learning capabilities of deep neural networks [25]. Masi *et al.* [26] first employed a convolution neural network (CNN) to cope with the pan-sharpening task and the network is dubbed as PNN. Then, an extended version of PNN was developed in [27] to improve the quality of the fused image. Wei *et al.* [28] introduced residual learning [29] into CNN to increase the depth of the network and then the fusion result is further enhanced owing to the high nonlinearity of the network. Wang *et al.* [30] presented a locally linear embedding residual network to exploit the geometric relationships in images. Moreover, Shao *et al.* [31] incorporated residual learning and generative adversarial network (GAN) [32] to enrich the spatial details in the fused image. GAN was also considered in [33] and different architectures were evaluated to obtain the desired HR MS image. Diao *et al.* [34] proposed a zero-reference GAN, which can fuse the LR MS and PAN images without training in advance. Deng *et al*. [35] explored the combinations of CNN and the schemes of CS- and MRA-based methods. Wang *et al*. [36] constructed a high-pass modification block to enrich the spatial details in LR MS images more efficiently. Following the concept of image super-resolution (SR), Cai *et al*. [37] proposed a progressive fusion approach, which consisted of SR, pan-sharpening, and residual modules. An elaborate network was constructed in [38] to fuse PAN and LR MS images, in which a saliency cascade network was used to distinguish the regions with different textures. A dilated deformable CNN was equipped in this method to promote the receptive fields. Besides, Qu *et al*. [39] presented an unsupervised pan-sharpening network to reduce the demand for the HR MS image, where the spatial details and gains were calculated by a self-attention network. A multiscale dense network [40] was designed to extract the spatial and spectral information, in which the concatenated multiscale blocks captured the subtle features in LR MS and PAN images. Besides, more advanced networks, such as transformer [41]-[43], are also applied to this field. The methods above-mentioned produce good fusion results, but they train specific models for different datasets or integrate all datasets for training. The formulations lead to repetitive training or the neglect of spectral response differences between different satellites. To efficiently use all datasets and consider the spectral responses of different satellites, we built UTSN, in which SEN is trained on the mixed dataset and SANs corresponding to different satellites are designed for spectral adjustment.

E. Transfer Learning

The proposed UTSN can achieve simultaneous training of all data from different satellites owing to the introduction of SANs. Besides, we can extend the proposed UTSN to the dataset from the new unseen satellite only by sharing the same SEN with other existing satellite datasets and training a corresponding SAN on a few-shot dataset. Thus, it can be viewed as a new satellite-specific SAN constructed for further adaption to the data from the new satellite. To the best of our knowledge, the proposed UTSN is the first attempt to consider integrated training on all data from different satellites and few-shot learning on limited data for better generalization. According to the formulation, the proposed UTSN belongs to transfer learning. Researchers have so far not considered transfer learning in the pan-sharpening field. But transfer learning has been applied to other remote sensing fields. For example, Wang *et al.* [44] designed an adaptive learning strategy, which transfers the knowledge from a pre-trained model for the classification of remote sensing scenes. Ma *et al.* [45] introduced dual-branch attention into transfer learning to alleviate the issues of inter- and intra-class differences. Besides, transfer learning is also used for the classification of targets or pixels in images [46]-[47].

III. Proposed Method

In this section, we introduce the proposed UTSN method and reformulate the training problem on the mixed dataset according to the proposed framework. Then, the architectures of SEN and SAN are described in detail.

A. Overall Framework

As introduced in Section I, we utilize a shared SEN to enhance the spatial details in the fusion image, and then use the satellite-specific SAN to adjust the spectral information in the intermediate image. The framework is presented in Fig. 3, where k is the index of the satellite sensor. K is the number of all considered satellite sensors. For images to be fused from different satellites, SEN is considered to merge the spatial information in LR MS and PAN images. Then, the output of SEN is further refined by SAN to adapt to the specific properties of the considered satellite sensor.

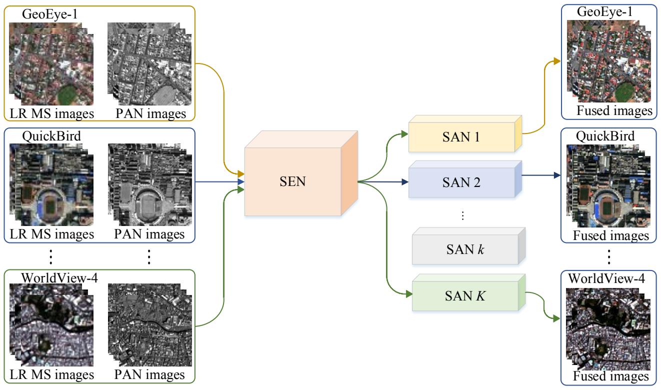


Fig. 3. Illustration of the framework of the proposed UTSN.

According to the description in Fig. 3, the training task of the proposed UTSN can be formulated as:

$$\begin{gathered}\{\theta_0,\theta_1,\theta_2,\ldots,\theta_K\} = \underset{\theta_0,\theta_1,\theta_2,\ldots,\theta_K}{\arg\min}\ \mathcal{L} \\ \mathcal{L}=\sum_{k=1}^{K}\sum_{n=1}^{N_k} L\left(\mathbf{R}_n^k, f\left(\mathbf{L}_n^k,\mathbf{P}_n^k;\theta_0,\theta_1,\theta_2,\ldots,\theta_K\right)\right)\end{gathered} \tag{1}$$

where θ_0 is the parameter of SEN. θ_k stands for the parameters of SAN for the *k*th satellite sensor. $\mathbf{L}_n^k$ and $\mathbf{P}_n^k$ are the nth pair of LR MS and PAN images from the *k*th satellite. $\mathbf{R}_n^k$ is the corresponding HR MS image. The number of training images from each satellite is denoted as N_k. $f(\cdot)$ represents the mapping of UTSN. Compared with existing methods that learn independent models for different satellites, the parameters in the proposed framework can be jointly learned by minimizing the loss function $\mathcal{L}$.

B. SEN

In UTSN, SEN is mainly responsible for the enhancement of spatial details in the fused images. To inject enough spatial details from PAN images, their feature maps can be directly combined with those of LR MS images. However, this would lead to local dissimilarities in the fused image [48]. Moreover, some spatial details may be injected repeatedly into the feature maps of the LR MS image. This would cause spatial artifacts in the fused images.

We build SEN to enhance the spatial information of LR MS images, as shown in Fig.4. In SEN, the up-sampled LR MS and PAN images are concatenated together to reconstruct the fused image by CNN 1. CNN 1 is composed of seven convolution

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.2023.3281602

layers in which the filter size is 3×3 with stride 1. In the first six layers, convolution and ReLU are regarded as the basic convolution block. The last layer of CNN 1 contains four filters with a size of 3×3. The number of filters in the last layer is equal to the number of bands in the MS image. Moreover, dense connections are introduced into CNN 1. Through dense connections, the feature maps from shallow layers are introduced into deep layers for the preservation of spatial information in the fused images. Batch normalization (BN) often used in DL-based methods is removed in the proposed SEN to improve its generalization ability [49]. As shown in Fig. 1, we expect spectral differences for the images from different satellites. BN tends to normalize the images from different satellites into the same distribution, which will degrade the fusion performance of SEN.

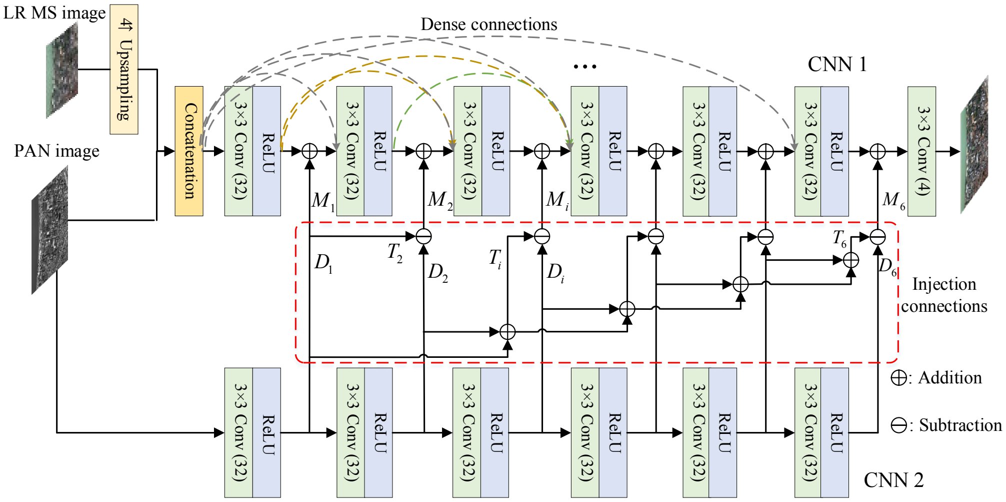


Fig. 4. Architecture of the proposed SEN.

In addition, we employ another CNN 2 with a similar architecture to focus on the extraction of spatial details in the PAN image. Then, these spatial details are injected into CNN 1 for the spatial enhancement of the fused image. To avoid excessive injection of spatial details from the PAN image, we devise the injection connections shown in Fig. 4 as the link between the two CNNs. The subtraction among feature maps from different layers is introduced into SEN. As shown in Fig. 4, injection connections for the *i*th layer can be written as:

$$M_i = D_i - T_i \tag{2}$$

where M_i denotes the features that will be injected into the *i*th layer of CNN 1. D_i is the output of the *i*th layer of CNN 2. $T_i = \sum_{q=1}^{i-1} D_q$ is the sum of feature maps from the 1st to the *i*-1th layers in CNN 2. For the injection in the first layer, T_1 will be zero. Through injection connections, the injected information at previous layers can be removed from the feature maps at the following layers. Then, spatial details in the fused image are enhanced well and the subtle structures in the fusion result are also preserved. Moreover, the spectral information of the LR MS image is also retained because only fine details are added to the feature maps for the reconstruction of the fused image.

SEN is a common network and thus is trained by the mixed dataset from all satellites. So, the shared SEN can be learned by:

$$\theta_0 = \arg\min_{\theta_0} \sum_{k=1}^{K} \sum_{n=1}^{N_k} L\left(\mathbf{R}_n^k, f_0\left(\mathbf{L}_n^k, \mathbf{P}_n^k; \theta_0\right)\right) \tag{3}$$

where the mapping of SEN is denoted by $f_0(\cdot)$ and θ_0 stands for the parameters of the mapping. Since most of the parameters of the proposed UTSN are concentrated in SEN, the feature extraction capability of SEN will be more powerful by leveraging on all data. Besides, all data from different satellites are integrated for the training of SEN. Although the data from different satellites have distinct spatial resolutions, the patch recurrence property demonstrates that spatial patterns in images tend to recur many times at different scales [50]-[51]. Moreover, the spatial information of small-scale targets in HR images also can be learned from large-scale ones in LR images. Thus, we use the mixed dataset containing the data from different satellites for the training of SEN.

C. SAN

SAN aims to adjust the spectral information of the result of SEN. For a specific satellite, SAN is trained to capture the spectral information of the MS image. SANs for different satellites share the same architecture shown in Fig. 5. Generally, MS images from different satellites have diverse spectral properties, which depend on the spectral responses of imaging sensors. For MS images, the spectral information is highly correlated to the dependencies or correlations among bands, which can be reflected by the relationships [52]. Thus, we introduce the attention mechanism [53] into the specific-satellite SAN to capture the correlations among the bands of the MS image. In Fig. 5, a global average pooling (GAP) is implemented on the feature map along the channel dimension to obtain a vector. Then, fully connected layers (FC) are considered to learn the interdependency of all elements in the vector. Finally, the interdependency learned from the vector is combined with each channel in the feature map. Through the channel attention module, the spectral information of MS images is captured in feature space. Then, SAN for the *k*th satellite is trained by:

$$\theta_k = \arg\min_{\theta_k} \sum_{n=1}^{N_k} L\left(\mathbf{R}_n^k, f_k\left(\mathbf{H}_n^k;\theta_k\right)\right) \tag{4}$$

where $\mathbf{H}_n^k$ is the output of SEN whose corresponding inputs are $\mathbf{L}_n^k$ and $\mathbf{P}_n^k$. $f_k(\cdot)$ is the mapping of the *k*th SAN with parameter θ_k. Through SAN, we can obtain a spectral adjustment mask. The mask is then multiplied with $\mathbf{H}_n^k$ element-by-element to generate the final fusion result $\mathbf{F}_n^k$:

$$\mathbf{F}_n^k = \mathbf{S}_n^k \odot \mathbf{H}_n^k \tag{5}$$

where $\mathbf{S}_n^k$ is the spectral adjustment mask. Despite the simple architecture of SAN, the nonlinear relationships among the bands of the MS image are extracted efficiently. Due to the lightweight structure, SAN can be trained on a limited amount of data to deal with the few-shot case. Specifically, only a relatively few pairs of LR MS and PAN images are needed to train a specific SAN for a given satellite. Then, the shared SEN trained on the datasets from other satellites is used to produce the inputs of the specific SAN. This reduces the cost for the collection of a large number of image pairs.

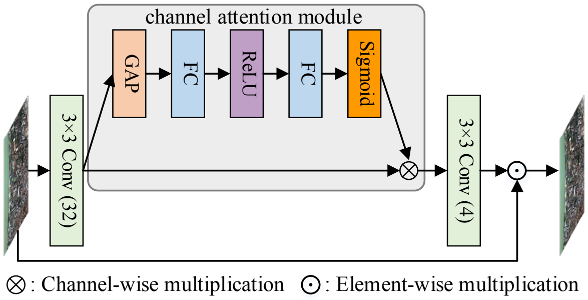


Fig. 5. Architecture of proposed SAN.

D. Loss Function

In the proposed UTSN, we train the model with L_2-norm, and the total loss is formulated as:

$$\mathcal{L} = \sum_{k=1}^{K}\sum_{n=1}^{N_k} \left\|\mathbf{H}_n^k - \mathbf{R}_n^k\right\|_F^2 + \left\|\mathbf{S}_n^k \odot \mathbf{H}_n^k - \mathbf{R}_n^k\right\|_F^2 \tag{6}$$

With the joint training of SEN and SAN, the spatial and spectral information for the dataset from each satellite can be learned efficiently. Finally, the fused image can be obtained by (7) for the *n*th image pair from the *k*th satellite.

$$\mathbf{F}_n^k = f_k\left(f_0\left(\mathbf{L}_n^k, \mathbf{P}_n^k\right)\right) \tag{7}$$

IV. Experimental Results

In this section, we evaluate the proposed UTSN using reduced-scale and full-scale datasets. First, we validate the effectiveness of the proposed method on three reduced-scale datasets. Then, fusion experiments are conducted on the three full-scale datasets to compare the performance of all methods. The compared methods are modulation transfer function (MTF) -GLP with context-based decision (MTF-GLP-CBD) [54], Gram-Schmidt (GS) [55], low-rank pan-sharpening (LRP) [20], pan-sharpening network (PanNet) [56], CNN-based pan-sharpening (PNN) [26], two-stream fusion network (TFNet) [57], GAN-based pan-sharpening (PSGAN) [33], GAN with multiple discriminators (M-GAN) [3], and gradient projection based pan-sharpening neural network (GPPNN) [4]. The first three methods are classified as traditional methods, while the latter methods are based on DL.

In the reduced-scale experiments, the results of all methods are assessed quantitatively by five reference-based metrics, such as Q4, SAM, universal image quality index (UIQI) [58], root mean squared error (RMSE), and ERGAS. Q4 and UIQI vary from 0 and 1. A larger Q4 or UIQI means better fusion performance. For SAM, RMSE, and ERGAS, their optimal value is 0, and smaller values imply an accurate reconstruction of the fused image. In addition, three no-reference metrics, D_λ, D_S, and quality w/no reference (QNR) [59], are considered to evaluate the results in the full-scale experiments. D_λ and D_S measure the spectral and spatial distortions of the fused image, respectively. QNR is a global metric, whose best value is 1.

Moreover, few-shot experiments are analyzed on the dataset of the Pléiades satellite that is not included in the satellite data used for the former experiments. Finally, ablation studies and network analysis are also introduced to provide a comprehensive evaluation of the proposed UTSN.

A. Datasets and Design of Experiments

For the training of the proposed UTSN, we build different datasets from three satellites, including QuickBird[1], GeoEye-1[2], and WorldView-4[2]. For these satellites, the acquired MS images are made up of four bands. Table I reports detailed information on these datasets. The ground sample distance (GSD) varies from 0.31m to 0.64m for different datasets. The sizes of PAN and LR MS images are 256×256 and 64×64, respectively. The PAN and LR MS image pairs for training are synthesized according to Wald's protocol [7]. The original PAN and MS images are down-sampled to produce the reduced-scale counterparts with a factor of *r*. Here, the down-sample ratio is generally 4. Then, the original MS image is viewed as the reference image for the supervised training. 80% and 10% of data are selected from each satellite dataset for the composition of the mixed training and validation datasets of SEN, respectively. For the training and validation of SAN, a satellite-specific dataset is considered by choosing the corresponding image pairs from the mixed training dataset and validation dataset. Then, the remaining data are prepared for the test. For testing at full scale, each satellite dataset contains 15 LR MS and PAN image pairs without reference images. Fusion experiments are conducted on reduced-scale and full-scale datasets. For the full-scale experiment, the model trained on the reduced-scale dataset is directly employed to fuse LR MS and PAN images at full scale because there are no reference images for training in the full-scale dataset.

[1] http://glcf.umiacs.umd.edu/data/quickbird/
[2] https://resources.maxar.com/product-samples

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.2023.3281602

TABLE I. DETAILS OF THE THREE CONSIDERED DATASETS.

Dataset	Number of image pairs	Spatial resolution of the PAN image	Spatial resolution of the LR MS image	Time	Location
QuickBird	250	0.64 m GSD at nadir	2.56 m GSD at nadir	Sept. 30, 2008	Xi'an, China
GeoEye-1	302	0.46 m GSD at nadir	1.84 m GSD at nadir	Feb. 24, 2009	Hobart, Australia
WorldView-4	202	0.31 m GSD at nadir	1.24 m GSD at nadir	Apr. 5, 2017	Acapulco, Mexico

The DL-based methods compared in the section are trained on the three datasets from different satellites independently to produce better fusion results. So, each DL-based method has to be trained three times. For the proposed UTSN, the datasets from three satellites are collaboratively integrated as a mixed dataset for the training of SEN, and then the mixed dataset is divided to learn satellite-specific SANs. Traditional methods are executed on a computer with an Intel Core i7-6700 processor, 3.4 GHz, and 16G memory by MATLAB R2017a. The training and test of all DL-based methods are performed on the same device with one NVIDIA 2080Ti GPU by PyTorch. The settings of compared methods based on DL are derived from their literature recommendations. For the proposed UTSN, the learning rate is set as 0.0002. We use Adam optimizer [60] to train the model for 800 epochs with batch size 4. For the few-shot experiment, the number of epochs for a specific SAN is 500.

B. Experiments on Reduced-Scale Datasets

In this section, the experiments are conducted on the reduced-scale LRMS and PAN images from three satellites. Fig. 6 shows the fusion results of all methods on the GeoEye-1 dataset. We selected interesting regions from the fused images and magnified them for further visual analysis. Fig 6 also presents absolute error maps among the fused images and the reference image. From Fig. 6, we can see that the spatial details of the GS result in Fig. 6(e) are over-enhanced because the PAN image is regarded as the new spatial component directly. The result of LRP in Fig. 6(f) suffers from spectral distortions, especially in the vegetation area, which may result from the estimation error of the spectral degradation model in LRP. For the results of DL-based methods, spectral information is preserved well. But some spatial distortions appear in the result of PNN in Fig. 6(h) due to its simple architecture. Moreover, slight spectral distortions are found in the magnified region of the PSGAN result in Fig. 6(j). Obvious spectral distortions can be found in Fig. 6(k), which may be caused by the excessive constraint of the spectral loss. Compared with other results, the fused image derived by the proposed UTSN in Fig. 6(m) is better in terms of both spatial and spectral information. Moreover, from the absolute error maps of traditional methods in Fig. 6, one can see that there are larger reconstruction errors. DL-based methods have better reconstruction performance than other methods and the absolute errors of the proposed method are smaller than those of other methods. Table II lists the average evaluations of all methods on the simulated GeoEye-1 dataset and the best values are labeled in bold. From the table, we can observe that the proposed UTSN provides the best values in terms of Q4, UIQI, RMSE, and ERGAS. The best SAM is produced by TFNet and the second-best SAM value is from our proposed UTSN.

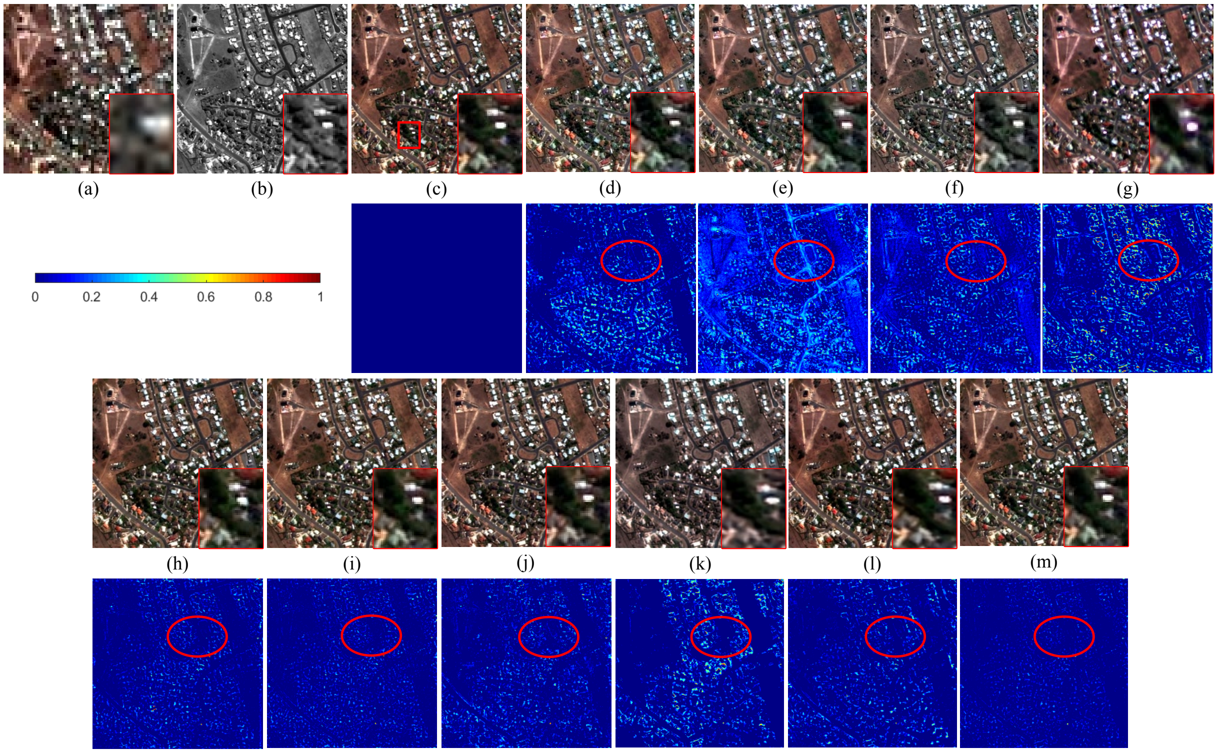
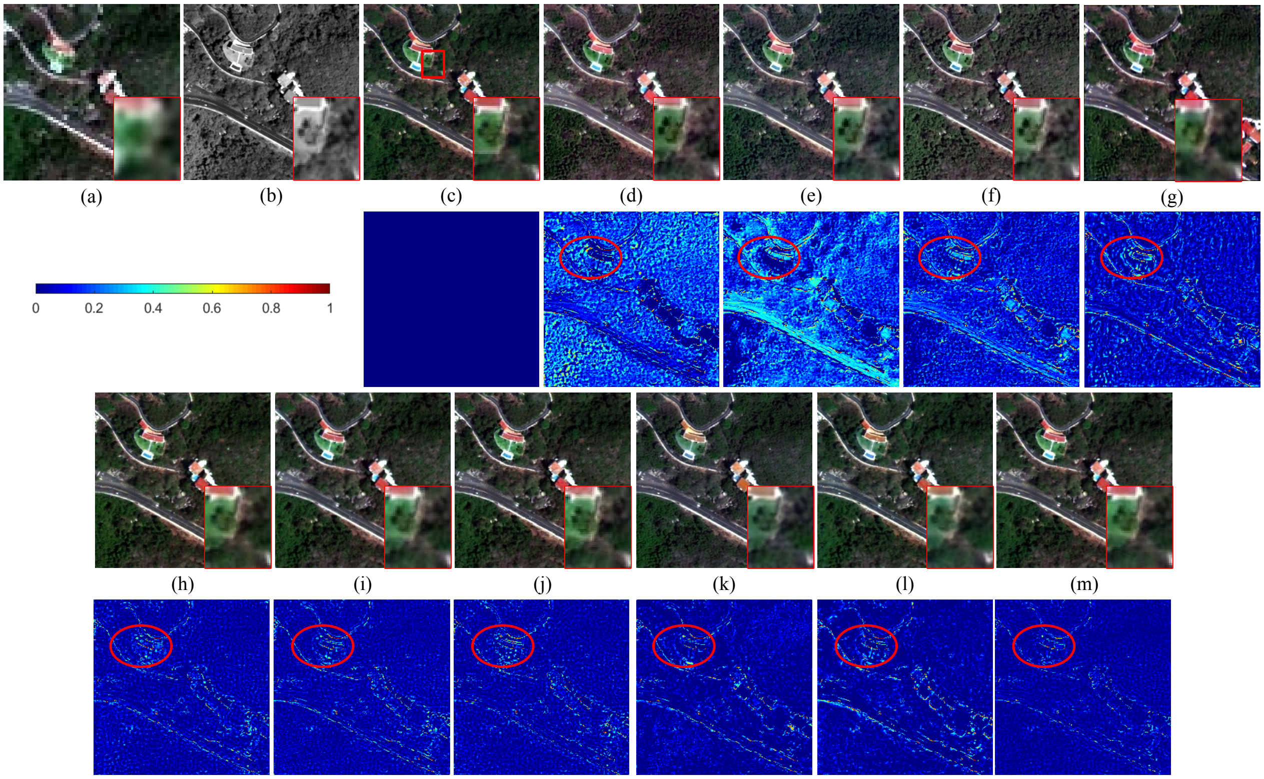


Fig. 6. Qualitative comparison of the fused images from different methods on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) MTF-GLP-CBD; (e) GS; (f) LRP; (g) PanNet; (h) PNN; (i) TFNet; (j) PSGAN; (k) M-GAN; (l) GPPNN; (m) Proposed UTSN.

The fusion results on the WorldView-4 dataset are displayed in Fig. 7. Similarly, Fig. 7 also demonstrates magnified regions and absolute error maps of the fused images for direct visual perception. From Fig. 7(d), spectral distortions can be observed in the result of MTF-GLP-CBD. The color of the vegetation areas in Fig. 7(d) is unnatural, which may be caused by the improper context division. Similarly, the spectral information of the vegetation areas in the result of LRP is distorted. For the PanNet result, the texture information of the tree areas is not consistent with that of the reference image. The reason for this is that the spatial details learned by PanNet are distorted. Results of other DL-based methods better preserve the spectral feature, and the fused image of the proposed method is closer to the reference image in terms of reconstruction precision. Moreover, the absolute error maps exhibit a similar trend. The absolute errors of MTF-GLP-CBD and GS are obvious. PNN, TFNet, PSGAN, and the proposed UTSN perform better in the vegetation regions but produce larger differences in the edges of road regions. When compared with other methods, the absolute errors of UTSN are closer to 0. The numerical metrics are reported in Table III. It shows that the proposed UTSN produces the best values for all metrics.

TABLE II. QUANTITATIVE EVALUATIONS ON GEOEYE-1 DATASET

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
Q4	0.8048	0.7672	0.7435	0.7983	0.8186	0.8271	0.8186	0.8026	0.8073	**0.8274**
SAM	4.7013	4.6604	5.3090	6.1171	3.1198	**2.7926**	3.3697	4.5152	4.0189	2.8076
UIQI	0.9527	0.9404	0.9101	0.9533	0.9789	0.9825	0.9751	0.9538	0.9692	**0.9837**
RMSE	25.4749	24.9112	49.7777	36.3701	16.5597	14.9036	18.2434	24.0138	19.7152	**14.2591**
ERGAS	1.6543	1.6064	3.2810	2.3642	1.0688	0.9676	1.1850	1.5532	1.2790	**0.9274**


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

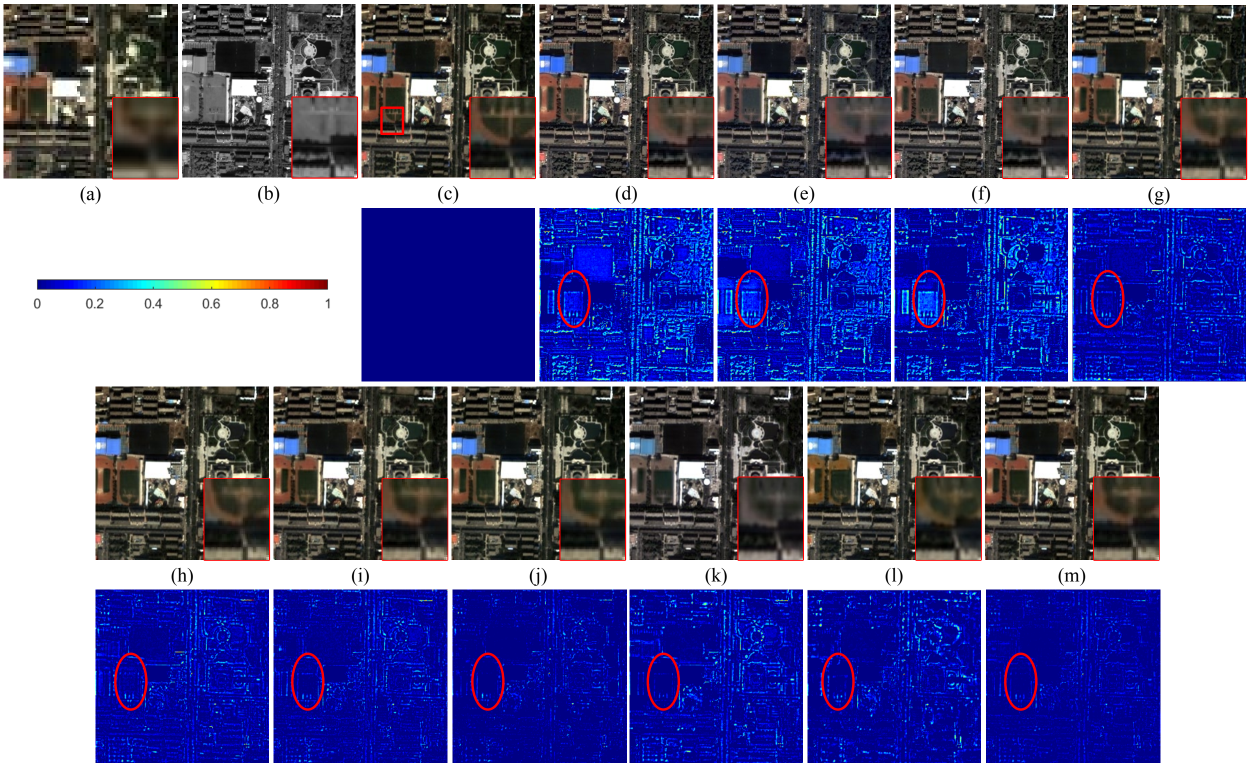
(h) (i) (j) (k) (l) (m)

Fig. 7. Qualitative comparison of the fused images from different methods on the WorldView-4 dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) MTF-GLP-CBD; (e) GS; (f) LRP; (g) PanNet; (h) PNN; (i) TFNet; (j) PSGAN; (k) M-GAN; (l) GPPNN; (m) Proposed UTSN.

TABLE III. QUANTITATIVE EVALUATIONS ON WORLDVIEW-4 DATASET

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
Q4	0.8289	0.7953	0.8009	0.8677	0.8723	0.8767	0.8679	0.8679	0.8472	**0.8795**
SAM	4.1402	3.9265	4.4397	4.9550	3.0100	3.0589	3.2032	3.3853	3.8474	**2.7910**
UIQI	0.9414	0.9325	0.9332	0.9588	0.9646	0.9655	0.9639	0.9577	0.9452	**0.9722**
RMSE	55.3963	54.8852	68.6143	53.5092	41.7399	42.2686	43.4857	47.0192	54.0462	**37.7357**
ERGAS	1.9345	1.8790	2.3565	1.7376	1.3988	1.4233	1.4409	1.6086	1.7837	**1.2640**

Fig. 8 shows the fusion results of all considered methods on an urban area in the QuickBird dataset. We also show the interesting regions and absolute error maps in Fig. 8. The edges of some objects are distorted in the results of MTF-GLP-CBD, GS, and LRP. For instance, we can see that the spatial details are lost in the magnified regions of Figs. 8(d)-(f). Moreover, spectral distortions appear in the result of GS because the spatial component synthesized by GS is not matched with the PAN image. For the result of PNN in Fig. 8(h), some blurring effects can be seen. The performance of PNN may be constrained by the shallow structure of the network. Compared with the reference image, slight spectral differences can be seen in the TFNet result. Moreover, M-GAN produces some spectral distortions in its result. The result of GPPNN in Fig. 8(l) suffers from some spatial blurring effects, especially in the enlarged area. The result of the proposed UTSN in Fig. 8(m) shows that the spatial and spectral information in the fused image is more similar to that of the reference image in Fig. 8(c). From the absolute error maps, one can see that larger errors tend to appear on the edges of building areas for all methods. This is because sharp variations in these areas make accurate reconstruction more difficult. Moreover, the absolute error map of the proposed UTSN is more similar to that of the reference image. For the quantitative assessments, Table VI shows that the proposed UTSN achieves the best values, which reflects the high fidelity of the fused image in terms of spatial and spectral features.



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

(h) (i) (j) (k) (l) (m)

Fig. 8. Qualitative comparison of the fused images from different methods on the QuickBird dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) MTF-GLP-CBD; (e) GS; (f) LRP; (g) PanNet; (h) PNN; (i) TFNet; (j) PSGAN; (k) M-GAN; (l) GPPNN; (m) Proposed UTSN.

TABLE IV. QUANTITATIVE EVALUATIONS ON QUICKBIRD DATASET

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
Q4	0.9039	0.8970	0.9145	0.9428	0.9465	0.9468	0.9491	0.9402	0.9373	**0.9496**
SAM	2.8042	2.7874	3.0420	2.9874	1.8703	2.0373	1.6968	2.5240	2.2680	**1.6823**
UIQI	0.9484	0.9367	0.9411	0.9794	0.9852	0.9829	0.9873	0.9705	0.9765	**0.9882**
RMSE	32.1656	29.9777	30.7646	29.0654	15.6712	16.9283	15.1797	22.9652	19.7050	**15.1373**
ERGAS	1.5182	1.4043	1.4209	1.4293	0.7281	0.7867	0.7053	1.0561	0.9179	**0.7041**

C. *Experiments on Full-Scale Datasets*

In this section, we analyze the fusion results of all methods on the full-scale LRMS and PAN images from three satellites. Fig. 9 demonstrates the fusion results of all methods on the GeoEye-1 dataset. Compared with the LRMS image in Fig. 9(a), the fused images of all compared methods provide abundant spatial details. However, we can observe some spectral distortions from the MTF-GLP-CBD and LRP results in Figs. 9(c) and 9(e). Moreover, some noise can be seen from the zoomed regions in Figs. 9(c) and 9(e). Similarly, the result of PSGAN in Fig. 9(i) is affected by the noise and its color information is not consistent with that of other fused images. Some spatial details are blurred in the PanNet result. The spectral information in the PNN result is also distorted because of the limited capacities of nonlinear learning. In addition, some spatial artifacts are observed from the result of GPPNN in Fig. 9(k), especially in the enlarged region. Compared with other methods, the result of the proposed UTSN method is superior in terms of spectral information, which reflects the better performance of SAN. For the values of the no-reference metrics in Table V, the proposed method obtains the best D_{λ} and QNR. Moreover, the difference between the proposed UTSN and PSGAN in terms of D_S is small.

Fig. 10 displays the fused images obtained by the compared methods on the WorldView-4 dataset. Some blurring effects can be observed in the results of MTF-GLP-CBD and LRP in Figs. 10(c) and 10(e). Obvious spectral distortions arise in the result of PNN in Fig. 10(g), which may be caused by the shallow architecture of PNN. We can find some spatial effects in the magnified areas of the fused images of TFNet and PSGAN. Particularly, the spatial information is sharply distorted in the building areas of Fig. 10(i). In the selected region of Fig. 10(k), we can find that the spatial information is destroyed, which may result from improper unfolding. Compared with other methods, one can see that our proposed UTSN better preserves the spatial structures in both smooth and edge areas. The corresponding quantitative results are reported in Table VI. Compared with other DL-based methods, The D_λ value of the proposed method is also close to its best counterpart. The best D_S and QNR are from M-GAN, but the values of the two metrices the proposed UTSN are close to those of M-GAN.

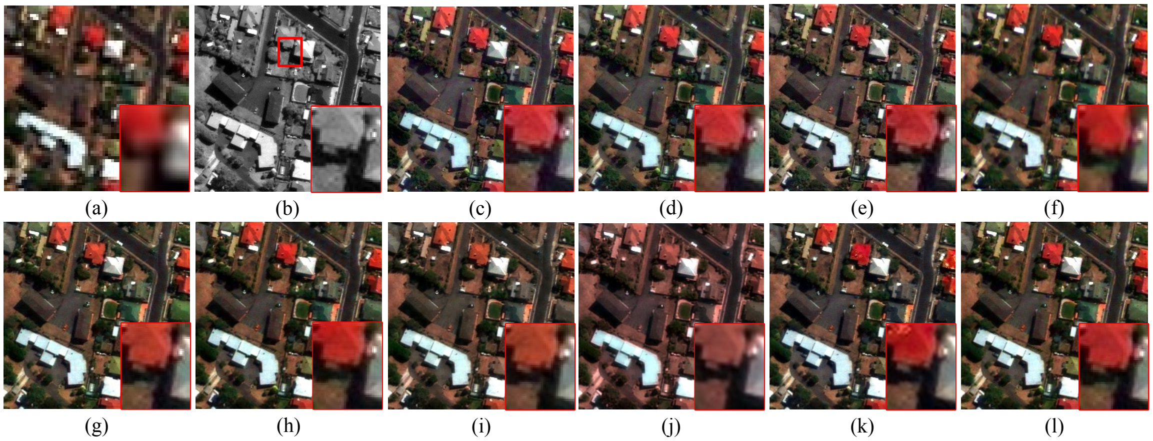

(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k) (l)

Fig. 9. Qualitative comparison of the fused images from different methods on the GeoEye-1 dataset. (a) LR MS image; (b) PAN image; (c) MTF-GLP-CBD; (d) GS; (e) LRP; (f) PanNet; (g) PNN; (h) TFNet; (i) PSGAN; (j) M-GAN; (k) GPPNN; (l) Proposed UTSN.

TABLE V. QUANTITATIVE EVALUATIONS ON GEOEYE-1 DATASET

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
D_λ	0.1495	0.0845	0.0869	0.0644	0.0824	0.0769	0.0936	0.1308	0.0924	**0.0609**
D_S	0.0755	0.0560	0.0930	0.0436	0.0447	**0.0377**	0.0396	0.0656	0.0468	0.0385
QNR	0.7864	0.8642	0.8282	0.8949	0.8767	0.8883	0.8707	0.8124	0.8653	**0.9030**

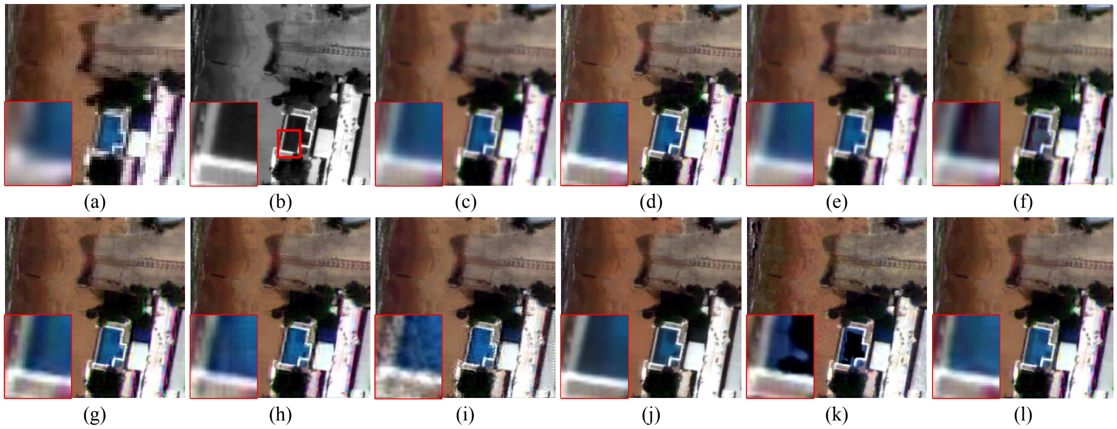

(a) (b) (c) (d) (e) (f)
(g) (h) (i) (j) (k) (l)

Fig. 10. Qualitative comparison of the fused images from different methods on the WorldView-4 dataset. (a) LR MS image; (b) PAN image; (c) MTF-GLP-CBD; (d) GS; (e) LRP; (f) PanNet; (g) PNN; (h) TFNet; (i) PSGAN; (j) M-GAN; (k) GPPNN; (l) Proposed UTSN.

TABLE VI. QUANTITATIVE EVALUATIONS ON WORLDVIEW-4 DATASET

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
D_λ	0.1036	0.1090	0.1156	0.1020	0.1017	0.0725	0.0748	0.0774	0.0931	**0.0715**
D_S	0.1312	0.1280	0.1337	0.1643	0.0872	0.0829	0.1066	**0.0657**	0.0774	0.0911
QNR	0.7793	0.7774	0.7668	0.7532	0.8205	0.8511	0.8275	**0.8623**	0.8366	0.8446

The fusion results of the QuickBird dataset are reported in Fig. 11. In the source images used in Fig. 11, there are some vegetation and road areas. The results of MTF-GLP-CBD and GS in Figs. 11(c)-(d) show that the spectral features of vegetation areas are distorted. The results of TFNet and PSGAN have inferior performance on the color of road areas, which is unnatural and not consistent with that in Fig. 11(a). Besides, some spatial artifacts are introduced into the result of PSGAN, especially in the car areas. These artifacts may result from the misalignment in the feature domain. For the result of M-GAN in Fig. 11(j), a performance similar to Fig. 8(k) is found, and some spectral distortions arise. The fusion result of the proposed UTSN in Fig. 11(l) shows a better performance in terms of spatial details. Table VII reports the full-scale quantitative metrics. The proposed method behaves better in terms of D_S and QNR, while the value of D_λ is very close to the best value from PSGAN.

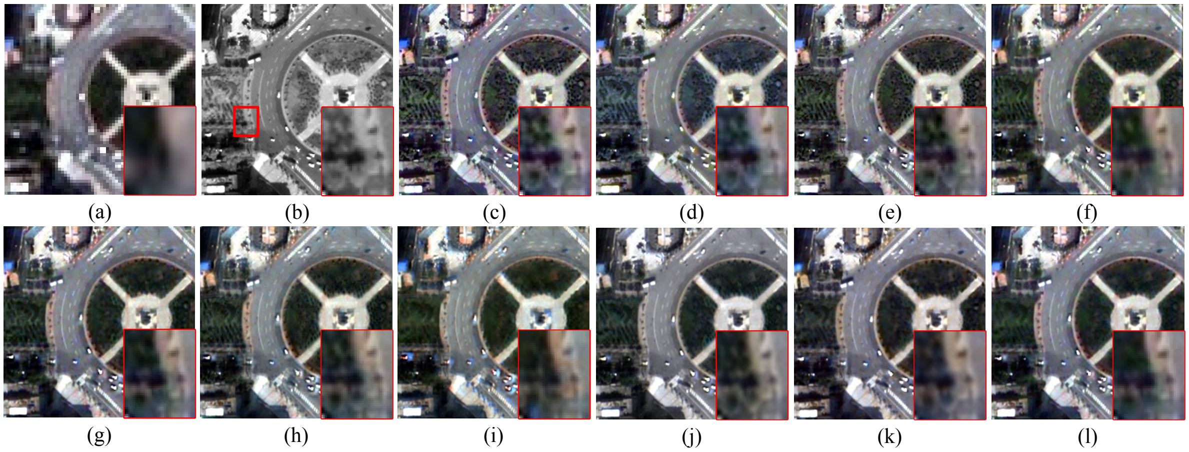


Fig. 11. Qualitative comparison of the fused images from different methods on the QuickBird dataset. (a) LR MS image; (b) PAN image; (c) MTF-GLP-CBD; (d) GS; (e) LRP; (f) PanNet; (g) PNN; (h) TFNet; (i) PSGAN; (j) M-GAN; (k) GPPNN; (l) Proposed UTSN.

TABLE VII. Quantitative Evaluations on QuickBird Dataset

Metric	MTF-GLP -CBD	GS	LRP	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
D_λ	0.0888	0.0842	0.1357	0.0632	0.0707	0.0636	**0.0599**	0.0856	0.0711	0.0686
D_S	0.0597	0.0996	0.1568	0.0441	0.0385	0.0714	0.0911	0.0541	0.0556	**0.0363**
QNR	0.8593	0.8267	0.7317	0.8964	0.8950	0.8696	0.8547	0.8668	0.8790	**0.8975**

D. Visualization of Intermediate Results

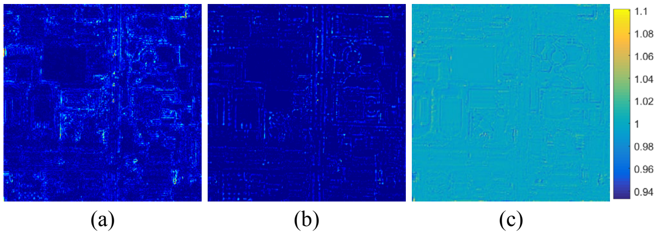


Fig. 12. Visualization of intermediate results. (a) Absolute error map of the result of SEN; (b) the absolute error map of the result of UTSN; (c) Spectral adjustment mask $\mathbf{S}_n^k$.

In this section, we use the images in Figs. 8(a) and 8(b) from the QuickBird dataset as an example to validate the effect of SAN. In Fig. 12(a), the absolute error map is derived from the difference between the output of SEN and the reference image in Fig. 8(c). Fig. 12(b) shows the absolute error map of the output of UTSN, which is exactly the difference map between the fused image in Fig. 8(k) and the reference image in Fig. 8(c). Compared with the error map in Fig.12(a), we can see that the reconstruction errors in Fig. 12(b) are smaller through the spectral adjustment of SAN. Moreover, the spectral adjustment mask $\mathbf{S}_n^k$ is also displayed in Fig. 12(c). We can see that the adjustment values in Fig. 12(c) are close to 1. Thus, the result of SEN is refined by the spectral adjustment mask from SAN.

E. Few-shot Case

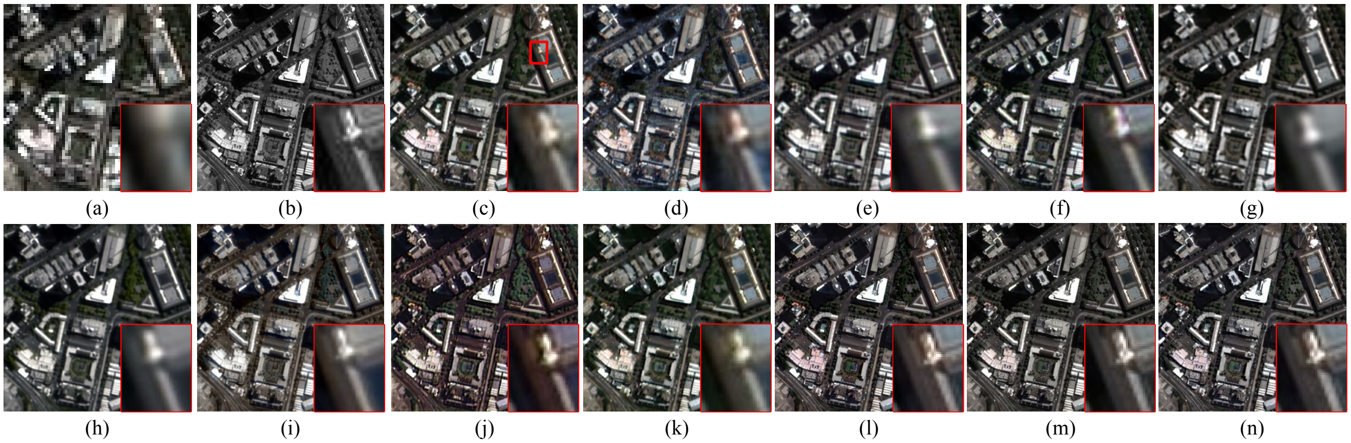
In this section, the experiments are conducted to validate the effectiveness of the proposed UTSN in few-shot cases. We use the Pléiades dataset[3] as the few-shot test dataset. The Pléiades dataset is captured from Melbourne, Australia. In the few-shot case, only few pairs of LRMS and PAN images from the Pléiades dataset are collected. Thus, the SEN trained on the other satellites listed in Table I is shared with the Pléiades satellite. Then, a new satellite-specific SAN is introduced after training on the few-shot dataset of the Pléiades satellite. SAN is light-weighted and can be trained well on the limited dataset. Specifically, we set the number of few-shot image pairs from the Pléiades dataset I as 1, 2, 10, and 15 for the validation of the proposed UTSN. Then, I pairs of LRMS and PAN images are employed to learn a specific SAN for the Pléiades satellite. For the test, the LRMS and PAN images from the Pléiades satellite are first fed into the SEN trained on the mixed dataset from the three existing satellites. The result of SEN is then refined by the SAN of the Pléiades satellite. Besides, for a more comprehensive comparison, we also train the DL-based methods mentioned above on the few-shot dataset of the Pléiades satellite. Here, 15 image pairs are used for the training of these DL-based methods.

The reduced-scale fusion results from the Pléiades dataset are shown in Fig. 13. From Fig. 13, we can find that DL-based

[3] https://earth.esa.int/eogateway/catalog/pleiades-esa-archive

methods trained on the limited dataset cannot fuse LR MS and PAN images well. Their results suffer from some blurring effects and spectral losses. For the proposed UTSN, we only train the satellite-specific SAN in the few-shot case using *I* pairs of LRMS and PAN images from the Pléiades satellite. One can see that the proposed UTSN can produce comparable fusion results with only one shot (*I*=1) when compared with other DL-based methods. Moreover, with more training images, the proposed UTSN can produce more abundant spatial details in the fused images. Table VIII provides the objective evaluations, which are consistent with the analysis in Fig. 13. Although DL-based methods trained on 15 image pairs can produce plausible results, they cannot be trained sufficiently, which limits their performance. Because of the lightweight model, SAN is trained better and produces better fusion results.

Fig. 14 illustrates the fusion results of all methods on the full-scale Pléiades dataset. From Fig. 14(i), it can be found that serious spectral distortions occur when LR MS and PAN images are directly fused by SEN trained on the mixed dataset from other satellites. This is due to the differences between the spectral responses of different satellites. However, it should be noted that SEN can enhance the spatial details of LR MS images efficiently although the data from the Pléiades satellite dataset is not used for the training of SEN. Therefore, SEN can deal with images from different spatial resolutions. After training a new SAN on the few-shot dataset from the Pléiades satellite, the spectral information in Fig. 14(i) is adjusted to the characteristics of the Pléiades satellite. The results in Table IX show that the QNR value of the proposed UTSN increases by increasing the number of training images. The spectral metric decreases first and then increases by increasing *I*.

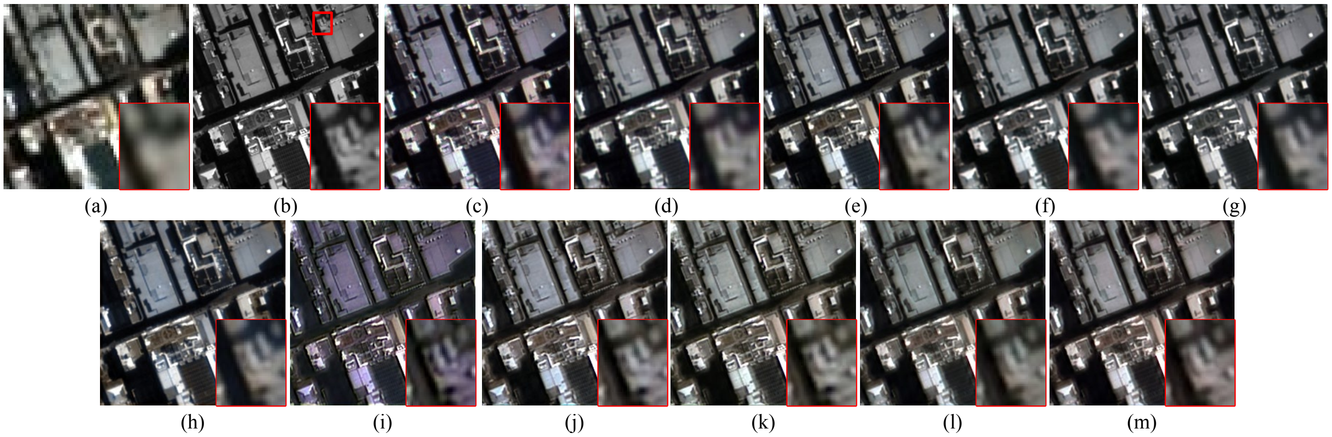

(a) (b) (c) (d) (e) (f) (g)

(h) (i) (j) (k) (l) (m) (n)

Fig. 13. Qualitative comparison of the fused images on the few-shot Pléiades dataset. (a) LR MS image; (b) PAN image; (c) Reference image; (d) PanNet; (e) PNN; (f) TFNet; (g) PSGAN; (h) M-GAN; (i) GPPNN; (j) SEN; (k) *I*=1; (l) *I*=2; (m) *I*=10; (n) *I*=15.

TABLE VIII. QUANTITATIVE EVALUATIONS IN THE FEW-SHOT CASE ON THE PLÉIADES DATASET

Metric	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	SEN	*I*=1	*I*=2	*I*=10	*I*=15
Q4	0.8783	0.9138	0.9110	0.9178	0.9144	0.9107	0.8484	0.9186	0.9194	0.9177	**0.9198**
SAM	6.4638	5.0016	4.6051	4.5950	4.7643	5.6569	8.3811	4.4392	**4.3567**	4.6836	4.5481
UIQI	0.9667	0.9647	0.9734	0.9734	0.9761	0.9700	0.8958	0.9755	0.9742	0.9753	**0.9769**
RMSE	42.8475	44.4884	36.8599	37.1269	36.8489	39.6629	80.8292	35.8874	36.8174	35.9082	**34.5568**
ERGAS	1.9259	1.9010	1.5618	1.6379	1.6037	1.6881	3.4995	1.5421	1.5756	1.5486	**1.4864**


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

(h) (i) (j) (k) (l) (m)

Fig. 14. Qualitative comparison of the fused images on the few-shot Pléiades dataset. (a) LR MS image; (b) PAN image; (c) PanNet; (d) PNN; (e) TFNet; (f) PSGAN; (g) M-GAN; (h) GPPNN; (i) SEN; (j) *I*=1; (k) *I*=2; (l) *I*=10; (m) *I*=15.

TABLE IX. QUANTITATIVE EVALUATIONS IN THE FEW-SHOT CASE ON PLÉIADES DATASET

Metric	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	SEN	*I*=1	*I*=2	*I*=10	*I*=15
D_λ	0.0453	0.0473	0.0624	0.0525	**0.0119**	0.0260	0.0674	0.0252	0.0243	0.0239	0.0277
D_S	0.0443	0.0321	0.0149	**0.0114**	0.0547	0.0561	0.0575	0.0355	0.0358	0.0357	0.0308
QNR	0.9125	0.9221	0.9236	0.9367	0.9340	0.9194	0.8790	0.9402	0.9408	0.9412	**0.9423**

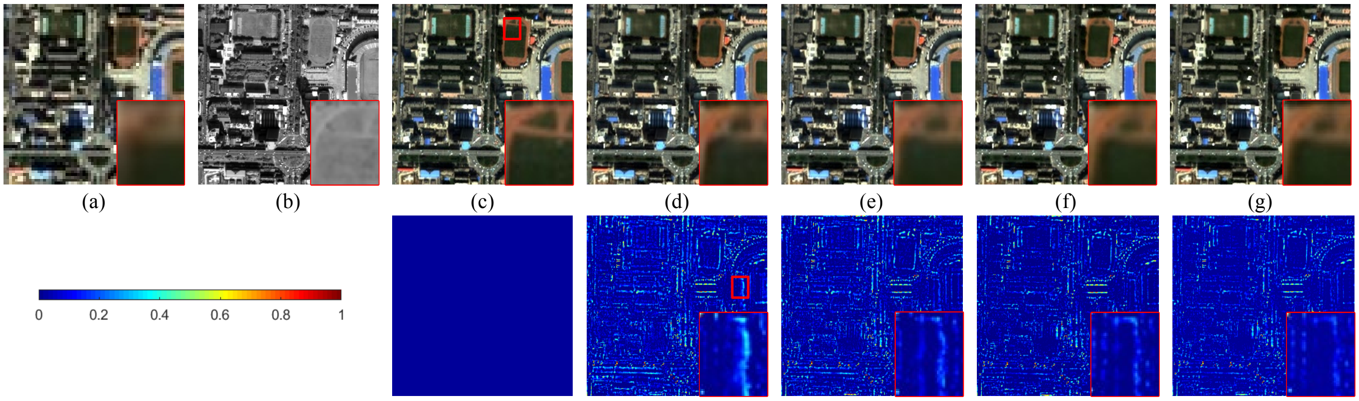


Fig. 15. Ablation study of the contribution of different modules. (a) LR MS image; (b) PAN image; (c) Reference image; (d) w/o dense connections; (e) w/o injection connections; (f) w/o channel attention module; (g) complete UTSN.

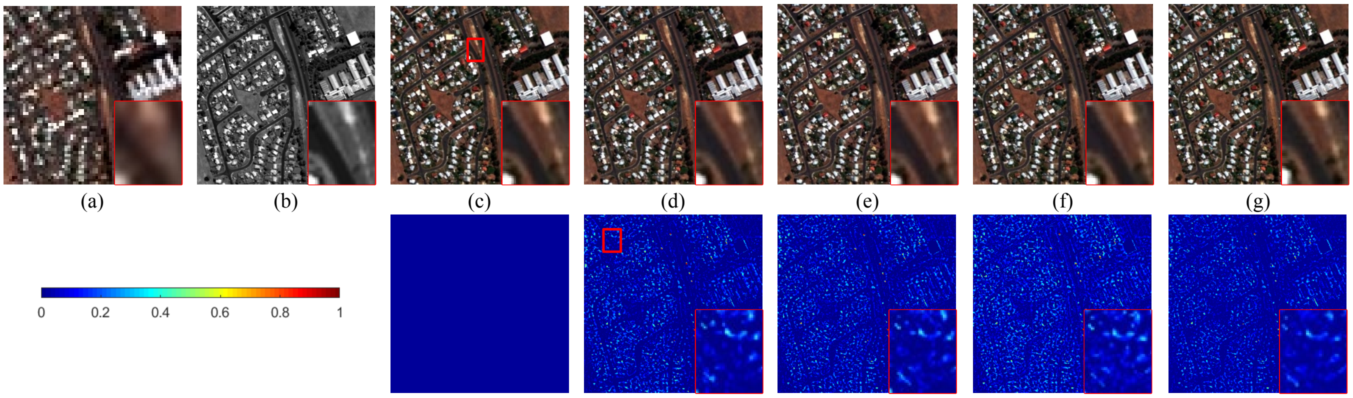


Fig. 16. Influence of the number of basic convolution blocks in SEN. (a) LR MS image; (b) PAN image; (c) Reference image; (d) *J*=3; (e) *J*=4; (f) *J*=5; (g) *J*=6 (UTSN).

F. Ablation Study

In this section, the ablation study is carried out on the QuickBird dataset to assess the effectiveness of different modules in the proposed UTSN. Specifically, we remove dense connections, injection connections, and the channel attention module from the proposed UTSN. Fig. 15 displays the fusion results of different configurations. The absolute error maps between the reference image and the fused images are also given in the second row of Fig. 15. From the figure, we can see that the ablation results are similar. However, we can observe that the reconstruction errors are smaller from the second row of Fig. 15 when dense connections, injection connections, and the channel attention module are introduced into the proposed UTSN. The numerical differences are given in Table X. When dense or injection connections are removed, the spatial details in the fused image cannot be enhanced well. Without the channel attention module, the spectral information in the fused image is degraded. The complete UTSN improves the quantitative performance is improved with the introduction of different modules, which verify the effectiveness of these modules.

TABLE X. QUANTITATIVE EVALUATIONS OF ABLATION STUDY ON QUICKBIRD DATASET

Metric	w/o dense connections	w/o injection connections	w/o channel attention module	complete UTSN
Q4	0.9474	0.9474	0.9468	**0.9496**
SAM	1.9952	1.8187	1.8058	**1.6823**
UIQI	0.9847	0.9857	0.9856	**0.9882**
RMSE	16.0677	15.4793	15.3956	**15.1373**
ERGAS	0.7420	0.7185	0.7148	**0.7041**

G. Investigation on Network Architecture

In this section, we analyze the influence on the fusion results of the number of basic convolution blocks *J* in SEN. Fig. 16 illustrates the fusion results and absolute error maps of different blocks on the GeoEye-1 dataset. One can see that by increasing *J*, the fusion results in Fig. 16 become clearer and the spectral information is better preserved. The reconstruction errors are also reduced because a deeper network means better capabilities of nonlinear approximation. Table XI shows that the quantitative metrics become better with larger *J*. Although the proposed UTSN with more basic convolution blocks produces better fusion results, the increasing parameters and computational complexity cannot be ignored. Thus, we use SEN with 6 basic convolution blocks to achieve the fusion of LRMS and PAN images.

TABLE XI. INFLUENCES OF THE NUMBER OF BASIC CONVOLUTION BLOCKS ON GEOEYE-1 DATASET

Metric	*J*=3	*J*=4	*J*=5	*J*=6
Q4	0.8228	0.8264	0.8270	**0.8274**
SAM	3.2252	2.8440	2.8348	**2.8076**
UIQI	0.9795	0.9830	0.9832	**0.9837**
RMSE	16.3050	14.6045	14.5208	**14.2591**
ERGAS	1.0581	0.9485	0.9433	**0.9274**

H. Model Size and Complexity

Table XI reports the model size of all DL-based pan-sharpening methods mentioned above. It can be observed that the model sizes of PanNet and PNN are smaller than that of the proposed UTSN. However, the pan-sharpening performance of the proposed UTSN is better than those of the PanNet and the PNN. Compared with other methods, the TFNet and the PSGAN contain more parameters to be trained. Sufficient data have to be collected for the training of these models. For the proposed UTSN, the model is composed of SEN and SAN. However, most of the parameters of UTSN are from SEN which is trained on the mixed dataset from all satellites. Only the SAN parameters are learned from the satellite-specific dataset. From Fig. 5, we can find that the structure of SAN is simple and only two convolution layers and two FC layers contain learnable parameters. Only 3412 parameters are induced in SAN. Because of the small model size, the training of SAM can be achieved on the few-shot datasets, which improves the flexibility and generalization of the proposed UTSN.

Table XIII lists the training time and test time of all DL-based methods. From Table XIII, we can find that the training time of our proposed UTSN is longer than that of other methods. However, it should be noted that the proposed UTSN is trained on the integrated dataset including all datasets reported in Table I. For compared DL-based methods, they are trained three times independently on the three datasets in Table I. Therefore, the training time of our method is less than the total time of other DL-based methods, such as M-GAN and GPPNN, on the three datasets. Thanks to the formulation of UTSN, the datasets from different satellites can be integrated for training.

TABLE XII. MODEL SIZES OF DL-BASED PAN-SHARPENING METHODS.

Method	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN	
							SEN	**SAN**
#Para.	0.15M	0.08M	2.36M	2.62M	15M	0.12M	0.25M	3412

TABLE XIII. TRAINING TIME AND TEST TIME OF DL-BASED PAN-SHARPENING METHODS.

Dataset	Time	PanNet	PNN	TFNet	PSGAN	M-GAN	GPPNN	Proposed UTSN
QuickBird	**Training (h)**	23.58	13.87	14.15	30.99	33.85	38.49	96.43
GeoEye-1	**Training (h)**	28.37	13.92	14.76	31.63	34.11	36.85	
WorldView-4	**Training (h)**	26.07	13.94	14.75	31.89	33.82	28.15	
QuickBird	**Test (s)**	1.31	0.35	0.87	0.66	0.77	0.98	1.03
GeoEye-1	**Test (s)**	1.42	0.44	0.88	0.72	0.82	1.01	1.11
WorldView-4	**Test (s)**	1.84	0.52	0.89	0.78	0.81	1.12	1.09

V. CONCLUSIONS

Deep neural networks can boost the pan-sharpening performance efficiently. However, for obtaining satisfying results, they have to be trained on satellite-specific datasets repetitively, which requires a large amount of data. In this paper, we have proposed a unified framework to leverage the data from different satellites collaboratively for pan-sharpening. Specifically, the data from different satellites are employed to train a shared SEN, which is designed to enrich the spatial details of the fused image. Then, a specific SAN is trained for each satellite to capture the spectral information in MS images. Equipped by the shared SEN and satellite-specific SAN, the proposed UTSN produces better fusion results than other DL-based methods. Moreover, the experimental results also demonstrate that in the few-shot case, the proposed UTSN can adapt to a new satellite with only a few, (e.g., 15), pairs of LRMS and PAN images. In the proposed work, we only consider the satellites collecting 4-band MS images. The proposed UTSN cannot be trained on an integrated dataset including 4-band MS images and 8-band MS images because the numbers of bands in MS images are not equal. For example, WorldView-2 and WorldView-4 provide 8-band MS images and 4-band MS images, respectively. So, the datasets from the two satellites cannot be combined into a mixed dataset. For future work, we will consider more efficient networks to integrate 4-band MS images and 8-band MS images for training.

REFERENCES

[1] L. Bergamasco, F. Bovolo, L. Bruzzone, “A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2147-2162, 2023.

[2] Y. Chen, L. Bruzzone, “A self-supervised approach to pixel-level change detection in bi-temporal RS images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 4413911, 2022.

[3] A. Gastineau, J. Aujol, Y. Berthoumieu, C. Germain, “Generative adversarial network for pansharpening with spectral and spatial discriminators,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 4401611, 2022.

[4] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, "Deep gradient projection networks for pan-sharpening," in *Proc. IEEE CVPR*, Nashville, TN, USA, Jun. 2021, pp. 1366-1375.

[5] 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.

[6] 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.

[7] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens*., vol. 63, no. 6, pp. 691-699, Jun. 1997.

[8] K. Zhang, F. Zhang, W. Wan *et al*., "Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead," *Inf. Fusion*, vol. 93, pp. 227-242, 2023.

[9] L. Deng, G. Vivone, M. Paoletti *et al*., "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE. Geosci. Remote Sens. Mag*., vol. 10, no. 3, pp. 279-315, Sept. 2022.

[10] T. Tu, P. Huang, C. Hung, and C. Chang, "A fast intensity hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett*., vol. 1, no. 4, pp. 309-312, Oct. 2004.

[11] J. Duran, A. Buades, Restoration of pansharpened images by conditional filtering in the PCA domain, *IEEE Geosci. Remote Sens. Lett*., vol. 16, no. 3, pp. 442-446, Mar. 2019.

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

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

[14] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens*., vol. 46, no. 5, pp. 1323-1335, May 2008.

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

[16] R. Restaino, G. Vivone, P. Addesso, and J. Chanussot, "A pan-sharpening approach based on multiple linear regression estimation of injection coefficients," *IEEE Geosci. Remote Sens. Lett*., vol. 17, no. 1, pp. 102-106, Jan. 2020.

[17] G. Vivone, S. Marano, and J. Chanussot, "Pan-sharpening: Context-based generalized Laplacian pyramids by robust regression," *IEEE Trans. Geosci. Remote Sens*., vol. 58, no. 9, pp. 6152-6167, Sept. 2020.

[18] H. Aanæs, J. Sveinsson, A. Nielsen, T. Bøvith, and J. Benediktsson, "Model-based satellite image fusion," *IEEE Trans. Geosci. Remote Sens*., vol. 46, no. 5, pp. 1336-1346, May 2008.

[19] 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.

[20] 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.

[21] J. Wright, A. Ganesh, S. Rao, Y. Peng, Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. 22nd NIPS*, 2009, pp. 2080-2088.

[22] F. Zhang, H. Zhang, K. Zhang, and Y. Xing *et al*., "Exploiting low-rank and sparse properties in strided convolution matrix for pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 14, pp. 2649-2661, 2021.

[23] 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.

[24] P. Liu, L. Xiao, "Multicomponent driven consistency priors for simultaneous decomposition and pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens*., vol. 12, no. 11, pp. 4589-4605, Nov. 2019.

[25] K. Zhang, A. Wang, F. Zhang, W. Diao, J. Sun, and L. Bruzzone, "Spatial and spectral extraction network with adaptive feature fusion for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, 2022, Art. no. 5410814.

[26] G. Masi, D. Cozzolino, L. Verdoliva and G. Scarpa, "Pan-sharpening by convolutional neural networks," *Remote Sens*., vol. 8, pp. 594, 2016.

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

[28] Y. Wei, Q. Yuan, H. Shen, L. Zhang, "Boosting the accuracy of multispectral image pan-sharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett*., vol. 14, no. 10, pp. 1795-1799, Oct. 2017.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, Jun. 2016, pp. 770-778.

[30] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, G. Cheng "Pan-sharpening via deep locally linear embedding residual network," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5409413, 2022.

[31] Z. Shao, Z. Lu, M. Ran, L. Fang *et al*., "Residual encoder-decoder conditional generative adversarial network for pan-sharpening," *IEEE Geosci. Remote Sens. Lett*., vol. 17, no. 9, pp. 1573-1577, Sept. 2020.

[32] I. Goodfellow *et al*., "Generative adversarial nets", In *Proc. NIPS*, Montréal, Dec. 2014, Canada, pp. 2672-2680.

[33] 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.

[34] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, L. Bruzzone, "ZeRGAN: Zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Neural Netw. Learn. Syst*., early access, 2022, doi: 10.1109/TNNLS.2021.3137373.

[35] L. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pan-sharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 8, pp. 6995-7010, Aug. 2021.

[36] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, J. Ma, "Pan-sharpening via high-pass modification convolutional neural network," in *Proc. IEEE ICIP*, Sept. 2021, pp. 1714-1718.

[37] J. Cai and B. Huang, "Super-resolution-guided progressive pan-sharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 6, pp. 5206-5220, Jun. 2021.

[38] L. Zhang, J. Zhang, J. Ma, and X. Jia, "SC-PNN: Saliency cascade convolutional neural network for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 11, pp. 9697-9715, Nov. 2021.

[39] Y. Qu, R. Baghbaderani, H. Qi, C. Kwan, "Unsupervised pan-sharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 4, pp. 3192-3208, Apr. 2021.

[40] J. Peng, L. Liu, J. Wang, and E. Zhang *et al*., "PSMD-Net: A novel pan-sharpening method based on a multiscale dense network," *IEEE Trans. Geosci. Remote Sens*., vol. 59, no. 6, pp. 4957-4971, Jun. 2021.

[41] K. Zhang, Z. Li, F. Zhang, W. Wan, J. Sun, "Pan-sharpening based on transformer with redundancy reduction," *IEEE Geosci. Remote Sens. Lett*., vol. 19, pp. 5513205, 2022.

[42] X. Meng, N. Wang, F. Shao, S. Li, "Vision transformer for pansharpening," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 1-11, 2022.

[43] F. Zhang, K. Zhang, and J. Sun, "Multiscale spatial-spectral interaction transformer for pan-sharpening," *Remote Sens*., vol. 14, no. 7, p. 1736, Apr. 2022.

[44] W. Wang, Y. Chen, P. Ghamisi, "Transferring CNN with adaptive learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5533918, 2022.

[45] M. Ma, W. Ma, L. Jiao *et al*., , "Transfer representation learning meets multimodal fusion classification for remote sensing images," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5632415, 2022.

[46] H. Lang, G. Yang, C. Li, J. Xu, "Multisource heterogeneous transfer learning via feature augmentation for ship classification in SAR imagery," *IEEE Trans. Geosci. Remote Sens*., vol. 60, pp. 5228814, 2022.

[47] H. Wang, X. Wang, Y. Cheng, "Graph meta transfer network for heterogeneous few-shot hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens*., vol. 61, pp. 5501112, 2023.

[48] 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.

[49] 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.

[50] Y. Li, X. Fu, Z. Zha. Cross-patch graph convolutional network for image denoising. In *Proc. IEEE ICCV*, pp. 4651-4660, 2021.

[51] S. Zhou, J. Zhang, W. Zuo, C. Loy. Cross-scale internal graph neural network for image super-resolution. In *Proc. 34th NeurIPS*, 2021, pp. 1-11.

[52] B. Arad, R. Timofte, R. Yahel, *et al.*, "NTIRE 2022 spectral recovery challenge and data set," in *IEEE CVPRW*, 2022, pp. 863-881.

[53] D. Lei, P. Chen, L. Zhang, W. Li, "MCANet: A multidimensional channel attention residual neural network for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5411916, 2022.

[54] L. Alparone *et al.*, "Comparison of pan-sharpening algorithms: Outcome of the 2006 GRS-S data fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

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

[56] 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.

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

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

[59] 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.

[60] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015, pp. 1-11.