A Knowledge Optimization-driven Network with Normalizer-Free Group ResNet Prior for Remote Sensing Image Pan-sharpening

Jiang He, Qiangqiang Yuan, *Member, IEEE,* Jie Li, *Member, IEEE,* and Liangpei Zhang, *Fellow, IEEE*

***Abstract*—Multispectral images play a crucial role in environmental monitoring or ecological analysis for their large scope, quick acquisition, and big data. With the rapid development of technology and increasing demand, very high-resolution multispectral images have attracted a lot of attention these days. However, due to sensor equipment and the imaging environment, the spatial resolution of multispectral images is always restricted. With the help of panchromatic images, pan-sharpening is a very important technique to enhance the spatial details of multispectral images. In this study, we proposed a knowledge optimization-driven pan-sharpening network with normalizer-free group ResNet prior, called PNXnet, which is unfolded from a physical knowledge optimization-driven variational model. We solved the memory overhead brought by the traditional ResNet relying on batch normalization. Results on four sensors show that high quantitative indexes and natural visual effects have verified the reliability of PNXnet. Focusing on the NIR band where spatial details are hard to be injected, we compared the Normalized Difference Vegetation Index (NDVI) generated from the fused results, the estimated NDVI shows a high consistency to the ground truth with R^2 above 0.91. Besides, we also compared the model generation. Furthermore, low model complexity and quicker computational speed make the daily application of PNXnet possible.**

***Index Terms*—Pan-sharpening, Deep learning, Knowledge unfolding, Normalization, Satellite imagery**

I. Introduction

REMOTE sensing technique, with the merit of huge observation scope, abundant information, and fixed revisit period, has been of extraordinary significance to many fields, including hydrometeorology [1], agriculture [2]–[4], and environmental monitoring [5]–[7]. Meanwhile, images collected by remote sensing satellites always represent Earth's surface from two aspects, namely spectral and spatial dimensions. Spectral information is beneficial to recognize ground objects by representing physical property while spatial detail makes great sense to finer application, both of which are of great importance [8].

Manuscript received March 6, 2022; revised April 19, 2022 and June 5, 2022; accepted 24 June, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 41922008, Grant 62071341, and Grant 61971319; in part by the Hubei Science Foundation for Distinguished Young Scholars under Grant 2020CFA051; and in part by the Fundamental Research Funds for the Central Universities under Grant 531118010209. *(Corresponding authors: Qiangqiang Yuan)*

J. He, Q. Yuan, and J. Li are with the School of Geodesy and Geomatics, Wuhan University, Hubei, 430079, China (e-mail: jiang_he@whu.edu.cn; yqiang86@gmail.com; jli89@sgg.whu.edu.cn).

L. Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, Hubei, 430079, China (e-mail: zlp62@whu.edu.cn)

Due to the technical limitation of sensors, it is difficult for satellites to capture images with high resolution in both spatial and spectral dimensions [9]–[11]. However, most satellites collect data in two modalities: high-resolution panchromatic (PAN) images with low spectral resolution and low-resolution multispectral (MS) images with high spectral resolution [12]–[14]. To combine the advantage of both data and produce high-resolution products for further applications, pan-sharpening becomes more and more popular in remote sensing.

In the past few decades, many traditional methods have been developed to solve pan-sharpening, which can be divided into four categories: *1) component substitution-based methods.* The main idea of component substitution-based methods is to substitute the PAN images for the low-resolution spatial component of MS images. Note that the spatial components are always extracted by methods based on intensity-hue-saturation (IHS) [15], principal component analysis (PCA) [16], Gram-Schmidt transformation (GS) [17], and Brovey transformation [18]. *2) multi-resolution analysis-based methods.* In these methods, MS and PAN images are decomposed into multiple scales, and the spatial information of PAN images is injected into the same-scale MS images. Laplacian pyramids [19], wavelets [20], contourlet [21], and curvelet transformations [22] are all classified into this group. *3) hybrid methods.* This category of methods combines the advantages of both component substitution and multi-resolution analysis methods by merging wavelet-based methods with IHS or PCA methods based on the idea of improving spatial details of the fused image. Substitute Wavelet Intensity (SWI) [23], Additive Wavelet Luminance Proportional (AWLP) [24], and GS-Wavelet [25] are some typical approaches in this group. *4) variational optimization-based methods.* Considering pan-sharpening task as an ill-posed optimization problem which is searching the best estimation of the ideal high-resolution MS images, methods based on variational optimization are also proposed to fuse PAN and MS images, such as P+XS [26], a new pan-sharpening algorithm based on Total Variation (TV) [27], and Local Gradient Constraints (LGC) [28].

Benefitting from the strong nonlinear learning ability, deep learning is also utilized to address pan-sharpening and has achieved good performance [29]–[34]. Inspired by a three-layered convolutional neural network proposed in single image super-resolution (SRCNN), Masi *et al*. regarded the pan-sharpening task as a special form of image super-resolution and proposed a pan-sharpening neural network (PNN) [35]. Combining GS transform with spatial super-resolution CNN,

TABLE I: Characteristics of satellite data used in this study.

	QuickBird	WorldView-2	Gaofen-2	Gaofen-1
MS Channel Number	4	8	4	4
Spatial Resolution	0.61 m / 2.44 m	0.46 m / 1.85 m	0.81 m / 3.24 m	2 m / 8 m
Temporal Resolution	1 to 6 days	1.1 to 3.7 days	5 days	4 days
Swath Width	16.5 km	16.4 km	45 km	60 km
Available Time	2001 to 2014	2009 to now	2014 to now	2013 to now
Acquisition Area	Shenzhen, China	San Francisco, USA	Nanning, China	Qujing, China
	Nanchang, China			Zhaotong, China
	Yichang, China			Nantong, China

Zhong *et al*. [36] put forward a new framework to enhance the spatial details in the fused images. As the deeper networks are, the stronger learning ability they have, residual learning is also employed to improve CNN-based pan-sharpening [37], [38]. Wei *et al*. [39] employed a global residual skip to improve the spatial details. Yang *et al*. [40] extracted finer textures from high-pass features using ResNet [41]. In 2018, considering multi-scale features in remote sensing images, Yuan and Wei *et al*. [42] further proposed a multi-scale and multi-depth CNN for image pan-sharpening. To further improve CNN's modeling capability for pan-sharpening, several strategies are put forward, such as pyramid network [43], adaptive weight [44], gradient prior [45], two-stream network [46], *etc*.

With great learning capacity from data, deep learning has also been used to achieve unsupervised pan-sharpening [47]–[51]. Luo *et al*. [47] proposed a new loss function where the input MS and PAN images are used to enhance the spatial constrains and spectral consistency, respectively. Ciotola *et al*. [49] further improved this training strategy with a target-adaptive operating modality. Seo *et al*. [48] combined unsupervised learning with registration learning to implicitly learn the registration between PAN and MS images. Zhou *et al*. [50] designed a generative adversarial network based on auto-encoder and perceptual loss to achieve unsupervised pan-sharpening. Liu *et al*. [51] employed a two-stream generator with a dual discriminator to extract features from the PAN and MS images.

Though CNN-based algorithms have achieved great successes in pan-sharpening, the fact is there are still some problems to be solved. One is that ResNets are always utilized with batch normalization (BN), and the other is the interpretability of deep learning. As for BN, on the one hand, without BN, the training of CNN would be unstable, while, on the other hand, BN incurs computation memory overhead and breaks the independence of distributions between training examples within a batch. Recently, several works have tried to replace BN. A part of them is using alternative normalization [52], and another line of these works looks for eliminating layers that normalize hidden activations entirely [53]. Nevertheless, they more or less degraded generalization or added compute costs in tests [54].

As for the interpretability of deep learning, most researchers are trying to combine CNNs with physical model-driven methods [55], [56], where CNNs always play a role as priors. While they mostly break the end-to-end running mode of CNNs, which ensures that deep learning can achieve flexible and generalizable applications.

In this paper, an end-to-end physical optimization-driven CNN with group ResNet prior is proposed and a normalizer-free strategy that keeps stable generalization with low compute costs is also presented, which is called PNXnet. To keep the similar standard initializations as BN, we build a normalizer-free ResNet by directly introducing two variables to simulate means and variances in ResNet, which decrease the memory cost brought by BN layers. Furthermore, rather than alternately running a variational model and CNN, an optimization-based pan-sharpening model is unfolding into CNN with normalizer-free ResNet-based multispectral prior. The proposed CNN updates the fused MS image in an end-to-end manner. The contributions are as follows.

- Focusing on pan-sharpening, this paper unfolds a variational model into end-to-end CNN with the help of knowledge optimization, which brings physical interpretability to deep learning-based algorithms and keeps the data-driven training mode.
- Neither directly employing BN layers in ResNet nor completely abandoning normalization, we introducing two variables to simulate means and variances layer by layer, which could achieve similar standard initializations as BN layers and doesn't require more memory cost to save intermediate features. Furthermore, group convolutions are also utilized to reduce the network parameters.
- Data captured by four satellites are tested, including WorldView-2, QuickBird, Gaofen-2, and Gaofen-1, which prove that the proposed model can handle various data from different satellites. Besides, focusing on the NIR band where spatial details are hard to be injected, we also compared the NDVI generated from the fused results.
- The proposed PNXnet achieves better performance with the state-of-the-art methods but acquires less parameters and running time.

The remaining part of the paper is organized as follows. Section 2 describes the degradation model and unfolds the variational pan-sharpening algorithm into knowledge optimization-driven PNXnet. Section 3 shows the experiments on data from four satellites and presents some discussions. Finally, conclusions are given in Section 4.

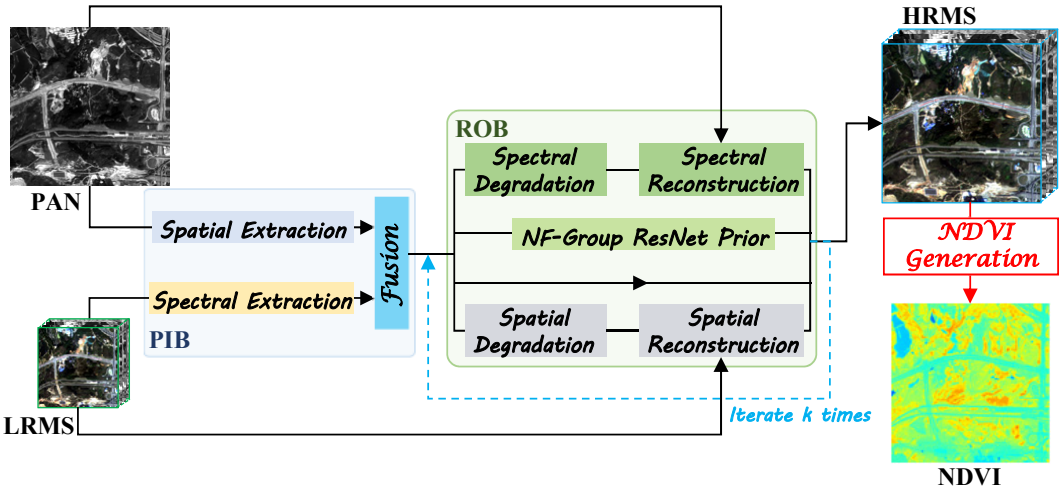


Fig. 1: The framework of the proposed PNXnet.

II. Material and Methods

A. Data sources

1) Satellite remote sensing data: In this study, remote sensing data collected by four satellites are used, including QuickBird, Gaofen-2, WorldView-2, and Gaofen-1, as shown in Table I.

QuickBird satellite was launched on October 18, 2001. it acquires a PAN channel (450-900 nm) and MS images with four channels in the visible and Near-InfraRed (NIR) wavelength range: Blue (450-520 nm), Green (520-600 nm), Red (630-690 nm) and NIR (760-900 nm). The spatial resolution of the PAN channel is between 61 (at nadir) and 72 cm (25° off-nadir), and the MS images are with a resolution between 2.44 and 2.88 m (25° off-nadir).

On October 6, 2009, the WorldView-2 satellite was launched by Maxar Technologies, then known as DigitalGlobe, which collects data with one PAN channel (450-800 nm) and eight MS channels. The acquired PAN channel is with a resolution of 0.46 m and the MS channels are with 1.85 m resolution. Besides the common Blue (450-510 nm), Green (510-580 nm), Red (630-690 nm), the MS channels also involve Coastal-Blue (400-450 nm), Yellow (585-625 nm), Red-Edge (705-745 nm), NIR1 (770-895 nm), and NIR2 (860-1040 nm).

Equipped with two PAN/MS cameras, the Gaofen-2 satellite was launched on August 19, 2014, and capable of collecting images with a ground sampling distance of 0.81 m in PAN channel (450-900 nm) and 3.24 m in the four MS channels, including Blue (450-520 nm), Green (520-590 nm), Red (630690 nm), and NIR (770-890 nm). And Gaofen-1 satellite, which was launched on April 26, 2013, shares all the similar characteristics to Gaofen-2 except the coarser spatial resolution, 2 m for the PAN channel and 8 m for the MS images.

2) Data Location: In this paper, remote sensing data from eight cities in China and USA are collected for study. Table I lists the location of the selected data from different satellites.

The QuickBird data are selected from three cities in China, including Shenzhen, Nanchang, and Yichang. Data in Shenzhen and Nanchang covers the urban area, where roads and buildings are the main parts in Shenzhen while a variety of land-use can be found in Nanchang covering lakes, croplands, rivers, and buildings. And data from Yichang mainly encompasses mountains, buildings, and rivers. As for the Gaofen-2 satellite, we choose Nanning which includes very diverse topography as the study area, including vegetation, land, and waters. In the USA, WorldView-2 data covering San Francisco is selected in this study. San Francisco is a coastal city in Northern California with an area of 121.4 km^2, which is one of the world's top travel destinations with various buildings, hills, bays, trees, and urban. For the Gaofen-1 satellite, we select the images covering Qujing, Zhaotong, and Nantong, in China to build the fourth data set. Qujing and Zhaotong are both cities in Southwest China and are covered by mountains. And Nantong is in East China near the ocean.

The selected images acquired by four satellites cover a variety of topography makes great sense to fully verify model generalization. All the results and discussions are based on the mentioned-above data sets.

B. Methods

PNXnet is proposed to fuse the rich spectral information in MS images and the fine spatial details in PAN images for higher-resolution multispectral data in this study. Inspired by solving the optimization problem based on a physical degradation model, we roughly fuse the low-resolution MS images and high-resolution PAN images by a physical inverse block (PIB), and then we feed the intermediate results and original images into the recurrent optimization-driven blocks (ROB), which equipped with normalizer-free ResNet for prior representation and group convolution for model lightweight. Furthermore, the whole network is end-to-end. Details of our proposed PNXnet are illustrated in Fig. 1.

1) Proposed Knowledge Optimization-driven Pan-sharpening Networks: In pan-sharpening, our objective is to recover the ideal MS images with high spatial resolution from low-resolution MS images and high-resolution PAN images. We assume that $X \in R^{W \times H \times C}$ represents the desired high-resolution MS images, where C is the number

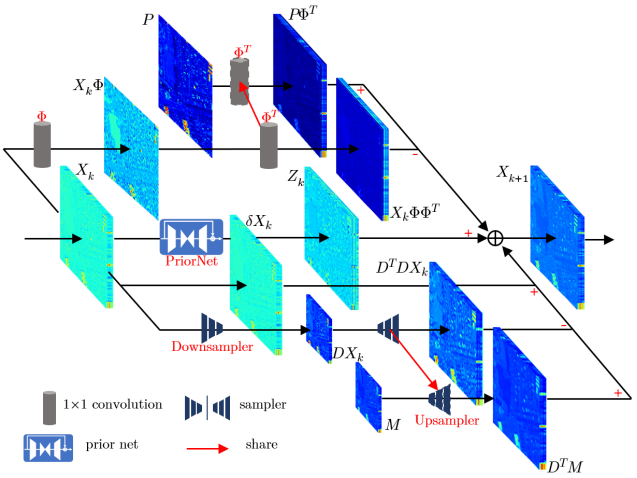


Fig. 2: The proposed ROB. We replace all matrices in Eq. 4 with CNN-based modules but follow the same data flow.

of the spectral channels, and W and H are the width and height, respectively, $P \in R^{W\times H\times 1}$ represents the PAN image with the same spatial resolution as X but only one band, and $M \in R^{w\times h\times C}$ denotes the low-resolution MS image. According to the satellite imaging, the degradation model is established as:

$$\begin{aligned} P &= X\Phi \\ M &= DX \end{aligned} \tag{1}$$

With the help of the degradation model, pan-sharpening is to figure out the closest approximation of X, which can be regarded as an optimization problem, *i.e.*,

$$\hat{X} = \arg\min_X \frac{1}{2}\left\|P - X\Phi\right\|_2^2 + \frac{1}{2}\left\|M - DX\right\|_2^2 + \lambda\mathcal{R}(X) \tag{2}$$

Employing the half-quadratic splitting method with an auxiliary variable Z, a new cost function is derived:

$$\begin{aligned} \mathcal{L}_\mu(X,Z) = \frac{1}{2}\|P-X\Phi\|^2 + \frac{1}{2}\|M-DX\|^2 + \mu\|Z-X\|^2 + \lambda\mathcal{R}(Z) \\ s.t.\ Z = X \end{aligned} \tag{3}$$

where μ is a penalty parameter. Based on the half-quadratic splitting method, Eq (3) can be split into two subproblems which allows us to address the data fidelity term and the prior term separately:

$$\begin{cases} \hat{X} = \arg\min_X \frac{1}{2}\left\|P-X\Phi\right\|^2 + \frac{1}{2}\left\|M-DX\right\|^2 + \mu\left\|Z-X\right\|^2 \\ \hat{Z} = \arg\min_Z \frac{1}{2}\left\|Z-X\right\|_2^2 + \frac{\lambda}{\mu}\mathcal{R}(Z) \end{cases} \tag{4}$$

Considering the X-subproblem, the approximation would be updated by the gradient descent algorithm:

$$\begin{aligned} \hat{X}_{k+1} &= X_k - \epsilon\left(X_k\Phi\Phi^T + D^TDX_k - P\Phi^T - D^TM + \mu X_k - \mu Z_k\right) \\ &= \delta X_k - \epsilon X_k\Phi\Phi^T - \epsilon D^TDX_k + \epsilon P\Phi^T + \epsilon D^TM + \epsilon\mu Z_k \end{aligned} \tag{5}$$

where $\delta = 1-\epsilon\mu$, and ϵ is the optimization stride. As shown in Fig. 2, we unfold the variational solution into CNN, namely recurrent optimization-driven blocks.

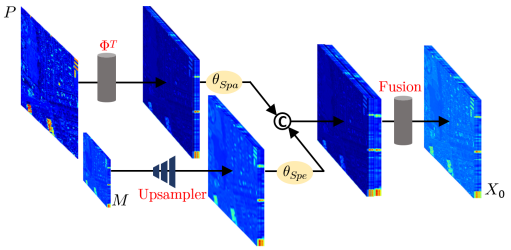


Fig. 3: The proposed Physical Inverse Block.

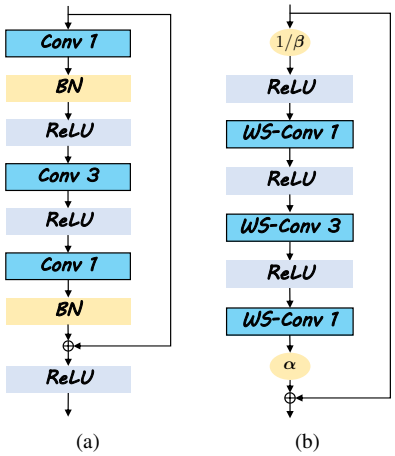


Fig. 4: Two ResNet-based blocks. Left is the original residual block with BN, called BN-ResB. Right is the normalizer-free residual block, called NF-ResB, where "WS-Conv" denotes weight-scaled convolutions. The number after "Conv" is the kernel size.

With the same calculation process as the spectral response function, 1×1 convolutions are employed to perform Φ and Φ^T. Meanwhile, D is replaced with a down-sampler, which can be decomposed into 2D convolutions and average-pooling operators, and D^T is performed by 2D deconvolutions. PriorNet is the proposed Normalizer-free ResNet for prior representation to solve Z-subproblem. All hyper-parameters are adaptively learned by channel attention [57], [58]. To initialize the X_0 more closed to ideal X, we propose a physical inverse block to extract and fuse the spectral information and spatial information from M and P, as shown in Fig. 3.

In formulation, PIB can be represented as:

$$X_0 = \mathrm{Fusion}(\mathrm{Cat}(P\Phi^T\theta_{Spa}, D^TM\theta_{Spe})) \tag{6}$$

where $\mathrm{Cat}(\cdot)$ denotes concatenation, $\mathrm{Fusion}(\cdot)$ represents a 1×1 convolution to fuse the extracted spectral information and spatial information. θ_{Spa} and θ_{Spe} are two hyperparameters to control the information weight scale in subsequent fusion, which are learnable in training. With X_0 initialized, the ROB can update X_k recurrently until the results are optimal, which is supervised by an ℓ_1 loss function, defined as $\left|\hat{X} - X\right|$.

To update X_k, the most important problem is to update Z_k involved the prior term as shown in Eq (4). Details are described in the next sub-subsection.

TABLE II: Quantitative assessment on the QuickBird and WorldView-2 data. The best performance is shown in **bold** and the second best is underlined.

	QuickBrid					WorldView-2				
Methods	CC	PSNR	SSIM	SAM	ERGAS	CC	PSNR	SSIM	SAM	ERGAS
BDSD	0.9351	41.8275	0.9431	2.2678	2.0736	0.9262	33.3169	0.8931	7.4913	5.2809
PRACS	0.9511	44.2434	0.9566	1.8110	1.4983	0.9265	33.7069	0.8690	6.5751	5.2193
GSA	0.9389	42.2266	0.9444	2.0716	1.9624	0.9398	33.9905	0.8990	6.4442	4.7628
ATWT-M3	0.9355	42.6250	0.9392	2.3004	1.8965	0.9115	31.8137	0.8283	6.8969	6.1430
MTF-GLP-HPM	0.9422	42.9411	0.9534	1.8531	1.7386	0.9268	33.4101	0.8977	5.9326	10.5219
AWLP	0.9239	41.5724	0.9382	2.0083	1.9605	0.9403	34.0049	0.9069	6.0190	4.7664
TV	0.9485	41.9099	0.9477	2.1589	1.9592	0.9162	32.2562	0.8435	8.4302	5.6334
PanNet	0.9436	43.7540	0.9638	1.8252	1.4921	0.9361	33.8900	0.8970	6.1223	4.7344
DRPNN	0.9302	42.5424	0.9443	2.0893	1.7648	0.9431	34.6462	0.9122	5.9716	4.4121
MSDCNN	0.9542	44.6686	0.9633	1.6680	1.4062	0.9443	34.7514	0.9143	6.0456	4.4042
ResTFNet	0.9558	44.8494	**0.9658**	1.6222	1.3814	0.9608	36.0258	0.9415	**4.7843**	3.7341
PNXnet	**0.9622**	**45.4862**	0.9651	**1.5361**	**1.3112**	**0.9618**	**36.1242**	**0.9417**	4.8701	**3.6651**

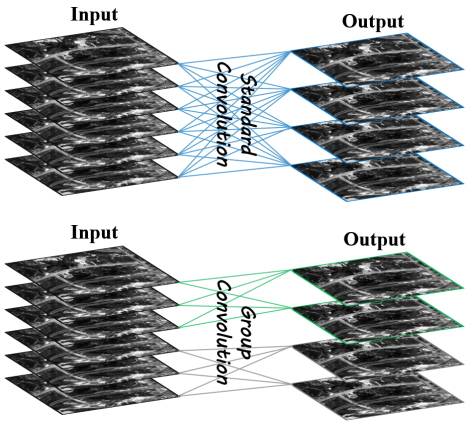


Fig. 5: Difference between standard convolution and group convolution.

2) Normalizer-Free ResNet for Prior Representation: In traditional optimization algorithms, the prior in pan-sharpening is always used with regularization, such as total variation, non-Gaussianity, and nonlocal self-similarity. Actually, the Z-subproblem with priors in Eq (4) is a proximal optimization problem, which should be solved by a proximal operator. Recently, with an ideal performance of modeling nonlinearity, ResNets are proved to be capable to learn priors implicitly [59], as shown in the left of Fig. 4a.

Let x_i^l present the i-th batch features after l-th residual block, and $x_i^{l+1} = x_i^l + f(x_i^l)$ always holds in residual blocks, where $f(\cdot)$ denotes the residual branch. Focusing on the variance of the training example in ResNet, we can find that the variance of the activations before and after the residual block satisfy:

$$\mathbf{Var}(x_i^{l+1}) = \mathbf{Var}(x_i^l) + \mathbf{Var}(f(x_i^l)) \tag{7}$$

In BN-ResB, with the help of Batch Normalization, $\mathbf{Var}(f(x_i^l))$ is very close to 1. So, if $\mathbf{Var}(x_i^0)$ is assumed to be 1, the variance $\mathbf{Var}(x_i^l) \approx l$, which ensures the controllability of BN and becomes easier to be initialized [60].

With BN and residual mapping, ResNets truly achieve good performance in image restoration. Meanwhile, BN also incurs computation memory overhead and breaks the independence of distributions between training examples within a batch. To replace BN in ResNet, in this paper, we utilize NF-ResB to achieve prior representation, as shown in Fig. 4b. Firstly, the weights of original convolution are all scaled-standardized:

$$\hat{W}_{i,j} = \gamma \cdot \frac{W_{i,j} - \mu_{W_{i,\cdot}}}{\sigma_{W_{i,\cdot}}\sqrt{N}} \tag{8}$$

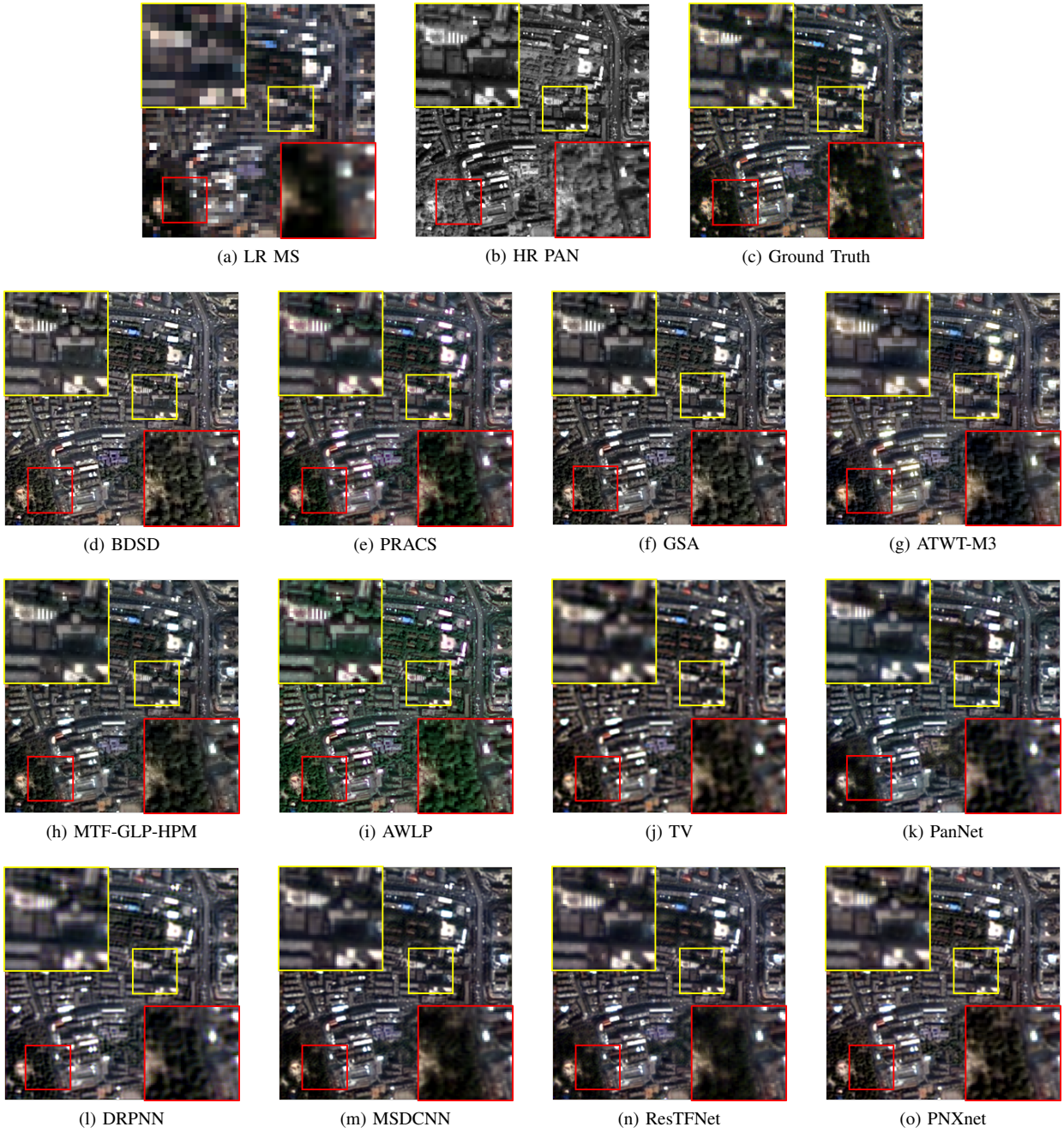
where N is the number of weights, $\sigma_{W_{i,\cdot}}$ and $\mu_{W_{i,\cdot}}$ are the standard deviation and mean of the i-th row in the kernel, γ is a constant. For networks with ReLU as activation functions, it implies that the outputs $g(x) = \max(x, 0)$ will be sampled from the rectified Gaussian distribution with variance ${\sigma_g}^2 = (1 - 1/\pi)/2$. A weight-scaled convolution with ReLU can be written as $z = \hat{W}g(x)$. With scaled-standardized weights, the variance $\mathbf{Var}(z) = \gamma^2 {\sigma_g}^2$. And when we set $\gamma = 1/\sigma_g = \sqrt{2}/\sqrt{1 - 1/\pi}$, the variance $\mathbf{Var}(\hat{W}g(x)) = 1$ always holds, which also indicates that if we build a residual branch $f(\cdot)$ with weight-scaled convolutions, the variance $\mathbf{Var}(f(x)) = \mathbf{Var}(x)$ is satisfied too.

With good variance preserving of weight-scaled convolutions, we change the ResNet from $x_i^{l+1} = x_i^l + f(x_i^l)$ into $x_i^{l+1} = x_i^l + \alpha f(x_i^l/\beta_l)$, where α denotes the rate of variance growth, and β_l is fixed as $\sqrt{\mathbf{Var}\left(x_i^l\right)}$. In this way, the variance $\mathbf{Var}\left(x_i^{l+1}\right)$ will be changed into:

$$\begin{aligned} \mathbf{Var}(x_i^{l+1}) &= \mathbf{Var}(x_i^l) + \alpha^2 \mathbf{Var}(f(x_i^l/\beta_l)) \\ &= \mathbf{Var}(x_i^l) + \alpha^2 \mathbf{Var}(x_i^l)/{\beta_l}^2 \\ &= \mathbf{Var}(x_i^l) + \alpha^2 \end{aligned} \tag{9}$$

With the same assumption $\mathbf{Var}(x_i^0) = 1$, the variance in NF-ResNet satisfy:

$$\mathbf{Var}(x_i^l) = 1 + l\alpha^2 \tag{10}$$



(a) LR MS (b) HR PAN (c) Ground Truth

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

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

(l) DRPNN (m) MSDCNN (n) ResTFNet (o) PNXnet

Fig. 6: Pan-sharpening results on QuickBird images. Enlarged views are shown in yellow and red boxes. The first row shows input data and ground truth. Row 2 to 4 present the results of comparison methods and the proposed PNXnet.

Eq (10) has a similar variance growth with BN-ResNet which is just multiplied with α^2. With such a strategy, we could build a ResNet without BN but with similar variance growth as BN-ResB, which keeps the strong prior learning ability of ResNet but does not introduce more computational cost. In this study, we employ the normalizer-free ResNet consisting of three NF-ResBs to achieve the prior representation implicitly.

3) Group Convolution for Model Lightweight: CNN-based prior representation achieves superior to single traditional prior with explicit formulas, while it also costs more computation. Let x be I-channel input features to a convolutional layer, and z denotes the corresponding O-channel output features. Standard convolutional layers can be defined as:

$$z^j = \sum_{i=1}^{I} x^i * W^{i,j} \tag{11}$$

where $1 \leq j \leq O$, $W^{i,j}$ denotes the convolutional kernel with the size of $k \times k$. So, the number of W to gain z from x is $I \times O$, and the number of weights is $I \times O \times k \times k$. When we build a very deep CNN, superabundant convolutions

(a) LR MS (b) HR PAN (c) Ground Truth

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

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

(l) DRPNN (m) MSDCNN (n) ResTFNet (o) PNXnet

Fig. 7: Pan-sharpening results on WorldView-2 images. Enlarged views are shown in yellow box. The first row shows input data and ground truth. Row 2 to 4 present the results of comparison methods and the proposed PNXnet.

will generate heavy computational costs, especially employing deep CNN iteratively [61]–[63].

As for a group convolutional layer, it takes I/G input channels to produce O/G output channels at each time, and differences between standard convolution and group convolution are shown in Fig. 5. Note that there are G-group convolutions in one layer. Group convolution can be defined as:

$$z_g^j = \sum_{i=1}^{I/G} x^i * \widetilde{W}_g^{i,j} \tag{12}$$

where $1 \leq g \leq G$, $\widetilde{W}_g^{i,j}$ denotes the convolutional kernel with the size of $k \times k$ to produce the g-th group feature. So, the number of $\widetilde{W}_g$ to gain z_g from x_g is $\frac{I}{G} \times \frac{O}{G}$, and the total number of weights is $\frac{I}{G} \times \frac{O}{G} \times k \times k \times G$, which is reduced by a factor G than standard convolution. In this study, because of recursively performing optimization stages, group convolution is utilized to reduce the model weights, where G is set to 4.

III. Results and Discussion

As mentioned in Section II-A, data acquired by QuickBird, WorldView-2, Gaofen-2, and Gaofen-1 sensors are used in this

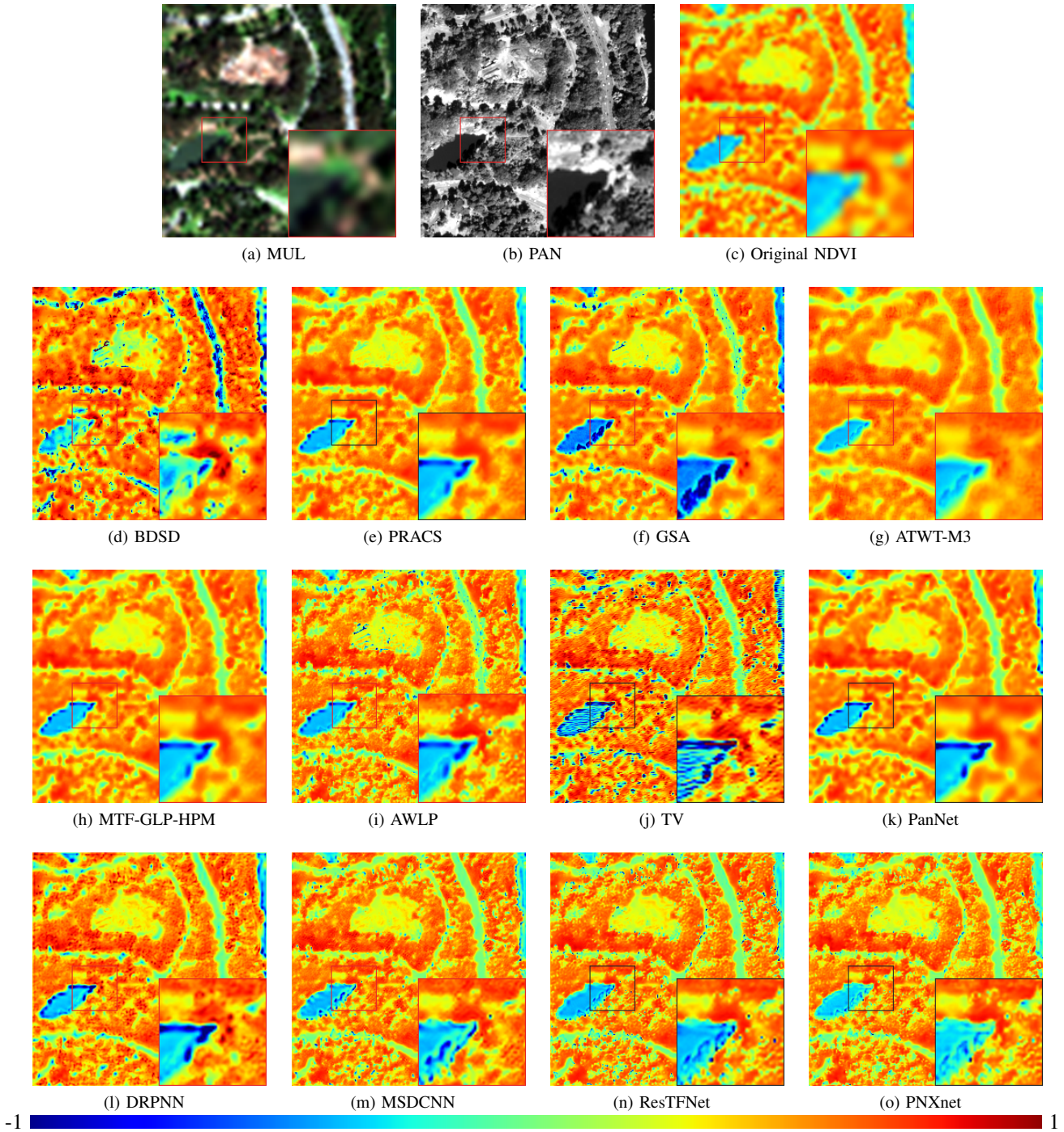


Fig. 8: NDVI products generating from WorldView-2 images. The minimum of NDVI is -1 and the maximum is 1.

paper. In detail, we utilize images collected by QuickBird and WorldView-2 to verify the pan-sharpening performance. And then WorldView-2 data are subsequently used to generate high-resolution NDVI. Furthermore, we also compared the model generalization from Gaofen-2 to Gaofen-1 data. Moreover, all experiments are after radiance calibration and atmospheric correction.

In this study, we chose seven classical traditional algorithms and four state-of-the-art deep learning-based methods, including BDSD [64], PRACS [65], GSA [66], ATWT-M3 [67], MTF-GLP-HPM [68], AWLP [24], TV [27], PanNet [40], DRPNN [39], MSDCNN [42], and ResTFNet [46].

Furthermore, there are two types of testing carried out in this study, including reduced-resolution testing under Wald's protocol [69] and full-resolution testing. For the reduced-resolution testing, five quantitative quality metrics are utilized to evaluate the pan-sharpening performance from spatial and spectral domains, including correlation coefficient (CC), mean peak signal-to-noise ratio (mPSNR) in decibel units, mean structural similarity (mSSIM) [70], spectral angle mapper (SAM) [71]

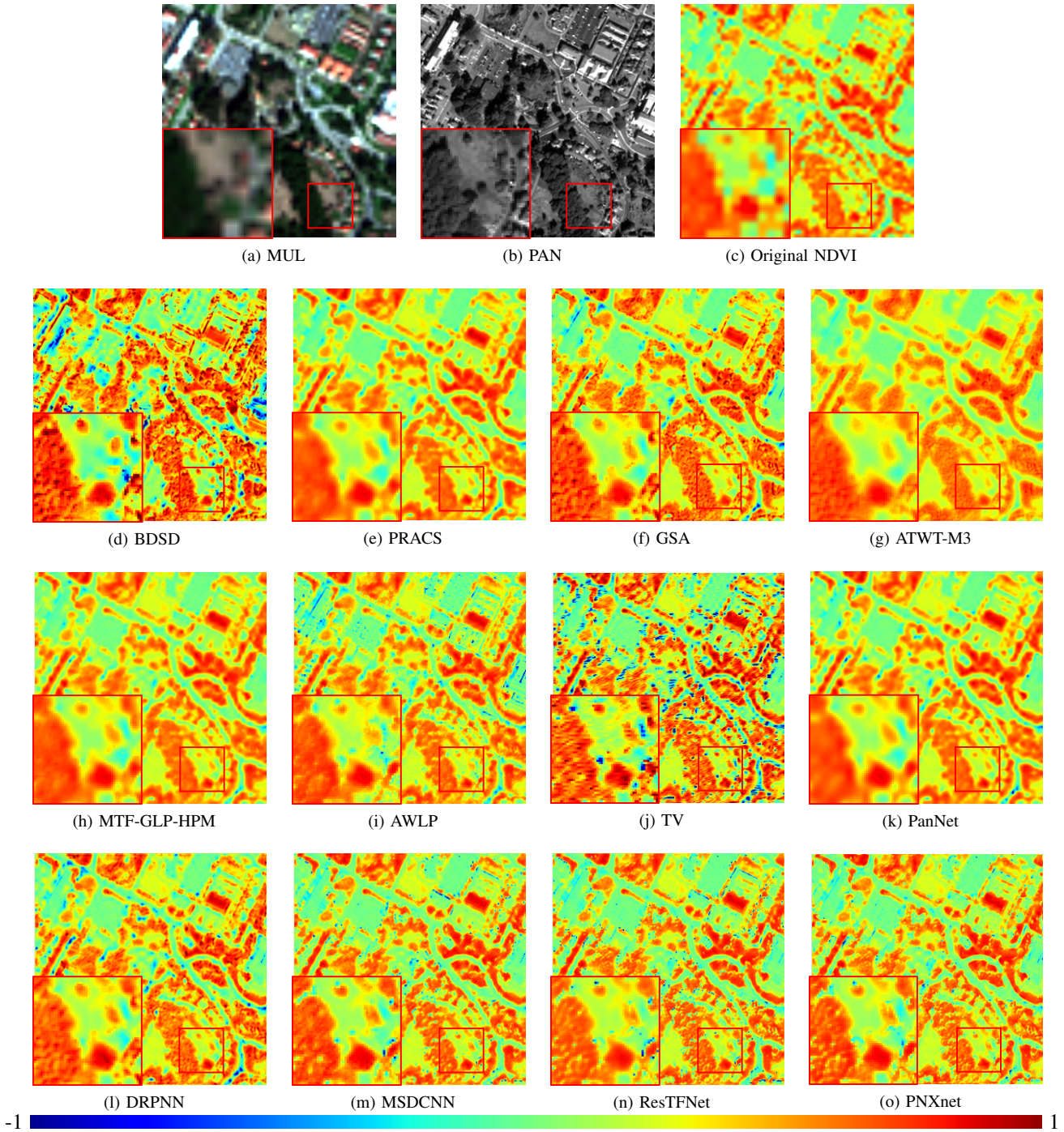
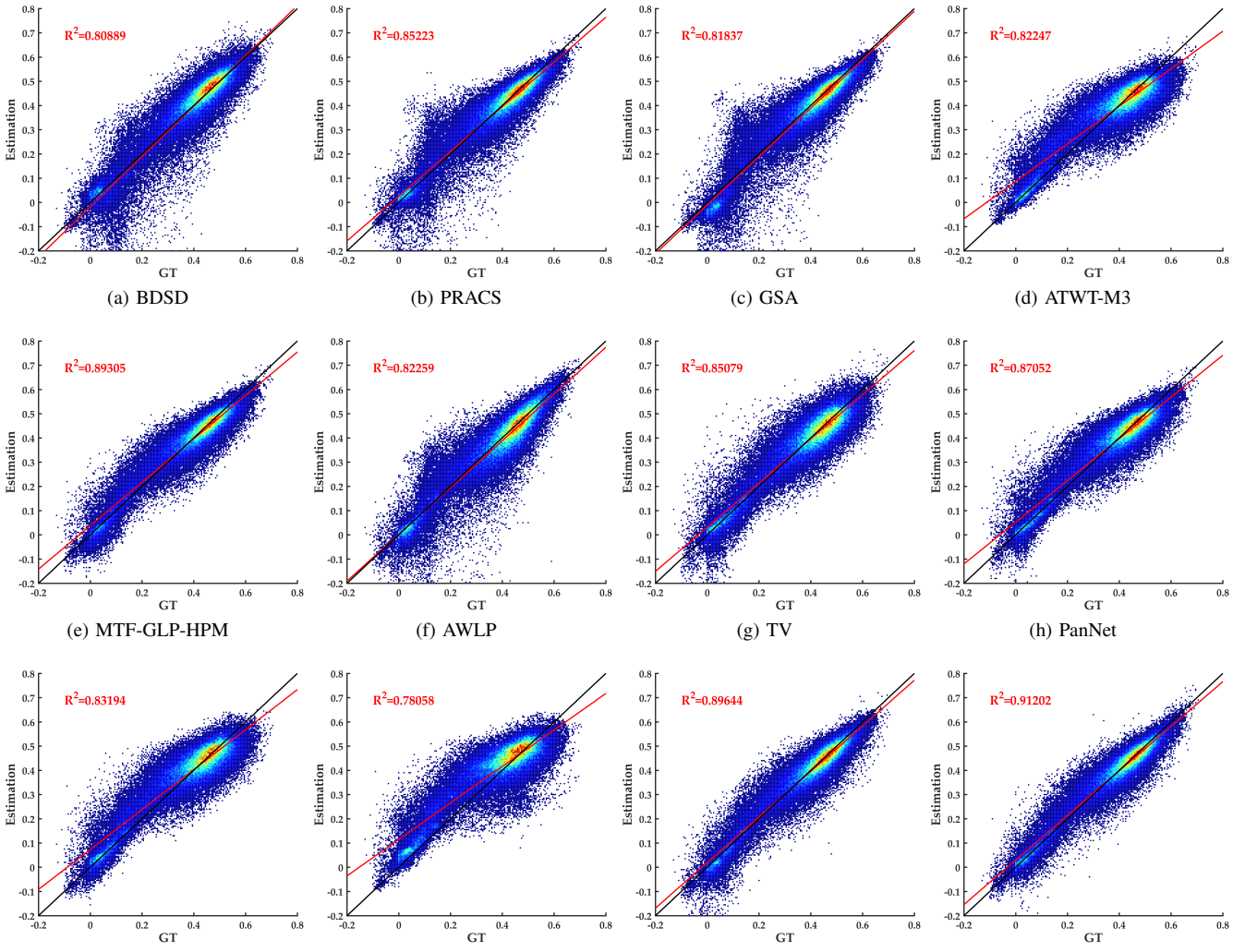


Fig. 9: NDVI products generating from WorldView-2 images. The minimum of NDVI is -1 and the maximum is 1.

in degree, and Erreur Relative Global Adimensionnelle de Synthèse (ERGAS). For the full-resolution testing, the spectral distortion index D_λ [72], spatial distortion index D_s [72], D_ρ [73], quality with no reference (QNR) [72], and hybrid quality with no reference (HQNR) [74] are introduced to characterize fusion performance. We only draw the reduced-resolution testing on QuickBird and WorldView-2 data. As for Gaofen-2 and Gaofen-1 data, we train model on Gaofen-2 data under Wald's protocol and achieve both reduced and full resolution testing on Gaofen-2 and Gaofen-1 data. Note that, models are not retrained on Gaofen-1 data. In this way, we can discuss the model generalization about comparison methods.

We build the proposed PNXnet with 9 ROBs, which shows the best performance and fast computational speed. And Adam optimization algorithm is employed to train PNXnet with l1 loss function and learning rate of 0.01. All CNN-based methods are trained by Pytorch framework running in the Windows 10 environment with 32 GB RAM and one Nvidia RTX 2080 GPU, and all traditional methods are performed with MATLAB with an Intel CPU (Core i7-8700 @ 3.20



(a) BDSD (b) PRACS (c) GSA (d) ATWT-M3

(e) MTF-GLP-HPM (f) AWLP (g) TV (h) PanNet

(i) DRPNN (j) MSDCNN (k) ResTFNet (l) PNXnet

Fig. 10: Relations between the different estimated NDVIs and the ground truth. Black lines denote the ideal relationship $y = x$, and red lines illustrate the linear regression results. The color illustrates the density of samples. Goodness of fit R^2 is displayed at the top left.

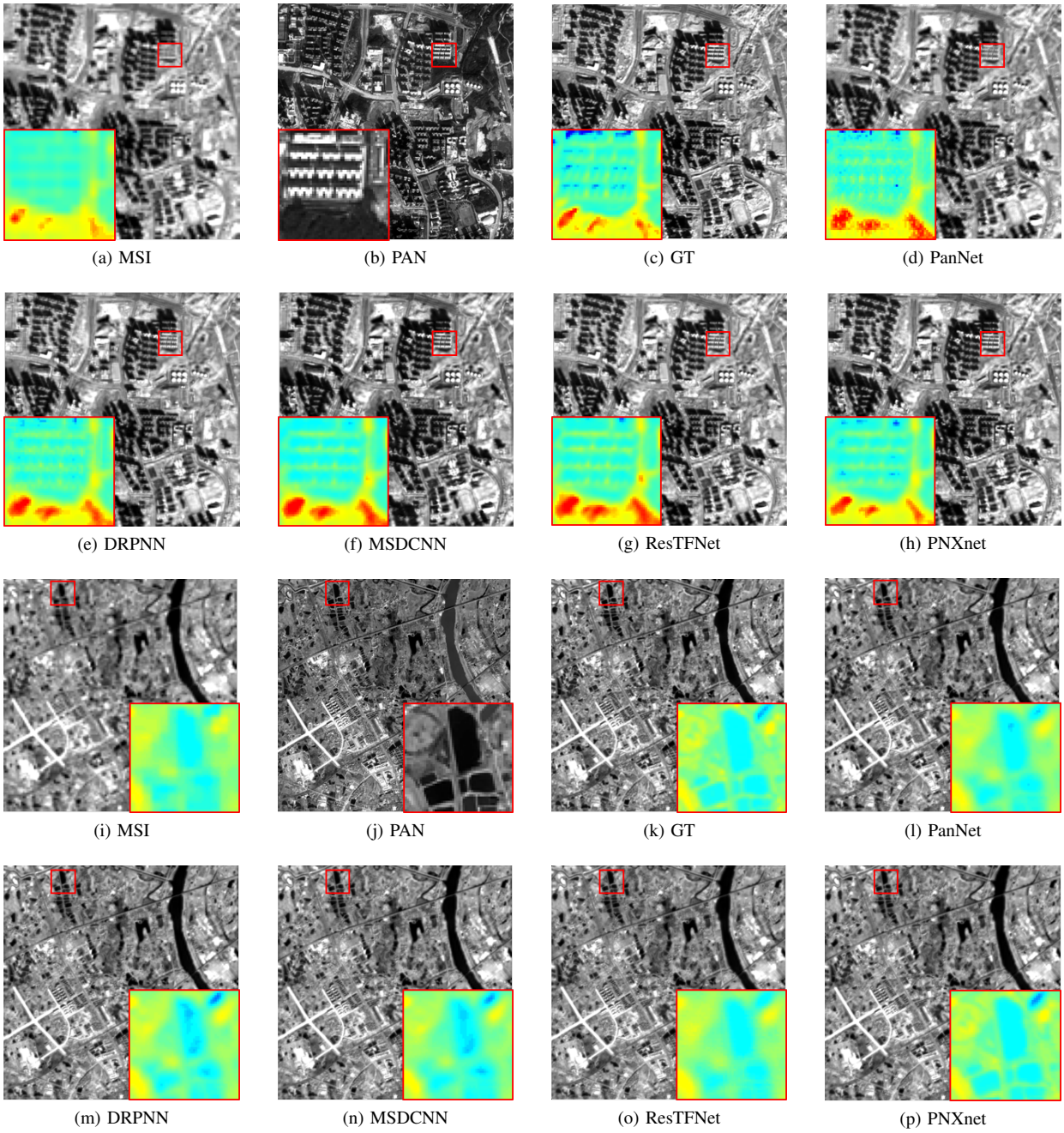
GHz).

A. Pan-sharpening results on QuickBird and WorldView-2 sensors

1) Reduced-resolution testing: Table II reports the quantitative results on the QuickBird and WorldView-2 data, where the proposed PNXnet achieves the best performance in both two data sets. Compared with traditional methods, deep learning-based methods have a comparative advantage in reducing spectral distortion, which performs with lower SAMs. MSDCNN shows unstable quantitative results in spatial or spectral domains, sometimes better than PanNet, and sometimes worse. With individual feature extraction modules for MS and PAN images, ResTFNet can obtain good results close to PNXnet, while it also suffers numerous model parameters and large computational costs. In traditional methods, PRACS and AWLP perform well in fusing spatial details, while MTF-GLP-HPM can keep better spectral fidelity.

Fig. 6 and Fig. 7 illustrate the visual comparisons randomly selected in QuickBird and WorldView-2 images, respectively. On QuickBird images, we show the details of buildings and vegetation in the red and yellow boxes. Traditional methods except for TV successfully integrate enough spatial details from PAN images to MS images. Compared with traditional methods, deep learning-based algorithms shows better spatial details except for MSDCNN. However, PanNet and DRPNN show more spectral distortion than ResTFNet and the proposed PNXnet. Compared ResTFNet with PNXnet, the buildings are restored well while the texture of vegetation cannot be injected sufficiently.

On WorldView-2 data, similar conclusions can be drawn. Besides, PanNet and MSDCNN perform well, although their results on QuickBird images look blurry. Moreover, ATWT-M3 produces fused images with paler colors, which seem to be covered with mist. BDSD excessively integrates edges or textures and shows untruthfulness. Significantly, comparing



(a) MSI (b) PAN (c) GT (d) PanNet

(e) DRPNN (f) MSDCNN (g) ResTFNet (h) PNXnet

(i) MSI (j) PAN (k) GT (l) PanNet

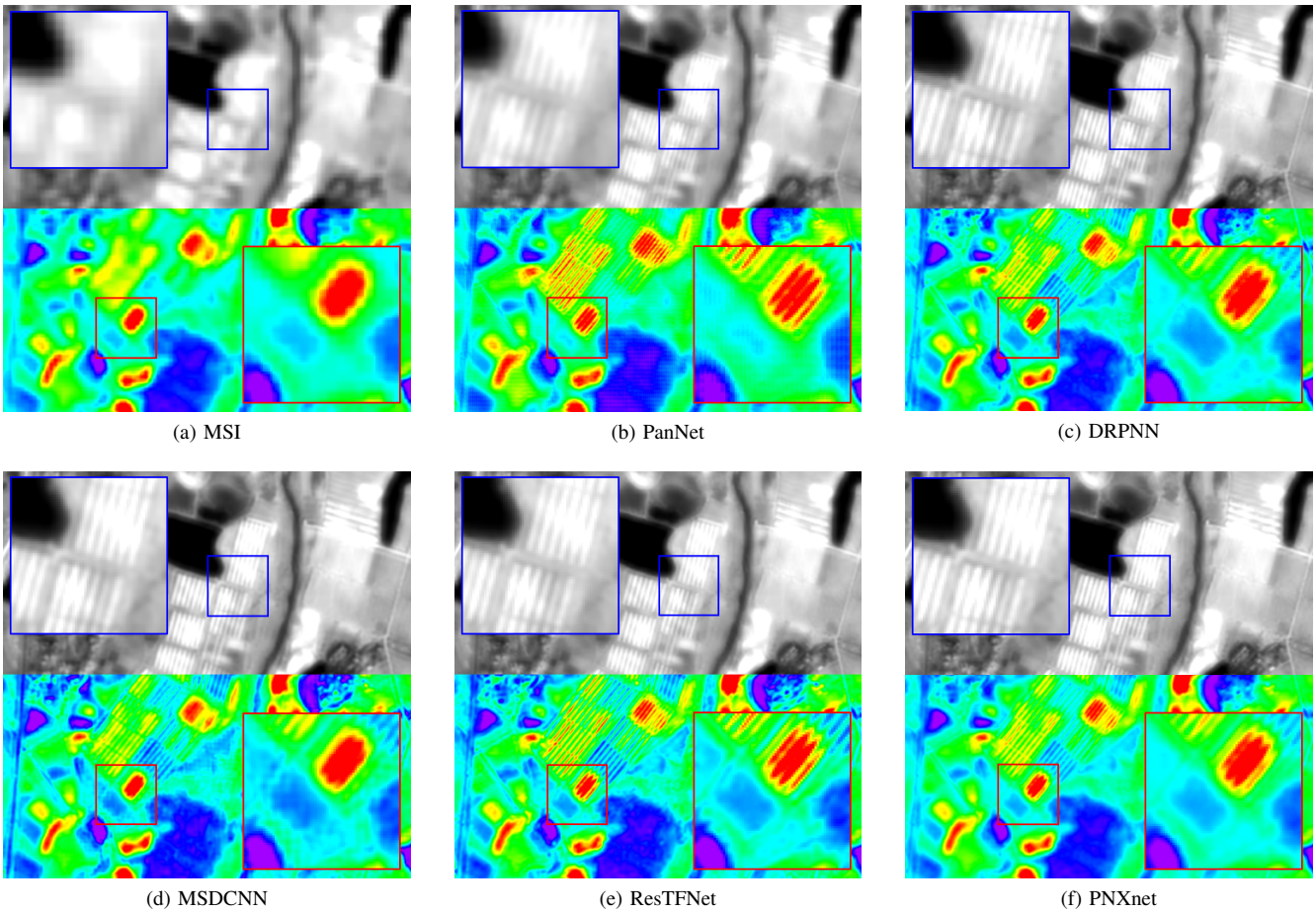
(m) DRPNN (n) MSDCNN (o) ResTFNet (p) PNXnet

Fig. 11: Visual results of NDVI and NIR band of Gaofen-2 and Gaofen-1 data. NDVIs are shown in the enlarged zooms. (a)-(h) present results on Gaofen-2 data and (i)-(p) illustrate results on Gaofen-2 data.

Fig. 7b and Fig. 7c, we can find that, for WorldView-2 sensors, there is a little time difference between PAN and MS imaging, represented by the different location of cars, which may explain why all quantitative metrics on WorldView-2 data all deteriorate.

2) Full-resolution testing: Limited by the article length, we put all results into *Supplementary Material*[1]. Table S1 lists the full-resolution quantitative results on QuickBird and WorldView-2 sensors. Moreover, Fig. S1 and Fig. S2 display the visual results in the full-resolution testing. For the injection of the spatial detail, GSA shows the great superiority beyond other pan-sharpening methods, and PNXnet shows more spatial details in deep learning-based methods. For the spectral fidelity, the proposed PNXnet keeps the best consistency with the original multispectral images. Seeking a balance between spatial details and spectral fidelity, the proposed PNXnet achieves the best performance in deep learning-based methods on both two data sets.

[1]Reduced-resolution and full-resolution testings involve model generalization, which will be systematically discussed on Gaofen-series data sets in subsequent sub-section.



(a) MSI (b) PanNet (c) DRPNN

(d) MSDCNN (e) ResTFNet (f) PNXnet

Fig. 12: Visual results of NDVI and NIR band on Gaofen-1 data in the full-resolution testing. The top half is NIR and the bottom half is NDVI. Details are shown in the enlarged areas.

B. Application on high-resolution NDVI using pan-sharpened WorldView-2 images

NDVI is a common and effective product to identify crop phenological characteristics for reflecting crop growth process within a year [75], which make great sense for agricultural monitoring. The formula is as follows:

$$NDVI = \frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}} \tag{13}$$

where ρ_{NIR} and ρ_{Red} mean surface reflectances of NIR and Red channel. In this way, we can sequentially generate the NDVI using pan-sharpened high-resolution MS images through multiple pan-sharpening algorithms. It's noted that there are two NIR channels in the WorldView-2 data, i.e., NIR1 (770-895 nm) and NIR2 (860-1040 nm). As mentioned in [76], the central wavelength of NIR regions is about 800 nm. So we utilize the NIR1 channel as ρ_{NIR} when dealing with WorldView-2 data.

Obtaining high-resolution MS images through multiple pan-sharpening algorithms, in this paper, we produce NDVI following Eq (13). Fig. 8 and Fig. 9 present the NDVI generated from the fused WorldView-2 images. As can be seen, PNXnet and ResTFNet can obtain the best NDVI. In traditional algorithms, AWLP, as well as MTF-GLP-HPM, produces the ideal NDVI with a finer spatial difference and better global consistency. TV suffers severe artifacts and ATWT-M3 presents blurring effects. Moreover, BDSD and GSA present too many spatial details, which may be due to the over-reliance on PAN images. As for deep learning-based methods, they all produce visually satisfactory NDVIs, except PanNet. PanNet shows little spatial details. Besides, DRPNN obtains higher values compared with the original low-resolution NDVI, shown with more red color. MSDCNN shows good spatial details, but ResTFNet and PNXnet produce finer details. Furthermore, compared with the proposed PNXnet, ResTFNet shows some artificial blue spots, which should be recognized as shadows.

To further evaluate the generated high-resolution NDVI, we show the scatter plots of the different estimated NDVIs and the ground truth in Fig. 10, where R^2, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) are also displayed. It can be found that the proposed PNXnet gets the highest consistency with the ground truth, showing R^2 above 0.91, RMSE under 0.05, and MAE under 0.04. In traditional algorithms, MTF-GLP-HPM achieves the best fitting with the ground truth, which illustrates the same conclusion as visual results. Besides, BDSD, GSA, and AWLP tend to estimate NDVIs with a lower value. As for deep learning-based methods, all models are inclined to produce higher NDVIs,

TABLE III: Quantitative comparison of deep learning-based methods on Gaofen-2 and Gaofen-1 images, including reduced-resolution and full-resolution testing. Models are all trianed on Gaofen-2 data set and transformed to Gaofen-1 sensor. The best performance is shown in **bold** and the second best is underlined.

Sensors	Models	Reduced Resolution					Full Resolution				
		CC	PSNR	SSIM	SAM	ERGAS	D_λ	D_s	D_ρ	QNR	HQNR
Gaofen-2	PanNet	0.9757	42.5536	0.9676	2.2735	1.6732	0.3482	0.3391	0.3350	0.4308	0.6171
	DRPNN	0.9793	42.4807	0.9760	2.9592	1.8487	0.0178	0.0403	0.2807	0.9426	0.9115
	MSDCNN	0.9792	42.6150	0.9763	2.7362	1.7741	0.3162	0.1145	0.5046	0.6056	0.8516
	ResTFNet	0.9832	42.7154	0.9784	2.5056	1.6961	**0.0161**	0.0392	**0.2565**	0.9453	0.9050
	PNXnet	**0.9838**	**44.8437**	**0.9800**	**1.7421**	**1.4068**	0.0295	**0.0132**	0.2728	**0.9577**	**0.9373**
Gaofen-1	PanNet	0.9762	37.5012	0.9706	0.7743	1.1364	0.1693	0.2316	0.3137	0.6383	0.2693
	DRPNN	0.9766	41.7804	0.9735	1.1209	0.7477	0.0778	0.0955	0.2988	0.8341	0.8041
	MSDCNN	0.9656	39.3684	0.9612	1.1244	0.9100	0.3392	0.1340	0.5069	0.5722	0.4521
	ResTFNet	0.9598	39.2505	0.9658	1.4544	0.9923	**0.0627**	0.0960	0.2584	0.8473	0.7503
	PNXnet	**0.9783**	**42.7794**	**0.9755**	**0.7673**	**0.7004**	0.0741	**0.0473**	**0.2468**	**0.8820**	**0.7947**

TABLE IV: Quantitative results in ablation study of the proposed modules and strategies. The best performance is shown in **bold** and the second best is underlined.

Methods	ROB	NF-ResB	CC	PSNR	SSIM	SAM	ERGAS
ResTFNet	✗	✗	0.98315	42.7154	0.9784	2.5056	1.6961
PNXnet w/o NF-ResB	✓	✗	0.98343	44.7436	0.9795	1.9902	1.4540
The proposed PNXnet	✓	✓	**0.98380**	**44.8437**	**0.9800**	**1.7421**	**1.4068**

especially MSDCNN. ResTFNet also obtains good results, while its fitting model looks more scattered than the proposed PNXnet.

C. Model generalization from Gaofen-2 to Gaofen-1 satellite

Generalization ability plays a great role in deep learning-based model performance. In this paper, we utilize two similar sensors, Gaofen-2 and Gaofen-1, to compare the model generalization of five deep learning-based models. All models are trained on the same data-set that comes from Gaofen-2 images and tested on both Gaofen-2 and Gaofen-1 images. Results are reported in Table III.

In the reduced-resolution testing, the proposed PNXnet obtains the best quantitative results on both Gaofen-2 and Gaofen-1 sensors. As we can see, ResTFNet achieves the second-best on Gaofen-2 data set, while its results on Gaofen-1 images get worse. On the contrary, DRPNN shows no advantage on Gaofen-2 images, but when transferred into Gaofen-1 data set, the rank of its quantitative results is improved. Moreover, in the full-resolution testing, ResTFNet performs well on spectral maintaining as it obtains the best D_λ, and the spatial fidelity of PNXnet is superior.

Because the NIR band is particularly important in generating NDVI, we also select images randomly from Gaofen-2 and Gaofen-1 data sets and present the visual results of the NIR band in Fig. 11, where NDVI is shown in the enlarged area. Generally, spatial details in PAN images are hardly injected into the NIR band because of their different imaging mechanisms. So, it is obvious that, on Gaofen-2 images, PanNet, DRPNN, and MSDCNN get blurry results and ResTFNet shows distinct color differences with the ground truth. Nevertheless, PNXnet fuses the details into the NIR band as well as keeps the original color information. Besides, other methods tend to overestimate NDVI as showing more red area, while PNXnet keeps good fidelity with ground truth. Moreover, transferred to Gaofen-1 data, all models show performance degradation on improving spatial details, such as small lakes in the red box. In this situation, although MSDCNN and DRPNN could inject some spatial details, they suffer lower NDVI somewhere. However, PNXnet shows good spatial details as well as stable NDVI generation.

Similar conclusions can be found from the results in the full-resolution testing as shown in Fig. 12. The NIR bands of PanNet results suffer from unclear textures and structures. DRPNN shows better stability than MSDCNN for injecting more details into NDVIs. ResTFNet truly acquires more textures, however, it shows obvious color inconsistency with the original NDVI. The proposed PNXnet generates the NIR with enough spatial details as well as keeps high consistency with the original NDVI in tone.

Comparing the results between reduce-resolution and full-resolution testing, it can be observed that when the model trained on the simulated low-resolution data set is employed to solve pan-sharpening in real resolution, ResTFNet and PNXnet are more likely qualified because they don't show severe performance degradation. In summary, PNXnet shows not only good model generalization but also superior fusion effect whether in pan-sharpening or in NDVI generation, which is more suitable for daily application.

D. Ablation study and model complexity

In the proposed PNXnet, there are two main modules, including ROB and NF-group ResNet prior. To verify their effectiveness respectively, an ablation study on PNXnet is carried on in this subsection, and ResTFNet is chosen as the baseline.

Table IV lists the quantitative results in the ablation study. As we can see, the physical knowledge optimization-driven framework can greatly improve the fusion effect as the first two rows show, especially in spectral maintaining. It is noted that *PNXnet w/o NF-ResB* utilized the ResNet without BN as prior, while the residual units utilized in ResTFNet utilized BN. Under this background, Table IV illustrates that the NF ResNet haves surpassed ResNet with BN as well as ResNet without BN.

TABLE V: Comparisons on model complexity and computational speed between five deep learning-based models.

	Params/K	FLOPs/G	Time/s	Convergence
PanNet	78.920	4.799	0.0161	225 epochs
DRPNN	3673.881	224.161	0.0381	300 epochs
MSDCNN	228.556	13.932	0.0237	300 epochs
ResTFNet	2366.312	28.814	0.0250	189 epochs
PNXnet	309.977	17.357	0.0311	87 epochs

Furthermore, comparisons on model complexity and computational speed between deep learning-based models are reported in Table V. We counted the model parameter number (Params) in Kilos, floating-point operations (FLOPs) in Giga, running time in seconds, and the convergence in epochs, where the previous two are used to measure model complexity and the rest represents computational speed.

PanNet acquires the least parameters and FLOPs, which also costs the running time. Due to the 7×7 kernel size in all convolutional layers, DRPNN costs large Params, FLOPs, and the slowest running speed. Although ResTFNet costs many parameters, the residual blocks boost its running speed. However, ResTFNet still needs many epochs to achieve convergence. Seeing PNXnet, we can find that physical knowledge optimization-driven framework and group convolution reduce Params and FLOPs with a little cost of running time, and NF-ResB improves the convergence speed.

IV. Conclusion

This study shows a new approach to integrating spatial details in PAN images with spectral information in MS images and its application in generating high-resolution NDVI. By unfolding a variational pan-sharpening variational model into CNN with normalizer-free group ResNet prior, the proposed PNXnet fuses the low-resolution MS images and high-resolution PAN images to obtain high-resolution MS images and further produces high-resolution NDVIs. Experimental results illustrate that the fused MS images are visually satisfying and highly reliable, and the estimated NDVI shows a high consistency to the ground truth with R2 above 0.91. Moreover, the good model generation for pan-sharpening also shows a good advantage for generating high-resolution NDVI. Furthermore, fewer parameters, low FLOPs, and quicker computational speed make the daily application of PNXnet possible.

References

[1] Y. Li, S. Martinis, and M. Wieland, "Urban flood mapping with an active self-learning convolutional neural network based on terrasar-x intensity and interferometric coherence," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 178–191, 2019.

[2] L. Deng, Z. Mao, X. Li, Z. Hu, F. Duan, and Y. Yan, "Uav-based multispectral remote sensing for precision agriculture: A comparison between different cameras," *ISPRS journal of photogrammetry and remote sensing*, vol. 146, pp. 124–136, 2018.

[3] M. C. A. Picoli, G. Camara, I. Sanches, R. Simões, A. Carvalho, A. Maciel, A. Coutinho, J. Esquerdo, J. Antunes, R. A. Begotti *et al.*, "Big earth observation time series analysis for monitoring brazilian agriculture," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 328–339, 2018.

[4] Q. Zhang, Q. Yuan, J. Li, Y. Wang, F. Sun, and L. Zhang, "Generating seamless global daily AMSR2 soil moisture (SGD-SM) long-term products for the years 2013-2019," *Earth Syst. Sci. Data*, vol. 13, no. 3, pp. 1385–1401, Mar. 2021.

[5] U. Rajasekar and Q. Weng, "Urban heat island monitoring and analysis using a non-parametric model: A case study of indianapolis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 86–96, 2009.

[6] M. Piles, A. Camps, M. Vall-llossera, I. Corbella, R. Panciera, C. Rudiger, Y. H. Kerr, and J. Walker, "Downscaling smos-derived soil moisture using modis visible/infrared data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 9, pp. 3156–3166, 2011.

[7] A. Mui, Y. He, and Q. Weng, "An object-based approach to delineate wetlands across landscapes of varied disturbance with high spatial resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 109, pp. 30–46, 2015.

[8] Q. Zhang, Q. Yuan, J. Li, Z. Li, H. Shen, and L. Zhang, "Thick cloud and cloud shadow removal in multitemporal imagery using progressively spatio-temporal patch group deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 162, pp. 148–160, Apr. 2020.

[9] Z. Wu, W. Zhu, J. Chanussot, Y. Xu, and S. Osher, "Hyperspectral anomaly detection via global and local joint modeling of background," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3858–3869, 2019.

[10] Z. Wu, J. Sun, Y. Zhang, Y. Zhu, J. Li, A. Plaza, J. A. Benediktsson, and Z. Wei, "Scheduling-guided automatic processing of massive hyperspectral image classification on cloud computing architectures," *IEEE Transactions on Cybernetics*, vol. 51, no. 7, pp. 3588–3601, 2020.

[11] Z. Wu, J. Sun, Y. Zhang, Z. Wei, and J. Chanussot, "Recent developments in parallel and distributed computing for remotely sensed big data processing," *Proceedings of the IEEE*, vol. 109, no. 8, pp. 1282–1305, 2021.

[12] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2022.

[13] Y. Xiao, Q. Yuan, J. He, Q. Zhang, J. Sun, X. Su, J. Wu, and L. Zhang, "Space-time super-resolution for satellite video: A joint framework based on multi-scale spatial-temporal transformer," *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, p. 102731, 2022.

[14] Q. Zhang, Q. Yuan, Z. Li, F. Sun, and L. Zhang, "Combined deep prior with low-rank tensor SVD for thick cloud removal in multitemporal images," *ISPRS J. Photogramm. Remote Sens.*, vol. 177, pp. 161–173, Jul. 2021.

[15] 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, 1990.

[16] P. Kwarteng and A. Chavez, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens*, vol. 55, no. 1, pp. 339–348, 1989.

[17] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Jan. 4 2000, uS Patent 6,011,875.

[18] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques," *Remote Sensing of Environment*, vol. 22, no. 3, pp. 343–365, 1987.

[19] P. J. Burt and E. H. Adelson, "The laplacian pyramid as a compact image code," in *Readings in computer vision*. Elsevier, 1987, pp. 671–679.

[20] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," in *Wavelets and statistics*. Springer, 1995, pp. 281–299.

[21] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Transactions on image processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[22] J.-L. Starck, E. J. Candès, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Transactions on image processing*, vol. 11, no. 6, pp. 670–684, 2002.

[23] M. González-Audícana, J. L. Saleta, R. G. Catalán, and R. García, "Fusion of multispectral and panchromatic images using improved ihs and pca mergers based on wavelet decomposition," *IEEE Transactions on Geoscience and Remote sensing*, vol. 42, no. 6, pp. 1291–1299, 2004.

[24] 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 Transactions on Geoscience and Remote Sensing*, vol. 43, no. 10, pp. 2376–2385, 2005.

[25] F. D. Javan, F. Samadzadegan, S. Mehravar, A. Toosi, R. Khatami, and A. Stein, "A review of image fusion techniques for pan-sharpening of high-resolution satellite imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 171, pp. 101–117, 2021.

[26] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for p+ xs image fusion," *International Journal of Computer Vision*, vol. 69, no. 1, pp. 43–58, 2006.

[27] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 318–322, 2013.

[28] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 265–10 274.

[29] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[30] W. Yao, Z. Zeng, C. Lian, and H. Tang, "Pixel-wise regression using u-net and its application on pansharpening," *Neurocomputing*, vol. 312, pp. 364–371, 2018.

[31] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "Psgan: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[32] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[33] R. Dian, S. Li, A. Guo, and L. Fang, "Deep hyperspectral image sharpening," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 11, pp. 5345–5355, 2018.

[34] M. Zhou, X. Fu, J. Huang, F. Zhao, A. Liu, and R. Wang, "Effective pan-sharpening with transformer and invertible neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

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

[36] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sensing and Imaging*, vol. 17, no. 1, pp. 1–16, 2016.

[37] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*. IEEE, 2017, pp. 1–4.

[38] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 11, no. 5, pp. 1656–1669, 2018.

[39] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1795–1799, 2017.

[40] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5449–5457.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[42] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.

[43] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5549–5563, 2019.

[44] J. Liu, Y. Feng, C. Zhou, and C. Zhang, "Pwnet: An adaptive weight network for the fusion of panchromatic and multispectral images," *Remote Sensing*, vol. 12, no. 17, p. 2804, 2020.

[45] H. Zhang and J. Ma, "Gtp-pnet: A residual learning network based on gradient transformation prior for pansharpening," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 223–239, 2021.

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

[47] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 4295–4310, 2020.

[48] S. Seo, J.-S. Choi, J. Lee, H.-H. Kim, D. Seo, J. Jeong, and M. Kim, "Upsnet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images," *IEEE Access*, vol. 8, pp. 201 199–201 217, 2020.

[49] M. Ciotola, S. Vitale, A. Mazza, G. Poggi, and G. Scarpa, "Pansharpening by convolutional neural networks in the full resolution framework," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

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

[51] H. Zhou, Q. Liu, and Y. Wang, "Pgman: An unsupervised generative multiadversarial network for pansharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 6316–6327, 2021.

[52] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[53] S. De and S. L. Smith, "Batch normalization biases residual blocks towards the identity function in deep networks," *arXiv preprint arXiv:2002.10444*, 2020.

[54] A. Brock, S. De, and S. L. Smith, "Characterizing signal propagation to close the performance gap in unnormalized resnets," *arXiv preprint arXiv:2101.08692*, 2021.

[55] K. Zhang, W. Zuo, S. Gu, and L. Zhang, "Learning deep cnn denoiser prior for image restoration," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3929–3938.

[56] R. Dian, S. Li, and X. Kang, "Regularizing hyperspectral and multispectral image fusion by cnn denoiser," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1124–1135, 2020.

[57] J. He, Q. Yuan, J. Li, and L. Zhang, "Ponet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images," *Information Fusion*, vol. 80, pp. 205–225, 2022.

[58] J. He, J. Li, Q. Yuan, H. Shen, and L. Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[59] D. Ren, W. Zuo, D. Zhang, L. Zhang, and M.-H. Yang, "Simultaneous fidelity and regularization learning for image restoration," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 284–299, 2019.

[60] S. De and S. L. Smith, "Batch normalization biases residual blocks towards the identity function in deep networks," *arXiv preprint arXiv:2002.10444*, 2020.

[61] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2752–2761.

[62] T. Zhang, G.-J. Qi, B. Xiao, and J. Wang, "Interleaved group convolutions," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4373–4382.

[63] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[64] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal mmse pan sharpening of very high resolution multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 1, pp. 228–236, 2007.
[65] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 1, pp. 295–309, 2010.
[66] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of ms + pan data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 10, pp. 3230–3239, 2007.
[67] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The arsis concept and its implementation," *Photogrammetric engineering and remote sensing*, vol. 66, no. 1, pp. 49–61, 2000.
[68] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 5, pp. 591–596, 2006.
[69] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric engineering and remote sensing*, vol. 63, no. 6, pp. 691–699, 1997.
[70] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[71] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (sips)-interactive visualization and analysis of imaging spectrometer data," *Remote sensing of environment*, vol. 44, no. 2-3, pp. 145–163, 1993.
[72] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 193–200, 2008.
[73] G. Scarpa and M. Ciotola, "Full-resolution quality assessment for pansharpening," *Remote Sensing*, vol. 14, no. 8, p. 1808, 2022.
[74] M. M. Khan, L. Alparone, and J. Chanussot, "Pansharpening quality assessment using the modulation transfer functions of instruments," *IEEE transactions on geoscience and remote sensing*, vol. 47, no. 11, pp. 3880–3891, 2009.
[75] Y. Chen, R. Cao, J. Chen, X. Zhu, J. Zhou, G. Wang, M. Shen, X. Chen, and W. Yang, "A new cross-fusion method to automatically determine the optimal input image pairs for ndvi spatiotemporal data fusion," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 5179–5194, 2020.
[76] S. H. Yueh, R. Shah, M. J. Chaubell, A. Hayashi, X. Xu, and A. Colliander, "A semiempirical modeling of soil moisture, vegetation, and surface roughness impact on cygnss reflectometry data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

Jiang He received the B.S. degree in remote sensing science and technology from faculty of geosciences and environmental engineering in Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree in School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include hyperspectral super-resolution, image fusion, quality improvement, remote sensing image processing and deep learning.

Qiangqiang Yuan received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently a Professor. He published more than 90 research papers, including more than 70 peer-reviewed articles in international journals such as the IEEE Transactions Image Processing and the IEEE Transactions on Geoscience and Remote Sensing. His current research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was the recipient of the Youth Talent Support Program of China in 2019, and the Top-Ten Academic Star of Wuhan University in 2011. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is on the editor board of nine international journals, and has frequently served as a Referee for more than 50 international journals for remote sensing and image processing.

Jie Li received the B.S. degree in sciences and techniques of remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011 and 2016.

He is currently an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include image quality improvement, image super-resolution reconstruction, data fusion, remote sensing image processing, sparse representation and deep learning.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was the Head of the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Remote Sensing Division, Wuhan University. He is currently the "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China. He was a Principal Scientist for the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He has more than 700 research articles and six books. He is the holder of 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is also an Executive Member (Board of Governor) of the China National Committee of International Geosphere–Biosphere Program and the China Society of Image and Graphics and a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award and the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS). He regularly serves as the Co-Chair of the series SPIE conferences on multispectral image processing and pattern recognition, conference on Asia remote sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor for the International Journal of Ambient Computing and Intelligence, International Journal of Image and Graphics, International Journal of Digital Multimedia Broadcasting, Journal of Geo-Spatial Information Science, and Journal of Remote Sensing; and a Guest Editor for Journal of Applied Remote Sensing and Journal of Sensors. He is also serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.