Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection

Jiang He [a], Qiangqiang Yuan [a,*], Jie Li [a], Yi Xiao [a], Liangpei Zhang [b]

[a] School of Geodesy and Geomatics, Wuhan University, Hubei, 430079, China
[b] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, Hubei, 430079, China

ARTICLE INFO

Keywords:
Pan-sharpening
Unsupervised fusion
Spectral super-resolution
Multispectral images
Self-learning
Remote sensing

ABSTRACT

Pan-sharpening is a very productive technique to enhance the spatial details of multispectral images with the aid of panchromatic images. Nowadays, deep learning-based pan-sharpening has scored tremendous achievements. However, strict requirement for training image pairs and low generalization hamper the development of supervised pan-sharpening with limited samples. Unsupervised image fusion is an effective technique to fuse images without adequate ground truth as training samples. Existing unsupervised pan-sharpening methods are commonly based on the image generator model, suffering from unsatisfactory spatial details. In this study, we proposed a self-supervised pan-sharpening method with the aid of spectral super-resolution named sSRPNet. Coarsening-scale self-learning exploits the internal information in multispectral images at a coarsening scale and trains the initial fusion model without other labels. Spectral super-resolution injection explores the missing spatial details in the initial fused images and recovers it. Degradation self-learning introduces strong spectral constraints with original multispectral images. Both reduced-resolution and full-resolution experiments on three datasets have proved the superiority of sSRPNet. Furthermore, the proposed spectral super-resolution injection can be implemented on any existing pan-sharpening algorithms, improving their performance.

1. Introduction

For diverse observation purposes, remote sensing satellites are equipped with various image sensors, such as visible light cameras, multispectral (MS) scanners, and hyperspectral imagers. With the merits of large scale, continuous observation, and big data, remote sensing images have been widely used in various fields (Liao et al., 2023; Colomina and Molina, 2014), such as agriculture (Huang et al., 2018), hydro-meteorology (Xu et al., 2020; He et al., 2022b), and environmental monitoring (Wang et al., 2021; Mulverhill et al., 2023). Predictably, vigorous technological development has led to increasing demand for capturing a more significant number of remote sensing images with more accurate information (Li et al., 2022; Wu et al., 2023a).

Due to the technical limitation of sensors and environmental interference, it is impossible to acquire remote sensing data satisfying all needs. For instance, panchromatic (PAN) images involve rich spatial details. However, without color information, they are inconsistent with human visual habits. On the other hand, multispectral (MS) images are captured with many spectral bands that represent more physical radiation properties, while they usually suffer from low spatial resolution. To fully use satellite images and produce high-resolution multispectral images for further applications, pan-sharpening is a common post-imaging processing technique in remote sensing (Wu et al., 2023c).

Over the past few decades, many methods have been proposed to achieve pan-sharpening. The main algorithms for traditional pan-sharpening can be divided into four categories: *(1) component substitution-based methods.* Methods based on component substitution aim at replacing the low-resolution spatial component of MS images with PAN images. Moreover, the spatial components are always extracted by methods based on intensity-hue-saturation (IHS) (Carper et al., 1990), Brovey transformation (Gillespie et al., 1987), principal component analysis (PCA) (Kwarteng and Chavez, 1989), and Gram–Schmidt transformation (GS) (Laben and Brower, 2000). *(2) multi-resolution analysis-based methods.* In this type of methods, MS and PAN images are decomposed into various resolutions, and the spatial details in PAN images are injected into the same-level MS features. Laplacian pyramids (Burt and Adelson, 1987), curvelet (Starck et al., 2002), wavelets (Nason and Silverman, 1995), and contourlet

[*] Corresponding author.
E-mail addresses: jiang_he@whu.edu.cn (J. He), qqyuan@sgg.whu.edu.cn (Q. Yuan), jli89@sgg.whu.edu.cn (J. Li), xiao_yi@whu.edu.cn (Y. Xiao), zlp62@whu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.isprsjprs.2023.09.003
Received 1 June 2023; Received in revised form 30 July 2023; Accepted 2 September 2023

transformations (Do and Vetterli, 2005) are some classical decomposition in this group. *(3) hybrid methods.* These methods try to combine the strengths of both component substitution and multi-resolution analysis methods. The main idea is to improve the spatial details of the fused image at multiple scales. Substitute Wavelet Intensity (SWI) (González-Audícana et al., 2004), Additive Wavelet Luminance Proportional (AWLP) (Otazu et al., 2005), and GS-Wavelet (Javan et al., 2021) are all hybrid methods. *(4) optimization-based methods.* Considering the spatial and spectral degradation in remote sensing imaging, the variational optimization-based methods regard pan-sharpening as an inverse problem and build different cost functions to search for the best estimation of the ideal high-resolution MS images. The optimization-based methods that fuse PAN and MS images by different constraints include P+ XS (Ballester et al., 2006), Total Variation (TV) (Palsson et al., 2013), $\ell 1/2$ gradient prior (Zeng et al., 2016), filter estimation (Vivone et al., 2014), and local gradient constraints (Fu et al., 2019).

Deep learning, as a famous data-driven technology, is capable of exploring the nonlinear relationship between different feature domains in an implicit manner (Zhong et al., 2016; Rao et al., 2017; Wei et al., 2017; Yang et al., 2017; Yuan et al., 2018; Xiao et al., 2022), which also sets off a global craze of convolutional neural networks (CNN)-based pan-sharpening (Zhong et al., 2016; Rao et al., 2017; Wu et al., 2023b). It can directly build a mapping from low-resolution MS images to high-resolution MS images with the help of PAN image. Masi et al. regarded the pan-sharpening task as a particular form of image super-resolution and utilized a three-layer convolutional neural network (PNN) to address pan-sharpening (Masi et al., 2016). As the deeper networks achieve a more robust learning ability, residual learning is employed to improve the depth of CNN and achieves better performance (Shao and Cai, 2018). Wei et al. (2017) introduced a global residual skip to improve the spatial details. Yang et al. (2017) employed high-pass filters before ResNet to extract better textures. In 2018, Yuan et al. (2018) further proposed a multi-scale and multi-depth CNN for multi-scale features in remote sensing images. To further improve the modeling capability of CNN, there have been many works, including pyramid networks (Zhang et al., 2019), adaptive weights (Liu et al., 2020a), the gradient prior (Zhang and Ma, 2021), two-stream networks (Liu et al., 2020b), the deep unrolling (He et al., 2022c), generative adversarial network (GAN) (Liu et al., 2020c; Shao et al., 2019; Gastineau et al., 2021), *etc.*

With good learning capacity from data, deep learning-based methods have achieved great superiority over traditional pan-sharpening (Zhang et al., 2022; Deng et al., 2022; Guo et al., 2022). On the other hand, the data-driven training mode also limits the applications of CNN-based pan-sharpening and brings the big challenge of achieving pan-sharpening without ground truth. To achieve unsupervised pan-sharpening, many researchers have made tremendous efforts by proposing various strategies. Most of these works are devoted to exploring the potential of GAN in unsupervised pan-sharpening (Li et al., 2021; Zhou et al., 2020). Ma et al. (2020) proposed two discriminators to achieve spatial and spectral constraints, respectively. Ozcelik et al. (2020) trained their GAN with reduced-resolution multispectral images through gray-scale transformation and spatial downsampling. Zhou et al. (2021) designed a two-stream generator with a dual discriminator. Furthermore, they introduce a hybrid loss based on the cycle consistency to improve the performance (Zhou et al., 2022). Xu et al. (2023) introduced a medium scale into the resolution gap and regarded pan-sharpening as a two-step fusion task. The GAN-based methods are very effective in addressing image processing without ground truth. Nevertheless, they generate images rather than fuse the existing spatial–spectral information and thus produce some artifacts.

Some other works tried to improve loss functions and performed more binding constraints (Liu et al., 2023; Xiong et al., 2020). Luo et al. (2020) proposed a new loss function where the input MS and PAN images are used to enhance the spatial constraints and spectral consistency, respectively. Ciotola et al. (2022) further introduced a target-adaptive operating modality. Qu et al. (2020) presented a self-attention mechanism with sparse constraint and detail reconstruction error to achieve unsupervised pan-sharpening. Ni et al. (2022) learned degradation processes using multiple CNN blocks to improve the constraints. Moreover, some works introduce extra prior knowledge into unsupervised pan-sharpening. Seo et al. (2020) combined unsupervised learning with registration learning to learn the alignment between PAN and MS images implicitly. Wang et al. (2022b) introduced meta-learning into the supervised method and achieved unsupervised adaptation. Uezato et al. (2020) proposed a guided deep decoder network to perform feature refinement between PAN and multispectral domains.

However, in these methods, the spatial details in PAN images cannot be perfectly merged into multispectral images. They commonly regard unsupervised pan-sharpening as an image generation task and does not make full use of PAN images. As we all know, spectral super-resolution (sSR) is a low-level image restoration task that only increases the spectral resolution and keeps the whole spatial details in the original images (He et al., 2022). Thus, in this paper, we are trying to introduce spectral super-resolution into pan-sharpening and inject more spatial details for unsupervised pan-sharpening (sSRPNet). Furthermore, rather than GAN, we proposed an efficient unsupervised pan-sharpening framework with dual-stage self-learning, consisting of coarsening-scale self-learning (CSSL) and degradation self-learning. Our contributions are as follows:

- Embedding spectral super-resolution into pan-sharpening, the proposed sSRPNet utilizes spectral super-resolution to inject more spatial details of PAN image into the fused multispectral images.
- Rather than using an image generation model, we design a dual-stage self-learning framework to achieve unsupervised pan-sharpening. Coarsening-scale self-learning is used for internal information extraction, and degradation self-learning is employed for binding constraints.
- Data captured by three satellites are tested, including WorldView-2, QuickBird, and Gaofen-2, proving that the proposed model can handle various satellite data in both reduced-resolution and full-resolution testings.
- The proposed sSRPNet can be applied to other state-of-the-art pan-sharpening algorithms and achieve further improvement.

The remaining part of the article is organized as follows. Section 2 describes the proposed sSRPNet, including the spectral super-resolution injection block and dual-stage self-learning. Section 3 shows the experiments on data from four satellites and presents some discussions. Finally, conclusions are given in Section 4.

2. Methodology

Considering the steady ability of spectral super-resolution to maintain spatial details, we introduce spectral super-resolution to inject the high-resolution details into the initial fused MS images. Meanwhile, a dual-stage self-learning framework is designed to achieve pan-sharpening without ground truth. The proposed sSRPNet is depicted in Fig. 1. With coarsening-scale self-learning, the high-resolution PAN and low-resolution MS images are initially fused. Then the spectral super-resolution injection module explores the missing spatial details with the guidance of the original PAN image. Furthermore, degradation self-learning improves spectral fidelity.

2.1. Pan-sharpening

Let $\mathbf{X} \in \mathbb{R}^{W\times H\times C}$ represent the ideal multispectral image, where W and H are the width and height, respectively, and C is the spectral channel number. $\mathbf{P} \in \mathbb{R}^{W\times H\times 1}$ denotes the captured PAN image with

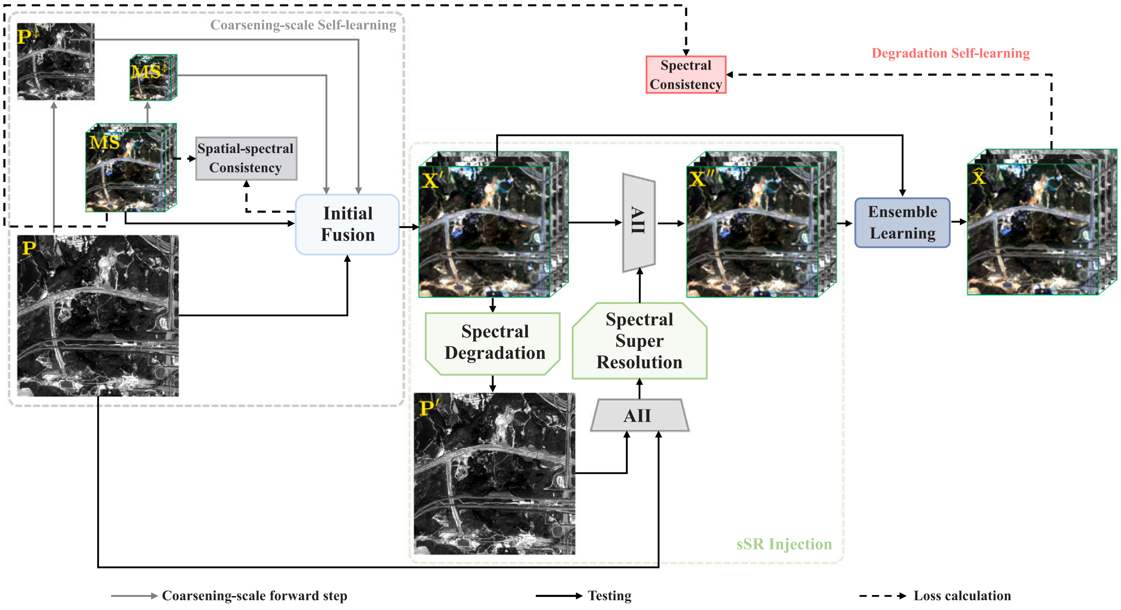

Coarsening-scale Self-learning
P↓
MS↓
MS
Spatial-spectral Consistency
Initial Fusion
P
X′
AII
X″
Ensemble Learning
X̂
Spectral Consistency
Degradation Self-learning
Spectral Degradation
Spectral Super Resolution
P′
AII
sSR Injection
Coarsening-scale forward step
Testing
Loss calculation

Fig. 1. The framework of the proposed sSRPNet with spectral super-resolution injection and dual self-learning. Spectral super-resolution injection recovers more spatial details. Coarsening-scale self-learning keeps the spatial and spectral consistency with the help of coarsening-scale MS images. Degradation self-learning enhances the spectral constraint.

the same spatial resolution but consists of only one band, $\mathbf{MS} \in \mathbb{R}^{w \times h \times C}$ is the captured low-resolution multispectral image, where $w = W/r$ and $h = H/r$ are the width and height of the degraded MS images. r is the resolution ratio between the ideal and the degraded MS images, usually four in pan-sharpening.

As we all know, the captured **MS** and **P** retain different information of the ideal multispectral image **X**, *i.e.* **MS** involves the rich spectral information. In contrast, **P** collects the finer spatial details. The final goal of pan-sharpening is to integrate **MS** and **P** for an optimal estimation $\hat{\mathbf{X}}$ of the ideal MS image, which requires the help of fusion model F. The mathematical formulation of pan-sharpening can be defined as:

$$\hat{\mathbf{X}} = \mathsf{F}(\mathbf{MS},\ \mathbf{P}) \tag{1}$$

2.2. Dual-stage self-learning

Data-driven deep learning is a good solution for pan-sharpening. However, existing deep learning-based pan-sharpening algorithms usually require ground truth to train CNNs. In practice, there are not always accessible high-resolution multispectral images. Thus, unsupervised pan-sharpening is of great significance. Self-learning is an excellent way to exploit the internal information in the collected data. However, classical self-learning suffers from the problem of over-fitting (Xiao et al., 2023). In this paper, we designed a dual-stage self-learning strategy to achieve unsupervised pan-sharpening, including coarsening-scale self-learning and degradation self-learning. Coarsening-scale self-learning exploits the coarse spatial information to train the initial fusion model, and degradation self-learning performs a supervisor that keeps the spectral fidelity.

2.2.1. Coarsening-scale self-learning

In remote sensing image processing, up-sampling means resampling images into a finer scale, while down-sampling is utilized to resample images into a coarser scale. Let $\mathbf{MS}^{\downarrow} \in \mathbb{R}^{\frac{w}{r} \times \frac{h}{r} \times C}$ denote the down-sampled MS image and $\mathbf{P}^{\downarrow} \in \mathbb{R}^{w \times h \times C}$ represent the degraded PAN image. In coarsening-scale self-learning, the image pair $\{\mathbf{MS}^{\downarrow}, \mathbf{P}^{\downarrow}\}$ involves the coarsening-scale information of **MS**. For reliable initial fused results, the original MS image **MS** is used as the ground truth, and the image pair $\{\mathbf{MS}^{\downarrow}, \mathbf{P}^{\downarrow}\}$ is used as input to train the initial fusion model. The formulation is:

$$\theta = \arg\min_{\theta} \left\| \mathsf{F}_{\theta}\left(\mathbf{MS}^{\downarrow},\ \mathbf{P}^{\downarrow}\right) - \mathbf{MS} \right\|_{p} \tag{2}$$

where θ is parameters of the initial fusion model F_{θ}, and $\|\cdot\|_p$ means the selected norm, usually L_1 norm and L_2 norm. The model parameters would be iteratively optimized following Eq. (2) for fully extracting internal information.

After coarsening-scale self-learning, the original image pair $\{\mathbf{MS}, \mathbf{P}\}$ is fed into the self-trained F_{θ} and obtain the initial fused image $\mathbf{X}'$:

$$\mathbf{X}' = \mathsf{F}_{\theta}(\mathbf{MS},\ \mathbf{P}) \tag{3}$$

Sequentially, spectral super-resolution injection extracts the missing spatial details and injects it into $\mathbf{X}'$ to generate the final results $\hat{\mathbf{X}}$.

2.2.2. Degradation self-learning

While the spatial details are injected, the final fused results $\hat{\mathbf{X}}$ should also keep high spectral consistency with the original multispectral domain. In unsupervised pan-sharpening, no ground truth can be used as the target to build spectral constraints. This paper utilized degradation self-learning to maintain spectral consistency with **MS**.

Employing a convolution layer Conv_{de} with a stride r as an estimator to simulate the spatial degradation, we can down-sample the final fused results $\hat{\mathbf{X}} \in \mathbb{R}^{W \times H \times C}$ to the same size of original $\mathbf{MS} \in \mathbb{R}^{w \times h \times C}$. Let $\hat{\mathbf{X}}^{\downarrow}$ denotes the down-sampled image, the degradation can be formulated as:

$$\hat{\mathbf{X}}^{\downarrow} = \mathrm{Conv}_{de}\left(\hat{\mathbf{X}}\right) \tag{4}$$

As we all know, if the fused image $\hat{\mathbf{X}}$ is the ideal optimal estimation of the target image **X**, their degraded images, *i.e.* $\hat{\mathbf{X}}^{\downarrow}$ and **MS** are also highly similar. After spectral super-resolution injection, $\hat{\mathbf{X}}$ contains fine spatial details while the spectral information might be distorted. The degradation self-learning fully uses the original multispectral image **MS**. It builds a strong degradation constraint with spectral angle mapper (SAM) loss (He et al., 2022a) to guarantee better spectral fidelity.

$$\mathcal{L}_{degra} = \left\| \mathrm{Conv}_{de}\left(\mathrm{sSRI}\left(\mathbf{X}',\ \mathbf{P}\right)\right) - \mathbf{MS} \right\|_{1} + \lambda \mathcal{L}_{SAM} \tag{5}$$

where $\mathrm{sSRI}(\cdot)$ represents the spectral super-resolution injection model, λ is the spectral constraint parameter, and $\mathscr{L}_{SAM}$ denotes the SAM loss as following:

$$\mathscr{L}_{SAM} = \sum_{i=1}^{WH} \cos^{-1}\left(\frac{\langle \hat{\mathbf{X}}_i^{\downarrow}, \mathbf{MS}_i\rangle}{\left\|\hat{\mathbf{X}}_i^{\downarrow}\right\|_2 \|\mathbf{MS}_i\|_2}\right) \tag{6}$$

where $\mathbf{MS}_i$ is the spectral vector of the original multispectral image in the ith pixel, $\hat{\mathbf{X}}_i^{\downarrow}$ is the spectral vector of the degraded fused image in the ith pixel, $\langle\cdot\rangle$ represents the inner product, and $\|\cdot\|_2$ is the L_2 norm.

2.3. Spectral super-resolution injection

Spatial details are crucial in pan-sharpening. However, traditional methods might destroy the spectral fidelity during fusion or cannot extract enough spatial textures. Supervised pan-sharpening algorithms based on deep learning balance spectral fidelity and spatial details well. Nevertheless, their performance depends more on the quality of training datasets. For unsupervised pan-sharpening, CNN-based methods lack the constraints of ground truth and tend to neglect the tiny textures and details in PAN images.

To our knowledge, spatial details between PAN and MS images differ in spatial resolution and spectral domain. Simultaneously addressing two types of information gaps is extremely thorny. Spectral super-resolution aims to improve the spectral resolution of the input data, which can maintain spatial information concurrently.

In this paper, a module named spectral super-resolution injection is proposed to overcome the insufficient spatial constraints in unsupervised pan-sharpening. In sSR injection, the spectral information in the initial fused result $\mathbf{X}'$ is degraded into the PAN domain, eliminating the interference of spectral difference and bringing more benefit for exploring missing spatial details. Subsequently, spectral super-resolution is introduced to transform the missing spatial details from the PAN domain into the MS domain. Finally, ensemble learning is adopted to integrate the informative knowledge in the multiple fused results.

2.3.1. Spectral degradation

As the spatial details between high-resolution PAN images and low-resolution MS images are different in both spatial resolution and spectral domain, to better obtain the missing spatial information in the initial fused result, $\mathbf{X}'$ should be aligned to PAN images in the spectral domain. Moreover, the spectral relationship between MS and PAN sensors should be considered when achieving spectral alignment.

In remote sensing, PAN images can be regarded as the linear weighted sum of MS bands according to spectral response functions, and the processing of linear weighted summation is also commonly known as spectral degradation. The degradation model between MS and PAN images can be formulated as:

$$\mathbf{P} = \mathbf{X}\cdot\boldsymbol{\Phi} \tag{7}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{C\times 1}$ is the spectral response function. For unsupervised pan-sharpening, the spectral response function is unknown. Traditional methods downsample high-resolution PAN images and figure out the weights between $\mathbf{MS}$ and $\mathbf{P}^{\downarrow}$ by using linear regression in preprocessing. In CNNs, the convolution layer with 1×1 kernels equals the linear weighted summation. Considering the string adaptive learning ability, we employed a 1×1 convolution layer Conv_{sd} as spectral degradation:

$$\mathbf{P}' = \mathrm{Conv}_{sd}\left(\mathbf{X}'\right) \tag{8}$$

With the help of spectral degradation, sSRPNet aligns the spatial details in the initial fused results to the original PAN image. Thus, the ignored spatial details can be explored, which is conducive to the further improvement of spatial details.

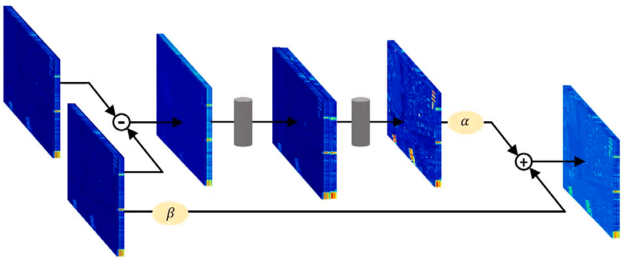


Fig. 2. The proposed Adaptive Information Injection module in spectral super-resolution injection, where the gray cylinders denotes convolution layers. α and β are two learnable parameters.

2.3.2. Spectral super-resolution

Obtaining the spectral-aligned spatial information $\mathbf{P}'$, we hope the model focuses on the missing spatial details compared with the original PAN image $\mathbf{P}$. Thus, $\mathbf{P}'$ and $\mathbf{P}$ are firstly fed into an *Adaptive Information Injection* (AII) module as shown in Fig. 2. The missing spatial details can be integrated into the results $\widetilde{\mathbf{P}}$. The formulation can be written as:

$$\begin{aligned}\widetilde{\mathbf{P}} &= \mathrm{AII}\left(\mathbf{P}', \mathbf{P}\right)\\ \mathrm{AII}\left(\mathbf{P}', \mathbf{P}\right) &= \alpha * \mathrm{Conv}_2\left(\mathrm{Conv}_1\left(\mathbf{P}' - \mathbf{P}\right)\right) + \beta * \mathbf{P}'\end{aligned} \tag{9}$$

where α and β are two learnable adaptive weights, Conv_1 and Conv_2 are two 3×3 convolution layers. α and β represent the weights of original information and enhanced difference information to generate the results $\widetilde{\mathbf{P}}$.

Employing the AII module, spectral super-resolution injection can regain the missing spatial information of $\mathbf{P}'$ with the help of original PAN images. Then, $\widetilde{\mathbf{P}}$ will be fed into a spectral super-resolution module. Spectral super-resolution aims to enhance spectral information and almost does not destroy spatial details. With spectral super-resolution module, $\widetilde{\mathbf{P}}$ will be restored into the MS domain to generate $\mathbf{X}''$ as follows:

$$\widetilde{\mathbf{X}} = \mathsf{F}_{ssr}\left(\widetilde{\mathbf{P}}\right) \tag{10}$$

where F_{ssr} represents the spectral super-resolution module, an arbitrary existing sSR model can be employed here. Processing images only in the spectral domain, F_{ssr} can remain the spatial details in $\widetilde{\mathbf{P}}$ completely.

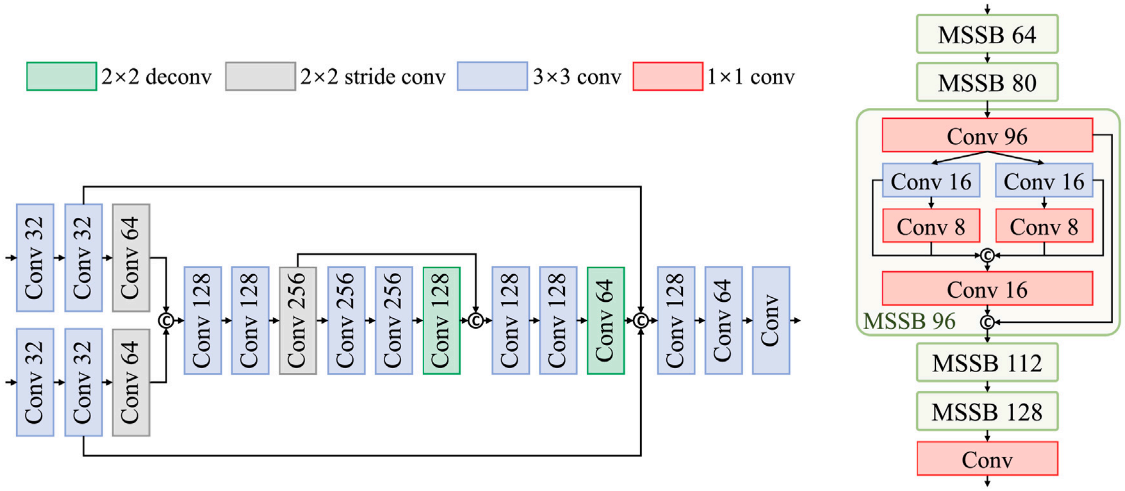
With spectral degradation, the spatial information in the initial fused result has been spectral-aligned to the high-resolution PAN image. Then, the missing spatial details are explored, and spectral super-resolution is utilized to transform the missing spatial information into the MS domain. Obtained enhanced spatial details, the sSR injection further utilizes an AII module to integrate $\mathbf{X}'$ and $\widetilde{\mathbf{X}}$, which significantly increases spatial details in the final results $\mathbf{X}''$:

$$\mathbf{X}'' = \mathrm{AII}\left(\mathbf{X}', \widetilde{\mathbf{X}}\right) \tag{11}$$

Through spectral super-resolution injection, the initial fused results are spectral-aligned to the high-resolution PAN image firstly, and the missing spatial details are retrieved. With the help of spectral super-resolution, the missing spatial details in the PAN domain are transformed into the MS domain and furthered merged into initial results.

2.3.3. Ensemble learning

Deep learning-based algorithms can achieve ideal performance during supervised learning when training datasets are sufficient and complete. With few ground truth as training labels, however, unsupervised learning-based methods always suffer three problems: the statistical problem, the computational problem, and the representation problem (Dietterich et al., 2002; Wang et al., 2022a; Sagi and Rokach, 2018).



(a) Initial fusion model **(b)** Spectral super-resolution module

Fig. 3. The initial fusion model and spectral super-resolution module used in our study, where the number after ‘Conv’ denotes the output channel number of the convolution layer. (a) is a two-stream fusion model with multi-scale feature fusion. (b) is a spectral supre-resolution module with multi-depth spectral–spatial blocks, where the number after ‘MSSB’ is the output channel number of the first convolution layer in this MSSB.

Lack of training data for constraints, there may be several different potential solutions for unsupervised pan-sharpening, but the algorithm can only choose one of these to output (Dong et al., 2020). There is a risk that the chosen solution will not predict the fused results well, and here comes the statistical problem.

The computational problem arises when the algorithm cannot guarantee that the found solution is the best. With various optimization algorithms, including gradient descent and adaptive moment estimation (Adam), unsupervised pan-sharpening can get stuck in local minima and hence fail to find the best solution.

Finally, without any ground truth as training datasets, unsupervised pan-sharpening can hardly obtain any prior that represents a good approximation to the target, called the representational problem.

In this paper, we utilize ensemble learning to overcome three problems that unsupervised pan-sharpening suffers. The most straightforward way is a weighted combination of several potential solutions:

$$\text{Ensemble}\,(\mathscr{X}) = \sum_{i=1}^{k}\left(a_k * \hat{\mathbf{X}}^k\right) \tag{12}$$

where $\mathscr{X} = \{\hat{\mathbf{X}}^1, \dots, \hat{\mathbf{X}}^k\}$ is the set of potential solutions, a_k denotes the weight of $\hat{\mathbf{X}}^k$.

Considering multiple solutions, a weighted sum can reduce the risk of occasional cases and mitigate the statistical problem. Secondly, a combination of several different local minima can reduce the risk of choosing the wrong local minimum to output. Finally, the weighted sum of potential solutions expands the solution space that can be represented, which can also address the representational problem well.

In pan-sharpening, multiple equally-good estimations are around the targets but with tiny differences in different pixel positions. The weighted combination can reduce the bias and variance of estimations and produce more accurate predictions.

2.4. *Initial fusion model & spectral super-resolution module*

We employed a two-stream fusion model with multi-scale feature fusion as the initial fusion model and a spectral super-resolution module with multi-depth spectral–spatial blocks as the spectral super-resolution module in sSRPNet. For easy access to reproduce the proposed sSRPNet, details are as follows.

2.4.1. *Initial fusion model*

In this paper, we utilized a two-stream fusion model with multi-scale feature fusion (TFNet) (Liu et al., 2020b) as the initial fusion model F_θ, as shown in Fig. 3a. Treating MS features and PAN features separately, TFNet shows good performance in supervised pan-sharpening. In TFNet, there are three image scales throughout the network. Fusion between features with the same scale but different depths ensures feature reuse and maintains the memory of shallow features. We removed all the batch-normalization layers, which have been proved to be detrimental to image restoration.

2.4.2. *Spectral super-resolution module*

Seeking a balance between performance and efficiency, we employed a CNN (HSCNN+) with multi-depth spectral–spatial blocks (MSSB) (Shi et al., 2018) as F_{ssr}, as shown in Fig. 3b. There are five MSSBs in the employed sSR module. MSSB extracts spectral features with 1 × 1 convolutions and further extracts spatial features with 3 × 3 convolutions. Multi-depth spectral features and spatial features are concatenated for feature enhancement. Although the channel number of middle features is up to 128, with feature dimension reduction in MSSB, HSCNN+ still achieves fast running time, which keeps a good balance between sSR effort and running speed.

2.5. *Loss functions*

There are two self-learning stages in the proposed sSRPNet, *i.e.* coarsening-scale self-learning and degradation self-learning. The coarsening-scale self-learning loss is the L_1 distance between the coarsening-scale fusion results and the original multispectral image:

$$\mathscr{L}_{sc} = \left\|\mathsf{F}_\theta\left(\mathbf{MS}^{\downarrow},\ \mathbf{P}^{\downarrow}\right) - \mathbf{MS}\right\|_1 \tag{13}$$

The degradation self-learning loss consists of the L_1 distance between the degraded fused result and the original multispectral image and the SAM loss:

$$\mathscr{L}_{degra} = \left\|\hat{\mathbf{X}}^{\downarrow} - \mathbf{MS}\right\|_1 + \lambda\mathscr{L}_{SAM} \tag{14}$$

To simplify the $\mathscr{L}_{SAM}$ and boost the training process, the SAM loss in Eq. (6) equals to a modified L_2 loss:

$$\mathscr{L}_{SAM} = \cos^{-1}\left(1 - \frac{1}{2}\left\|\text{Unit}\left(\hat{\mathbf{X}}^{\downarrow}\right) - \text{Unit}\left(\mathbf{MS}\right)\right\|_2^2\right) \tag{15}$$

where $\mathrm{Unit}(\cdot)$ denotes the pixel-by-pixel vector unitization:

$$\mathrm{Unit}\,(\mathbf{MS}) = \frac{\mathbf{MS}}{\sqrt{\langle \mathbf{MS},\ \mathbf{MS}\rangle}} \tag{16}$$

$\mathscr{L}_{sc}$ and $\mathscr{L}_{degra}$ are mutually independent. In the practical application, we calculate the loss and optimize the model until convergence in two stages.

3. Experimental results

In this section, we show some experiments to verify the superiority of the proposed sSRPNet, including reduced-resolution and full-resolution experiments on three datasets. We also present the ablation study about the proposed sSRPNet. Moreover, we also deploy the proposed sSR injection into other supervised networks and some traditional methods.

3.1. Experimental setting

3.1.1. Experimental datasets

In this paper, three datasets are used for reduced-resolution experiments and full-resolution experiments, including QuickBird (QB), WorldView-2 (WV2), and Gaofen-2 (GF2).

QuickBird acquires a PAN channel (450 to 900 nm) corresponding to a MS image with four channels from the visible to Near-InfraRed (NIR) wavelength range: Blue (450 to 520 nm), Green (520 to 600 nm), Red (630 to 690 nm) and NIR (760 to 900 nm). The spatial resolution of the PAN channel is 61 cm at the sub-satellite point, and the resolution of MS channels is 2.44 m. The QB data used in this paper is selected from the urban area of Shenzhen, China, where roads and buildings are the main parts.

WorldView-2 satellite captures data with one PAN channel (450 to 800 nm) and eight MS channels. The acquired PAN channel has a resolution of 0.46 m, while the MS channels have a resolution of 1.85 m. Besides the standard Blue (450 to 510 nm), Green (510 to 580 nm), and Red (630 to 690 nm), the MS channels also involve Coastal-Blue (400 to 450 nm), Yellow (585 to 625 nm), Red-Edge (705 to 745 nm), NIR1 (770 to 895 nm), and NIR2 (860 to 1040 nm). The WV2 data selected in this study covers San Francisco, USA, one of the world's top travel destinations with various buildings, hills, bays, and trees.

Equipped with two PAN/MS cameras, the Gaofen-2 satellite is capable of collecting images with a ground sampling distance of 0.81 m in the PAN channel (450 to 900 nm) and 3.24 m in the four MS channels, including Blue (450 to 520 nm), Green (520 to 590 nm), Red (630 to 690 nm), and NIR (770 to 890 nm). We choose GF2 data in Nanning for our experiments, which includes diverse topography as the study area, including vegetation, land, and waters.

The selected images acquired by three satellites cover a variety of topography and make great sense to verify model performance thoroughly. All the results and discussions are based on the mentioned-above datasets.

3.1.2. Comparison methods

In this study, we choose seven traditional algorithms and three state-of-the-art deep learning-based methods, including BDSD (Garzelli et al., 2007), Adaptive Component Substitution with Partial Replacement (PRACS) (Choi et al., 2010), Adaptive Gram–Schmidt transformation (GSA) (Aiazzi et al., 2007), ATrous WaveleT (ATWT)-M3 (Ranchin and Wald, 2000), Modulation Transfer Functions-Generalized Laplacian Pyramid-High-Pass Modulation (MTF-GLP-HPM) (Aiazzi et al., 2006), AWLP (Otazu et al., 2005), TV (Palsson et al., 2013), PanNet (Yang et al., 2017), Multi-Scale and Depth Convolutional Neural Network (MSDCNN) (Yuan et al., 2018), and TFNet (Liu et al., 2020b). LDP-Net (Ni et al., 2022) is a unsupervised pan-sharpening method without GAN.

Traditional algorithms can quickly achieve unsupervised pan-sharpening without any training dataset, while deep learning-based algorithms are data-driven. We deploy these methods with the proposed CSSL, training them without other datasets. Using the same training process ensures the fairness of comparisons.

3.1.3. Quantitative metrics

Two types of testing are carried out in this study, including reduced-resolution testing under Wald's protocol (Wald et al., 1997) and full-resolution testing.

For the reduced-resolution testing, six quantitative quality metrics are utilized to evaluate the pan-sharpening performance from spatial and spectral domains, including *Correlation Coefficient* (CC), *Root Mean Squared Error* (RMSE), *mean Peak Signal-to-Noise Ratio* (mPSNR) in decibel units, *mean Structural SIMilarity* (mSSIM) (Wang et al., 2004), *Spectral Angle Mapper* (SAM) (Kruse et al., 1993) in degree, and *Erreur Relative Global Adimensionnelle de Synthèse* (ERGAS). The higher values for CC, mSSIM, and mPSNR show better image quality. On the contrary, lower RMSE, SAM, and ERGAS show less image distortion.

For the full-resolution testing, the spectral distortion index D_λ, spatial distortion index D_s, and *Quality with No Reference* (QNR) are introduced to characterize fusion performance (Alparone et al., 2008). Lower values of D_λ and D_s lead to higher QNR and denotes better performance.

3.1.4. Implementation details

In this paper, the Adam optimization algorithm is employed to train sSRPNet. The learning rates of coarsening-scale and degradation self-learning are 0.002 and 0.001, respectively. The learning rate in coarsening-scale self-learning will be reduced by a scheduler named *ReduceLROnPlateau* when mPNSR has stopped improving. It reads the calculated mPSNR, and the learning rate would be reduced if no improvement is seen for six epochs. The learning rate in degradation self-learning is fixed. All CNN-based methods are trained by Pytorch framework running in the Linux environment with 64 GB RAM and one Nvidia RTX A5000 GPU, and all traditional methods are performed with MATLAB with an Intel CPU (Core i7-8700 @ 3.20 GHz).

3.2. Results

We conduct reduced-resolution and full-resolution experiments on three dataset, including QuickBird, WorldView-2 and Gaofen-2. The reduced-resolution experiments are under Wald's protocol (Wald et al., 1997), where the original MS images are used as target and the down-sampled MS and PAN images are used as input. In the full-resolution experiments, there is no ground truth and full-resolution data are used to generate the fused images. Moreover, the running time of different methods in the full-resolution testing are also compared in this part.

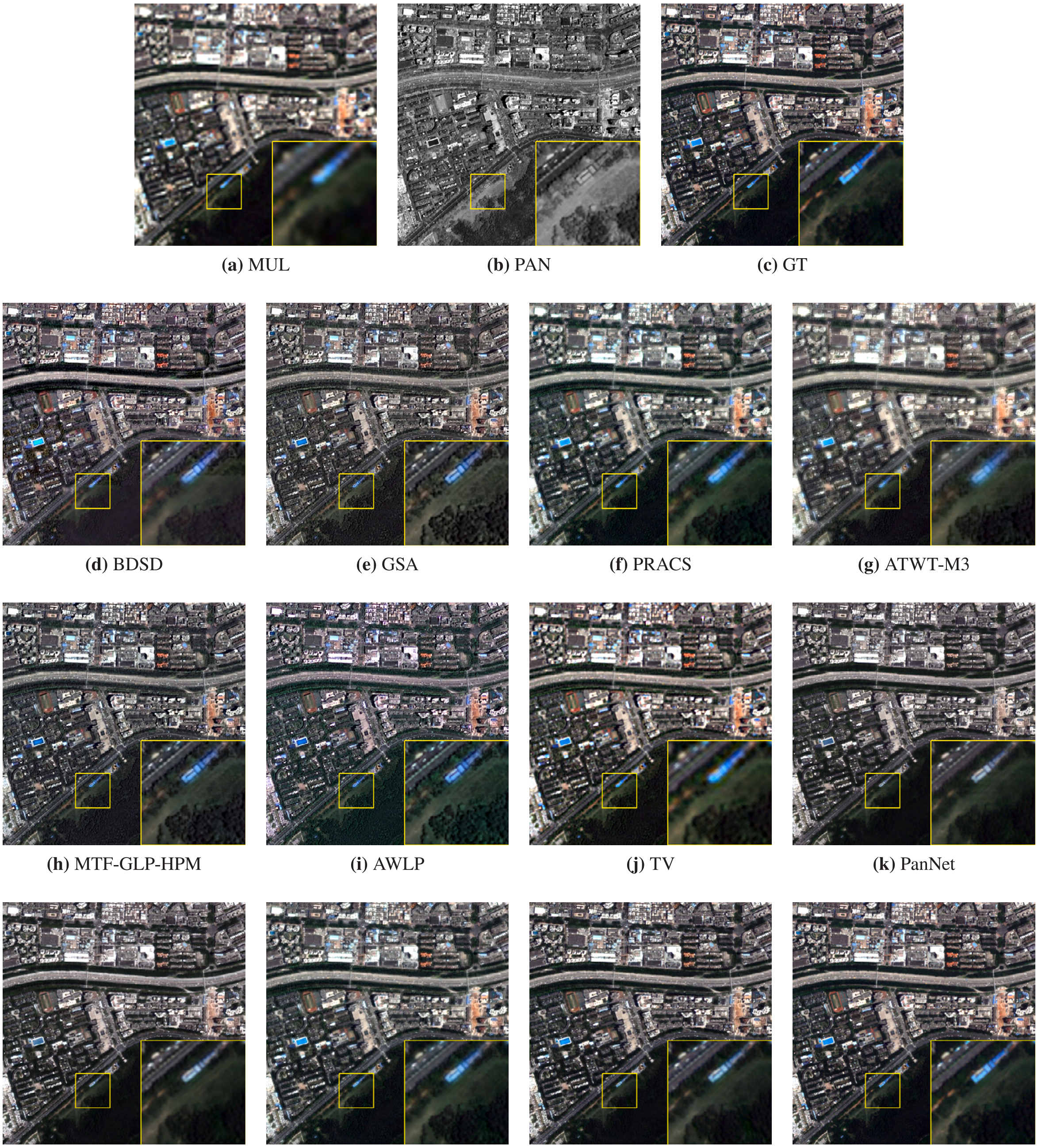
3.2.1. Results on QuickBird dataset

Table 1 reports the quantitative results on the QuickBird dataset, where the best performance is shown in bold and the second best are underlined. Moreover, Fig. 4 displays the visual results in true-color synthesis.

Reduced-resolution experiments. As listed in Table 1, on the QuickBird dataset, BDSD presents the worst performance in spectral fidelity and spatial maintenance in the reduced-resolution testing. Among traditional pan-sharpening algorithms, PRACS outperforms all the others. TV shows the second-best performance and even a slight improvement in SAM than PRACS. The results of deep learning-based methods are better than most traditional model-based algorithms. MSDCNN got the worst performance, while the value of mSSIM is still higher than all traditional pan-sharpening algorithms. TFNet with CSSL has achieved great superiority with even a 3.6 dB improvement on mPSNR than BDSD. Moreover, the SAM also decrease by 1.19. This improvement could be attributed to the two-stream framework that

Table 1
Quantitative assessment on the QuickBird data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution						Full-resolution			time/s
	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS	D_λ	D_s	QNR	
BDSD	0.9242	0.0305	30.5169	0.8478	4.4679	3.9596	0.0579	0.0717	0.8746	0.4039
GSA	0.9242	0.0285	31.0654	0.8099	4.4626	3.8718	0.1398	0.2185	0.6722	1.0459
PRACS	0.9461	0.0214	33.8796	0.8675	3.3744	2.6525	0.0481	0.1133	0.8441	1.9477
ATWT-M3	0.9317	0.0248	32.4974	0.8397	3.8768	3.1345	0.0477	0.0637	0.8917	7.3220
MTF-GLP-HPM	0.9341	0.0261	31.8357	0.8507	3.6952	3.5152	0.1435	0.1930	0.6911	0.9696
AWLP	0.9234	0.0296	31.0402	0.8282	3.4563	3.7308	0.1309	0.1743	0.7176	2.3520
TV	0.9424	0.0233	33.0411	0.8640	3.3349	2.9332	0.0321	0.1046	0.8666	44.3935
PanNet	0.9466	0.0215	33.5651	0.8975	3.4067	2.7409	0.0774	**0.0443**	0.8818	2.7599
MSDCNN	0.9407	0.0223	33.3518	0.8764	3.4416	2.8240	0.0419	0.0782	0.8832	5.7804
LDPNet	0.9419	0.0225	33.3191	0.8694	3.5755	2.8559	0.0352	0.0587	0.9082	43.1455
TFNet	0.9520	0.0202	34.1408	0.8990	3.2742	2.5772	0.0227	0.0537	0.9248	9.1690
sSRPNet	**0.9549**	**0.0194**	**34.5452**	**0.9034**	**3.2085**	**2.4490**	**0.0215**	0.0476	**0.9319**	36.0505



(a) MUL (b) PAN (c) GT

(d) BDSD (e) GSA (f) PRACS (g) ATWT-M3

(h) MTF-GLP-HPM (i) AWLP (j) TV (k) PanNet

(l) MSDCNN (m) LDPNet (n) TFNet (o) **sSRPNet**

Fig. 4. Visual results of reduce-resolution experiments on QuickBird dataset. The band 3, 2, and 1 are integrated for true-color displays.

Table 2
Quantitative assessment on the WorldView-2 data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution						Full-resolution			time/s
	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS	D_λ	D_s	QNR	
BDSD	0.9132	0.0332	30.0298	0.8722	9.5575	5.9012	0.0791	0.0979	0.8308	0.7221
GSA	0.9321	0.0302	30.8290	0.8822	8.1906	5.3249	0.0632	0.0640	0.8768	1.4844
PRACS	0.9162	0.0317	30.7024	0.8370	8.7213	5.9334	0.0214	0.0289	0.9503	5.3244
ATWT-M3	0.8895	0.0413	28.2366	0.7613	8.7565	7.1070	0.0702	0.0324	0.8996	13.9391
MTF-GLP-HPM	0.9196	0.0320	30.3951	0.8500	7.7999	5.6071	0.0565	0.0473	0.8988	1.9673
AWLP	0.9276	0.0302	30.8332	0.8750	8.1005	5.3396	0.0547	0.0426	0.9051	4.5546
TV	0.9131	0.0334	29.8584	0.8362	9.7600	5.8295	0.0642	0.0360	0.9021	89.6807
PanNet	0.9215	0.0316	30.5205	0.8623	8.7232	5.6280	0.0297	0.0378	0.9336	4.6480
MSDCNN	0.9206	0.0321	30.3005	0.8578	7.9002	5.5887	0.0186	**0.0244**	0.9574	11.0148
LDPNet	0.9246	0.0306	30.8589	0.8596	8.1171	5.5694	0.0125	0.0311	0.9567	52.5556
TFNet	0.9330	0.0296	31.0155	0.8822	7.5128	5.2054	0.0037	0.0333	0.9631	11.5462
sSRPNet	**0.9373**	**0.0286**	**31.3189**	**0.8896**	**7.1923**	**5.0344**	**0.0031**	0.0284	**0.9686**	38.6365

can better exploit MS and PAN features. LDPNet shows good spatial fidelity while bad spectral maintaining. The proposed sSRPNet presents the best performance among all comparison methods and shows great superiority to TFNet, which verifies the effectiveness of sSR injection. With degradation learning to spectral consistency constraint, the SAM value decreases by 0.07. Similar conclusions can be found in the visual results, as shown in Fig. 4. The fusion results generated by GSA are influenced more by PAN images, representing over-sharpening. PRACS and TV keep good spectral fidelity to the ground truth, while the TV results suffer from blur. AWLP performs great spectral distortion, and the ATWT-M3 results seem covered by a thin mist. It can be noticed that the colors of blue buildings fused by PanNet, MSDCNN, LDPNet and TFNet are changed, which may be due to the weak constraint to spectra. On the contrary, sSRPNet can keep better spectral fidelity with the ground truth.

Full-resolution experiments. We also conduct full-resolution testing on the QuickBird dataset. As reported in the last four columns of Table 1, ATWT-M3 and BDSD outperform other model-based methods, obtaining a better QNR higher than 0.87. GSA performs the highest D_s, indicating its unsuccessful extraction of spatial details. MTF-GLP-HPM got the highest D_λ. A possible explanation might be that the buildings on QuickBird data are small and dense, which increases the difficulty of spatial feature extraction. In deep learning-based methods, PanNet surprisingly got the lowest D_s while the spectral loss is the highest. LDPNet got a not-bad QNR among all methods. TFNet performs good quantitative results except D_s. With the aid of sSR injection and the degradation learning on the spectral domain, sSRPNet presents the 11.4% decrease on D_s and achieves the highest QNR up to 0.9313, which verifies the benefits of the proposed algorithm. Furthermore, comparing the running time in full-resolution testing, traditional methods achieve great superiority in computational speed. Moreover, the time of data-driven methods includes training and test time. sSRPNet achieves the best fusion performance at a similar time cost to variational optimization models.

3.2.2. Results on WorldView-2 dataset

The second dataset used in this study is from the WorldView-2 satellite. Table 2 lists the quantitative results, and Fig. 5 displays the visual results.

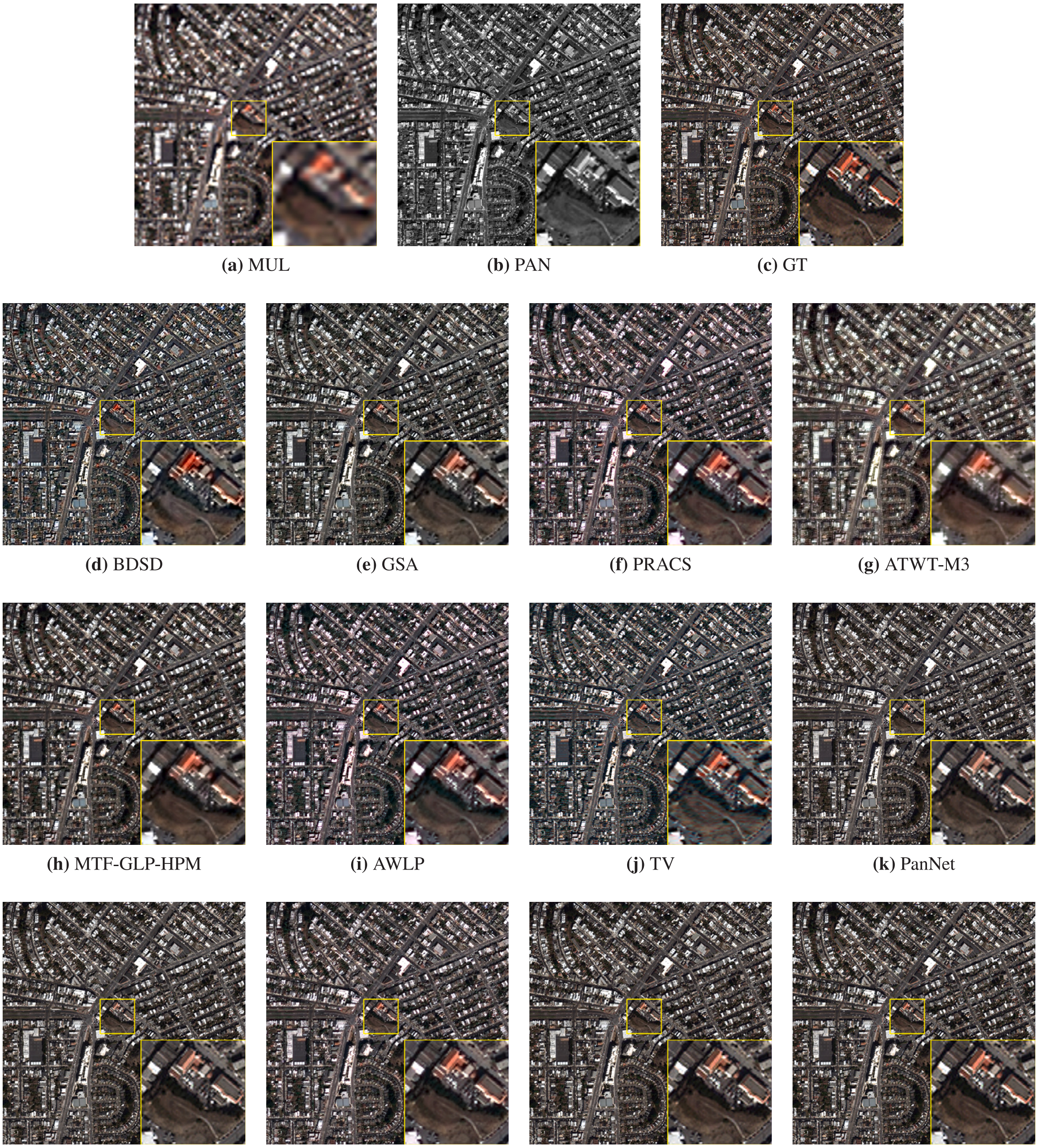
Reduced-resolution experiments. The quantitative results of reduced-resolution experiments on the WorldView-2 dataset also elucidate the advantage of the proposed sSRPNet. Specifically, data-driven methods have a comparative advantage in reducing spectral distortion and obtaining lower SAMs than model-driven methods. MSDCNN shows fluctuated quantitative results, better than PanNet in the spectral domain and worse in the spatial domain. LDPNet got higher CC and mPSNR with also higher SAM. With individual feature extraction modules for MS and PAN images, TFNet can obtain good results. In traditional methods, GSA and AWLP perform well in fusing spatial details, presenting higher mPSNR and mSSIM. MTF-GLP-HPM can keep better spectral fidelity, obtaining a low value of SAM. In Fig. 5, results generated by BDSD and GSA perform poorly on vegetation, showing lighter green on the small hill in the zoom-in area. TV generates some artifacts. From visual results, MTF-GLP-HPM outperforms other model-driven algorithms. ATWT-M3 suffers the same problem as it is on the QuickBird dataset. In data-driven methods, the brick-red buildings of PanNet and MSDCNN show pale red, indicating the unstable spectral maintenance to diverse land covers. These results can be related to the weak spectral constraints of unsupervised pan-sharpening. Compared with TFNet and LDPNet, the proposed sSRPNet maintains good ground truth consistency in both spatial and spectral domains. Nevertheless, compared with the improvement on the QuickBird dataset, the improvement on the WroldView-2 dataset stagnates, demonstrating that the more bands in MS images, the less help spectral super-resolution can provide. The performance of sSR would be strongly influenced by the spectral resolution gap (He et al., 2023).

Full-resolution experiments. As reported in Table 2, the full-resolution results on the WorldView-2 dataset all obtain high QNR, except BDSD. PRACS outperforms other model-driven algorithms vastly, especially in spectral maintenance. TV can integrate enough spatial information while performing poorly in keeping spectral fidelity. In data-driven algorithms, MSDCNN presents the lowest D_s but high D_λ. LDPNet got lower D_λ. sSRPNet achieves the highest QNR, profiting from the excellent spectral maintaining with D_λ as low as 0.0031. Moreover, the D_s of sSRPNet is only 0.004 higher than MSDCNN. Compared with the results generated by TFNet, the improvement is evident, which strongly verifies the benefit of sSR injection.

3.2.3. Results on Gaofen-2 dataset

The last dataset is generated from images captured by the Gaofen-2 satellite. We also conduct reduced and full-resolution tests on the Gaofen-2 dataset. Quantitative results are reported in Table 3, and the visual results in reduced-resolution experiments are depicted in Fig. 6. Moreover, the visual results of full-resolution experiments are also given in Fig. 7, where images are pseudo-color enhanced using bands 4, 3, and 2.

Reduced-resolution experiments. Compared with the other two datasets, the quantitative results on the Gaofen-2 dataset advance incredibly. Among model-driven algorithms, TV presents CC up to 0.9851 while obtaining the lowest mPSNR and the highest SAM. AWLP decreases SAM to 1.7691. Meanwhile, its CC is lower than most algorithms. ATWT-M3 performs best in quantitative indexes evaluating spatial details among model-driven algorithms, presenting the lowest ERGAS as 1.9398. In data-driven algorithms, except PanNet, all methods obtain a high value of mPSNR over 40. This result may be explained by the fact that coarsening-scale self-learning helps data-driven methods better explore the internal information from a coarsening scale. LDPNet got lower SAM than PanNet and MSDCNN while lower CC and mPSNR, which indicates that LDPNet cannot learn spatial and spectral degradation well simultaneously. TFNet can only achieve somewhat



(**a**) MUL (**b**) PAN (**c**) GT

(**d**) BDSD (**e**) GSA (**f**) PRACS (**g**) ATWT-M3

(**h**) MTF-GLP-HPM (**i**) AWLP (**j**) TV (**k**) PanNet

(**l**) MSDCNN (**m**) LDPNet (**n**) TFNet (**o**) **sSRPNet**

Fig. 5. Visual results of reduced-resolution experiments on WorldView-2 dataset. The band 5, 3, and 2 are integrated for true-color displays.

Table 3
Quantitative assessment on the Gaofen-2 data. The best performance are shown in **bold** and the second best are underlined.

Methods	Reduced-resolution						Full-resolution			time/s
	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS	D_λ	D_s	QNR	
BDSD	0.9382	0.0147	36.7443	0.9220	2.6796	2.9619	0.0439	0.0742	0.8852	0.3964
GSA	0.9594	0.0116	38.7976	0.9382	2.0583	2.2794	0.0560	0.0943	0.8550	1.0055
PRACS	0.9239	0.0125	38.1810	0.9026	2.0698	2.3156	0.0460	0.1667	0.7950	1.8487
ATWT-M3	0.9666	0.0088	41.5168	0.9584	2.2409	1.9398	0.0614	0.0319	0.9087	6.8032
MTF-GLP-HPM	0.9587	0.0107	39.6281	0.9442	1.8201	2.2352	0.1013	0.0901	0.8177	0.9112
AWLP	0.9479	0.0122	38.3235	0.9288	**1.7691**	2.3637	0.1006	0.0876	0.8206	2.2504
TV	**0.9851**	0.0155	36.6299	0.9681	4.0424	3.1703	0.0285	0.1006	0.8737	43.0714
PanNet	0.9515	0.0111	39.2144	0.9444	2.4907	2.2622	0.0772	0.0341	0.8913	2.6947
MSDCNN	0.9632	0.0085	41.9825	0.9670	2.5824	1.7650	0.0715	0.0510	0.8812	5.8283
LDPNet	0.9433	0.0111	39.2334	0.9271	2.0159	2.1528	0.0341	0.1128	0.8575	76.4444
TFNet	0.9627	0.0098	40.2858	0.9516	1.9620	1.9900	0.0222	0.0590	0.9201	16.4734
sSRPNet	0.9771	**0.0071**	**43.3256**	**0.9723**	1.8321	**1.5473**	**0.0160**	**0.0233**	**0.9611**	57.0433

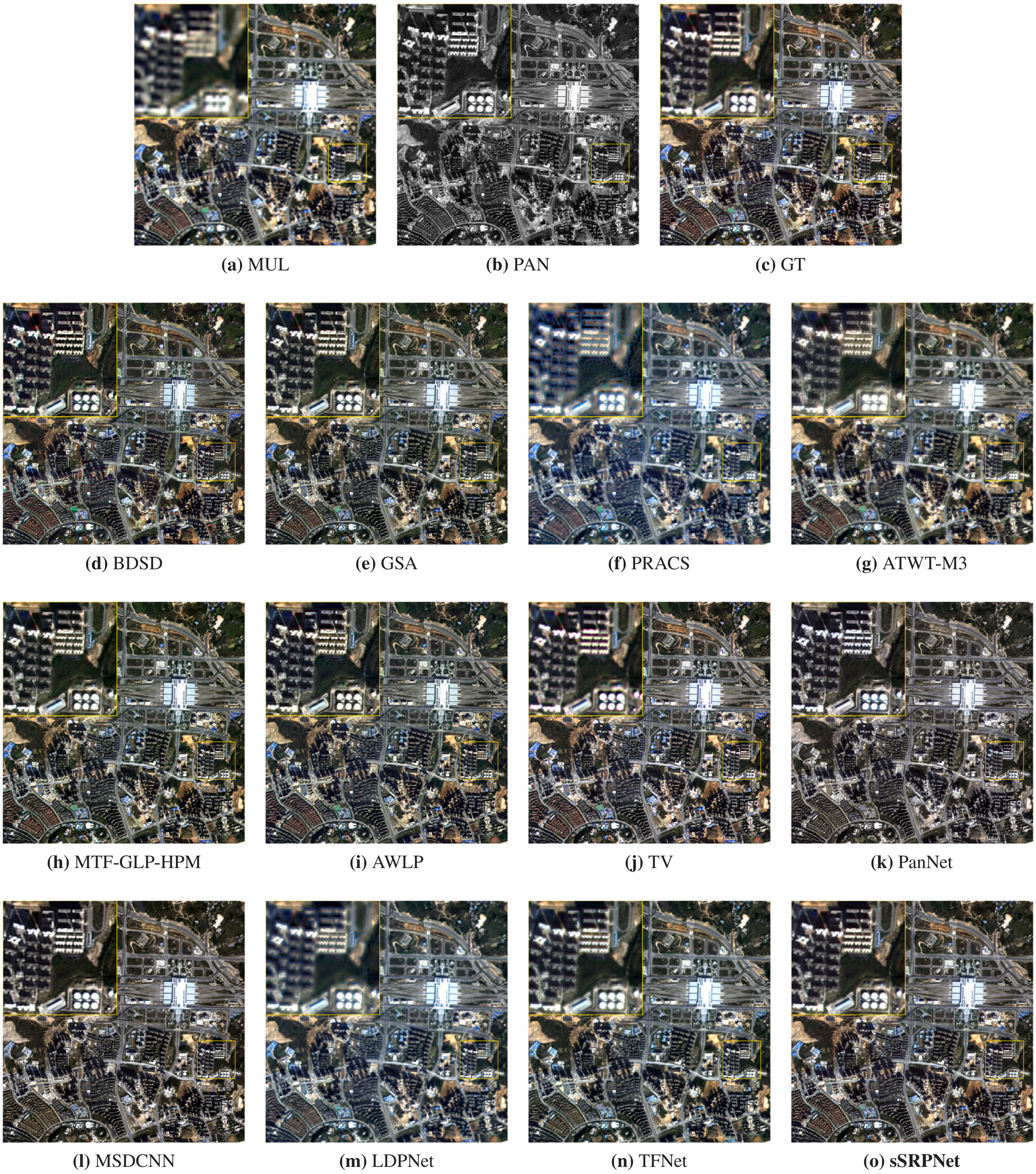


Fig. 6. Visual results of full-resolution experiments on Gaofen-2 dataset. The band 3, 2, and 1 are integrated for true-color displays.

satisfactory spectral maintenance, while its CC, mPSNR, and mSSIM are not high enough. With the aid of sSR injection, sSRPNet keeps the low SAM and improves the quantitative metrics evaluating spatial details, strongly verifying its capacity to extract the missing spatial details. Visual results in Fig. 6 also prove the earlier conclusions. TV suffers from blur and low color consistency to the ground truth. PRACS generates severe artifacts, which may be due to inaccurate local instability adjustment parameters. In data-driven algorithms, PanNet suffers severe color distortion, and TFNet shows insufficient spatial details. LDPNet suffers serious artifacts. Among all the comparison methods, our proposed sSRPNet achieves the best performance, which two reasons can explain. Firstly, sSR injection provides more spatial details for pan-sharpening. Secondly, dual-stage self-learning ensures the unsupervised pan-sharpening and strong spectral constraint.

Full-resolution experiments. In the quantitative results of full-resolution testing on the Gaofen-2 dataset, the proposed sSRPNet outperforms all the comparison methods and achieves a high QNR up to 0.9611. It should be noticed that TFNet presents a middling D_s which is even higher than ATWT-M3, although TFNet achieves good D_λ. After employing sSR injection and degradation self-learning, the value of D_s drops to half of the original D_s, demonstrating that spectral super-resolution can inject fine spatial details for pan-sharpening. We also give the visual results in this part, and images are in false color, where the band combination is band 4, 3, and 2. As shown in Fig. 7, BDSD, ATWT-M3, LDPNet and TV present some blur in buildings. PRACS and ATWT-M3 fail to inject finer spatial details. PanNet and MSDCNN suffer strong color distortion. With the help of spectral super-resolution, the proposed sSRPNet shows more spatial details than TFNet.

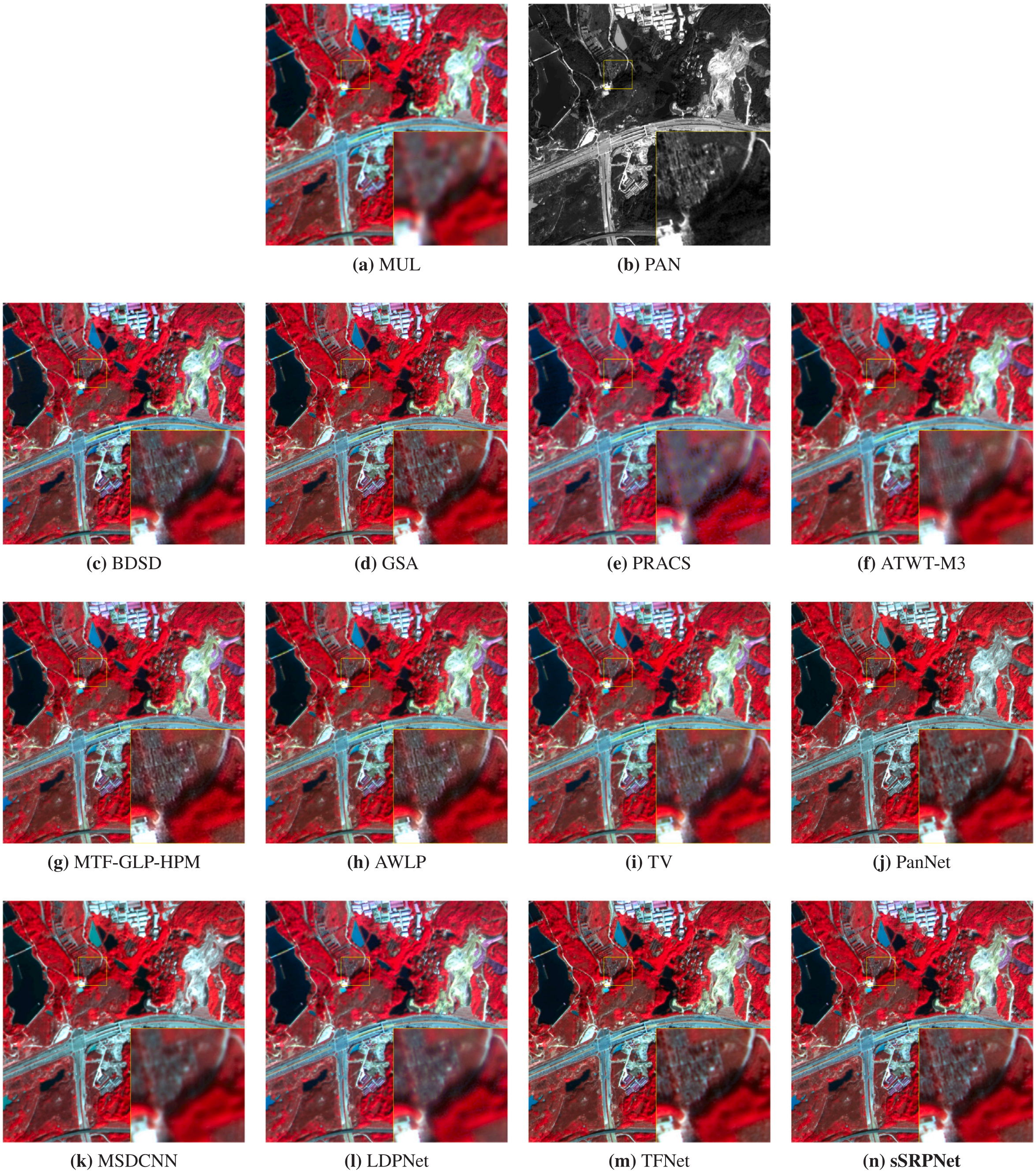


Fig. 7. Visual results of full-resolution experiments on Gaofen-2 dataset. The band 4, 3, and 2 are integrated for false-color displays.

3.3. Ablation study

The proposed sSRPNet consists of two main stages: initial fusion and sSR injection. sSR injection consists of two AII modules and a sSR module. We present an ablation study to verify the effectiveness of the proposed modules. Moreover, the modified SAM loss function is also discussed in this part. We select the most classical deep learning-based method PanNet as the baseline. PanNet is a supervised algorithm, and without an extra training dataset, PanNet can hardly perform well. To keep the fairness of experiments, we train it with CSSL.

Table 4 reports the quantitative results of the ablation study on the Gaofen-2 dataset. The best performance is shown in bold, and the second best is underlined. Compared with PanNet, IFN with CSSL reaches higher quantitative metrics, which reveals that separately addressing MS and PAN features is more beneficial to pan-sharpening.

When the sSR module is used in sSRPNet, spectral degradation and degradation self-learning should also be utilized. Moreover, when AII modules are removed, the features are fused by a concatenation operation followed by a 1 × 1 convolution. Without any AII modules, results generated by *sSRPNet w/o AII* obtain lower spatial fidelity to ground truth and more spectral distortion, indicating that ordinary convolutions can hardly explore the missing spatial details in the initial results. Comparing the third and the fourth line, the AII module after the sSR module seems to be more significant in sSRPNet, which may arise from the fact that AII2 is at the end of sSRPNet and can be more easily optimized by back-propagation gradients.

Significantly, with sSR injection, all the metrics about the spatial domain reach higher values, while the index of spectral error, SAM, gets worse. On the one hand, this phenomenon suggests that sSR injection can provide more spatial details for unsupervised pan-sharpening. On

Table 4
Ablation study of the proposed sSRPNet, including sSR Injection and SAM loss. We choose PanNet as the baseline.

Methods	IFN	sSR Injection			SAM Loss	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS
		AII1	sSR	AII2							
PanNet	✗	✗	✗	✗	✗	0.9515	0.0111	39.2144	0.9444	2.4907	2.2622
IFN_CSSL	✓	✗	✗	✗	✗	0.9627	0.0098	40.2858	0.9516	1.9620	1.9900
sSRPNet w/o AII	✓	✗	✓	✗	✗	0.9624	0.0101	40.0122	0.9592	2.4062	2.0938
sSRPNet w/o AII2	✓	✓	✓	✗	✗	0.9702	0.0091	41.2032	0.9603	2.6262	1.9729
sSRPNet w/o AII1	✓	✗	✓	✓	✗	0.9694	0.0085	41.6701	0.9659	2.2307	1.7947
sSRPNet w/o SAM	✓	✓	✓	✓	✗	0.9737	0.0079	42.3722	0.9700	2.0017	1.6947
sSRPNet	✓	✓	✓	✓	✓	**0.9768**	**0.0071**	**43.4302**	**0.9720**	**1.8947**	**1.5541**

Table 5
Quantitative assessment of different sSR modules in sSRPNet on the Gaofen-2 data.

Modules	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS	time/s
SSDCNN	0.9737	0.0076	42.7271	0.9687	1.9696	1.6417	114
GDNet	0.9744	0.0075	42.9003	0.9713	1.8915	1.6100	507
CanNet	0.9697	0.0083	41.8590	0.9629	1.9623	1.7509	110
PoNet	0.9718	0.0080	42.2404	0.9650	1.9576	1.7077	563
HSCNN+	0.9768	0.0071	43.4302	0.9720	1.8947	1.5541	184

Table 6
The improvement that sSR injection helps the existing pan-sharpening achieve on Gao-fen 2 dataset. The percentages denotes the decrease or the increase that sSR injection achieves, where red values present increases and blue values present decreases.

Initial Fusion	W/O sSR Injection						With sSR Injection					
	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS	CC	RMSE	mPSNR	mSSIM	SAM	ERGAS
BDSD	0.9382	0.0147	36.7443	0.9220	2.6796	2.9619	0.9447 (+0.7%)	0.0122 (−16.8%)	38.3387 (+4.3%)	0.9417 (+2.1%)	2.3102 (−13.8%)	2.4844 (−16.1%)
GSA	0.9594	0.0116	38.7976	0.9382	2.0583	2.2794	0.9659 (+0.7%)	0.0095 (−17.8%)	40.4920 (+4.4%)	0.9530 (+1.6%)	1.9577 (−4.9%)	1.9209 (−15.7%)
PRACS	0.9239	0.0125	38.1810	0.9026	2.0698	2.3156	0.9281 (+0.5%)	0.0123 (−1.4%)	38.2602 (+0.2%)	0.9071 (+0.5%)	2.0719 (+0.1%)	2.3011 (−0.6%)
ATWT-M3	0.9666	0.0088	41.5168	0.9584	2.2409	1.9398	0.9664 (−0.0%)	0.0088 (−0.2%)	41.5353 (+0.0%)	0.9595 (+0.1%)	2.2386 (−0.1%)	1.9385 (−0.1%)
MTF-GLP-HPM	0.9587	0.0107	39.6281	0.9442	1.8201	2.2352	0.9687 (+1.0%)	0.0083 (−22.2%)	42.0533 (+6.1%)	0.9607 (+1.7%)	1.8910 (+3.9%)	1.8407 (−17.6%)
AWLP	0.9479	0.0122	38.3235	0.9288	1.7691	2.3637	0.9706 (+2.4%)	0.0082 (−32.7%)	41.9633 (+9.5%)	0.9606 (+3.4%)	1.8871 (+6.7%)	1.7285 (−26.9%)
TV	0.9851	0.0155	36.6299	0.9681	4.0424	3.1703	0.9811 (−0.4%)	0.0066 (−57.5%)	44.0635 (+20.3%)	0.9772 (+0.9%)	1.6884 (−58.2%)	1.4411 (−54.5%)
PanNet	0.9515	0.0111	39.2144	0.9444	2.4907	2.2622	0.9643 (+1.3%)	0.0088 (−20.7%)	41.5094 (+5.9%)	0.9629 (+2.0%)	2.3503 (−5.6%)	1.8863 (−16.6%)
MSDCNN	0.9632	0.0085	41.9825	0.9670	2.5824	1.7650	0.9642 (+0.1%)	0.0081 (−4.7%)	42.6466 (+1.6%)	0.9689 (+0.2%)	2.3537 (−8.9%)	1.7114 (−3.0%)

the other hand, it indicates that the spectral information generated by the sSR module is not ideal. Employing the modified SAM loss function, the proposed sSRPNet achieves better spectral maintenance and finer spatial details.

3.4. Influence of different sSR modules

To discuss the influence of different sSR modules in sSRPNet, we select some classical spectral super-resolution models and compare their performance on the Gaofen-2 dataset. Table 5 lists the quantitative results. The used sSR modules include SSDCNN (Chen et al., 2022), GDNet (Zhu et al., 2021), CanNet (Can and Timofte, 2018), PoNet (He et al., 2022d) and HSCNN+ (Shi et al., 2018).

The selected methods are all lightweight and fast-running. The fastest method is CanNet, while also obtaining the lowest quantitative assessment. SSDCNN and PoNet show considerable results. Among them, SSDCNN converges faster, while PoNet can hardly be optimal with a small iteration. GDNet obtains ideal results and even the best spectral fidelity. However, GDNet also takes a long time to reach the best performance. Compared with other sSR modules, HSCNN+ achieves the best spatial fidelity and the second-best spectral maintenance with an acceptable computational speed. Thus, we employed HSCNN+ as the sSR module in the proposed sSRPNet.

3.5. sSR injection helps the existing pan-sharpening a lot

In this part, to investigate the potential of the proposed sSR injection for improving the existing pan-sharpening results, we conduct further experiments on the Gaofen-2 dataset. We deploy sSR injection after the existing pan-sharpening methods, and the quantitative results are listed in Table 6. We calculate the changing percentages of the assessment indexes after sSR injection is adopted, where red values present increases and blue values present decreases.

We can quickly conclude that sSR injection can improve the results generated by existing pan-sharpening algorithms, where the increase of mPSNR and mSSIM is evident. Moreover, RMSE and ERGAS also decrease obviously. Among these methods, some algorithms that do not perform well in spatial enhancement are improved even more, such as BDSD, AWLP, and TV. Results thoroughly verify the spatial enhancement ability of the proposed sSR injection. Most algorithms' performance on spatial details and spectral fidelity is enhanced with sSR injection, except AWLP and MTF-GLP-HPM. The possible reason may be that AWLP and MTF-GLP-HPM have presented ideal SAM and can hardly be further improved without other extra priors. Remarkably, TV with sSR injection can present the mPSNR up to 44.0635, and the SAM decreases to 1.6884, which is even better than the proposed sSRPNet. However, outstanding performance brings high time costs, leading to low practicability.

In brief, the proposed sSR injection can be applied to other pan-sharpening algorithms and improve their performance, especially for integrating spatial details. In practice, sSR injection considers the fused image as the initial result, and the further improvement is plug-and-play.

4. Conclusions

Different low-level vision tasks focus on different aims and sometimes might benefit each other. It is known that fine spatial details are highly significant for unsupervised pan-sharpening. Furthermore, spectral super-resolution can only improve spectral resolution without spatial information loss. In this paper, we proposed a dual-stage self-learning pan-sharpening network with spectral super-resolution injection. Coarsening-scale self-learning explores the internal information at a coarsening scale and achieves initial unsupervised pan-sharpening. After down-sampling the spectral information of the initial results to the PAN domain, spectral super-resolution injection extracts the missing spatial details and recovers better MS images by realigning details into the MS domain. Finally, degradation self-learning keeps the strong spectral constraints. In experiments, the proposed sSRPNet shows remarkable advantages in both reduced-resolution and full-resolution experiments on three satellite datasets. As a plug-and-play module, sSR injection can be easily applied to the existing pan-sharpening algorithms with considerable improvement. Playing a role as spectral alignment, spectral super-resolution helps the proposed sSRPNet maintain fine spatial details. Meanwhile, we also find that the spectral fidelity would decreases in some cases. Thus, introducing spatial super-resolution for better spectral fidelity would also be considered in our future work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 42230108, Grant 61971319, and Grant 62071341; and in part by the Fundamental Research Funds for the Central Universities under Grant 2042023KFYQ04 and Grant 2042023KF1007.

References

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M., 2006. MTF-tailored multiscale fusion of high-resolution MS and pan imagery. Photogramm. Eng. Remote Sens. 72 (5), 591–596.

Aiazzi, B., Baronti, S., Selva, M., 2007. Improving component substitution pansharpening through multivariate regression of MS + pan data. IEEE Trans. Geosci. Remote Sens. 45 (10), 3230–3239.

Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M., 2008. Multispectral and panchromatic data fusion assessment without reference. Photogramm. Eng. Remote Sens. 74 (2), 193–200.

Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B., 2006. A variational model for P+XS image fusion. Int. J. Comput. Vis. 69 (1), 43–58.

Burt, P.J., Adelson, E.H., 1987. The Laplacian pyramid as a compact image code. In: Readings in Computer Vision. Elsevier, pp. 671–679.

Can, Y.B., Timofte, R., 2018. An efficient CNN for spectral reconstruction from RGB images. arXiv preprint arXiv:1804.04647.

Carper, W., Lillesand, T., Kiefer, R., 1990. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. Photogramm. Eng. Remote Sens 56 (4), 459–467.

Chen, W., Zheng, X., Lu, X., 2022. Semisupervised spectral degradation constrained network for spectral super-resolution. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Choi, J., Yu, K., Kim, Y., 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE Trans. Geosci. Remote Sens. 49 (1), 295–309.

Ciotola, M., Vitale, S., Mazza, A., Poggi, G., Scarpa, G., 2022. Pansharpening by convolutional neural networks in the full resolution framework. IEEE Trans. Geosci. Remote Sens.

Colomina, I., Molina, P., 2014. Unmanned aerial systems for photogrammetry and remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 92, 79–97.

Deng, L.J., Vivone, G., Paoletti, M.E., Scarpa, G., He, J., Zhang, Y., Chanussot, J., Plaza, A., 2022. Machine learning in pansharpening: A benchmark, from shallow to deep networks. IEEE Geosci. Remote Sens. Mag. 2–38.

Dietterich, T.G., et al., 2002. Ensemble learning. In: The Handbook of Brain Theory and Neural Networks, vol. 2, (no. 1), MIT press Cambridge, MA, USA, pp. 110–125.

Do, M.N., Vetterli, M., 2005. The contourlet transform: An efficient directional multiresolution image representation. IEEE Trans. Image Process. 14 (12), 2091–2106.

Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q., 2020. A survey on ensemble learning. Front. Comput. Sci. 14, 241–258.

Fu, X., Lin, Z., Huang, Y., Ding, X., 2019. A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10265–10274.

Garzelli, A., Nencini, F., Capobianco, L., 2007. Optimal MMSE pan sharpening of very high resolution multispectral images. IEEE Trans. Geosci. Remote Sens. 46 (1), 228–236.

Gastineau, A., Aujol, J.F., Berthoumieu, Y., Germain, C., 2021. Generative adversarial network for pansharpening with spectral and spatial discriminators. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

Gillespie, A.R., Kahle, A.B., Walker, R.E., 1987. Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques. Remote Sens. Environ. 22 (3), 343–365.

González-Audícana, M., Saleta, J.L., Catalán, R.G., García, R., 2004. Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition. IEEE Trans. Geosci. Remote Sens. 42 (6), 1291–1299.

Guo, Q., Li, S., Li, A., 2022. An efficient dual spatial–spectral fusion network. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

He, J., Li, J., Yuan, Q., Shen, H., Zhang, L., 2022a. Spectral response function-guided deep optimization-driven network for spectral super-resolution. IEEE Trans. Neural Netw. Learn. Syst. 33 (9), 4213–4227.

He, D., Shi, Q., Liu, X., Zhong, Y., Zhang, L., 2022b. Generating 2 m fine-scale urban tree cover product over 34 metropolises in China based on deep context-aware sub-pixel mapping network. Int. J. Appl. Earth Obs. Geoinf. 106, 102667.

He, J., Yuan, Q., Li, J., Xiao, Y., Liu, D., Shen, H., Zhang, L., 2023. Spectral super-resolution meets deep learning: Achievements and challenges. Inf. Fusion.

He, J., Yuan, Q., Li, J., Xiao, Y., Liu, X., Zou, Y., 2022. Dster: a dense spectral transformer for remote sensing spectral super-resolution. Int. J. Appl. Earth Obs. Geoinf. 109, 102773.

He, J., Yuan, Q., Li, J., Zhang, L., 2022c. A knowledge optimization-driven network with normalizer-free group ResNet prior for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

He, J., Yuan, Q., Li, J., Zhang, L., 2022d. PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images. Inf. Fusion 80, 205–225.

Huang, Y., Chen, Z.x., Tao, Y., Huang, X.z., Gu, X.f., 2018. Agricultural remote sensing big data: Management and applications. J. Integr. Agric. 17 (9), 1915–1931.

Javan, F.D., Samadzadegan, F., Mehravar, S., Toosi, A., Khatami, R., Stein, A., 2021. A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery. ISPRS J. Photogramm. Remote Sens. 171, 101–117.

Kruse, F.A., Lefkoff, A., Boardman, J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., 1993. The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44 (2–3), 145–163.

Kwarteng, P., Chavez, A., 1989. Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. Photogramm. Eng. Remote Sens. 55 (1), 339–348.

Laben, C.A., Brower, B.V., 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6, 011, 875.

Li, J., Sun, W., Jiang, M., Yuan, Q., 2021. Self-supervised pansharpening based on a cycle-consistent generative adversarial network. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Li, Z., Zhang, H., Lu, F., Xue, R., Yang, G., Zhang, L., 2022. Breaking the resolution barrier: A low-to-high network for large-scale high-resolution land-cover mapping using low-resolution labels. ISPRS J. Photogramm. Remote Sens. 192, 244–267.

Liao, L., Xiao, J., Yang, Y., Ma, X., Wang, Z., Satoh, S., 2023. High temporal frequency vehicle counting from low-resolution satellite images. ISPRS J. Photogramm. Remote Sens. 198, 45–59.

Liu, J., Feng, Y., Zhou, C., Zhang, C., 2020a. Pwnet: An adaptive weight network for the fusion of panchromatic and multispectral images. Remote Sens. 12 (17), 2804.

Liu, X., Liu, Q., Wang, Y., 2020b. Remote sensing image fusion based on two-stream fusion network. Inf. Fusion 55, 1–15.

Liu, Q., Meng, X., Shao, F., Li, S., 2023. Supervised-unsupervised combined deep convolutional neural networks for high-fidelity pansharpening. Inf. Fusion 89, 292–304.

Liu, Q., Zhou, H., Xu, Q., Liu, X., Wang, Y., 2020c. PSGAN: A generative adversarial network for remote sensing image pan-sharpening. IEEE Trans. Geosci. Remote Sens. 59 (12), 10227–10242.

Luo, S., Zhou, S., Feng, Y., Xie, J., 2020. Pansharpening via unsupervised convolutional neural networks. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4295–4310.

Ma, J., Yu, W., Chen, C., Liang, P., Guo, X., Jiang, J., 2020. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. Inf. Fusion 62, 110–120.

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by convolutional neural networks. Remote Sens. 8 (7), 594.

Mulverhill, C., Coops, N.C., Achim, A., 2023. Continuous monitoring and sub-annual change detection in high-latitude forests using harmonized landsat sentinel-2 data. ISPRS J. Photogramm. Remote Sens. 197, 309–319.

Nason, G.P., Silverman, B.W., 1995. The stationary wavelet transform and some statistical applications. In: Wavelets and Statistics. Springer, pp. 281–299.

Ni, J., Shao, Z., Zhang, Z., Hou, M., Zhou, J., Fang, L., Zhang, Y., 2022. LDP-Net: An unsupervised pansharpening network based on learnable degradation processes. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 5468–5479.

Otazu, X., González-Audícana, M., Fors, O., Núñez, J., 2005. Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods. IEEE Trans. Geosci. Remote Sens. 43 (10), 2376–2385.

Ozcelik, F., Alganci, U., Sertel, E., Unal, G., 2020. Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANs. IEEE Trans. Geosci. Remote Sens. 59 (4), 3486–3501.

Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., 2013. A new pansharpening algorithm based on total variation. IEEE Geosci. Remote Sens. Lett. 11 (1), 318–322.

Qu, Y., Baghbaderani, R.K., Qi, H., Kwan, C., 2020. Unsupervised pansharpening based on self-attention mechanism. IEEE Trans. Geosci. Remote Sens. 59 (4), 3192–3208.

Ranchin, T., Wald, L., 2000. Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation. Photogramm. Eng. Remote Sens. 66 (1), 49–61.

Rao, Y., He, L., Zhu, J., 2017. A residual convolutional neural network for pan-shaprening. In: 2017 International Workshop on Remote Sensing with Intelligent Processing. RSIP, IEEE, pp. 1–4.

Sagi, O., Rokach, L., 2018. Ensemble learning: A survey. Wiley Interdiscipl. Rev.: Data Min. Knowl. Discov. 8 (4), e1249.

Seo, S., Choi, J.-S., Lee, J., Kim, H.-H., Seo, D., Jeong, J., Kim, M., 2020. UPSNet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images. IEEE Access 8, 201199–201217.

Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (5), 1656–1669.

Shao, Z., Lu, Z., Ran, M., Fang, L., Zhou, J., Zhang, Y., 2019. Residual encoder–decoder conditional generative adversarial network for pansharpening. IEEE Geosci. Remote Sens. Lett. 17 (9), 1573–1577.

Shi, Z., Chen, C., Xiong, Z., Liu, D., Wu, F., 2018. Hscnn+: Advanced cnn-based hyperspectral recovery from rgb images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 939–947.

Starck, J.L., Candès, E.J., Donoho, D.L., 2002. The curvelet transform for image denoising. IEEE Trans. Image Process. 11 (6), 670–684.

Uezato, T., Hong, D., Yokoya, N., He, W., 2020. Guided deep decoder: Unsupervised image pair fusion. In: 16th European Conference on Computer Vision, Vol. 12351. ECCV 2020, Springer, pp. 87–102.

Vivone, G., Simões, M., Dalla Mura, M., Restaino, R., Bioucas-Dias, J.M., Licciardi, G.A., Chanussot, J., 2014. Pansharpening based on semiblind deconvolution. IEEE Trans. Geosci. Remote Sens. 53 (4), 1997–2010.

Wald, L., Ranchin, T., Mangolini, M., 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogramm. Eng. Remote Sens. 63 (6), 691–699.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, Y., Yuan, Q., Li, T., Zhu, L., Zhang, L., 2021. Estimating daily full-coverage near surface O3, CO, and NO2 concentrations at a high spatial resolution over China based on S5P-TROPOMI and GEOS-FP. ISPRS J. Photogramm. Remote Sens. 175, 311–325.

Wang, Y., Yuan, Q., Zhou, S., Zhang, L., 2022a. Global spatiotemporal completion of daily high-resolution TCCO from TROPOMI over land using a swath-based local ensemble learning method. ISPRS J. Photogramm. Remote Sens. 194, 167–180.

Wang, D., Zhang, P., Bai, Y., Li, Y., 2022b. MetaPan: Unsupervised adaptation with meta-learning for multispectral pansharpening. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Wei, Y., Yuan, Q., Shen, H., Zhang, L., 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci. Remote Sens. Lett. 14 (10), 1795–1799.

Wu, Z.C., Huang, T.Z., Deng, L.J., Huang, J., Chanussot, J., Vivone, G., 2023a. LRTCFPan: Low-rank tensor completion based framework for pansharpening. IEEE Trans. Image Process. 32, 1640–1655.

Wu, Z.C., Huang, T.Z., Deng, L.J., Vivone, G., 2023b. A framelet sparse reconstruction method for pansharpening with guaranteed convergence. Inverse Probl. Imaging 17 (6), 1277–1300.

Wu, J., Lin, L., Zhang, C., Li, T., Cheng, X., Nan, F., 2023c. Generating sentinel-2 all-band 10-m data by sharpening 20/60-m bands: A hierarchical fusion network. ISPRS J. Photogramm. Remote Sens. 196, 16–31.

Xiao, Y., Yuan, Q., He, J., Zhang, Q., Sun, J., Su, X., Wu, J., Zhang, L., 2022. Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer. Int. J. Appl. Earth Obs. Geoinf. 108, 102731.

Xiao, Y., Yuan, Q., Jiang, K., He, J., Wang, Y., Zhang, L., 2023. From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution. Inf. Fusion 96, 297–311.

Xiong, Z., Guo, Q., Liu, M., Li, A., 2020. Pan-sharpening based on convolutional neural network by using the loss function with no-reference. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 897–906.

Xu, L., Chen, N., Moradkhani, H., Zhang, X., Hu, C., 2020. Improving global monthly and daily precipitation estimation by fusing gauge observations, remote sensing, and reanalysis data sets. Water Resour. Res. 56 (3), e2019WR026444.

Xu, Q., Li, Y., Nie, J., Liu, Q., Guo, M., 2023. UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network. Inf. Fusion 91, 31–46.

Yang, J., Fu, X., Hu, Y., Huang, Y., Ding, X., Paisley, J., 2017. PanNet: A deep network architecture for pan-sharpening. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5449–5457.

Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L., 2018. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 11 (3), 978–989.

Zeng, D., Hu, Y., Huang, Y., Xu, Z., Ding, X., 2016. Pan-sharpening with structural consistency and $\ell 1/2$ gradient prior. Remote Sens. Lett. 7 (12), 1170–1179.

Zhang, L., He, J., Yang, Q., Xiao, Y., Yuan, Q., 2022. Data-driven multi-source remote sensing data fusion: Progress and challenges. Acta Geodaetica et Cartographica Sin. 51 (7), 1317–1337.

Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019. Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans. Geosci. Remote Sens. 57 (8), 5549–5563.

Zhang, H., Ma, J., 2021. GTP-PNet: A residual learning network based on gradient transformation prior for pansharpening. ISPRS J. Photogramm. Remote Sens. 172, 223–239.

Zhong, J., Yang, B., Huang, G., Zhong, F., Chen, Z., 2016. Remote sensing image fusion with convolutional neural network. Sens. Imaging 17 (1), 1–16.

Zhou, H., Liu, Q., Wang, Y., 2021. Pgman: An unsupervised generative multiadversarial network for pansharpening. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 6316–6327.

Zhou, H., Liu, Q., Weng, D., Wang, Y., 2022. Unsupervised cycle-consistent generative adversarial networks for pan sharpening. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Zhou, C., Zhang, J., Liu, J., Zhang, C., Fei, R., Xu, S., 2020. PercepPan: Towards unsupervised pan-sharpening based on perceptual loss. Remote Sens. 12 (14), 2318.

Zhu, Z., Liu, H., Hou, J., Jia, S., Zhang, Q., 2021. Deep amended gradient descent for efficient spectral reconstruction from single RGB images. IEEE Trans. Comput. Imaging 7, 1176–1188.