ZeRGAN: Zero-Reference GAN for Fusion of Multispectral and Panchromatic Images

Wenxiu Diao, Feng Zhang, Jiande Sun, Yinghui Xing, Kai Zhang, and Lorenzo Bruzzone, *Fellow, IEEE*

***Abstract*— In this article, we present a new pansharpening method, a zero-reference generative adversarial network (ZeRGAN), which fuses low spatial resolution multispectral (LR MS) and high spatial resolution panchromatic (PAN) images. In the proposed method, zero-reference indicates that it does not require paired reduced-scale images or unpaired full-scale images for training. To obtain accurate fusion results, we establish an adversarial game between a set of multiscale generators and their corresponding discriminators. Through multiscale generators, the fused high spatial resolution MS (HR MS) images are progressively produced from LR MS and PAN images, while the discriminators aim to distinguish the differences of spatial information between the HR MS images and the PAN images. In other words, the HR MS images are generated from LR MS and PAN images after the optimization of ZeRGAN. Furthermore, we construct a nonreference loss function, including an adversarial loss, spatial and spectral reconstruction losses, a spatial enhancement loss, and an average constancy loss. Through the minimization of the total loss, the spatial details in the HR MS images can be enhanced efficiently. Extensive experiments are implemented on datasets acquired by different satellites. The results demonstrate that the effectiveness of the proposed method compared with the state-of-the-art methods. The source code is publicly available at https://github.com/RSMagneto/ZeRGAN.**

***Index Terms*— Generative adversarial network (GAN), image fusion, multispectral image, panchromatic (PAN) image, zero-reference training.**

Manuscript received 21 January 2021; revised 13 June 2021 and 22 September 2021; accepted 18 December 2021. Date of publication 4 January 2022; date of current version 30 October 2023. This work was supported in part by the Natural Science Foundation of China under Grant 61901246; in part by the China Postdoctoral Science Foundation under Grant 2019TQ0190 and Grant 2019M662432; in part by the Open Fund of National Engineering Laboratory for Integrated Aerospace-Ground-Ocean Big Data Application Technology, China, under Grant 20200208; in part by the China Scholarship Council under Grant 202008370035; and in part by the Natural Science Foundation for Distinguished Young Scholars of Shandong Province under Grant JQ201718. *(Corresponding authors: Jiande Sun; Kai Zhang.)*

Wenxiu Diao, Feng Zhang, and Jiande Sun are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (e-mail: diaowx0920@163.com; fengzhangpl@163.com; jiandesun@hotmail.com).

Yinghui Xing is with the National Engineering Laboratory for Integrated Aerospace-Ground-Ocean Big Data Application Technology and the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xyh_7491@nwpu.edu.cn).

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

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

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2021.3137373.

Digital Object Identifier 10.1109/TNNLS.2021.3137373

I. Introduction

In RECENT years, a large number of remote sensing images have been collected by different Earth observation satellites, such as QuickBird, GeoEye-1, and WorldView-2. These satellites can acquire low spatial resolution multispectral (LR MS) images and panchromatic (PAN) images simultaneously. At present, the captured images have been extensively and successfully used in target discovery [1], land-cover analysis [2], and environmental monitoring [3]. However, it is difficult to obtain high spatial and spectral resolution MS (HR MS) images for the abovementioned satellites due to their inherent tradeoff between spatial and spectral resolutions [4]. On the one hand, LR MS images contain abundant spectral information but a lower spatial resolution than PAN images. On the other hand, PAN images are made up of only one high spatial resolution band. Therefore, image fusion, also called pansharpening [5], is applied to produce the fused HR MS images, by integrating the spectral information in LR MS images with the spatial details in PAN images together.

In the last two decades, various algorithms have been proposed and developed to cope with the pansharpening task. They can be classified into four types: 1) component substitution (CS) methods; 2) multiresolution analysis (MRA) methods; 3) degradation model (DM)-based methods; and 4) deep neural network (DNN)-based methods. Due to the fast implementation and simple principles, CS methods have been widely used. They project the interpolated LR MS images into a new domain to estimate suitable spatial components. PAN images are then employed to substitute the spatial components, and the fused HR MS images are generated by the corresponding inverse projection. For example, typical transformations that have been used in this context are intensity–hue–saturation (IHS) transformation [6], principal component analysis (PCA) [7], and Gram–Schmidt (GS) transformation [8]. Besides, the band-dependent spatial detail (BDSD) algorithm [9] was proposed to estimate the gain parameters more accurately. However, significant spectral distortions are produced in the fusion results because a global transformation is considered among images.

MRA-based methods assume that the spatial information to be added into the LR MS images is acquired from the PAN image, which is referred to as Amélioration de la Résolution Spatiale par Injection de Structures (ARSIS) [10]. In these methods, the extraction of spatial details and calculation of injection gains have significant impacts on the fusion results. The spatial details are extracted by many MRA tools, such as contourlet [11] and generalized Laplacian pyramid [12].

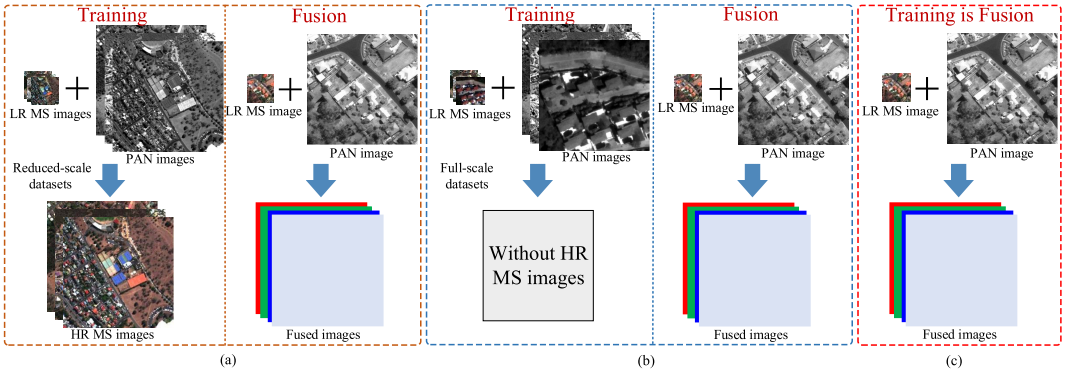

Fig. 1. Training and fusion settings in different pansharpening methods based on DNN. (a) Paired training. (b) Unpaired training. (c) Zero-reference fusion. The previous DNN-based pansharpening methods require a large-scale dataset. By comparing them, our model is directly trained for the test LR MS and PAN images.

Otazu *et al.* [13] proposed an additive wavelet luminance proportional (AWLP) method to estimate the high-frequency components in PAN images. Because only spatial details from PAN images are injected into LR MS images, the fusion results of MRA-based methods have a good performance in terms of spectral fidelity.

DM-based methods assume that the observed LR MS and PAN images are the degradation versions of HR MS images in spatial and spectral domains, respectively. For example, Li and Yang [14] reformulated the image fusion task as a compressed sensing problem [15] by regarding the spatial and spectral DMs as the measurement matrix. Moreover, the fusion model is regularized by other efficient priors, such as sparsity [16], nonnegativity [17], and low-rank prior [18]. Although these methods perform well in terms of spatial and spectral information preservation, their computational complexity is much higher than that of the former two types of methods.

Nowadays, DNNs have achieved great success in various fields [19]–[21] and have also been used in pansharpening. For example, Huang *et al.* [22] employed a stacked modified sparse denoising autoencoder for pansharpening. Masi *et al.* [23] presented a new pansharpening method based on a convolutional neural network (CNN), which is termed PNN and inspired by the superresolution model in [24]. In [25], PanNet was built by incorporating problem-specific priors with a residual network (ResNet) [26]. It can achieve better spectral and spatial preservations in the fused images. Subsequently, Fu *et al.* [27] introduced a grouped multiscale dilated network to improve the multiscale representation capability of spatial information. Zhang *et al.* [28] proposed a bidirectional pyramid network to inject the spatial details derived from PAN images into LR MS images level by level. In [29], a stacked sparse autoencoder was constructed on the grouped patches. According to the geometric structures of these patches, they are categorized and then fed into the autoencoder. Besides, generative adversarial networks (GANs) [30] have also been used to fuse LR MS and PAN images. For instance, Liu *et al.* [31] first utilized GAN to produce the fused image, which was then extended in [32]. Then, Ma *et al.* [33] adopted two discriminators to maintain the spatial and spectral information in the fused image, which can avoid the need for HR MS images during the training. Moreover, the bidiscriminator framework was utilized in MDSSC-GAN [34]. In this method, the first discriminator is fed by the luminance and the near-infrared band of images, while the input of the second discriminator is the concatenation of spectral components. In [35], an unsupervised GAN-based method was proposed, which includes supervised pretraining and unsupervised fine-tuning. However, pansharpening methods based on DNN may be prone to overfit the paired training data. Thus, the generalization capability is reduced when analyzing new data obtained by other satellites. In addition, two issues related to training data need to be considered. On the one hand, most of the pansharpening methods based on DNN usually use paired images for training, as shown in Fig. 1(a). However, HR MS images are not available in real scenarios. Accordingly, the paired images are composed of the reduced-scale versions of LR MS and PAN images after spatial degradation. Thus, the original LR MS images are directly viewed as reference data. However, spatial details in the full-scale images cannot be learned efficiently from the reduced-scale image pairs. On the other hand, some GAN-based pansharpening methods [33] are proposed for the explicit training on full-scale images, which contains LR MS and PAN images at the original scale. They are also named unpaired images, as displayed in Fig. 1(b). However, these networks need repetitive training to obtain desirable results because of the diverse distributions among the images from different satellites. Moreover, they require a large amount of training data.

Considering the two aspects mentioned above, we proposed a novel method based on GAN, zero-reference GAN (ZeRGAN), for the sharpening of LR MS images by PAN images. As shown in Fig. 1(c), ZeRGAN does not require any paired reduced-scale images or unpaired full-scale images for training. Thus, the fused images can be directly obtained by the multiscale generators after the optimization of the loss function of the proposed method is completed. In particular, we adopt a set of cascaded multiscale generators to progressively increase the spatial information in MS images while preserving spectral information. At each scale, residual learning is embedded into the generator to improve the spatial details in the intermediate HR MS images. Meanwhile, the corresponding discriminator at the same scale is employed to further distinguish the spatial information in the intermediate and real PAN images. Through spectral response filtering (SSF), the intermediate PAN images are generated from the intermediate HR MS images. In addition, to ensure the fusion performance of zero-reference training, we designed an unsupervised loss function containing

an adversarial loss, spatial and spectral reconstruction losses, a spatial enhancement loss, and an average constancy loss for the optimization of the generator and discriminator at each scale. The experimental results show that, even without any training set, ZeRGAN still has a competitive performance compared with the methods that rely on paired or unpaired images for training.

The main contributions of this article are summarized as follows.

1) ZeRGAN does not require any training data. The fusion of LR MS and PAN images is achieved by optimizing GAN and combining spatial and spectral DMs. A multiscale generator architecture is utilized to enhance the spatial details in LR MS images collaboratively.

2) A task-driven nonreference loss function is formulated for an efficient measure of spatial and spectral information in the fused image, which mitigates the need for a large amount of training data.

3) For spectral preservation, we introduce a new loss term, the average constancy loss, which assumes that the averages of the bands in LR MS images should be correspondingly equal to those of the bands in HR MS images.

The remainder of this article is organized as follows. Section II briefly introduces GAN. The proposed ZeRGAN method is described in detail in Section III. In Section IV, extensive experiments are implemented, and comparisons are presented. Conclusions are given in Section V.

II. Generative Adversarial Networks

Since GAN was proposed by Goodfellow *et al.* [30], a dramatic performance increase in various fields, such as image manipulation [36]–[39] and image synthesis [40], has been shown due to its powerful generative capacity. GAN mainly learns a generator G and a discriminator D through a min–max adversarial game. The generator G can learn the data distribution and create realistic samples to fool the discriminator D. On the contrary, the discriminator D aims to classify whether the samples are synthesized by the generator G or from real data. Then, the above two-player game can be mathematically formulated as

$$\min_G \max_D \; V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\big[\log D(x)\big] + \mathbb{E}_{z\sim p_z(z)}\big[\log(1-D(G(z)))\big] \quad (1)$$

where $p_{\text{data}}(x)$ is the distribution of the real sample x and $p_z(z)$ is the distribution of the sample z from latent space. When optimizing (1), the parameters of G (or D) are updated alternatively by fixing the parameters of D (or G). After training, the samples generated by G are difficult to be distinguished by D.

However, the original GAN suffers from training instability. Then, deep convolutional GAN (DCGAN) [41] was proposed to stabilize the training of GAN, in which both the generator and discriminator are made up of CNN. Mao *et al.* [42] penalized the distribution of the fake samples to be closer to that of the real data under the constraint of the least-squares loss. Wasserstein GAN (WGAN) [43] adopted the Wasserstein distance that has better theoretical properties to measure the discrepancy between distributions of real and fake data. However, the convergence of WGAN is slow and sometimes unstable. Thus, Gulrajani *et al.* [44] proposed WGAN-GP loss and introduced gradient penalty to guarantee the Lipschitz condition directly, whose objective function is

$$L_{\text{WGAN-GP}} = \mathbb{E}_{z\sim p_z(z)}[D(G(z))] - \mathbb{E}_{x\sim p_{\text{data}}(x)}[D(x)] + \lambda\mathbb{E}_{\hat{x}\sim p_{\hat{x}}(\hat{x})}\big[(\|\nabla_{\hat{x}}D(\hat{x})\|_2-1)^2\big] \quad (2)$$

where $p_{\hat{x}}(\hat{x})$ is sampled uniformly between the sample pairs from real and fake data. ∇ is the gradient operator. λ is a regularization parameter. Taking the training stability into consideration, the WGAN-GP loss is selected as the adversarial loss in our proposed method.

III. Proposed Method

The framework of ZeRGAN is presented in Fig. 2, in which generators and discriminators are mainly responsible for the injection and distinction of the realistic spatial information in HR MS images, respectively. More specifically, a series of generators are designed to enhance the spatial details of the intermediate HR MS images at different scales while preserving spectral information. Besides, the discriminator at each scale is further responsible for the consistency of spatial information in the intermediate PAN image and the real PAN image. The intermediate PAN image is produced from an intermediate HR MS image by SSF. In addition, real PAN images corresponding to different scales are synthesized by downsampling the original PAN image with different ratios. Moreover, an unsupervised loss is derived from both spatial–spectral models and priors to make the zero-reference training possible. Sections III-A–III-C present the details of the important components in ZeRGAN, i.e., the multiscale generator, the spatial discriminator, and the nonreference loss. Although many pansharpening methods based on GAN have been proposed, such as PSGAN [32] and MDSSC-GAN [34], these methods need a large number of paired images for training. The proposed ZeRGAN does not require any paired or unpaired data, which eliminates the issue of training data. Moreover, the existing GAN-based methods generally use a single generator to synthesize the fusion results. The multiscale framework consisting of multiple generators is proposed in ZeRGAN to produce the fused images from coarse to fine, in order to improve the fusion results.

A. Multiscale Generator

The multiscale generator is composed of L cascaded generators, in which the output of generator G_l at scale l is $\mathbf{H}_l \in R^{r_l M\times r_l N\times B}$, the input of the generator G_{l+1} at scale $l+1$. $M\times N\times B$ is the size of the original LR MS image $\mathbf{H}_0$. r_l is the spatial resolution ratio between $\mathbf{H}_0$ and $\mathbf{H}_l$. Thus, we can write the successive enhancement of the LR MS image at different scales as

$$\mathbf{H}_l = G_l(\mathbf{H}_{l-1}, \mathbf{P}_l) \quad (3)$$

where $\mathbf{H}_{l-1}$ and $\mathbf{P}_l$ are both fed into the generator G_l. $\mathbf{P}_l$ is generated from the original PAN image by downsampling. Note that the size of $\mathbf{P}_l$ is consistent with that of $\mathbf{H}_l$. Then, the desired HR MS image is $\mathbf{H}_L \in R^{r_L M\times r_L N\times B}$, the output of G_L at scale L. Generally, r_L is equal to 4 in the pansharpening task.

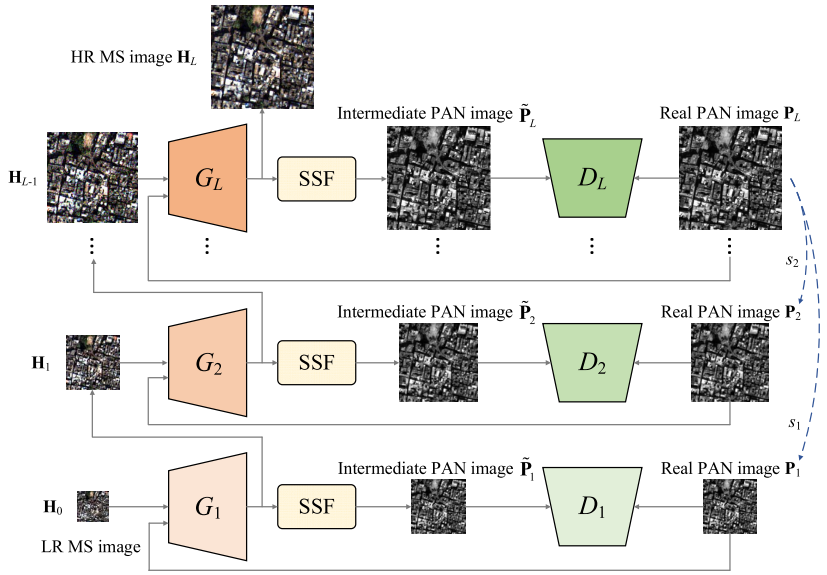

Fig. 2. Proposed ZeRGAN framework. Through multiscale generators $\{G_1, G_2, \ldots, G_L\}$, the LR MS image is interpolated scale by scale to the HR MS image with simultaneous spatial detail enhancement and spectral information preservation. Without direct analysis of spatial information in intermediate MS images, intermediate PAN images are acquired by SSF and then fed into discriminators $\{D_1, D_2, \ldots, D_L\}$, where the real PAN images are generated by downsampling the original PAN image with ratio s_l.

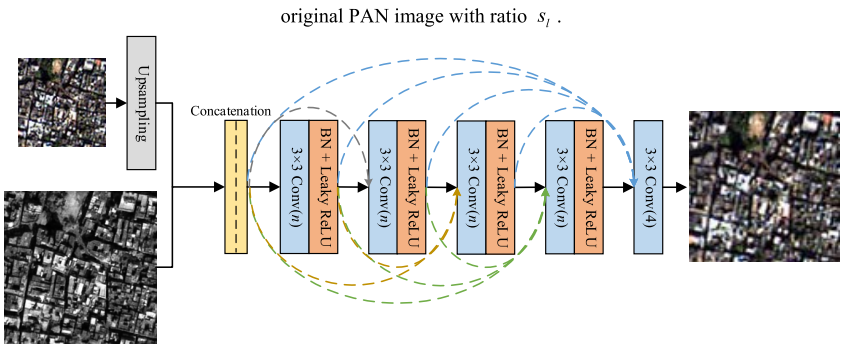

Fig. 3. Network architecture of the ZeRGAN generator. Conv(n): convolution layer containing n filter.

The architecture of generators at different scales is illustrated in Fig. 3. For the input of the generator, we first upsample directly the MS image $\mathbf{H}_{l-1}$ through the bicubic operator to the size of the real PAN image $\mathbf{P}_l$. Then, the upsampled MS image is concatenated with the PAN image together as the input of the generator. The generator is composed of five convolution layers. The filter size is 3×3 with stride 1. For the former four convolution layers, the number of filters is set as n. Four filters are used in the last convolution layer. The leaky ReLU activation function is applied in the former four convolution layers. Batch normalization (BN) is also concatenated to prevent vanishing gradient. To make full use of previous features, dense connections [45] are introduced into the generator. Through dense connections, the feature propagation in different layers can be strengthened so that the spatial details are injected into HR MS images efficiently.

B. Spatial Discriminator

In the architecture of ZeRGAN, discriminators $\{D_1, D_2, \ldots, D_L\}$ are designed to distinguish the spatial information in HR MS images indirectly due to the introduction of SSF. The SSF is modeled as

$$\tilde{\mathbf{P}}_l = \sum_{b=1}^{B} w_b \mathbf{H}_l^b \tag{4}$$

where $\mathbf{H}_l^b$ is the bth band of $\mathbf{H}_l$ and w_b is the fixed spectral response weight. The intermediate PAN image $\tilde{\mathbf{P}}_l$ is produced from the MS image $\mathbf{H}_{l-1}$ via SSF.

To efficiently capture the differences between distributions of the real and intermediate PAN images, a fully convolutional network is constructed, whose structure is illustrated in Fig. 4. The discriminator D_l consists of five convolution layers, where each of the first four layers contains n kernels with a size of 3×3. The last convolution layer contains only one filter with a size of 3×3. Besides, the first four convolution layers are connected with BN and leaky ReLU. A fully convolutional setting is used to efficiently model the spatial details in the images. All discriminators at all different scales share the same architecture in the proposed method.

C. Nonreference Loss Function

In ZeRGAN, the multiscale generators and discriminators are learned in sequence with an unsupervised training procedure. In the proposed method, the generated HR MS images are expected not only to fool discriminators but also to satisfy the degradation relationships with source images and other constraints. Thus, we cast additional losses on generators for efficient learning. The following losses are considered to train the proposed model.

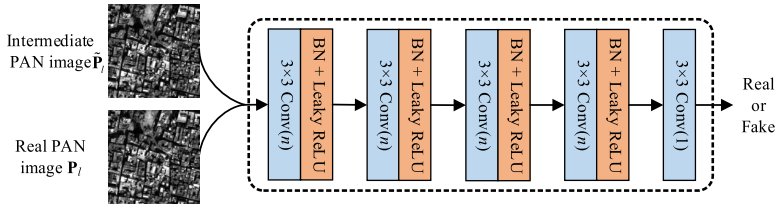

Fig. 4. Architecture of the discriminator.

Spatial and Spectral Reconstruction Losses: Generally, LR MS and PAN images are regarded as the spatial and spectral degradation results of the HR MS image, respectively. Concretely, the spatial and spectral observation models at scale l are defined as

$$\mathbf{H}_0 = \downarrow_{r_l} B(\mathbf{H}_l) + n_1 \tag{5}$$

$$\mathbf{P}_l = \sum_{b=1}^{B} w_b \mathbf{H}_l^b + n_2 \tag{6}$$

where $\downarrow_{r_l}$ stands for the downsampling operation with ratio r_l. $B(\cdot)$ is the blurring operation. The filter is bell-shaped and can be approximated by a Gaussian filter [46]. n_1 and n_2 are additive noises. Thus, the spatial and spectral degradation losses can be formulated as

$$L_{\text{sr}}^l = \alpha \left\| \mathbf{H}_0 - \downarrow_{r_l} B(\mathbf{H}_l) \right\|_F^2 + \beta \left\| \mathbf{P}_l - \sum_{b=1}^{B} w_b \mathbf{H}_l^b \right\|_F^2 \tag{7}$$

where α and β are the regularization parameters. The two constraints in (7) are a spatial fidelity term and a spectral fidelity term, respectively. Then, L_{sr}^l can preserve the spatial and spectral information with a tradeoff between the two terms, tuned by α and β values.

1) Spatial Enhancement Loss: For different bands in MS images, the edges or textures have obvious differences because of their spectral responses. The use of the same spatial enhancement strategy will result in artifacts on different bands. In the proposed method, the high-frequency information in $\mathbf{H}_l$ is supposed to follow the same spectral degradation relation as (6), which is similarly expressed as

$$\nabla \mathbf{P}_l = \sum_{b=1}^{B} w_b \nabla \mathbf{H}_l^b + n_3 \tag{8}$$

where the gradient operator ∇ is used for high-frequency information extraction. Then, the spatial enhancement can be achieved by

$$L_{\text{se}}^l = \delta \left\| \nabla \mathbf{P}_l - \sum_{b=1}^{B} w_b \nabla \mathbf{H}_l^b \right\|_F^2 \tag{9}$$

where δ is a regularization parameter. By minimizing the error in (9), spatial details in the HR MS can be further enhanced.

2) Average Constancy Loss: Inspired by color constancy loss in [47], we propose an average constancy loss to preserve the spectral information in HR MS images. It is assumed that the averages of the bands in LR MS images should be correspondingly equal to those of the bands in HR MS images. Through this assumption, the relations among bands of LR MS images can be inherited into those of HR MS images. Then, the average constancy loss is modeled as

$$L_{\text{mc}}^l = \gamma \sum_{b=1}^{B} \left(m_L^b - m_H^b \right)^2 \tag{10}$$

where m_L^b and m_H^b are the averages of the bth band of LR MS and HR MS images, respectively. γ is a regularization parameter.

3) Adversarial Loss: In the proposed method, an intermediate PAN image $\tilde{\mathbf{P}}_l$ is utilized to fool the discriminator D_l, which is generated from $\mathbf{H}_l$ by SSF. For training stability, WGAN-GP loss is considered in the proposed method

$$L_{\text{adv}}^l = \min_{G_l} \max_{D_l} \mathbb{E}\left[D_l(\mathcal{W}(G_l(\mathbf{H}_{l-1}))) \right] - \mathbb{E}[D_l(\mathbf{P}_l)] + \lambda \mathbb{E}\left[\left(\left\| \nabla D\left(\hat{\mathbf{P}}_l \right) \right\|_2 - 1 \right)^2 \right] \tag{11}$$

where $\mathcal{W}(\cdot)$ is the SSF operation for the synthesis of $\tilde{\mathbf{P}}_l$, and specifically, $\tilde{\mathbf{P}}_l = \sum_{b=1}^{B} w_b \mathbf{H}_l^b$. $\hat{\mathbf{P}}_l$ is synthesized by the real PAN image $\mathbf{P}_l$ and the intermediate PAN image $\tilde{\mathbf{P}}_l$ according to the formulation in WGAN-GP. Through the discriminator at each scale, the distribution of spatial information in $\tilde{\mathbf{P}}_l$ can be matched with that in $\mathbf{P}_l$.

All the abovementioned losses are incorporated together, and the total loss for the lth scale is summarized as

$$L_{\text{total}}^l = L_{\text{adv}}^l + L_{\text{sr}}^l + L_{\text{se}}^l + L_{\text{mc}}^l. \tag{12}$$

Finally, the fused images can be obtained from the output of the generator G_L after optimization of (12).

IV. Experiments

In this section, some state-of-the-art methods are used for qualitative and quantitative comparisons to validate the performance of the proposed method. The compared methods include BDSD [9], AWLP [13], PSGAN [32], SVT [48], VPLGC [49], PNN [23], PanNet [25], and MDSCC-GAN [34]. The experiments are also conducted for a comprehensive analysis of the proposed method.

A. Dataset and Implementation Details

1) Dataset: Datasets from WorldView-4, GeoEye-1, and WorldView-2 satellites are considered in our experiments. The data from GeoEye-1 are collected on February 24, 2009, from the urban area of Hobart, Australia. The ground resolutions of acquired LR MS and PAN images are 0.41 and 1.64 m at nadir, respectively. WorldView-4 offers 0.31-m PAN images and 1.24-m LR MS images at nadir. The images from WorldView-4 are acquired in Acapulco, Mexico, on April 5, 2017. For WorldView-2, the LR MS and PAN images are acquired on

April 11, 2012, in Sydney, Australia, with the spatial resolutions of 2.0 and 0.5 m. Since there are no available HR MS images for direct comparison with fusion results, we produce the reduced-scale LR MS and PAN images according to Wald's protocol [50]. Then, the fused images are compared with the original LR MS images. In the reduced-scale datasets, the sizes of the original LR MS image and PAN image are $256 \times 256 \times 4$ and 1024×1024, respectively. Then, $64 \times 64 \times 4$ and 256×256 are the sizes of the LR MS image and the PAN image to be fused. Finally, the fused image with size $256 \times 256 \times 4$ is compared with the ground truth, i.e., the original LR MS image. We utilize some reference-based evaluation indexes for quantitative comparisons, including Q4 [51], spectral angle mapper (SAM) [52], universal image quality index (UIQI) [53], root mean squared error (RMSE), and Erreur Relative Globale Adimensionnelle de Synthèse (ERGAS) [50]. For these indexes, Q4 and UIQI range from 0 to 1, and values close to 1 indicate better fusion results, while lower values denote better performance in terms of SAM and ERGAS. Besides, the experiments are also implemented on full-scale datasets. In this case, the sizes of LR MS and PAN images are $64 \times 64 \times 4$ and 256×256. Naturally, the size of the fused image is $256 \times 256 \times 4$. Then, no-reference indexes, such as quality w/o reference (QNR) [54], D_λ, and D_S, are used for quality evaluation. Higher QNR means better quality of the fused images, and its highest value is 1. For D_λ and D_S, the smaller the values, the better the fused.

2) Implementation Details: Our proposed method is implemented by PyTorch on one NVIDIA 2080Ti GPU. For the filter weight initialization, we use a Gaussian function with zero mean and 0.02 standard deviation. Moreover, bias is initialized by 0. Adam optimizer is used in the proposed method with a learning rate of 0.0005. At each scale, the number of epochs is set as 3000 for the generator. In the framework of ZeRGAN, the number of scales L is 4, and the ratio of the upsampling layer is $\sqrt{2}$ for the generator at each scale. The effects of L are analyzed in Section IV-F in detail. For the generators, at different scales, we set n as 32, 32, 64, and 64, respectively. Besides, the regularization parameters α, β, δ, and γ are set as 10, 300, 10, and 100, respectively. For the gradient penalty in L^l_{adv}, λ is equal to 0.1 according to the settings in [44].

B. Experiments on Reduced-Scale Datasets

In this part, we adopt three pairs of LR MS and PAN images for fusion performance analysis of the proposed method and the compared methods. Fig. 5 displays the LR MS and PAN images acquired by the WorldView-4 satellite and the fusion results of different methods. Moreover, a region is chosen and magnified for further qualitative evaluation. Its corresponding enlarged regions lie in the bottom right corner of each fused image. From Fig. 5, some spatial or spectral differences can be found between the fusion images and the reference image. For example, the spectral information is distorted in the building areas of the result of BDSD in Fig. 5(d). This is because the gain parameters are not estimated accurately. The result of SVT in Fig. 5(g) also shows spectral information loss. In Fig. 5(h), the image fused by VPLGC has good spectral quality but blurred spatial details. The blurring effects may be caused by insufficient gradient constraints. Some spectral distortions can be observed in Fig. 5(i) of PNN, where the color of the road in the magnified area is not consistent with that of the reference image. The fused image of PanNet has distorted spatial and spectral properties with the presence of blurring effects. Compared with other images, the proposed method provides better spatial details, as visible in the selected magnified region. This is due to the capability of the progressive framework to increase the spatial and spectral information of the fused image gradually. Table I shows the quantitative results of the fused images in Fig. 5. The best and second-best values are indicated by boldface and underline. One can see that the proposed method has the best performance in Q4, UIQI, RMSE, and ERGAS. For SAM, the best value is obtained by VPLGC. However, even if the SAM value of the proposed method is only the third best, it is still much better than those of many other methods.

Fig. 6 presents the fusion results of different methods on the GeoEye-1 dataset. For a better analysis, a region labeled by a red rectangle is magnified and then put on the bottom right corner of each fused image. By analyzing the results, we can see that the hues of the results of BDSD and AWLP in Fig. 6(d) and (e) have larger differences with the reference image than those of the other fusion results. For instance, some spectral artifacts arise in the color of the trees in the magnified area of Fig. 6(d). Moreover, obvious spectral distortions can also be seen in the soil of the fused image from PNN. The reason for the spectral distortions in Fig. 6(i) may be due to the fact that it is difficult to capture the spectral relationships among the bands of MS images due to the simple architecture in PNN. For the proposed ZeRGAN, the spectral quality is more consistent with the reference image because of the introduction of average constancy loss. In terms of spatial information, some blurring effects can be observed from the results of AWLP and VPLGC in Fig. 6(e) and (h), especially in the magnified regions. The fused image of SVT in Fig. 6(g) presents accurate spatial details but with some spectral differences in the tree areas because the spatial details from PAN images are injected into the bands of MS images independently. The spatial textures are enhanced efficiently in the result of the proposed method [see Fig. 6(l)], which proves the effectiveness of the spatial enhancement loss. Moreover, Table II lists quality indexes computed from the images in Fig. 6. The best results are labeled in bold, and the second-best values are highlighted by underlining. The proposed method behaves best in Q4, UIQI, RMSE, and ERGAS. ZeRGAN produces the best SAM, while the second SAM value is given by VPLGC.

In addition, we also used the eight-band LR MS and PAN images acquired from WorldView-2 to analyze the performance of all methods. The fusion results are displayed in Fig. 7. In Fig. 7(d) and (g), we can see that fusion results of BDSD and SVT suffer from some spectral distortions. For BDSD and SVT, eight bands in MS images lead to a bias in coefficient estimation. The result of PSGAN in Fig. 7(f) shows some blurry effects because the spatial enhancement loss is not considered in it. For the proposed method, the image in Fig. 7(l) exhibits clear spatial details due to the spatial enhancement loss that helps to preserve the edges in the fused images. Table III reports the quantitative result of different quality metrics. Because Q4 is defined for the

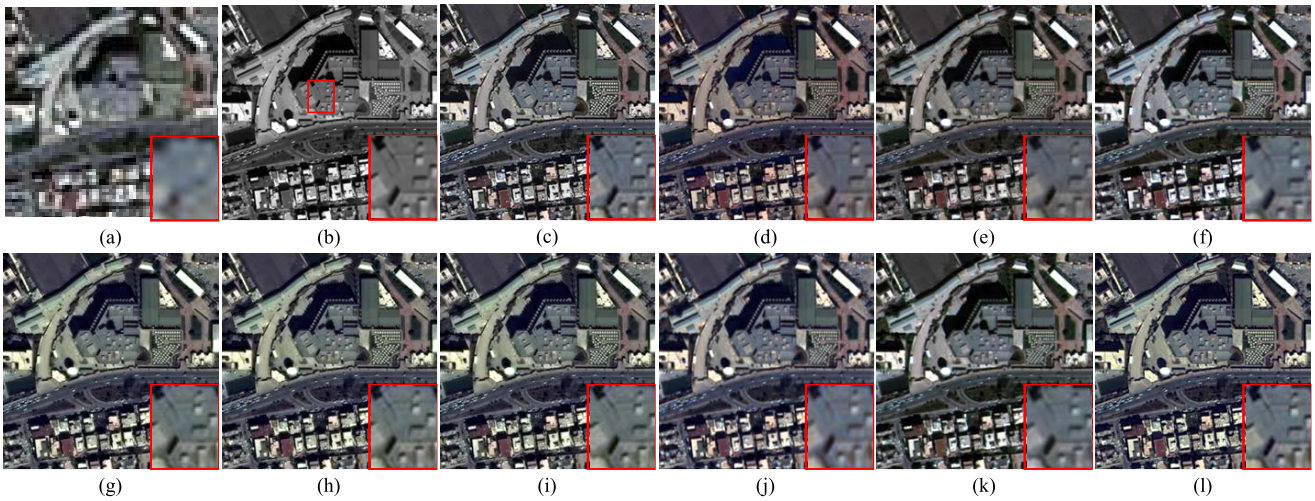

Fig. 5. Qualitative comparison of the fused images from different methods on the WorldView-4 dataset. (a) LR MS image. (b) PAN image. (c) Reference image. (d) BDSD. (e) AWLP. (f) PSGAN. (g) SVT. (h) VPLGC. (i) PNN. (j) PanNet. (k) MDSCC-GAN. (l) Proposed ZeRGAN.

TABLE I

QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 5 (WORLDVIEW-4 DATASET)

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
Q4	0.8895	0.8812	0.8717	0.8851	0.8090	0.8553	0.8792	0.8392	**0.8957**
SAM	6.0905	5.3589	11.9411	6.8045	**5.2167**	7.4585	6.3103	7.9123	5.5602
UIQI	0.9435	0.9329	0.9056	0.9425	0.9235	0.9269	0.9326	0.8974	**0.9491**
RMSE	82.4629	85.6250	128.8411	87.2965	99.6870	96.6047	88.0146	114.4116	**75.9055**
ERGAS	2.1462	2.2298	3.4236	2.2616	2.5786	2.5033	2.3303	2.9618	**1.9664**

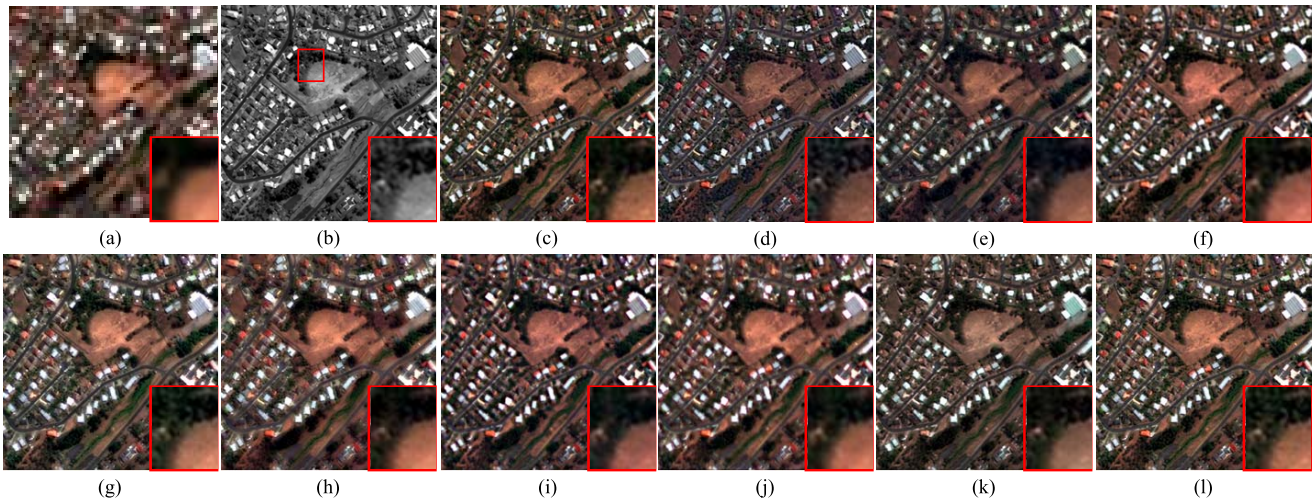

Fig. 6. Qualitative comparison of the fused images from different methods on the GeoEye-1 dataset. (a) LR MS image. (b) PAN image. (c) Reference image. (d) BDSD. (e) AWLP. (f) PSGAN. (g) SVT. (h) VPLGC. (i) PNN. (j) PanNet. (k) MDSCC-GAN. (l) Proposed ZeRGAN.

TABLE II

QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 6 (GEOEYE-1 DATASET)

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
Q4	0.7771	0.7998	0.7743	0.7842	0.7605	0.7599	0.7743	0.7979	**0.8031**
SAM	6.7184	5.3266	5.0859	5.7831	4.9623	7.2484	5.0859	5.4182	**4.8226**
UIQI	0.9271	0.9456	0.9289	0.9247	0.9232	0.9209	0.9289	0.9414	**0.9589**
RMSE	33.6297	26.5140	27.1398	29.0227	28.8286	34.8533	29.4064	31.3384	**22.2742**
ERGAS	2.0616	1.6134	1.8594	1.7770	1.7881	2.1485	1.8594	2.0375	**1.3722**

evaluation of four-band MS images, the values of Q4 are not given in Table III. It can be seen that the PanNet has the best performance in UIQI, RMSE, and ERGAS. The proposed method provides good performance close to the best methods.

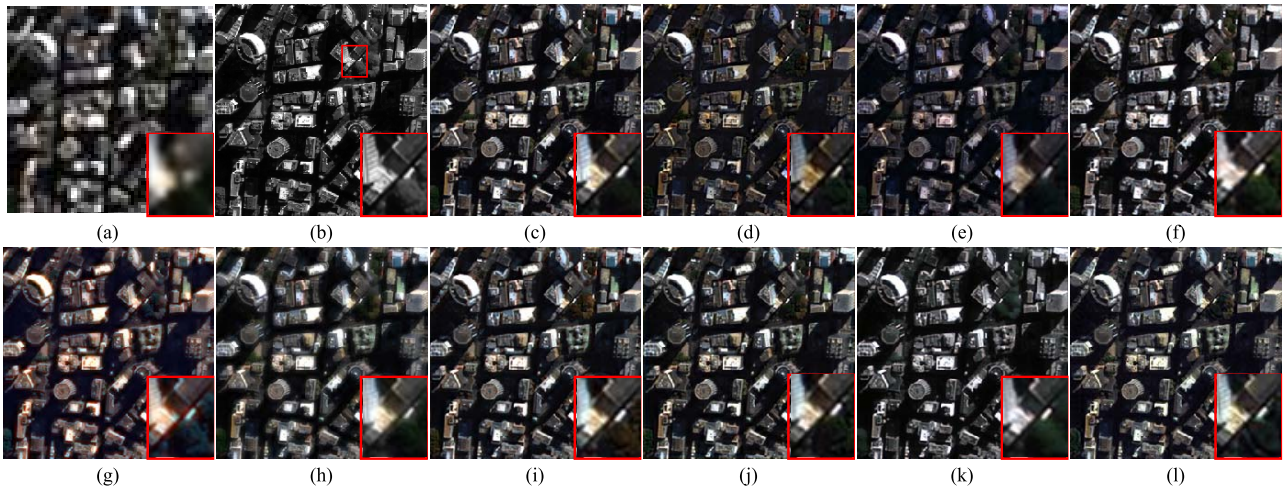

Fig. 7. Qualitative comparison of the fused images from different methods on the WorldView-2 dataset. (a) LR MS image. (b) PAN image. (c) Reference image. (d) BDSD. (e) AWLP. (f) PSGAN. (g) SVT. (h) VPLGC. (i) PNN. (j) PanNet. (k) MDSCC-GAN. (l) Proposed ZeRGAN.

TABLE III

QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 7 (WORLDVIEW-2 DATASET)

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
SAM	4.4929	3.9050	19.7816	8.0881	**3.6114**	5.1843	4.4540	13.8534	4.3381
UIQI	0.9720	0.9806	0.8534	0.9007	0.9665	0.9780	**0.9815**	0.9126	0.9732
RMSE	41.2874	33.6121	122.9689	67.8303	42.2555	35.9419	**32.3862**	81.5566	34.8409
ERGAS	1.6174	1.2970	5.0395	2.6830	1.5854	1.4034	**1.2367**	3.0717	1.2751

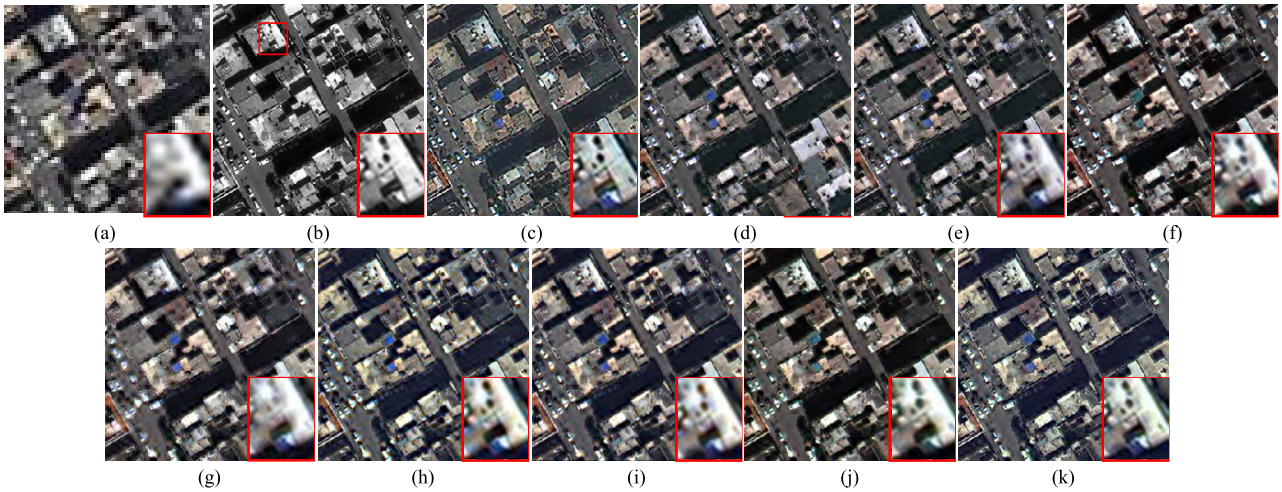

Fig. 8. Qualitative comparison of the fused images from different methods on the WorldView-4 dataset. (a) LR MS image. (b) PAN image. (c) BDSD. (d) AWLP. (e) PSGAN. (f) SVT. (g) VPLGC. (h) PNN. (i) PanNet. (j) MDSCC-GAN. (k) Proposed ZeRGAN.

TABLE IV

QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 8 (WORLDVIEW-4 DATASET)

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
D_λ	0.0352	0.0276	0.0491	0.0510	0.0253	0.1198	0.0464	0.0476	**0.0219**
D_S	0.0277	0.0229	0.0330	0.0361	0.0695	0.0788	0.0560	0.0427	**0.0196**
QNR	0.9381	0.9502	0.9195	0.9147	0.9069	0.8109	0.9001	0.9117	**0.9589**

C. Experiments on Full-Scale Datasets

The full-scale experiments are also conducted on three pairs of LR MS and PAN images from different satellites. Fig. 8 presents the fusion results of all methods on the WorldView-4 dataset. One can see some spatial artifacts in the result of BDSD [see Fig. 8(c)], which are due to the misestimated gain coefficients. In addition, some spatial information is oversmoothed in the fused image of VPLGC [see Fig. 8(g)]. Slight blurring effects can also be observed in the result of PNN in Fig. 8(h). For the fused images obtained by other methods in Fig. 8, the spatial information in Fig. 8(j)

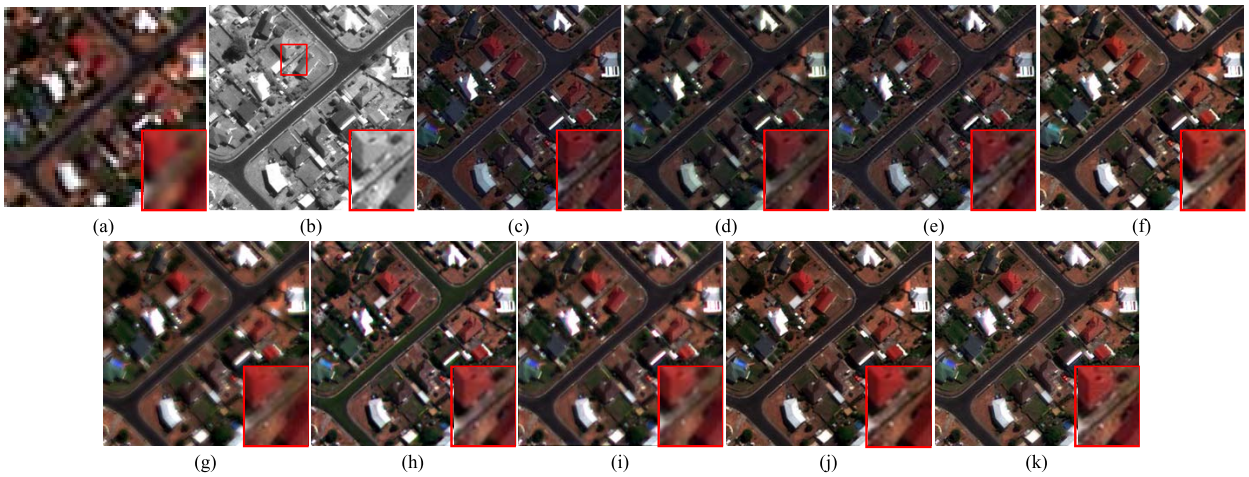

Fig. 9. Qualitative comparison of the fused images from different methods on the GeoEye-1 dataset. (a) LR MS image. (b) PAN image. (c) BDSD. (d) AWLP. (e) PSGAN. (f) SVT. (g) VPLGC. (h) PNN. (i) PanNet. (j) MDSCC-GAN. (k) Proposed ZeRGAN.

TABLE V

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

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
D_λ	0.0497	0.0806	0.0219	0.0658	0.0363	0.0326	0.0199	0.0179	**0.0178**
D_S	0.0749	0.0335	0.0233	0.0378	0.0825	0.0618	0.0882	0.0202	**0.0195**
QNR	0.8792	0.8886	0.9554	0.8989	0.8842	0.9076	0.8936	0.9623	**0.9631**

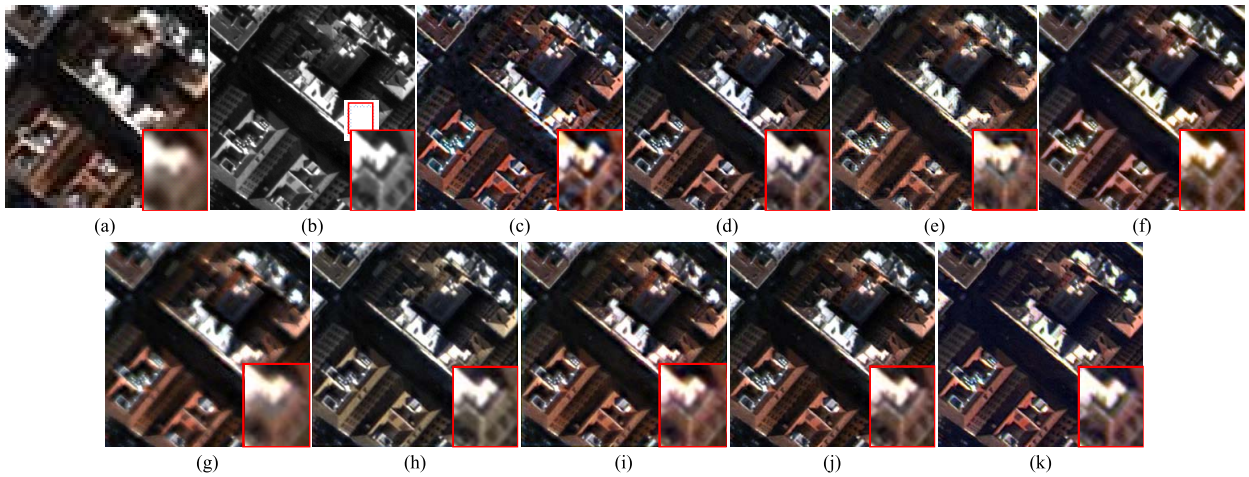

Fig. 10. Qualitative comparison of the fused images from different methods on the WorldView-2 dataset. (a) LR MS image. (b) PAN image. (c) BDSD. (d) AWLP. (e) PSGAN. (f) SVT. (g) VPLGC. (h) PNN. (i) PanNet. (j) MDSCC-GAN. (k) Proposed ZeRGAN.

is preserved better because MDSCC-GAN adopted SAM to constraint the spectral features in the fused image. As for the spectral information, unnatural color arises in the enlarged area of the result of PNN. Fig. 8(d) and (e) of AWLP and PSGAN provides different hues in the chosen regions compared with the results from other methods. For PanNet, the building areas of the result in Fig. 8(i) show spectral distortions, which are caused by the learned residuals by ResNet. On the contrary, we can observe that the spectral information in Fig. 8(k) is preserved efficiently by the proposed method because the spectral information is constrained by the spectral reconstruction and average constancy terms simultaneously. The quantitative index values of all methods are given in Table IV, in which the values in the top two places are labeled by boldface and underline, respectively. The best D_S is achieved by the proposed method, which has the best spatial performance. For D_λ, ZeRGAN also achieves the best value, which is followed by VPLGC. Besides, our method can provide the best QNR, which means a better overall effect.

Fig. 9 shows all fused images from the GeoEye-1 satellite dataset. For BDSD, AWLP, and PSGAN, the hues in Fig. 9(c)–(e) are darker than those of the other fusion results, which points out some spectral distortions. The result of VPLGC in Fig. 9(g) shows that the spatial details are oversmoothed in the magnified region because the gradient constraint suppresses the textures in the smooth areas. The color of the buildings in Fig. 9(h) looks unnatural and is not consistent with that in other fusion images. The fused image obtained by the proposed method [see Fig. 9(k)] presents better spectral information preservation. Spatial textures are also degraded in the result of PanNet [see Fig. 9(i)]. For example, the edges of the roof in the magnified region are smooth, and

TABLE VI
QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 10 (WORLDVIEW-2 DATASET)

Metric	BDSD	AWLP	PSGAN	SVT	VPLGC	PNN	PanNet	MDSCC-GAN	Proposed ZeRGAN
D_λ	0.0650	0.0494	0.0641	0.1185	0.0486	**0.0298**	0.0587	0.0331	0.0378
D_S	0.1024	0.0452	0.0807	0.0815	0.1047	0.0321	0.0495	0.0314	**0.0240**
QNR	0.8393	0.9076	0.8604	0.8096	0.8517	0.9391	0.8947	0.9366	**0.9392**

TABLE VII
QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 12

Metric	w/o spatial term	w/o spectral term	w/o spatial enhancement	w/o average constancy	Proposed ZeRGAN
D_λ	0.0655	0.5778	0.0590	0.0629	**0.0573**
D_S	0.2129	0.3863	0.0662	0.0642	**0.0639**
QNR	0.7356	0.2591	0.8787	0.8770	**0.8824**

the subtle textures are not visible. The result of the proposed method in Fig. 9(k) produces realistic spatial structures due to the adversarial loss between the real and intermediate PAN images. Table V illustrates the quantitative evaluations of all fused images in Fig. 9. It can be observed that the best values are all from the proposed method. From a global assessment, we can conclude that the proposed method behaves best.

Full-scale experiments are also conducted on the WorldView-2 dataset. All fused images are displayed in Fig. 10. From the figure, we can see that the fusion results of different methods present some differences in terms of spectral information. The spatial details of BDSD and AWLP [see Fig. 10(c) and (d)] are of high quality. The fused image of VPLGC in Fig. 10(g) is blurred, which is similar to the performance in Fig. 9(g). For PNN, the result in Fig. 10(h) suffers from significant spectral distortions, as PNN fails to capture the spectral information in MS images with more bands. The result of ZeRGAN in Fig. 10(k) shows that the spatial information is enhanced well because the spatial enhancement terms can preserve the subtle details. Moreover, the numerical values in Table VI also demonstrate that the proposed method can produce better fusion results.

D. Validation of Average Constancy Loss

For the validation of the constancy loss in (10), 3000 LR MS and HR MS image pairs were synthesized, where LR MS images were produced by blurring and decimation on HR MS images. For blurring, the modulation transfer function (MTF) was used for low-pass filtering, which is approximated through a 2-D independent Gaussian distribution. Fig. 11 shows the histograms of relative error between averages of LR MS and HR MS images. It is possible to observe that the band averages of LR MS images are approximately equal to those of HR MS images. Therefore, the loss in (10) can facilitate the reconstruction of the fused images.

E. Ablation Study

In this section, ablation experiments are implemented to analyze the effectiveness of each loss function in (11). The corresponding qualitative and quantitative results are reported

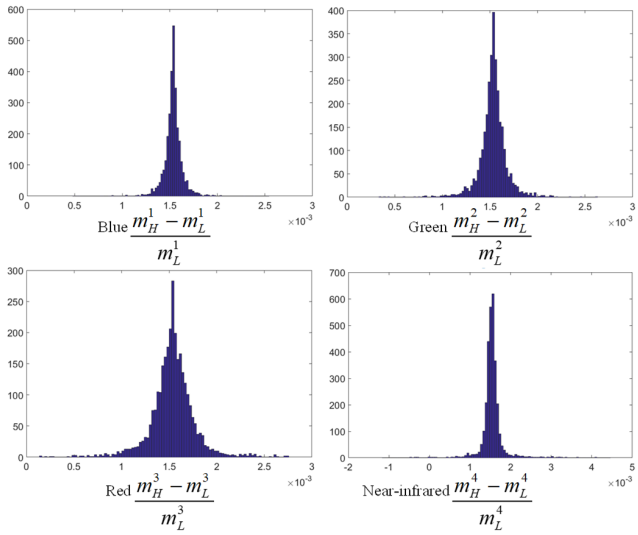

Fig. 11. Relative average errors for each band in HR MS image and LR MS image. Here, the MS images are made up of four bands: blue (B), green (g), red (R), and near-infrared (NIR) bands.

in Fig. 12 and Table VII, respectively. Moreover, we select the road region and magnify it for intuitive analysis. The magnified regions are put in the lower left corner of the fused results. For the fused image without spatial fidelity term in Fig. 12(c), we can see that the spatial information is blurry, and some edges or textures are smoothed. Fig. 12(d) displays the result without the spectral fidelity term. We can observe that the spectral features are sharply distorted. When spatial enhancement loss or average constancy loss are discarded [see Fig. 12(e) and (f)], some spectral distortions can be found. Fig. 12(g) shows that the result with all loss terms behaves better than any of the other fused images. Table V lists the numerical values of all results, which are consistent with the visual performance in Fig. 12.

F. Investigation on Network Architecture of the Generator

In the architecture of the generator, three important factors affect the fusion performance of ZeRGAN, such as scales, blocks, and dense connections. Here, we denote the structure displayed in Fig. 3 as one block. Then, we analyze the influences on D_λ, D_S, and QNR by setting different combinations of the three factors. Fig. 13 shows the variations of the three metrics with different architectures. Note that four scales, one block, and dense connections are used in ZeRGAN. From Fig. 13, we can find that the proposed ZeRGAN outperforms other architectures compared with different combinations of the three factors. With the introduction of dense connections, obvious improvements can be seen in terms of the metric values in Fig. 13. Although the computational complexity

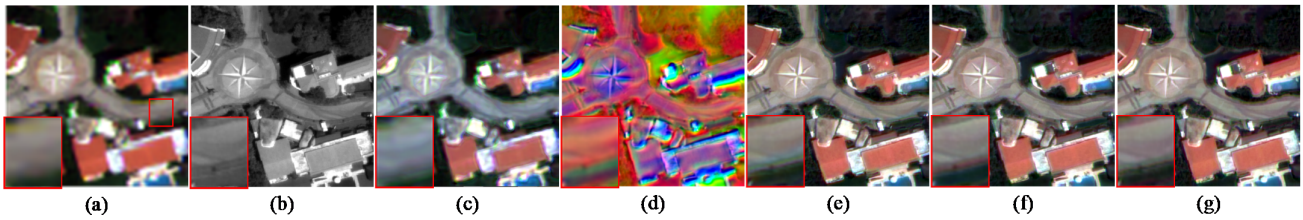

Fig. 12. Ablation study of the contribution of each loss term. (a) LR MS image. (b) PAN image. (c) W/o spatial term. (d) W/o spectral term. (e) W/o spatial enhancement. (f) W/o average constancy. (g) Complete ZeRGAN.

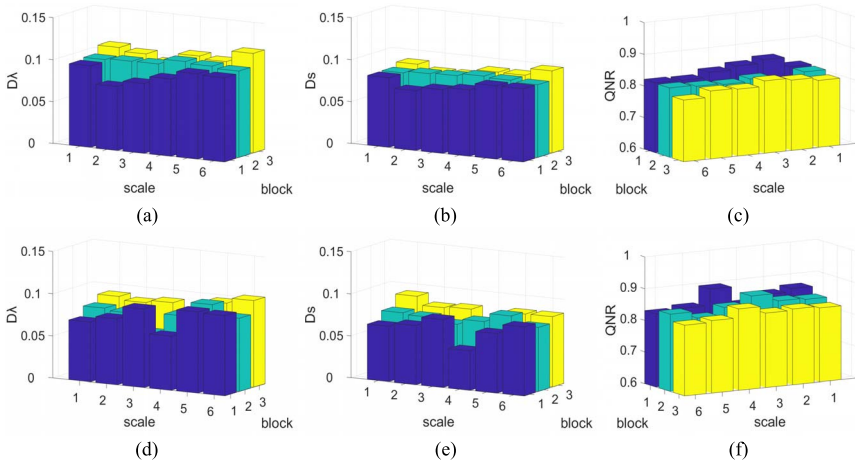

Fig. 13. Performance analysis of the fused images with different network structures. (a)–(c) Different network structures without dense connections. (d)–(f) Different network structures with dense connections.

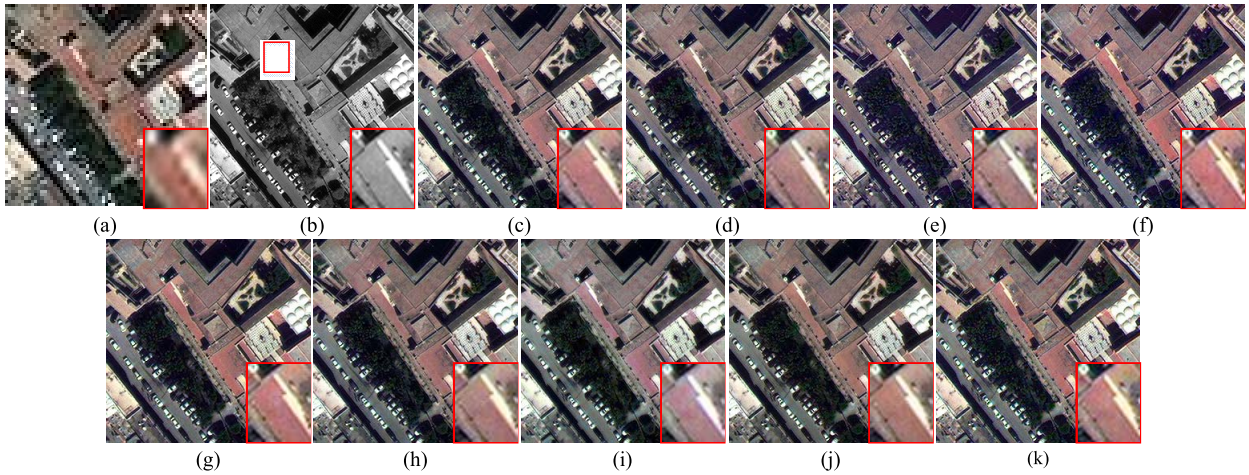

Fig. 14. Qualitative comparison of the fused images with different network structures. (a) LR MS image. (b) PAN image. (c) 4@1@w/o. (d) 4@2@W. (e) 4@3@W. (f) 1@1@W. (g) 2@1@W. (h) 3@1@W. (i) 5@1@W. (j) 6@1@W. (k) 4@1@W in proposed ZeRGAN.

increases dramatically with the increase in blocks, the metric values are not improved equally. With the increasing number of scales, the metric values generally become better first and then worse. Based on the above observations, we choose the settings in ZeRGAN for our experiments.

For further qualitative analysis, we select some fusion results of typical configurations in Fig. 13 and show them in Fig. 14. In Fig. 14, the configuration of ZeRGAN is denoted as 4@1@W for simplification; 4 and 1 denote the number of scales and the number of blocks, respectively. W implies that the dense connections are used in the networks, and w/o means that the dense connections are not introduced. The results of different networks are reported in Fig. 14. For further perception, an interesting region is chosen for comparison, whose enlarged versions are put in the lower right corners of the fused images. One can see that the results from different networks have different visual appearances. For spectral information, Fig. 14(c), (f), and (i) has obvious differences compared with the LR MS image and other fused images. Some spatial distortions can be found in Fig. 14(d) and (e). Moreover, we can see that the spectral information has a slight difference by increasing the number of scales. Although the visual differences are not evident among the fusion results in Fig. 14(c) and (k), the index values in Fig. 13 imply a great improvement by the dense connections. Thus, the architecture used in ZeRGAN is a good choice for modeling both spatial and spectral information.

G. Effect of Parameter Settings

In the proposed ZeRGAN, α, β, δ, and γ values affect the fusion results. α and β control the spatial and spectral degradation loss functions, respectively. For the spatial enhancement

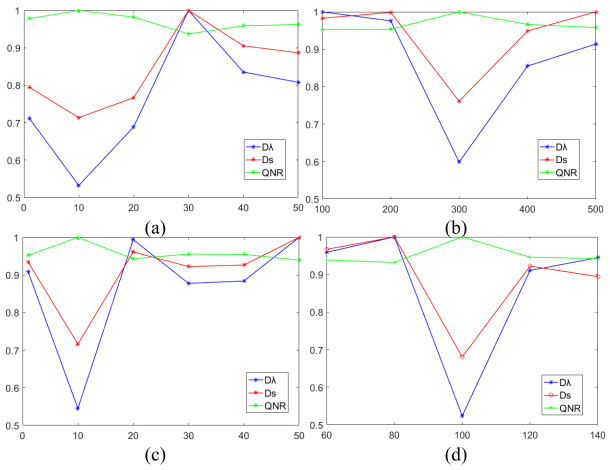

Fig. 15. Performance of ZeRGAN versus different parameters. (a) α. (b) β. (c) δ. (d) γ.

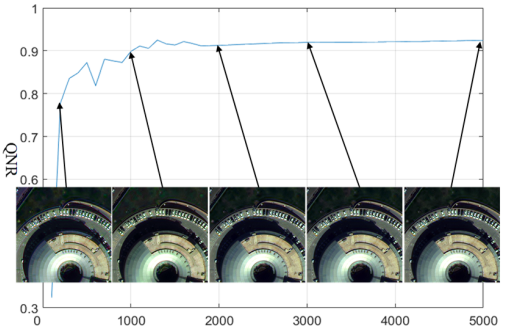

Fig. 16. Intermediate results of the fused images from the WorldView-4 satellite at epochs $t = 200$, 1000, 2000, 3000, and 5000.

loss, larger δ values introduce more spatial details. γ is responsible for the spectral preservation of the fused images. To investigate the influences of these parameter values on the fusion results, we considered the image pair in Fig. 14(a) and (b). The values of D_λ, D_S, and QNR from different fusion results are shown in Fig. 15, in which the values of different metrics are normalized. In the figure, D_λ and D_S are influenced by the variations of the parameters. The best values are provided in Fig. 15(a) when $\alpha = 10$. Moreover, we can also find similar trends in Fig. 5(b)–(d). Thus, α, β, δ, and γ are finally set as 10, 300, 10, and 100, respectively.

H. Visualization of Intermediate Results

Fig. 16 shows the intermediate results of the proposed fusion method on WorldView-4 images at epochs $t = 200$, 1000, 2000, 3000, and 5000 along with the QNR curve. When epoch $t = 200$, the intermediate image contains some spatial distortions, especially in the details of small targets. For the fused image at epoch $t = 1000$, obvious spectral distortions can be observed. For instance, the color of the roof in the image has some differences from the other intermediate images. The numerical values are unstable when t is smaller than 2000. By increasing the number of epochs, finer details and better spectral preservation can be found in the fused images, and a stable improvement in QNR is achieved. When t is larger than 3000, only a slight improvement is observed in QNR. Thus, taking the computational time into consideration, the number of epochs is set as 3000 finally.

The intermediate results from the GeoEye-1 satellite are also shown in Fig. 17. A similar trend in QNR can be observed in

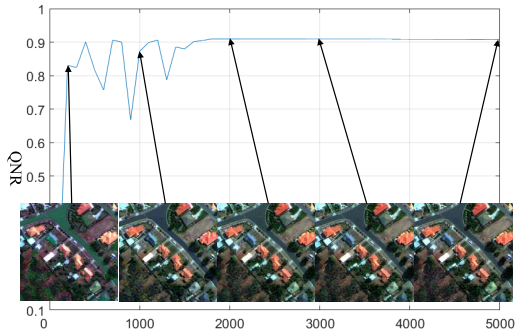

Fig. 17. Intermediate results of the fused images from the GeoEye-1 satellite at epochs $t = 200$, 1000, 2000, 3000, and 5000.

TABLE VIII

COMPUTATIONAL TIME COMPARISON AND MODEL SIZE OF ALL METHODS

Method	Training time (h)	Test time (s)	Number of parameters
BDSD	—	0.0061	—
AWLP	—	0.1398	—
PSGAN	31.56	1.56	3M
SVT	—	1.5849	—
VPLGC	—	16.0159	—
PNN	14.01	1.4483	0.08M
PanNet	21.09	9.1758	0.15M
MDSCC-GAN	63.39	1.5413	15M
ZeRGAN	—	6420.20	0.9M

Fig. 17. When the number of epochs t is smaller than 2000, QNR grows unsteadily. Some spectral distortions are involved in the intermediate images at epoch $t = 200$. The spatial details are blurred in the images when $t = 1000$. By increasing the number of epochs, the spectral and spatial information in the fused images is gradually corrected up to 3000. Thus, this is the number of epochs selected for the GeoEye-1 dataset.

I. Computational Time and Model Size Analysis

In this section, we compare the computation time of all methods and analyze the parameter numbers of the methods based on DNN. For the traditional methods, experiments are conducted by using MATLAB R2017a on a computer equipped with an Intel[1] Core[2] i7-6700 processor, 3.4 GHz, and 16-GB memory. The DNN-based methods are trained and tested by PyTorch on an NVIDIA 2080Ti GPU with Intel[1] Core[2] i7-9700 processor, 3.0 GHz, and 128-GB memory.

In Table VIII, traditional methods have a better performance in terms of running time. Besides, the proposed method takes a long time to fuse the LR MS and PAN images compared to other methods based on DNN because the minimization of the loss function is iteratively implemented in ZeRGAN to improve the quality of the fused images. For other DNN-based methods, there is no further optimization during the test. Thus, the complexity of ZeRGAN is higher than that of other methods based on DNN. However, the proposed method does not need any training in advance, which mitigates the critical requirements to have a large amount of training data. However, ZeRGAN can be used in the fusion of LR MS and PAN images acquired by any satellites, also when there are no

[1]Registered trademark.
[2]Trademarked.

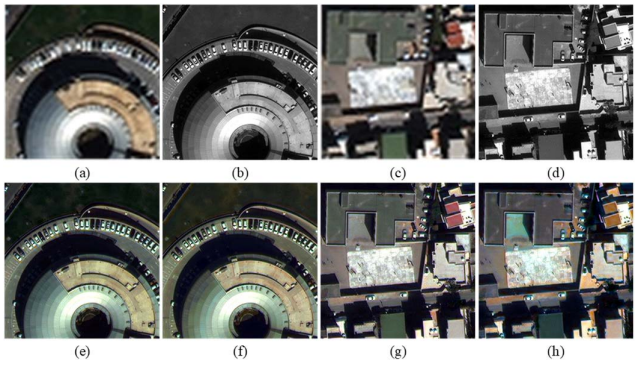

Fig. 18. Analysis of the generalization of the multiscale generator trained from a specific LR MS and PAN image pair. Image pair 1 (a) LR MS and (b) PAN. Image pair 2 (c) LR MS and (d) PAN. (e) Fused image of pair 1 from ZeRGAN. (f) Fused image of pair 1 from generator 2. (g) Fused image of pair 2 from ZeRGAN. (h) Fused image of pair 2 from generator 1.

paired or unpaired images for the training of the existing DNN-based methods. This is due to the generalization capability of the proposed method that does not need any training data. The model size of DNN-based methods is also reported in Table VIII. MDSCC-GAN contains about 15M parameters, where the proposed method has only about 0.9M parameters. Due to the introduction of discriminators, the model size of GAN-based methods is larger than those of the methods based on other kinds of DNN.

J. Analysis of the Multiscale Generator in ZeRGAN

In ZeRGAN, we use only the LR MS and PAN images to be fused for network optimization, and the fusion results can be directly generated after the loss function optimization in (12) is completed. In other words, the fusion of LR MS and PAN images is achieved during the training. In fact, we can also obtain a multiscale generator trained on a single pair of LR MS and PAN images. Naturally, we want to see whether the model trained on pair 1 of LR MS and PAN images can be directly used on the fusion of pair 2 of LR MS and PAN images. More specifically, LR MS and PAN images in pair 2 are directly fed into the multiscale generator trained on the source image pair 1 to generate the fusion image, which is exactly the test process in most of the pansharpening methods based on DNN shown in Fig. 1(a) and (b).

For the generalization analysis, two pairs of LR MS and PAN images from WorldView-4 are displayed in Fig. 18(a) and (d). Fig. 18(e) and (g) presents the fusion results of the two pair images obtained by ZeRGAN. We term the multiscale generator trained on image pair 1 as generator 1, while generator 2 denotes the multiscale generator derived from image pair 2. The fusion result of image pair 1 on generator 2 is shown in Fig. 18(f), while the fused image of image pair 2 on generator 1 is depicted in Fig. 18(h). From Fig. 18, we can see qualitatively that the fusion results obtained from their own model are superior to the fused images produced by the generators trained on the other image pair. For example, the spectral information in Fig. 18(e) is preserved better than that in Fig. 18(f). However, they have competitive performances in terms of spatial details. A similar trend can also be found in Fig. 18(g) and (h).

Table IX lists the quantitative index values of the results in Fig. 18(e)–(h). The fused image of pair 1 from generator 2 behaves better than the fused by ZeRGAN. However, the visual quality of Fig. 18(e) is better than that of Fig. 18(f), and the spectral information in Fig. 18(e) is more similar to Fig. 18(a). Moreover, the visual analysis remains consistent with the numerical performance for the results in Fig. 18(g) and (h) from image pair 2. Through this experiment, we can conclude that spatial details are enhanced well on the fused images from the generator trained on the other image pair, while their spectral information cannot be enhanced, as well as that on spatial details. This demonstrates the representation capacity of spatial details of the multiscale generator framework in ZeRGAN.

TABLE IX
QUANTITATIVE EVALUATIONS OF THE FUSED IMAGES IN FIG. 18

Metric	Fused image of pair 1 by ZeRGAN	Fused image of pair 1 by generator 2	Fused image of pair 2 by ZeRGAN	Fused image of pair 2 by generator 1
D_λ	0.0434	**0.0314**	**0.0270**	0.1724
D_S	0.0386	**0.0380**	**0.0434**	0.0723
QNR	0.9197	**0.9318**	**0.9307**	0.7678

It is worth noting that a large amount of training data is a huge burden for desirable fusion results in existing pansharpening methods based on DNN. For example, Yang *et al.* [25] used 18 000 image pairs, including upsampled LR MS, PAN, and HR MS images with size 64×64, for full-reference end-to-end training. 60 000 image pairs are employed for unsupervised network training and validation in [33], where the sizes of LR MS and PAN images are 32×32 and 128×128, respectively. On the contrary, the results in this part prove that plausible fusion results can be produced by the generator trained on only one full-scale image pair of LR MS and PAN images without any HR MS images as the reference images. This implies that a huge amount of training image pairs may be unnecessary for pansharpening, at least for spatial detail enhancement in the fused images. Furthermore, few-reference or zero-reference approaches, such as ZeRGAN in this article, can achieve satisfactory results.

V. CONCLUSION

In this article, we proposed a new pansharpening method, i.e., ZeRGAN, which establishes a multiscale framework based on GAN without any training data. It adopts a generator and a discriminator at each scale to capture the spectral and spatial information in the fused image, respectively. For each generator, five convolution layers and dense connections are used to efficiently enhance the spatial details in the intermediate MS images while preserving the spectral features. Then, the final fused image is progressively obtained by multiscale generators. To achieve the training without data, we design a nonreference loss function, which is composed of adversarial loss, spatial and spectral reconstruction losses, spatial enhancement loss, and average constancy loss. The average constancy loss behaves well in terms of spectral information preservation. Experiments demonstrate that our proposed method can obtain better fusion results compared with the state-of-the-art methods, which confirms the effectiveness of ZeRGAN. For further work, we will continue to advance the development of the pansharpening method based on DNN in the zero-reference case with more efficient spectral information preservation. Moreover, to reduce the running time

of the proposed method, we will investigate the relationships among the fusion performance and the multiscale framework, the network structure, and the iterative strategy to improve the usability of ZeRGAN in many operational cases.

REFERENCES

[1] P. Zhong, Z. Gong, and J. Shan, "Multiple instance learning for multiple diverse hyperspectral target characterizations," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 246–258, Jan. 2020.

[2] S. Jia, Z. Lin, B. Deng, J. Zhu, and Q. Li, "Cascade superpixel regularized Gabor feature fusion for hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1638–1652, May 2020.

[3] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local restricted convolutional neural network for change detection in polarimetric SAR images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 818–833, Mar. 2019.

[4] 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.

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

[6] W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, Apr. 1990.

[7] P. S. Chavez, Jr., S. C. Sides, and J. A. Anderson, "Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic," *Photogramm. Eng. Remote Sens.*, vol. 57, no. 3, pp. 295–303, Mar. 1991.

[8] 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, Jan. 4, 2000.

[9] 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.

[10] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 1, pp. 49–61, Jan. 2000.

[11] 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.

[12] J. Lee and C. Lee, "Fast and efficient panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 155–163, Jan. 2010.

[13] 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.

[14] 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.

[15] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Jan. 2006.

[16] X. He, L. Condat, J. M. Bioucas-Dias, J. Chanussot, and J. Xia, "A new pansharpening method based on spatial and spectral sparsity priors," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4160–4174, Sep. 2015.

[17] 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.

[18] 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.

[19] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.

[20] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

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

[22] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

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

[24] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

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

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] 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, doi: 10.1109/TNNLS.2020.2996498.

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

[29] Y. Xing, M. Wang, S. Yang, and L. Jian, "Pan-sharpening via deep metric learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 165–183, Nov. 2018.

[30] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.

[31] X. Liu, Y. Wang, and Q. Liu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 873–877.

[32] 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.*, early access, Dec. 24, 2020, doi: 10.1109/TGRS.2020.3042974.

[33] 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. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[34] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022, doi: 10.1109/TGRS.2021.3060958.

[35] C. Zhou, J. Zhang, J. Liu, C. Zhang, R. Fei, and S. Xu, "PercepPan: Towards unsupervised pan-sharpening based on perceptual loss," *Remote Sens.*, vol. 12, no. 14, p. 2318, Jul. 2020.

[36] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4570–4580.

[37] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1867–1888, Mar. 2020.

[38] D. Ren, K. Zhang, Q. Wang, Q. Hu, and W. Zuo, "Neural blind deconvolution using deep priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3341–3350.

[39] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[40] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion GAN for image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3648–3657.

[41] J. Li, J. Jia, and D. Xu, "Unsupervised representation learning of image-based plant disease with deep convolutional generative adversarial networks," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 9159–9163.

[42] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[43] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. ICML*, Aug. 2017, pp. 214–223.

[44] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. NIPS*, 2017, pp. 5767–5777.

[45] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[46] M. M. Khan, L. Alparone, and J. Chanussot, "Pansharpening quality assessment using the modulation transfer functions of instruments," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3880–3891, Nov. 2009.
[47] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1780–1789.
[48] S. Zheng, W.-Z. 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.
[49] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 10265–10274.
[50] 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.
[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] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summ. 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.
[53] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Aug. 2002.
[54] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.

Wenxiu Diao received the B.S. degree in computer science and technology from Shandong Normal University, Jinan, China, in 2019, where she is currently pursuing the M.S. degree in computer science and technology with the School of Information Science and Engineering.

Her research interests include image processing and deep learning.

Feng Zhang received the B.S. degree in electronic information engineering from Shandong Normal University, Jinan, China, in 2012, and the M.Sc. degree in electronic information engineering from Xidian University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree in computer science and technology with the School of Information Science and Engineering, Shandong Normal University.

Her main current research interests include deep learning and image processing.

Jiande Sun received the B.S. and Ph.D degrees in communication and information systems from Shandong University, Jinan, China, in 2000 and 2005, respectively.

From 2008 to 2009, he was a Visiting Researcher with the Institute of Telecommunications System, Technical University of Berlin, Berlin, Germany. From 2010 to 2012, he was a Post-Doctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and the State Key Laboratory of Digital-Media Technology, Hisense Group. From 2014 to 2015, he was a DAAD Visiting Researcher with the Technical University of Berlin and the University of Konstanz, Konstanz, Germany. From 2015 to 2016, he was a Visiting Researcher with the School of Computer Science, Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan. He has published more than 60 journal articles and conference papers. He is a coauthor of two books. His current research interests include multimedia content analysis, video hashing, gaze tracking, image/video watermarking, and 2-D-to-3-D conversion.

Yinghui Xing received the B.S. and Ph.D. degrees in circuit and systems from the School of Artificial Intelligence, Xidian University, Xi'an, China, in 2014 and 2020, respectively.

She is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an. Her research interests include remote sensing image processing, image fusion, and image super-resolution.

Kai Zhang was born in Shanxi, China, in 1992. He received the B.S. degree in electrical engineering and automation from the North University of China, Taiyuan, China in 2013, and the Ph.D. degree in circuit and systems from Xidian University, Xi'an, China, in 2018.

He is currently a Lecturer with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. He is also a Post-Doctoral Fellow with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research has been focused on multisource remote sensing image fusion, matrix factorization, and deep learning.

Lorenzo Bruzzone (Fellow, IEEE) received the Laurea (M.S.) degree (*summa cum laude*) in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1993 and 1998, respectively.

He is currently a Full Professor of telecommunications with the University of Trento, Trento, Italy, where he teaches remote sensing, radar, and digital communications. He is also the Founder and the Director of the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento. His research interests include remote sensing, radar, synthetic aperture radar (SAR), signal processing, machine learning, and pattern recognition. He promotes and supervises research on these topics within the frameworks of many national and international projects. He is the Principal Investigator of many research projects. Among the others, he is the Principal Investigator of the Radar for icy Moon Exploration (RIME) Instrument in the framework of the JUpiter ICy moons Explorer (JUICE) Mission of the European Space Agency (ESA) and the Science Lead for the High-Resolution Land Cover Project in the framework of the Climate Change Initiative of ESA. His articles are highly cited, as proven by the total number of citations (more than 37 000) and the value of the H-index (89).

Dr. Bruzzone has been a member of the Administrative Committee of the IEEE Geoscience and Remote Sensing Society (GRSS) since 2009, where he has been the Vice-President of Professional Activities since 2019. He has ranked first place in the Student Prize Paper Competition of the 1998 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Seattle, in July 1998. He was a recipient of many international and national honors and awards, including the recent IEEE GRSS 2015 Outstanding Service Award, the 2017 and 2018 IEEE IGARSS Symposium Prize Paper Awards, and the 2019 WHISPER Outstanding Paper Award. He was a guest coeditor of many special issues of international journals. He is the Co-Founder of the IEEE International Workshop on the Analysis of Multi-Temporal Remote-Sensing Images (MultiTemp) Series and is a member of the Permanent Steering Committee of this series of workshops. He was invited as a keynote speaker in more than 40 international conferences and workshops. Since 2003, he has been the Chair of the SPIE Conference on Image and Signal Processing for Remote Sensing. He has been the Founder of *IEEE Geoscience and Remote Sensing Magazine*, for which he has been the Editor-in-Chief from 2013 to 2017. He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He has been a Distinguished Speaker of the IEEE Geoscience and Remote Sensing Society from 2012 to 2016.