Self-Supervised Interactive Dual-Stream Network for Pansharpening

Qing Guo, *Senior Member, IEEE*, He Jia, and Shengsang Yang

***Abstract*—Pansharpening is crucial for obtaining high-resolution multispectral images. Existing deep learning-based pansharpening networks rely on supervised learning with external reference labels. Due to the lack of actual fusion results for labeling, simulated degraded data is used with the original multispectral image as the fusion result label. These process steps are cumbersome, which also leads to the problem of scale degradation, and the fusion relationship between data before and after the degradation cannot represent the real fusion relationship. To address these limitations, we propose a self-supervised interactive dual-stream network for pansharpening using real training datasets. Our approach incorporates a dual-stream network architecture, comprising a spatial scale enhancement stream and a spectral channel attention stream. Spatial and spectral features essential for fusion are extracted from the original panchromatic and multispectral images, respectively. Through interconnection at different levels, the network expands the search range in the feature space, enabling continuous interaction between spatial and spectral information during feature extraction and transmission. This ensures the injection of spatial features of varying scales into corresponding-scale spectral features, enhancing complementarity between features. Moreover, we introduce a novel joint spatial-spectral loss function, leveraging the original panchromatic and multispectral images themselves as self-supervised labels. Experimental results on diverse satellite datasets demonstrate the outstanding fusion performance of our method, as assessed through both subjective qualitative evaluation and objective quantitative evaluation. Furthermore, our proposed method exhibits exceptional generalization performance for full-scale remote sensing images, showcasing its practical value.**

***Index Terms*—Image fusion, pansharpening, remote sensing, self-supervised learning, two-stream interaction.**

I. Introduction

THERE are certain differences and complementarities between multispectral (MS) and panchromatic (PAN) images: MS images maintain high-resolution characteristics in the spectral domain, while PAN images have high-resolution characteristics in the spatial domain. However, due to the limitation of data transmission bandwidth and physical conditions of sensors, optical remote sensing images acquired from the same satellite can only maintain high-resolution characteristics in a single domain of space or spectrum, which prevents us from directly acquiring MS images with high spatial resolution (HRMS) characteristics. Pansharpening can solve this problem well. By integrating the spatial structure information of PAN images with high spatial resolution and the spectral information of MS images with high spectral resolution, it can fully combine the complementary information to overcome the defect of insufficient information in a single image, so that the fused image can depict more detailed spatial details while retaining the original spectral information as much as possible.

Traditional pansharpening methods are mainly divided into three categories: component substitution (CS) [1], [2], [3], multiresolution analysis (MRA) [4], [5], and variational optimization (VO) [6], [7], [8] methods. Among these, the CS and MRA approaches typically have two components: the first is the extraction of spatial information from PAN images, and the second is the injection of the extracted information into the up-sampled MS images. Intensity-Hue-Saturation (IHS) [9] and principal component analysis (PCA) [10] methods are the initial CS algorithms. The adaptive GS (Adaptive GS, GSA) [11] realizes the fusion of PAN and MS information through the combination of guided filter and Schmidt (Gram-Schmidt, GS) transformation. Then, by modeling the pixel values of each channel of PAN and MS, a partial replacement adaptive CS (PRACS) [12] is proposed. The fusion results obtained by the CS method usually have good spatial detail information, but due to the local nonsimilarity between PAN and MS images, the fusion results often face serious spectral distortion problems. Currently, common MRA methods include the smoothing filter-based intensity modulation (SFIM) [13], the generalized Laplacian pyramid (GLP) [14], and the additive wavelet luminance proportional (AWLP) [15]. Compared with the CS method, the MRA method is capable of multiscale decomposition and generally achieves better spectral quality.

The VO method is also called the model-based algorithm. P+XS [16] is the first application of the VO method in the field of pansharpening. Subsequently, Fasbender et al. [17] have proposed a Bayesian adaptive fusion method based on the statistical relationship between MS bands and PAN bands. Li and Yang [18] use the basis pursuit (BP) algorithm to reconstruct the fusion model through sparse representation, thus realizing the fusion

Manuscript received 9 July 2023; revised 5 September 2023; accepted 2 October 2023. Date of publication 10 October 2023; date of current version 6 November 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 61771470 and in part by the Leading Foundation on Frontier Sciences and Disruptive Technology Research of the Aerospace Information Research Institute, Chinese Academy of Sciences, under Grant E0Z218010F. *(Corresponding author: Qing Guo.)*

Qing Guo is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: guoqing@aircas.ac.cn).

He Jia and Shengsang Yang are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jiahe21@mails.ucas.ac.cn; yangshengsang21@mails.ucas.ac.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3323502

of PAN images and low-resolution MS images. Wu et al. [19] introduce a nonconvex regularization model (NC-FSRM) for pansharpening. Moreover, Wu et al. [20] innovatively apply the low-rank tensor completion-based framework to pansharpening, presenting LRTCPan. The VO method relies more on prior conditions, and requires a large number of iterative operations when solving functions, and its practicability needs to be improved.

In recent years, deep learning (DL) methods have sprung up due to their super nonlinear fitting ability and abstract feature extraction ability. At present, the pansharpening network based on DL mainly includes the network based on residual connection (detail injection-based CNN, DiCNN) [21], the network based on densely connection (multilevel dense connection network with feedback connections, MDCwFB) [22], based on generative adversarial network (generative adversarial network for Pansharpening, PSGAN) [23] and the transformer-based network (HyperTransformer) [24]. In the latest study, Zhou et al. [25] introduce a novel pansharpening network using cross-modality joint learning. Yan et al. [26] consider the pansharpening process as an alternating iterative reverse filtering procedure and design a straightforward and efficient network architecture. Most of the above-mentioned methods use a single-stream network architecture. Liu et al. [27] are inspired by the two-stream network [28]. Considering the information difference between MS and PAN, they have proposed a two-stream fusion network (TFNet). The dual-branch architecture is used to capture the feature information of MS and PAN, respectively, and then the features are spliced together to reconstruct the fusion results. Fang et al. [29] have proposed a parallel pyramid convolutional neural network (PPN) to achieve pansharpening. The network uses different branches to process spatial details and spectral details. In the two detail branches, the pyramid network structure is introduced to solve the weak correlation problem caused by scale differences. He et al. [30] propose a multiscale dual-domain guidance network (MSDDN) by fully exploring and utilizing the differentiated information in the spatial and frequency domains. In addition, bidirectional networks are also used in pansharpening. Zhou et al. [31] integrate pansharpening and degradation modeling using an invertible neural network, achieving two-directional closed-loop learning for low-resolution MS pan-sharpening and HRMS degradation.

The above-mentioned DL-based fusion methods mostly use supervised learning methods with external reference labels for training. Due to the lack of real HRMS images as reference labels, the simulated degraded datasets must be produced according to the Wald protocol [32]. Although this approach guarantees the ability of the model to describe spatial and spectral information to a certain extent, it also has the following disadvantages: 1) Reference labels need to be manually produced, the steps are cumbersome, and the scale disparity between simulated data and real data is ignored; 2) The degradation process in the Wald protocol is difficult to simulate the degradation process of real data, and the downsampling operation in the protocol is easy to cause the loss of spatial information of the PAN image.

To alleviate the problems in supervised learning, the authors in [33] and [34] have proposed the no-reference-based pansharpening method, respectively. Ni et al. [35] propose a self-supervised network based on learnable degradation processes (LDPNet). The spatial and spectral consistency is achieved by introducing the degenerate fuzzy process and the hybrid loss function. Moreover, Guo et al. [36] have presented a dual spatial-spectral fusion network (DSSN), which introduces gradient features as prior knowledge on the basis of a dual-branch architecture, thereby guiding the backbone network to extract richer spatial information. Inspired by these, this article proposes a self-supervised pansharpening algorithm that no longer requires fusion result labels, which uses a dual-stream network architecture to design the corresponding spectral channel attention stream and spatial scale enhancement stream for MS and PAN, respectively, to extract the spectral features and spatial features of both. In each stream, in order to improve the efficiency of feature extraction and enhance the complementarity between features, the transmission and multiplexing of features are realized in the form of level interconnection. In each level, the dense connection is used to encode and decode the input data, and the spatial features at different scales obtained from the encoding are injected into the spectral features at the corresponding scales to achieve the fusion of feature levels. In designing the loss function, this article introduces the spectral angle mapping function and KL divergence function, and establishes the joint spatial-spectral loss function with certain weight assignment rules to ensure that the original MS spectral features and the original PAN spatial features required for fusion are learned directly in the self-supervised mode. The contributions of this article are as follows.

1) We introduce a pivotal self-supervised pansharpening framework that harnesses the original images for training, eliminating the need for artificial labels. Our method not only dispenses with preprocessing but also effectively addresses the limitations of fusion frameworks reliant on simulated datasets. This method bridges the gap between simulation and real-world applications, ensuring that fusion outcomes align more closely with actual scenarios.
2) The self-supervised interactive dual-stream pansharpening (SIDP) network proposed in this article is based on two parallel streams—the spatial scale enhancement stream and the spectral channel attention stream, which process the spatial and spectral information, respectively. The concatenated multiplexing of features in streams and the continuous interaction of information between streams are realized by using the level interconnection, which improves the interaction efficiency of spatial and spectral information through a new information interaction mode, hence to reconstruct HRMS results.
3) In the process of network optimization, the joint spatial-spectral loss function consisting of the spatial constraint function and the spectral constraint function, is designed to get rid of the dependence of the network model on the reference label, and greatly reduce the workload of labeling while improving the fusion performance. The superiority of the proposed method in this article is effectively verified by using five different satellite sensor image data of GaoFen-1, GaoFen-2, WorldView-2, WorlView-3, and Pleiades to test and generalize the model.

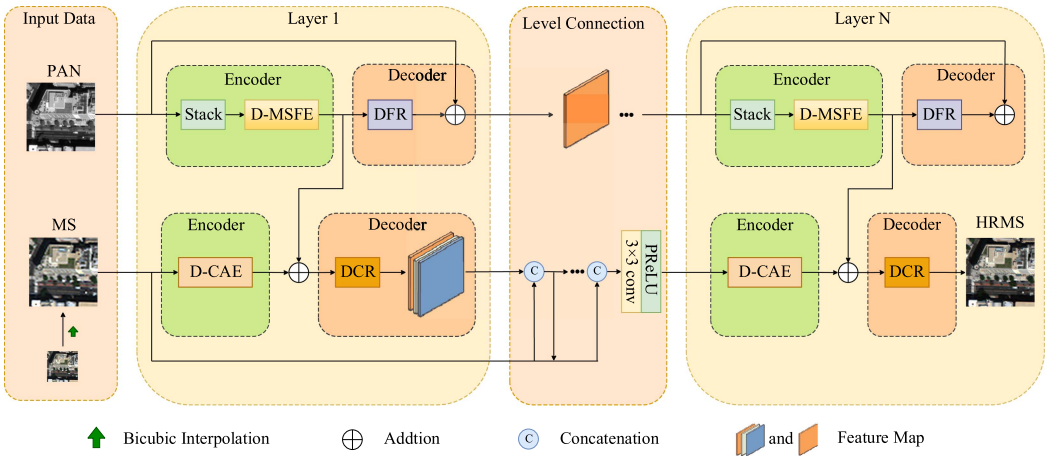


Fig. 1. Architecture of the proposed SIDP network. Stack denotes the expansion along channel dimension. Conv denotes the convolution operation. PReLU denotes the activation function.

The rest of this article is organized as follows. Section II introduces the specific implementation details of the SIDP model and the construction of the loss function. Section III comprehensively evaluates the SIDP method with the traditional and DL fusion methods, and gives the test and generalization results on different datasets. Section IV is the discussion of the SIDP method. Finally, Section V concludes this article.

II. Proposed Method

At first, the representation symbols of the data variables used in this article are unified. $M \in R^{m\times n\times c}$ and $P \in R^{M\times N}$ are used to represent the original MS image and the original PAN image, respectively, where m and n represent the width and height of the low spatial resolution MS image, M and N represent the width and height of the high spatial resolution PAN image, c represents the channel number of MS. $M\uparrow = R^{M\times N\times c}$ represents the MS after up-sampling according to the size of the PAN image. $\widehat{M} \in R^{M\times N\times c}$represents the fusion result HRMS. The spatial resolution scaling factor of MS and PAN is $r = M/m = N/n$.

A. Network Architecture

The architecture of the SIDP network proposed in this article is shown in Fig. 1. The network consists of two parts: the spatial scale enhancement stream and the spectral channel attention stream. The spatial stream contains two modules—the dense multiscale feature enhancement (D-MSFE) and the dense feature restore (DFR). The spectral stream includes another two modules—the dense channel attention enhancement (D-CAE) and the dense channel restore (DCR). The spatial stream and the spectral stream are used to extract spatial features and spectral features required for fusion, respectively. Moreover, in order to enhance the feature extraction capability of the network, a level interconnection method is used to promote the flow and transmission of features, and to enhance information exchange between two streams of network while realizing feature multiplexing.

The SIDP model can be expressed as follows:

$$\widehat{M} = f\left(M\uparrow, P; \Theta\right) \tag{1}$$

where $f(\cdot)$ represents the dual-stream fusion framework, and Θ represents the weight parameters of the model. First, the original P and M are used as the network inputs which are, respectively, input into two streams independently. P is stacked c times along the channel dimension to obtain $\widetilde{P} \in R^{M\times N\times c}$. M is interpolated by the bicubic interpolation operation to obtain $M\uparrow$. Afterward, through the D-MSFE module and the D-CAE module, $\widetilde{P}$ and $M\uparrow$ are, respectively, mapped to the high-dimensional feature space. The high-frequency spatial information contained in $\widetilde{P}$ and the spectral feature information contained in $M\uparrow$ are extracted. The extracted spatial information is injected into the corresponding spectral information to complete the information interaction. Then, through the DFR module and the DCR module, the dimensionality reduction operation is performed on the spatial features and the spectral features, to realize the low-dimensional reconstruction of the two features, and get the input for the next level. After N-layer feature iteration, the search range of network in the feature space can be effectively expanded, thereby improving the fusion accuracy and efficiency of spatial and spectral information, and completing the HRMS reconstruction operation.

B. Implementation of the SIDP Framework

1) Dense Connectivity: The backbone network of the SIDP model is the dense connectivity block [37], which uses the cross-channel stitching to establish connections between the features of all previous layers and the current convolutional layer. This means that there is a path between any two layers, which is conducive to the feature transfer. Assuming that the current layer is the ith convolutional layer, the feature map of this layer is

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}]) \tag{2}$$

herein, $[x_0, x_1, \ldots, x_{l-1}]$ denotes the stitching of features in the $0, \ldots, (l-1)$th layers, and $H_l(\cdot)$ denotes the convolution and

nonlinear mapping operations in the *l*th layer. This structure can effectively capture the nonlinear relationship in the original data and improve the information flow between convolutional layers. In order to comprehensively map the complex spectral and texture details in MS and PAN to the feature space, this article introduces the densely connected blocks in the feature extraction operation, which drives the spatial and spectral information to be freely propagated in the network, also enables the network to consider the complex relationship between multiple features at the same time, thus effectively fusing the shallow and deep features.

2) Spatial Enhancement and Channel Attention: In view of the different characteristics of spectral information and spatial information in the fusion process, after designing the backbone framework, it is necessary to use different feature mapping methods to represent them. This article introduces the large kernel attention (LKA) [38] mechanism to enhance the spatial feature extraction ability of the network, which is expressed as follows:

$$\text{Attention} = Conv_{1\times1}(Conv_{SC-D}(Conv_{SC}(F_{\text{in}}))) \tag{3}$$

$$F_{\text{out}} = \text{Attention} \otimes F_{\text{in}} \tag{4}$$

where $Conv_{SC}(\cdot)$ denotes the separable convolution, $Conv_{SC-D}(\cdot)$denotes the separable dilated convolution, and $Conv_{1\times1}(\cdot)$ denotes the 1×1 channel convolution. Different from traditional convolution forms, this module combines the convolution operation with the self-attention mechanism, which helps to expand the network's perceptual field of view and realize long-term dependency correlation between features. Then, the module is used in the spatial stream, which can effectively improve the extraction efficiency of local information and long-distance information of spatial features, and enhance the network's ability to express multiscale spatial information.

Unlike the spatial information, the representation of spectral information is very dependent on the relationship between feature channel bands. Therefore, a spatial-pooling squeeze-and-excitation (SSE) channel attention module is proposed to further enhance the channel correlation and self-adaptation between spectral features in the spectral stream. The module allows the network to automatically learn the importance weights of each channel and thus adaptively implements the feature selection. By adjusting the channel importance, the network focuses on the useful spectral information for the task and suppresses the irrelevant and redundant information. In addition, this module upgrades the 1×1 average pooling in the original squeeze-and-excitation (SE) [39] channel attention to the 4×4 average pooling, which ensures the spatial structure information is preserved while extracting spectral features. In addition, in order to avoid the feature fading between different layers in the network, the residual structure is introduced into the module, which can effectively improve the efficiency of feature extraction in network and restrain the gradient degradation problem that is prone to occur in deep networks.

3) Restore Mechanism: Dense connections are helpful for the transfer of feature information and the flow of gradients, but the feature splicing and dimension-up operations in the module greatly increase the number of network parameters. In addition, this article uses the level interconnection to realize continuous interaction and adaptive search of features.

Although this operation can extract information from different feature spaces well, it is easy to cause the problem of linear increase in the number of network parameters. To alleviate the problem of parameter number increasing, the dense restore module is proposed, which is also implemented using the densely connected block. However, unlike the usual densely connected operation, the channel dimension of the output feature is reduced exponentially with each convolutional layer of this restore block. This approach can preserve the main information as much as possible while reducing the feature dimension. Afterward, the dimensionality reduction features are used for information extraction and fusion at subsequent levels, thereby enhancing the fusion performance and improving the fusion efficiency of the network. The specific implementation details of each module in the network are shown in Fig. 2.

Specifically, in the spatial scale enhancement stream, the D-MSFE module consists of densely connected blocks and LKA modules. First, dense connections are used to map PAN into a high-dimensional feature space. The spatial structure information of PAN is extracted from high-dimensional features of different scales as much as possible. Then the extracted spatial features are sent to the LKA module for integration, so that the adaptive searchability of the network is enhanced in high-dimensional space. The DFR module includes two parts of dense restoration and channel averaging, to achieve dimensionality reduction while retaining the original spatial characteristics. Afterward, the D-MSFE module and the DFR module are connected in series to control the parameters of the network while extracting multiscale spatial features. This module can be expressed as follows:

$$P_i = f_{PR}(F_P^i) + P_{i-1}, (i \geq 1) \tag{5}$$

$$F_P^i = f_{PFE}(P_{i-1}), (i \geq 1). \tag{6}$$

In (5) and (6), $f_{PFE}(\cdot)$ denotes the D-MSFE module, $f_{PR}(\cdot)$denotes the DFR module, F_P^i denotes the spatial structure features extracted by the *i*th layer module in the spatial extraction stream, and P_{i-1} and P_i denote the spatial mapping of the *l*th layer input and output, respectively. When $i > 1$, P_{i-1} represents the mapped PAN image generated by the model. When $i = 1$, P_{i-1} represents the original PAN image.

Correspondingly, in the spectral channel attention stream, the D-CAE module is designed for spectral feature extraction, and the DCR module is designed for feature integration and feature dimensionality reduction. The D-CAE module consists of the dense connection and SSE. SSE serves to guide the learning objectives of D-CAE and directs the objectives of the module to the learning of spectral features. Then the dense connection block in the D-CAE module is replaced with the dense restore block to generate the DCR module, which can adaptively integrate the spatial features obtained from the spatial scale enhancement stream with the spectral features obtained in this module according to its own needs. The calculation process

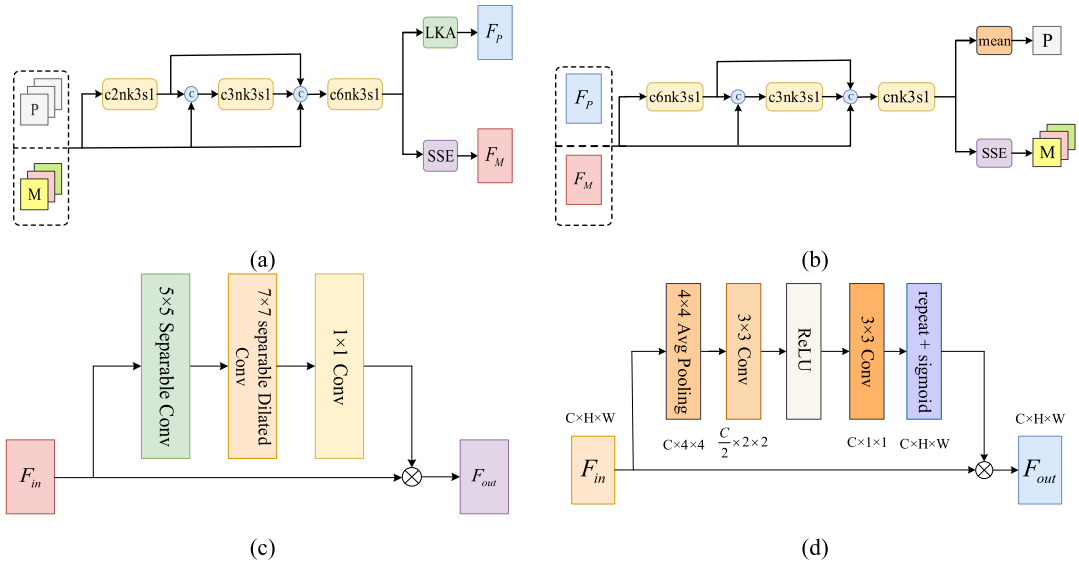


Fig. 2. Structural diagram of each module in SDIP. (a) D-MSFE and D-CAE. (b) DFR and DCR. (c) LKA. (d) SSE, where c3nk3s1 denotes a convolution layer with a 3 × 3 kernel size, 3n channels and 1 stride, n denotes a changeable channel dimension.

of the spectral channel attention stream is expressed as follows:

$$M_i^{\uparrow} = \begin{cases} f_{MR}(F_M^i + F_p^i) & ,i = 1 \\ f_{MR}(F_{M-cat}^i + F_p^i) & ,i > 1 \end{cases} \tag{7}$$

$$F_M^i = f_{MCE}(M_{i-1}^{\uparrow}), i = 1 \tag{8}$$

$$F_{M-cat}^i = f_{MCE}(f_{\mathrm{Re}-conv}(cat(M_1^{\uparrow}, M_2^{\uparrow}, \cdots, M_{i-1}^{\uparrow}))), i > 1 \tag{9}$$

where $f_{MCE}(\cdot)$ denotes the D-CAE module, $f_{MR}(\cdot)$ denotes the DCR module, and F_P^i is the spatial information extracted from the corresponding layer in the spatial scale enhancement stream, see (6). $f_{\mathrm{Re}-conv}(\cdot)$ denotes the cascade operation of activation and convolution, which is used to integrate the intermediate results collected from the different layers. F_M^i and F_{M-cat}^i represent the spectral feature information extracted by different layer modules in the spectral channel attention stream. $M_{i-1}^{\uparrow}$represents the mapped up-sampled MS image input by the lth layer. Correspondingly, $M_{i-1}^{\uparrow}$ represents the original up-sampled MS image when $i = 0$, $M_i^{\uparrow}$ is the high-resolution fusion result HRMS output by the network when $i = N$ and $M_i^{\uparrow} = \widehat{M}$ at this time. The algorithm flow is given in Algorithm 1.

C. Loss Function

To realize the fusion requirement in the self-supervised mode, a joint spatial-spectral loss function is designed in this article. Considering that the spectral information and spatial information of the fusion result come from MS and PAN, respectively, the spectral angle mapping (SAM) [40] function is introduced as the spectral metric function of the network, and the KL divergence function [41] is introduced as the spatial metric function. Then the spatial loss and spectral loss are combined to achieve the purpose of directly learning PAN spatial features and MS spectral features. The construction process of the loss function is shown in Fig. 3.

Algorithm 1: SIDP.

Input: MS: $M \in \mathbb{R}^{m \times n \times c}$, PAN: $P \in \mathbb{R}^{M \times N}$
Output: The Pansharpening HRMS: $\widehat{M} \in \mathbb{R}^{M \times N \times c}$
1: Begin:
2: for $epoch = 1 \rightarrow epochs$ do
3: The M is up-sampled by the bicubic interpolation.
4: for $i = 1 \rightarrow N$do
5: Copy P along the channel dimension.
6: Extract spatial features by (6).
7: Extract spectral features by (8)–(9).
8: Use (7) to complete the interaction of information.
9: end for
10: $\widehat{M} \Leftarrow \widehat{M^i} \in \mathbb{R}^{M \times N \times c}$
11: End

SAM is a widely used spectral evaluation index to assess the spectral loss between HRMS and MS and indicate the difference in spectral angle between the two. It is given by the following:

$$SAM(z_1, z_2) = \arccos\left(\frac{< z_1, z_2 >}{\| z_1\|_2 \cdot \| z_2\|_2}\right). \tag{10}$$

The SAM value calculated in (10) is the spectral angle of a single pixel in image, $< \cdot, \cdot >$ is the vector inner product symbol, $|| \cdot ||_2$ represents the l_2 norm, $\arccos(\cdot)$represents the arcsine function, $z_1 \in R^{1\times1\times c}$ and $z_2 \in R^{1\times1\times c}$represent the spectral vector at the specified pixel coordinates.

KL divergence is an indicator function to calculate the magnitude of similarity. The more similar the two probability distributions are, the smaller the KL scatter is. It is defined as follows:

$$D_{KL}(P||Q) = \sum_{i=1}^{N} (p(x_i) \log p(x_i) - p(x_i) \log q(x_i)) \tag{11}$$

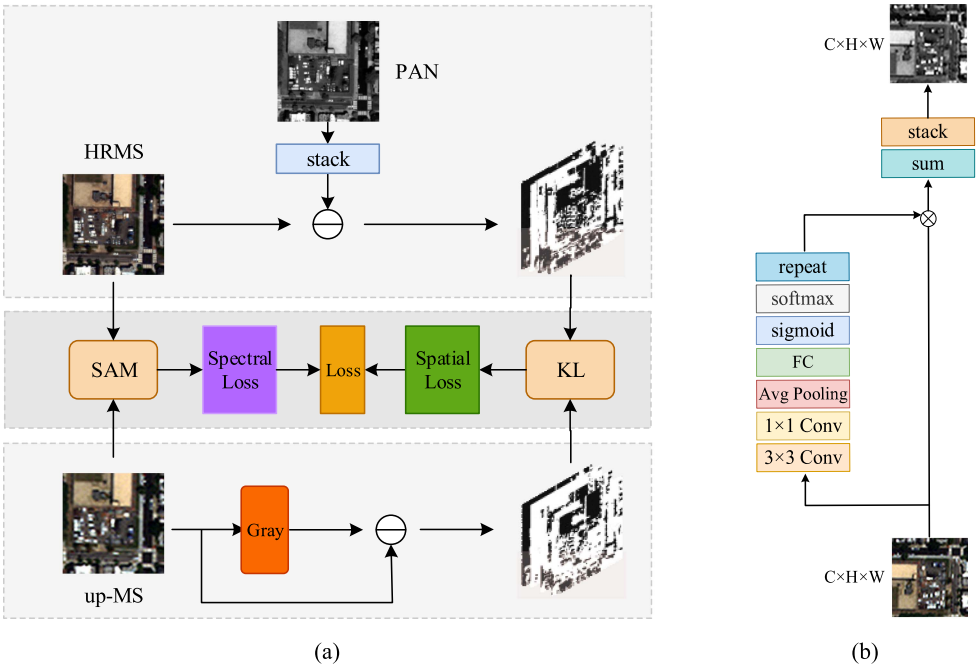


Fig. 3. Realization of the loss function. (a) Loss function construction diagram. (b) Grayscale degradation network.

where $p(x)$ is the probability distribution function of the real information, and $q(x)$ is the probability distribution function of the fitted information. By calculating the information entropy of the probability distribution of both, the information difference degree can be measured in a quantitative way.

1) Spectral Loss Function: The up-sampled MS and the original MS can be considered to contain the same spectral features. Spectral features are mapped into vector directions between image bands. Through minimizing the SAM value between HRMS and up-sampled MS, the spectral loss of the network can be controlled. The spectral loss function is expressed as follows:

$$L_{\text{spectral}} = SAM(\widehat{M}, M\uparrow). \tag{12}$$

2) Spatial Loss Function: Unlike the establishment process of spectral constraints, the connection between the predicted HRMS and the original PAN image cannot be directly established. However, the KL divergence function is used as the spectral loss function in the learnable degradation processes pansharpening (LDPNet) [35]. Inspired by this, this article designs the spatial loss based on the following three-point conventions: 1) The PAN image can be obtained by degrading the MS image of the same resolution; 2) The degraded grayscale image contains the spatial texture information of the source image; 3) The difference between MS and PAN images at different spatial resolutions should have a similar distribution. Thus, according to the establishment process of the loss function in LDPNet, the MS is degenerated into the grayscale image. The residual probability distribution map between the MS image and the grayscale image is established using the softmax function. The residual probability distribution between HRMS and PAN is fitted according to the above-mentioned residual probability distribution. After that, the KL divergence value between two probability distributions is minimized to constrain the spatial loss. The establishment process of the spatial loss is as follows:

$$M_G = f_G(M\uparrow) \tag{13}$$

$$res_{M\uparrow} = soft\max(M\uparrow - M_G) \tag{14}$$

$$res_{\widehat{M}} = soft\max(\widehat{M} - \widetilde{P}) \tag{15}$$

$$L_{\text{spatial}} = D_{KL}(res_{M\uparrow}, res_{\widehat{M}}) \tag{16}$$

where $f_G(\cdot)$ represents the grayscale degradation network, $M_G \in R^{M\times N\times c}$ is the grayscale image generated by the up-sampled MS through the degradation network, and $res_{M\uparrow}$ represents the probability distribution of the difference between the up-sampled MS image and the degraded grayscale image. $\widetilde{P} \in R^{M\times N\times c}$ is expanded from PAN along the channel dimension. $res_{\widehat{M}}$ represents the probability distribution of the difference between the fusion result HRMS and the original PAN.

3) Joint Spatial-Spectral Loss Function: The spectral loss function and the spatial loss function are weighted to construct the joint spatial-spectral loss function, which is expressed as follows:

$$L = \alpha \times L_{\text{spectral}} + \beta \times L_{\text{spatial}},\ (\alpha+\beta=1; \alpha>0; \beta>0) \tag{17}$$

where α and β are empirical values obtained through a large number of experimental tests. Unlike the general model that only uses a single loss function such as MSE or MAE, the joint loss function is more conducive to improve the learning efficiency of the network. The SAM function and KL function for controlling the spectral loss and spatial loss of the network, respectively, are used to guide the optimization process of the network accurately and effectively, thereby improving the overall fusion performance of the model.

TABLE I
BAND AND RESOLUTION PARAMETER OF REMOTE SENSING SATELLITE IMAGE

Satellite	Number of bands (PAN+MS bands)	Spatial resolution/(m)	
		MS	PAN
GaoFen-1	1 + 4	8	2
GaoFen-2	1 + 4	4	1
Pleiades	1 + 4	2	0.5
WorldView-2	1 + 8	1.6	0.4
WorldView-3	1 + 8	1.24	0.31

III. METHOD DATA AND RESULTS

A. Experiment Datasets

In order to evaluate the performance of the proposed SIDP method, different remote sensing satellite images are used to make training and testing data sets. Five satellite sensor images, GaoFen-1, GaoFen-2, Pleiades, and WordView-2 are used for experiments. The number of MS image bands from GaoFen-1, GaoFen-2, and Pleiades sensors is 4 (blue, green, red, and near-infrared). The number of MS bands from WorldView-2 and WorldView-3 sensors is 8 (blue, red edge, coast, green, red, yellow, near infrared 1, and near infrared 2). These data cover a rich variety of ground features, including housing buildings, trees, farmland, rivers, lakes, etc. The specific band and spatial resolution information are shown in Table I.

In real datasets, MS and PAN are cropped into image pairs with sizes of 64 × 64 and 256 × 256, respectively. Training datasets and verification datasets according to the ratio of 9:1 are randomly assigned. In order to compared with other methods that require simulated datasets, the corresponding simulated datasets are generated according to the Wald protocol. First, the Gaussian low-pass filter is used to filter the MS and PAN image pair. Then the PAN is down-sampled by bicubic interpolation. Finally, the corresponding MS image is first down-sampled and then up-sampled. Hence, the degraded MS and PAN image pair is obtained. The dimensions of MS and PAN of the test datasets are 128 × 128 and 512 × 512.

B. Experiment Details

The methods used in this article are all completed under the Linux system whose processor is Intel(R) Xeon(R) Gold 6278C CPU @2.60 GHz. Traditional methods use default parameters, while DL methods uniformly set hyperparameters. The DL training and testing use GPU acceleration, and the graphics card is GeForce RTX 3090. The batch size of the training is 16. Adam is selected as the optimizer of the network. The initial learning rate is set to 0.0005, and the training epochs are 100 rounds. In order to eliminate the variability of pixel value ranges between different satellite data, the preprocessing operation is performed on the input data. MS and PAN are normalized using a linear stretching method, which uniformly normalizes the pixel values to 0-1 for each band of the MS and PAN.

C. Experiment Results

Experimental results of the proposed method are compared with other three kinds of fusion results including the CS method—PCA [42] and GS [43], the MRA method—SFIM [13] and MTF_GLP [44], the VO method—NC-FSRM [19], the DL method—DRPNN [45], DiCNN [21], TFNet [27], LAGConv [46], MSDDN [30], and LDPNet [35]. Among these DL methods, LDPNet is the latest self-supervised method, while the remaining methods are all supervised approaches. In order to ensure that the compared DL methods achieve their optimal performance, we faithfully reproduced each method using its respective original training framework, without making any modifications to their training architectures. Both the subjective visual evaluation and the no-reference quantitative evaluation also are evaluated. The no-reference evaluation indicators include D_λ,D_s, and QNR. D_λ index is used to quantify the degree of spectral distortion of the fusion result.D_s index is used to determine whether the fusion result has the same spatial information as the original PAN. QNR is used to evaluate the overall quality of the fusion result, which is a comprehensive evaluation index including spectral evaluation and spatial evaluation.

1) Test results on WorldView-2 satellite data: The visual fusion comparison results of WorldView-2 are shown in Fig. 4. In order to verify the performance and highlight the advantages of the proposed method in this article, complex feature areas containing house buildings, trees, and greenery are selected for experiments. Overall, the spatial improvement of fusion results is very good, especially for DL methods. In the traditional method, the results of PCA not only have spectral distortion in the forest area, but also have blurred spatial texture compared with the original PAN image. Similarly, the results of MTF_GLP have information distortion with serious edge artifacts. It can be seen from the enlarged area that the results of NC-FSRM, DRPNN, DiCNN, and TFNet also have a little color difference compared with MS. In contrast, the proposed SIDP method in this article retains the original MS spectral information, while maximizing the spatial texture details of houses, vehicles, roads, and trees, which effectively improves the overall quality of image.

In order to effectively display the spatial details contained in the fusion result at the pixel level, the Laplace operator is used to obtain the gradient of the fusion result. The gradient effect diagram is shown in Fig. 5. It can be seen from Fig. 5 that GS, as a traditional CS method, improves the spatial performance of fusion results. TFNet and LAGConv DL methods also have ability to express spatial detail information. In the local enlarged image, only the gradient results of the NC-FSRM, MSDDN, and the proposed methods are consistent with the gradient result of the original PAN, and other methods have different degrees of spatial information loss. This is due to the dual-stream interaction strategy of the proposed SIDP method, which can better extract spatial features from the original PAN images and continuously inject information into the spectral stream during the forward feature extraction and transmission to complete information interaction.

In this article, 22 MS and PAN image pairs from the WorldView-2 satellite sensor are selected as the test dataset. The performance of each method is evaluated by calculating the average value of evaluation indicators such as AG, SCC, SSIM, ERGAS, D_λ, D_s, and QNR. AG, SCC, and D_s are used to evaluate the spatial fidelity of fusion results, while SSIM, ERGAS, and D_λ are used to evaluate the spectral quality of fusion results. The results of each index are shown in Table II.

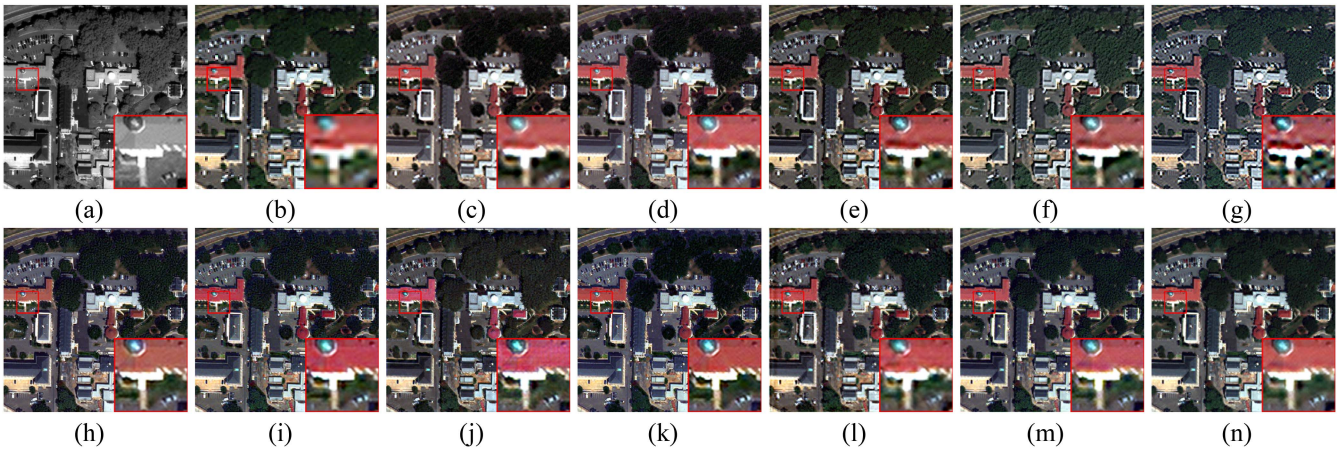


Fig. 4. Subjective visual comparison of fusion results on WorldView-2. (a) PAN. (b) MS. (c) PCA. (d) GS. (e) SFIM. (f) MTF_GLP. (g) NC-FSRM. (h) DRPNN. (i) DiCNN. (j) TFNet. (k) LAGConv. (l) MSDDN. (m) LDPNet. (n) Proposed.

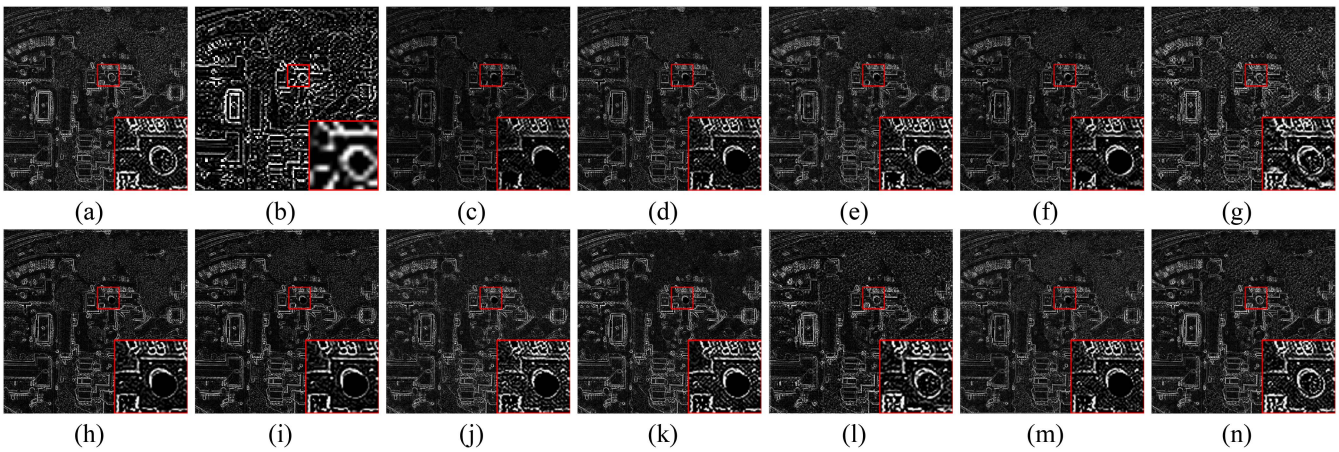


Fig. 5. Gradient comparison of fusion results using the Laplace operator on WorldView-2. (a) PAN. (b) MS. (c) PCA. (d) GS. (e) SFIM. (f) MTF_GLP. (g) NC-FSRM. (h) DRPNN. (i) DiCNN. (j) TFNet. (k) LAGConv. (l) MSDDN. (m) LDPNet. (n) Proposed.

TABLE II
RESULTS OF OBJECTIVE EVALUATION METRICS ON THE WORLDVIEW-2 DATASET

	Method	AG↑	SCC↑	SSIM↑	ERGAS↓	$D_\lambda \downarrow$	$D_s \downarrow$	QNR↑
I	PCA	24.9908	0.8430	0.8601	5.9949	0.0963	0.1150	0.8065
	GS	23.0659	0.8806	0.9212	6.4964	0.0321	0.0582	0.9121
	SFIM	25.2744	0.8383	0.9210	6.7591	0.0485	0.0512	0.9047
	MTF_GLP	24.3972	0.8737	0.9288	5.6941	0.0338	0.0392	0.9283
	NC-FSRM	25.9863	0.8614	0.9127	5.5903	0.0309	0.0425	0.9279
II	DRPNN	22.9181	0.8825	0.8846	5.7526	0.0370	0.0831	0.8832
	DiCNN	23.2194	0.8721	0.8911	5.1144	0.0456	0.0772	0.8811
	TFNet	23.8901	0.8917	0.9149	6.5144	0.0176	0.0521	0.9314
	LAGConv	26.4613	0.9101	0.9322	5.7438	0.0439	0.0326	0.9249
	MSDDN	26.9833	0.9402	0.9574	3.0165	0.0126	0.0381	0.9498
III	LDPNet	25.9008	0.9326	0.9569	3.2375	0.0163	0.0376	0.9467
	Proposed	**27.5454**	**0.9568**	**0.9680**	**2.9039**	**0.0070**	**0.0314**	**0.9618**
I Traditional methods			II Supervised methods			III Self-supervised methods		

The optimal values are highlighted in bold.

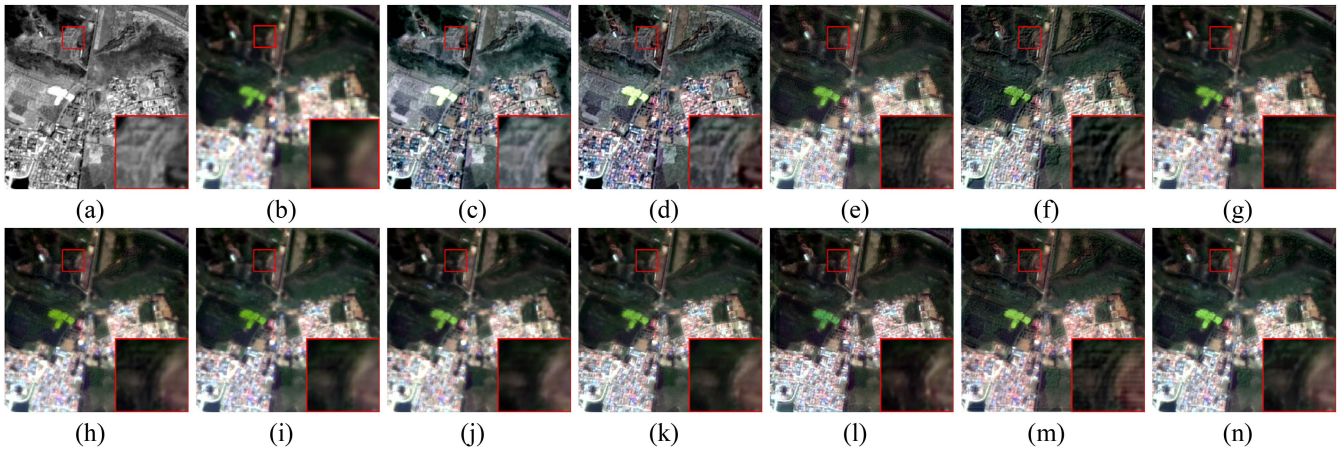


Fig. 6. Subjective visual comparison of fusion results on GaoFen-1. (a) PAN. (b) MS. (c) PCA. (d) GS. (e) SFIM. (f) MTF_GLP. (g) NC-FSRM. (h) DRPNN. (i) DiCNN. (j) TFNet. (k) LAGConv. (l) MSDDN. (m) LDPNet. (n) Proposed.

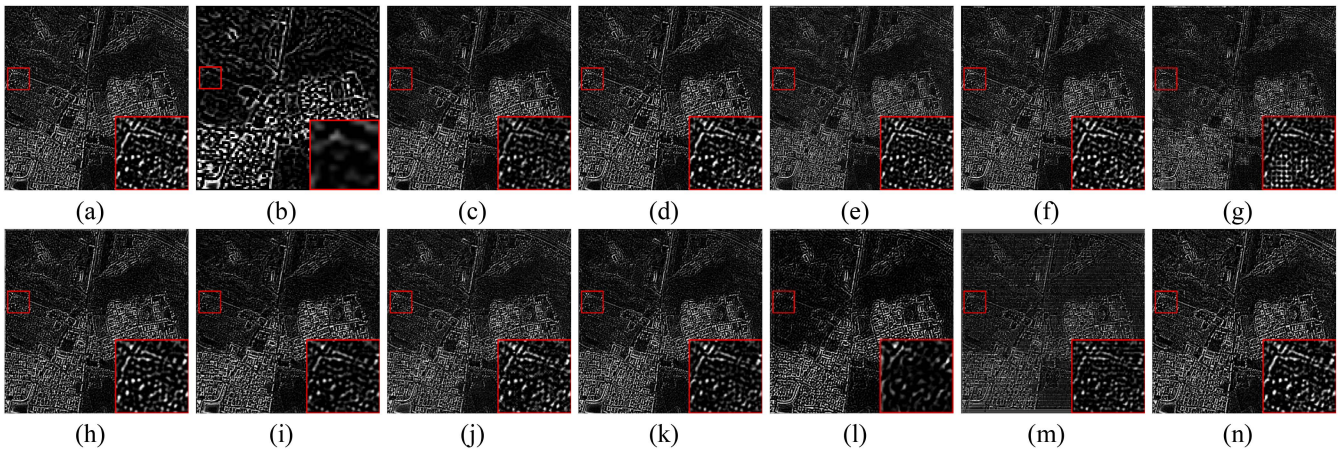


Fig. 7. Gradient comparison of fusion results using the Laplace operator on GaoFen-1. (a) PAN. (b) MS. (c) PCA. (d) GS. (e) SFIM. (f) MTF_GLP. (g) NC-FSRM. (h) DRPNN. (i) DiCNN. (j) TFNet. (k) LAGConv. (l) MSDDN. (m) LDPNet. (n) Proposed.

Overall, the index value of the DL method is better than that of the traditional method. For example, LAGConv, MSDDN, LDPNet, and the proposed SIDP method have excellent performance in the AG, SSIM, and QNR indicators. These methods not only enable the learning of spatial structural information required for fusion, but also effectively preserve the spectral information in the original MS data. However, traditional methods generally perform poorly in SCC, ERGAS, and D_λ. For instance, in PCA and SFIM methods, the range of SCC values falls between 0.8383 and 0.8430. However, the SIDP method has the optimal value of 0.9568 for this metric. The gap of performance is obvious. Compared with the PCA and DRPNN methods, the proposed method improves the SCC index by nearly 11.8%, and the QNR index by nearly 8.1%, which also objectively verifies the powerful spatial detail expression ability and stable spectral information retention ability of the SIDP method.

2) Test results on GaoFen-1 satellite data: Afterward, we use the GaoFen-1 satellite dataset to train and test the network. The fusion results of different methods are shown in Fig. 6. Different from the fusion results of WorldView-2 data, the fusion results of GaoFen-1 data have large differences in spectral preservation. Traditional methods such as PCA and GS suffer from the severe spectral distortion, which is difficult to apply to practical remote sensing applications. The overall spectral fidelity performance of DL methods such as Fig. 6(h)–(n) is better than that of traditional methods. In terms of spatial details, both SFIM and LDPNet methods have artifacts. In the partially enlarged area, it can be observed that the spatial details of Fig. 6(h) and (i) are blurred, and difficult to recognize the specific contour information of mountains and roads. Nevertheless, the proposed SIDP method has clear ground textures in the enlarged area.

The gradient map of the GaoFen-1 fusion result is shown in Fig. 7. It can be seen that the fusion result of each method contains rich gradient texture information. By enlarging the red-boxed area in the upper-left corner, Fig. 7(d)–(h) all exhibit good gradient details. Similarly, the gradient result of the SIDP method is basically consistent with those of PAN, which shows the result of the proposed method has good spatial detail clarity. In general, compared with other methods, the SIDP method has better spectral preservation performance and also fully combines

TABLE III
RESULTS OF OBJECTIVE EVALUATION METRICS ON THE GAOFEN-1 DATASET

	Method	AG↑	SCC↑	SSIM↑	ERGAS↓	D_λ ↓	D_s ↓	QNR↑
I	PCA	13.4531	0.9378	0.8479	3.1571	0.1053	0.0512	0.8494
	GS	18.6060	0.9441	0.8677	2.0751	0.0877	0.0465	0.8699
	SFIM	17.2206	0.9479	0.9328	1.9383	0.0330	0.0360	0.9324
	MTF_GLP	12.9302	0.9297	0.9263	1.4907	0.0701	0.0636	0.8715
	NC-FSRM	17.2465	0.9103	0.9336	1.3421	0.0414	0.0531	0.9077
II	DRPNN	14.0274	0.9316	0.9543	1.3069	0.0285	0.0704	0.9032
	DiCNN	13.2270	0.9335	0.9574	1.1838	0.0282	0.0662	0.9076
	TFNet	18.1336	0.9362	0.9644	0.9986	0.0276	0.0361	0.9374
	LAGConv	17.9990	0.9412	0.9524	1.3356	0.0269	0.0401	0.9341
	MSDDN	18.3283	0.9471	0.9615	0.9397	0.0242	0.0336	0.9430
III	LDPNet	15.1526	0.9099	0.9141	1.2580	0.0365	0.0371	0.9278
	Proposed	**19.7751**	**0.9526**	**0.9780**	**0.8461**	**0.0162**	**0.0277**	**0.9565**
I Traditional methods			II Supervised methods			III Self-supervised methods		

The optimal values are highlighted in bold.

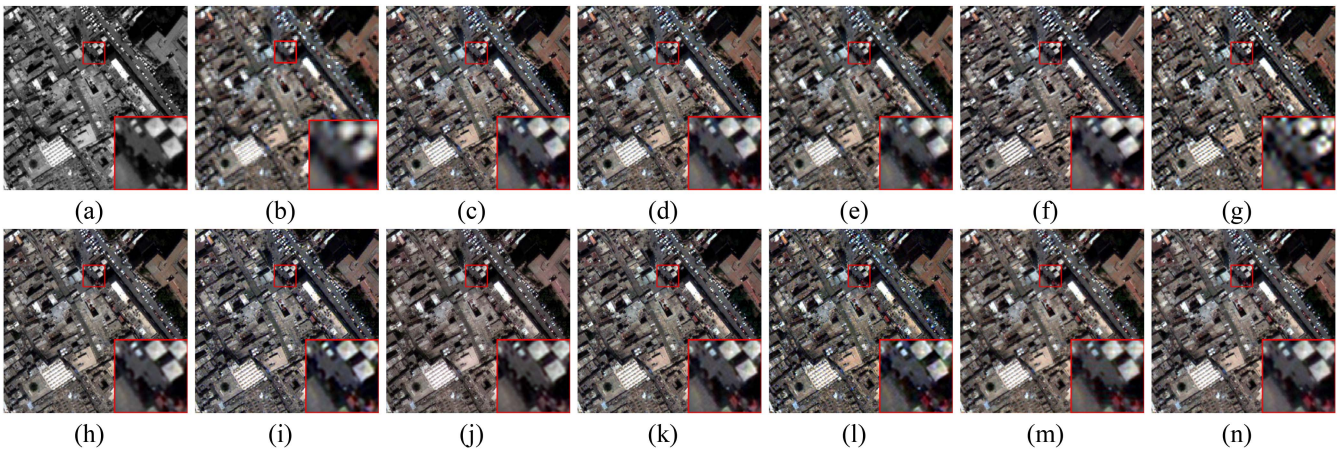


Fig. 8. Subjective visual comparison of fusion results on WorldView-3 Full Data. (a) PAN. (b) MS. (c) PCA. (d) GS. (e) SFIM. (f) MTF_GLP. (g) NC-FSRM. (h) DRPNN. (i) DiCNN. (j) TFNet. (k) LAGConv. (l) MSDDN. (m) LDPNet. (n) Proposed.

the spatial details of the PAN image, which helps to improve the recognition accuracy of different ground objects such as houses, mountains, and roads in the image.

Similarly, we select 22 MS and PAN image pairs on the GaoFen-1 satellite sensor as the test dataset for objective evaluation and quantitative analysis. The index comparison results are shown in Table III. The proposed method has achieved the best results in all index values. The SCC index values of all methods have reached above 0.9, which is basically consistent with the results shown in the above-mentioned gradient graph. This means that all the comparison methods in this article can effectively extract the spatial texture information of the original PAN image. But for spectral information, PCA and GS methods perform poorly in SSIM, ERGAS, and D_λ indicators, indicating that the spectral retention performance of these two approaches is insufficient. However, the DL method can well overcome the spectral distortion problem of GaoFen-1 results. For example, the index values of SSIM and ERGAS of the DL method are all greater than those of the traditional methods. In addition, the AG value of the proposed method is improved by 6.28% compared with the GS with the second highest performance, and the ERGAS is reduced by 15.3% compared with the TFNet with the second highest performance, which objectively verifies the superior performance of the SIDP method on the GaoFen-1 dataset.

3) Test results on WorldView-3 satellite data: We introduce the public WorldView-3 dataset [47] provided by Deng et al. for training and testing to enrich our experiments. Given that this training dataset is simulated data, both our method and the LDPNet method as the self-supervised framework, only use the input data for network training without using the corresponding labels. The other DL methods followed the originally specified training approach. Subsequently, we conducted tests using 20 full-resolution images provided by Deng et al. The test results and metrics are shown in Fig. 8 and Table IV, respectively.

From Fig. 8, it can be observed that PCA and GS exhibit remarkable visual performance. While SFIM and NCFSRM excel in the spectral fidelity, they tend to lose significant spatial information. LDPNet has spatial artifacts in the architectural region. LAGConv, MSDDN and SIDP methods effectively reconstruct the high-frequency edge region, presenting better spatial details and excellent spectral fidelity.

In addition, the quantitative metrics in Table IV align with the visual performance depicted in Fig. 8. Among traditional methods, GS demonstrates higher performance in metrics such

TABLE IV
RESULTS OF OBJECTIVE EVALUATION METRICS ON WORLDVIEW-3 FULL DATA

	Method	AG↑	SCC↑	SSIM↑	ERGAS↓	D_λ ↓	D_s ↓	QNR↑
I	PCA	42.4721	0.8916	0.9047	4.0633	0.0493	0.0516	0.9016
	GS	43.5380	0.9122	0.9031	4.0883	0.0467	0.0369	0.9181
	SFIM	38.6139	0.8924	0.8817	2.9526	0.0313	0.0687	0.9021
	MTF_GLP	40.2854	0.8890	0.9024	2.9754	0.0278	0.0785	0.8958
	NC-FSRM	38.7452	0.8673	0.8774	3.0956	0.0415	0.0845	0.8775
II	DRPNN	41.8296	0.9193	0.8976	4.6951	0.0372	0.0563	0.9085
	DiCNN	43.5471	0.8937	0.9122	4.9842	0.0435	0.0471	0.9114
	TFNet	40.7100	0.8815	0.8995	5.0932	0.0458	0.0579	0.8989
	LAGConv	43.7525	0.9253	0.9236	4.2210	0.0277	0.0554	0.9184
	MSDDN	45.0711	0.9387	0.9325	3.0811	0.0224	0.0392	0.9393
III	LDPNet	45.8969	0.9332	0.9289	2.9732	0.0219	0.0438	0.9352
	Proposed	**46.9511**	**0.9471**	**0.9432**	**2.8954**	**0.0146**	**0.0394**	**0.9465**
I Traditional methods			II Supervised methods			III Self-supervised methods		

The optimal values are highlighted in bold.

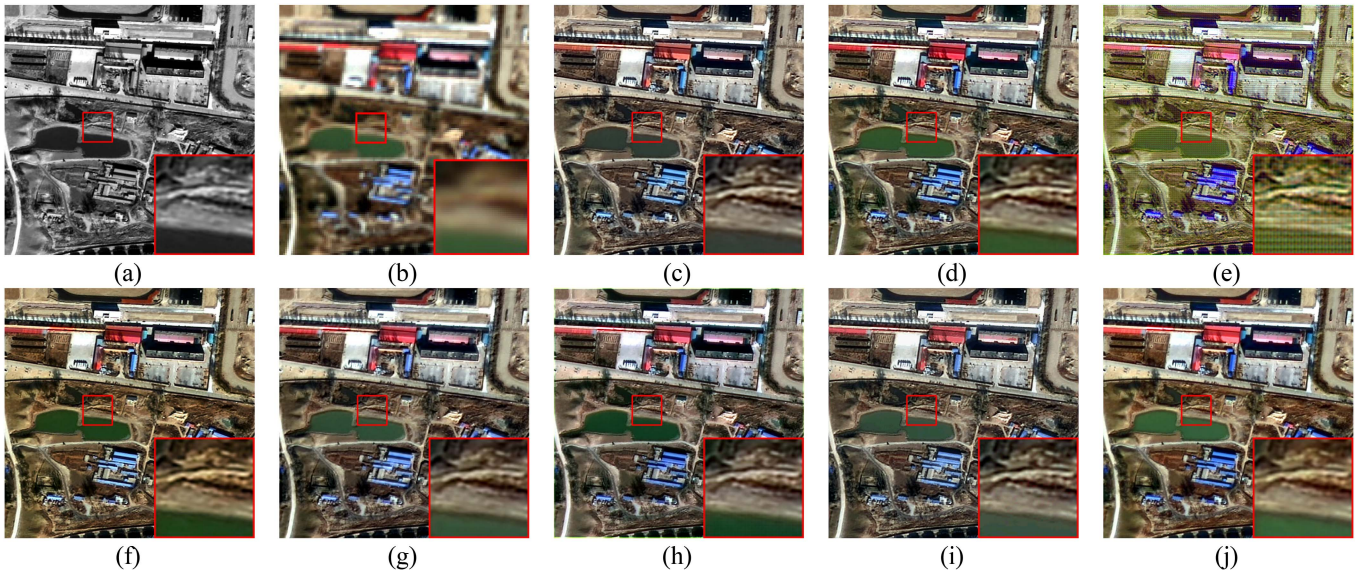


Fig. 9. Subjective visual comparison of generalization results on GaoFen-2. (a) PAN. (b) MS. (c) DRPNN. (d) DiCNN. (e) TFNet. (f) LAGConv. (g) MSDDN. (h) LDPNet. (i) DSSN. (j) Proposed.

as AG and SCC, while SFIM and MTF_GLP excel in spectral fidelity evaluation indicators such as ERGAS and D_λ. DRPNN and DiCNN, as DL methods, achieve a better balance between spectral and spatial aspects, displaying relatively stable overall performance. Notably, despite being trained on simulated data, our SIDP method still performs remarkably well across various metrics tests. This is mainly attributed to our model's ability to adaptively extract and interact information according to different inputs, thus effectively adapting to different data.

4) Pleiades and GaoFen-2 satellite data generalization results: In order to verify the generalization performance of the proposed SIDP method, the generalization experiment results are compared. The training completed network model on the WorldView-2 dataset is generalized directly to Pleiades and GaoFen-2 data for image fusion. The generalization results of various comparative methods of GaoFen-2 and Pleiades are shown in Figs. 9 and 10. Among the comparison methods, LDPNet and DSSN are the latest self-supervised methods.

Fig. 9 shows the fusion results of buildings, riversides, lakes, and other areas captured by the GaoFen-2 satellite sensor. The displayed area of the Pleiades satellite in Fig. 10 covers farmland, roads, shrubs, and other areas. Among the DL methods, the TFNet method has a serious problem of spectral distortion, which leads to the misjudgment of ground objects, and directly affects the accuracy of subsequent applications such as change detection and crop classification. Both LDPNet and DSSN also exhibit slight spectral distortion issues. On the contrary, the proposed method still shows excellent robust performance and visual effects in the face of complex land objects from heterogeneous satellite data, and no spatial distortion and spectral aberration are found in the presentation results of both GaoFen-2 and Pleaides images. This is mainly because the proposed model

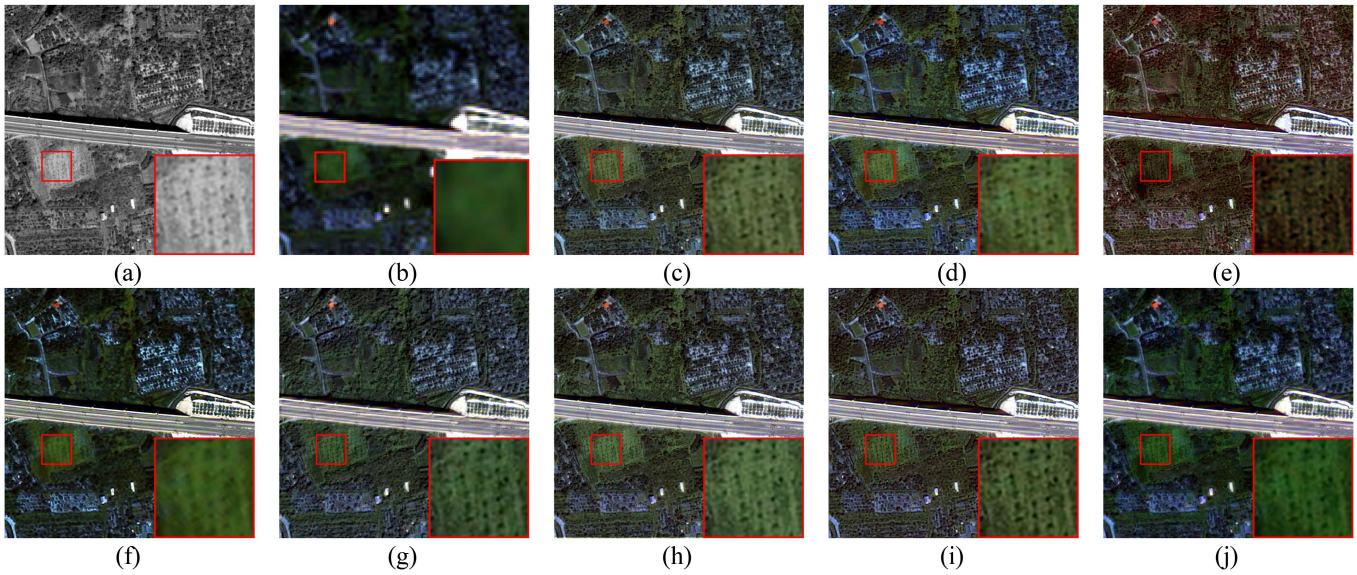


Fig. 10. Subjective visual comparison of generalization results on Pleiades. (a) PAN. (b) MS. (c) DRPNN. (d) DiCNN. (e) TFNet. (f) LAGConv. (g) MSDDN. (h) LDPNet. (i) DSSN. (j) Proposed.

TABLE V
KL AND SAM VALUES ON THE TRAINING DATASET OF WORLDVIEW-2 FOR DIFFERENT C AND N CASES

	KL						SAM				
	N=2	4	6	8	10		N=2	4	6	8	10
C=8	0.3780	0.3859	0.3443	0.3842	0.3776	C=8	2.1966	2.2207	2.0287	2.1389	2.1981
16	0.3220	0.3156	0.3338	0.3618	0.3182	16	1.5392	2.0687	1.9563	2.0229	2.0156
32	0.3221	0.2935	0.3248	0.3165	0.2901	32	1.8078	1.8808	1.8829	1.9571	1.8446
64	0.2753	**0.2496**	0.3060	0.2849	0.3298	64	1.8496	**0.9107**	1.8045	1.2442	1.7435
128	0.2776	0.2864	0.3129	0.3321	0.3492	128	1.7165	1.0224	1.8694	1.8930	1.9218

The optimal values are highlighted in bold.

in this article is based on real data for information extraction, and both spectral features and spatial features are derived from the original MS and PAN. The interaction between two features is enhanced by the feature reuse method of multilayer cascade, so that the network adaptively performs feature extraction and reorganization in each layer. Moreover, the learning target of the network is clarified in the loss function, so that the network purposefully learns and still has stable fusion performance with optimal robustness even on untrained image data.

D. Effect of Variable Parameters

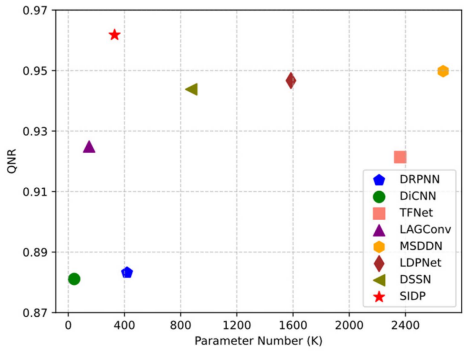
1) The Effect of Depth and Width: There are two variable parameters in the network structure of this article: the number of feature channels C and the depth of layers N. Both parameters play the important role in the model and affect the performance of the network. Consequently, these two parameters are compared in Table V, listing the KL and SAM values of the network on the training dataset under different C and N conditions. In Table V, fixing $\alpha = 0.5$ and $\beta = 0.5$, set C to 8, 16, 32, 64, and 128, respectively, and then set N to 2, 4, 6, 8, and 10 in sequence each time C is set. The results indicate that the increase in the number of network layers and channels does not imply an improvement in network performance. Because deeper and wider network models tend to bring more parameters to be trained, thus affecting the convergence speed of the network, increasing the training difficulty of the network, and making it difficult to obtain the optimal solution for the model. According to Table V, C at 64 and N at 4 are fixed.

TABLE VI
NO-REFERENCE EVALUATION VALUES UNDER DIFFERENT α AND β WEIGHTS

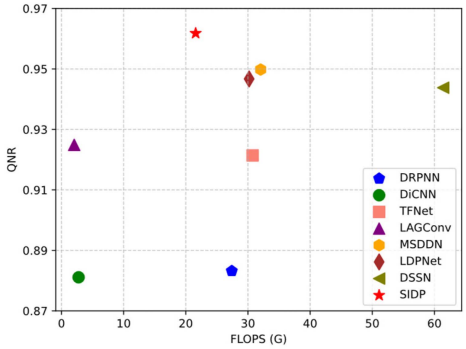
Loss weight		WorldView-2		
α	β	D_λ	D_s	QNR
0.1	0.9	0.1014	0.0572	0.8478
0.2	0.8	0.0825	0.0514	0.8708
0.3	0.7	0.0685	0.0495	0.8858
0.4	0.6	0.0663	0.0502	0.8872
0.5	0.5	0.0617	**0.0475**	**0.8940**
0.6	0.4	0.0730	0.0511	0.8800
0.7	0.3	**0.0576**	0.0522	0.8936
0.8	0.2	0.0642	0.0524	0.8873
0.9	0.1	0.0665	0.0507	0.8868

The optimal values are highlighted in bold.

2) Allocation of Loss Weights: Correspondingly, the weight distribution exponents α and β in the loss function reflect the proportion of spectral loss and spatial loss in the overall network



(a)



(b)

Fig. 11. Relationship between numbers of parameters, FLOPS and QNR performance metrics in models. (a) Numbers of parameters. (b) FLOPS. The higher positions in the graph are, the better performance of the model is. Meanwhile, positions further to the left indicate lower parameter count or faster computational speed.

loss. In order to make the network better balance the spectral quality and spatial quality during the fusion process, after fixing C and N, set α to 0.1, 0.2, …, 0.9, respectively, and β to 0.9, 0.8, …, 0.1 accordingly. Table VI lists the average QNR indicator values calculated from training dataset under different C and N. According to the optimal value in Table VI, the spectral loss weight is $\alpha = 0.5$, and the spatial loss weight is $\beta = 0.5$ in this article.

E. Calculation of Model Complexity

To assess the computational efficiency and complexity of the SIDP network model, we chose two pivotal metrics for verification: Floating point operations per second (FLOPS) and network parameter numbers. Specifically, we employ MS images with dimensions $4 \times 64 \times 64$ and PAN images with dimensions $1 \times 256 \times 256$ as inputs to the model. Subsequently, the metrics obtained from the input are compared with those of the aforementioned DL methods. The comparison results are illustrated in Fig. 11. In Fig. 11(a), the parameter count of the SIDP network is 329 K, which is notably smaller compared with other models such as TFNet with 2362 K parameters, MSDDN with 2668 K parameters, and LDPNet with 1585 K parameters. When compared with DRPNN which operates within a similar order of magnitude, the QNR metric exhibits a substantial improvement. This enhancement is attributed to the introduced restore mechanism in this study. Through the incorporation of dense connection operations, this module maintains the feature expression capability while substantially reducing the parameter count of subsequent convolutional layers.

In Fig. 11(b), the FLOPS metric of the SIDP model ranks third, following only DiCNN and LAGconv. This indicates that our proposed approach exhibits superior computational efficiency in terms of the computational complexity compared with current SOTA self-supervised pansharpening methods such as LDPNet and DSSN. Therefore, it can be concluded that our method achieves superior fusion results with lower hardware resource consumption.

F. Ablation Results

To further investigate the effects of the LKA and SSE modules used in the network, as well as the spatial and spectral loss functions on the performance of the SIDP model, this article conducts a series of ablation experiments. We divide the experiments into five groups, in which all networks are trained using the WorldView-2 dataset, tested and evaluated by objective metrics on the validation dataset. The results of the ablation experiments are shown in Table VII.

1) In the process of performing spatial feature extraction, this article uses the LKA module for spatial multiscale feature extraction. In the first experiment, the LKA module is directly removed and only the dense feature extraction is performed in the spatial scale enhancement stream. As can be seen from the first row of Table VII, the values representing spatial information features significantly increase (meaning a decrease in spatial quality), indicating that the LKA module introduced in the proposed SIDP does help to enhance the spatial feature representation capability of the network.
2) In the second ablation experiment of the network architecture, the SSE module is removed from the spectral channel attention stream. As seen in the second row of Table VII, all metric values are degenerated. This is because the SSE not only undertakes the task of spectral feature extraction, but also influences the integration of spectral and spatial information in the subsequent feature restore process. Moreover, this module has an important connectivity role in the model, which affects the overall performance of the network.
3) Subsequently, this article conducts the ablation experiment of the loss function. The results of the third group in Table VII show that it will make the index value deteriorate sharply and is difficult to achieve the purpose of optimizing the network, if only the spatial loss function is used.
4) In the fourth experiment, we delete the spatial loss function and only use the spectral loss function to train the network. The experimental results are also not optimistic and the spectral distortion quantization index value D_λ also deteriorates to a certain extent. Therefore, in the self-supervised model, in order to achieve the best performance of the network, both the spectral loss function and the spatial loss function are indispensable.

TABLE VII
COMPARISON OF MODEL PERFORMANCE UNDER DIFFERENT STRUCTURES AND LOSS FUNCTIONS

Group	Model		Loss		WorldView-2		
	LKA	SSE	$L_{spectral}$	$L_{spatial}$	D_λ	D_s	QNR
(I)	×	✓	✓	✓	0.0449	0.0547	0.9030
(II)	✓	×	✓	✓	0.0482	0.0528	0.9018
(III)	✓	✓	×	✓	0.2605	0.1116	0.6570
(IV)	✓	✓	✓	×	0.1216	0.4135	0.5162
SIDP	✓	✓	✓	✓	**0.0409**	**0.0451**	**0.9158**

The optimal values are highlighted in bold.

IV. DISCUSSION

In Section III, we test and compare the SIDP method with traditional methods and numerous SOTA DL methods on WorldView-2, GaoFen-1, and WorldView-3 satellite data (seen in Tables II–IV for objective metrics and in Figs. 4–8 for visual results). As can be seen from the presentation of the fusion results, the traditional methods fail to strike a good balance between spatial and spectral information preservation. The PCA method suffers from serious spectral distortion, while the MTF_GLP method suffers from artifacts in spatial aspects generally. In the gradient maps in Figs. 5 and 6, the gradient of SIDP is basically the same as that of the PAN image, which maximizes the preservation of spatial information of the PAN image. However, methods like DRPNN, DiCNN, TFNet, and LATGconv lose spatial information to varying degrees.

Analyzing from the network structural aspect, DRPNN and DiCNN methods simply utilize the residual structure for feature extraction, which makes the spectral and spatial features mixed together for learning, and is not conducive to feature extraction and reconstruction. Although TFNet adopts a dual-stream architecture, it ignores the complementarity between input images and lacks an effective loss optimization strategy. Distinctively, the SIDP method adopts a multilayer dual-stream interaction strategy. This strategy introduces the LKA module in the spatial stream, combines the convolution with the self-attention mechanism, effectively improves the extraction efficiency of the local spatial information, and continuously injects this information into the spectral stream for the information compensation between different layers.

In addition, analyzing from the loss function, most of the latest methods such as MSDDN and LAGConv use MSE and MAE as the loss function. These functions only consider individual pixel points in the image and do not take into account neighboring pixels as well as the relationship of pixel values between bands, which is not conducive to the optimization of the network. The reconstructed fused image may still have spatial distortion or spectral distortion. On the contrary, the SIDP method fully considers the characteristics of spectral and spatial information, and innovatively combines the SAM function and the KL divergence function to control the spectral and spatial losses of the network, respectively.

Furthermore, we conduct generalization experiments on GaoFen-2 and Pleiades satellite data (seen in Figs. 9 and 10). The TFNet spectral distortion is particularly severe. LAGConv has blurred texture details in the shrub in Fig. 10(k). However, the SIDP method has better performance. This is primarily due to the fact that we use a self-supervised framework that does not rely on the guidance of external labels and can adaptively extract and integrate the required information for fusion from the input data. Therefore, the SIDP method still demonstrates outstanding robustness when dealing with heterogeneous satellite data.

V. CONCLUSION

In this article, a SIDP method is proposed to solve the problem that the ideal fusion image does not exist as a reference label in the current DL pansharpening model. The proposed SIDP model no longer needs the ideal reference label of the fusion result, but directly builds the training of the model on the real dataset, and the original input image itself to be fused is the label. In the network architecture, in order to accurately and purely learn the spatial information and spectral information of the pansharpening, the two-stream network architecture including the spatial scale enhancement stream and the spectral channel attention stream is constructed. In the process of forward feature extraction and transmission, the multilayer cascading way is designed to continuously maintain the information interaction between spatial and spectral features, to inject spatial features of different scales into spectral features of corresponding scales, and realize spatial-spectral fusion at the feature level. In the design of the loss function, in order to achieve the self-supervised learning of the network, the spatial loss function KL and the spectral loss function SAM are designed considering the characteristics of spatial and spectral features to be learned by the network, respectively. The joint spatial-spectral loss function is composed by the weight assignment rule to guide the optimization of the network.

In order to verify the effectiveness of the proposed SIDP method, we compare and evaluate the SIDP method with traditional methods and the latest DL methods in terms of subjective vision and objective quantitative indicators. From the evaluation results, the SIDP method effectively integrates the spectral features and spatial features of the input data, which greatly improves the visual quality of the fusion results and achieves the optimum in the no-reference evaluation indexes such as D_λ, D_s, and QNR. In addition, the generalization effect on the GaoFen-2 and Pleiades datasets is also very satisfactory and far exceeds other comparison methods.

Moreover, we will further optimize the universal performance and explore the generalization performance of the model on satellite data such as Landsat-8 and GaoFen-6. In the future

research, our emphasis will be directed toward the fusion of multimodal heterogeneous satellite data. We will strive to further enhance the quality of fusion results by integrating various data types. Moreover, another research topic will combine the pansharpening techniques with tasks such as target detection and image classification to explore the universality of the fusion technique in engineering applications.

REFERENCES

[1] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015, doi: 10.1109/tgrs.2014.2361734.

[2] B. Xie, H. Zhang, and B. Huang, "Revealing implicit assumptions of the component substitution pansharpening methods," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 443, doi: 10.3390/rs9050443.

[3] R. J. Bhiwani and V. R. Pandit, "Image fusion in remote sensing applications: A review," *Int. J. Comput. Appl.*, vol. 120, no. 10, pp. 22–32, 2015, doi: 10.5120/21263-3846.

[4] A. Garzelli, "A review of image fusion algorithms based on the super-resolution paradigm," *Remote Sens.*, vol. 8, no. 10, 2016, Art. no. 797, doi: 10.3390/rs8100797.

[5] J. Jiao and L. Wu, "Image restoration for the MRA-based pansharpening method," *IEEE Access*, vol. 8, pp. 13694–13709, 2020, doi: 10.1109/access.2020.2965921.

[6] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2822–2834, Jul. 2013, doi: 10.1109/TIP.2013.2258355.

[7] M. Cheng, C. Wang, and J. Li, "Sparse representation based pansharpening using trained dictionary," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 293–297, Jan. 2014, doi: 10.1109/lgrs.2013.2256875.

[8] Y. Liu and Z. Wang, "A practical pan-sharpening method with wavelet transform and sparse representation," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, 2013, pp. 288–293.

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

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

[11] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + Pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[12] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2010.

[13] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, 2000.

[14] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.

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

[16] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P + XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, 2006, Art. no. 43.

[17] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[18] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011, doi: 10.1109/tgrs.2010.2067219.

[19] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, and G. Vivone, "A framelet sparse reconstruction method for pansharpening with guaranteed convergence," *Inverse Problems Imag.*, vol. 17, no. 6, pp. 1277–1300, 2023.

[20] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023, doi: 10.1109/TIP.2023.3247165.

[21] L. He et al., "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

[22] W. Li, M. Xiang, and X. Liang, "MDCwFB: A multilevel dense connection network with feedback connections for pansharpening," *Remote Sens.*, vol. 13, no. 11, 2021, Art. no. 2218.

[23] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2020.

[24] W. G. C. Bandara and V. M. Patel, "HyperTransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1767–1777.

[25] M. Zhou, J. Huang, F. Zhao, and D. Hong, "Modality-aware feature integration for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2022, Art. no. 5400312, doi: 10.1109/TGRS.2022.3232384.

[26] K. Yan et al., "Panchromatic and multispectral image fusion via alternating reverse filtering network," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 21988–22002, 2022.

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

[28] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, Art. no. 321052.

[29] S. Fang, X. Wang, J. Zhang, and Y. Cao, "Pan-sharpening based on parallel pyramid convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 453–457.

[30] X. He et al., "Multi-scale dual-domain guidance network for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5403213, doi: 10.1109/TGRS.2023.3273334.

[31] M. Zhou, J. Huang, D. Hong, F. Zhao, C. Li, and J. Chanussot, "Rethinking pan-sharpening in closed-loop regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 59, no. 4, pp. 3486–3501, Apr. 2023.

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

[33] Z. Xiong, Q. Guo, M. Liu, and A. Li, "Pan-sharpening based on convolutional neural network by using the loss function with no-reference," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 897–906, 2020, doi: 10.1109/JSTARS.2020.3038057.

[34] S. Li, Q. Guo, and A. Li, "Pan-sharpening based on CNN+ pyramid transformer by using no-reference loss," *Remote Sens.*, vol. 14, no. 3, 2022, Art. no. 624.

[35] J. Ni et al., "LDP-Net: An unsupervised pansharpening network based on learnable degradation processes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5468–5479, 2022, doi: 10.1109/JSTARS.2022.3188181.

[36] Q. Guo, S. Li, and A. Li, "An efficient dual spatial–spectral fusion network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5412913, doi: 10.1109/TGRS.2022.3222223.

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

[38] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Comput. Vis. Media*, vol. 9, no. 4, pp. 733–752, Jul. 2023.

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[40] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop*, vol. 1, Pasadena, CA, USA: AVIRIS Workshop, 1992.

[41] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Proc. IEEE 9th Int. Conf. Comput. Vis.*, 2003, pp. 487–493.

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

[43] C. A. Laben and B. V. Brower, *Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening*. San Francisco, CA, USA: Google Patents, 2000.

[44] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.
[45] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.
[46] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 1113–1121.
[47] L.-J. Deng, "PanCollection for remote sensing pansharpening," Jan. 2023, [Online]. Available: https://liangjiandeng.github.io/PanCollection.html

Qing Guo (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in optics from the Harbin Institute of Technology, Harbin, China, in 2006 and 2010, respectively.

From 2007 to 2009, she was an exchange Ph.D. Student with the Department of Electrical and Computer Engineering, University of Calgary, Calgary, AB, Canada. In 2010, she joined the Chinese Academy of Sciences, Beijing, China, where she is currently a Full Professor with the Aerospace Information Research Institute. From 2014 to 2015, she was a Visiting Scholar with the Institute for Geoinformatics and Remote Sensing, University of Osnabrück, Osnabrück, Germany. Her research interests focus on remote sensing information extraction and processing, including image fusion, cloud detection, deep learning, and landslide monitoring.

He Jia received the B.S. degree in electronic information engineering from Zheng Zhou University, Zhengzhou, China, in 2021. He is currently working toward the M.Sc. degree in electronic information with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China.

His research interests include image processing, image fusion, and deep learning.

Shengsang Yang received the B.Sc. degree in geology from the China University of Geosciences, Beijing, China, in 2021. She is currently working toward the M.Sc. degree in electronic information with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing.

Her research interests include machine learning and data processing.