A Progressive Implicit Neural Fusion Network for Multispectral Image Pansharpening

Yao Feng, Long Zhang, Yingwei Zhang, Xinguo Guo, Guangqi Xie, Chuang Liu, and Shao Xiang

***Abstract*—In the field of remote sensing, it is not feasible to obtain high spatial resolution multispectral (HRMS) images from a single satellite sensor. The existing methods use pansharpening techniques to obtain HRMS images by fusing panchromatic (PAN) and multispectral (MS) images. However, due to the scale difference between PAN and MS images, most pansharpening methods often use explicit sampling methods to integrate features at different scales. These explicit-based sampling techniques represent pixels as discrete points through predefined functions, rendering it challenging to fit the distribution among diverse modal data, this results in the loss of image texture details during the fusion process. Implicit neural networks can enhance the generative capability of images by incorporating pixel coordinate information, which is crucial for the fusion of remote sensing images with different spatial resolutions. Inspired by implicit neural representation, we propose a progressive implicit neural feature fusion network (PINFNet) for remote sensing images. A progressive implicit neural feature fusion is proposed; it establishes a coordinate modal relationship between the spatial and spectral information through the guidance of the high spatial features in PAN images. This enables the proposed PINFNet to progressively learn and integrate spatial and spectral information at different scales. Our method, as opposed to discrete sampling techniques, is capable of establishing a continuous representation between diverse modal data, which in turn preserves more texture detail information. Extensive experiments have shown that this approach outperforms state-of-the-art methods while maintaining high efficiency.**

***Index Terms*—Implicit neural representation (INR), multispectral (MS) image, panchromatic (PAN) image, pansharpening, remote sensing (RS).**

I. Introduction

IN THE remote sensing (RS) imaging process, a small instantaneous field of view (IFOV) can meet the requirements for signal-to-noise ratio (SNR) when spectral resolution is not taken into account. However, as the spectral resolution increases, it is necessary to process narrower spectral bands and employ larger IFOV to meet the SNR requirements. As a result, the spatial and spectral resolutions of RS images are constrained by each other, rendering it impracticable for a singular satellite sensor to acquire images with high spatial and spectral resolutions. However, the successful execution of subsequent tasks is contingent upon the utilization of high spatial resolution (HR) RS imagery. The fine-grained spatial texture information and high-fidelity spectral distribution can improve the accuracy of subsequent visual tasks, such as RS land classification [1], [2], [3], [4]. In order to facilitate image interpretation and understanding in subsequent tasks, the pansharpening technique is often employed to enhance the spectral and spatial resolution of RS images. Pansharpening is the process of combining a panchromatic (PAN) image with HR and a multispectral (MS) image with high spectral resolution to get an HR MS image [5], [6].

Pansharpening of MS images has been studied for decades. PAN images typically have HR but limited spectral information, whereas MS images typically have lower spatial resolution but richer spectral information. In order to leverage the advantages of both types of images, many scholars have conducted in-depth research on pansharpening techniques. The traditional methods for pansharpening include component substitution (CS) based [7], [8], [9], [10], [11], [12], [13], [14], [15], multiresolution analysis (MRA) based [16], [17], [18], [19], [20], [21], [22], [23], and variational optimization (VO) based methods [24], [25], [26], [27]. These methods extract the features manually from PAN and MS images. Due to the limited representational capabilities, it is difficult to obtain complex relationships between heterogeneous data. Furthermore, traditional methods rely too heavily on hand-made fusion rules and cannot adapt to intricate sharpening scenarios. In general, DL-based pansharpening methods can achieve better sharpening performance than traditional methods [28], [29], [30], [31], [32], [33], [34], [35]. The existing DL-based methods can broadly be categorized into single-branch and double-branch structures. The former involves stitching the upsampled MS (UPMS) image and PAN image in the channel dimension, and subsequently exploring the complementary information within the original image pairs by stacking numerous residual blocks or attention blocks; the latter employs two distinct branches to extract the features of the MS and PAN images, respectively, and subsequently integrating the features within the two branches to obtain the HRMS image. In

Manuscript received 3 May 2024; revised 13 July 2024; accepted 1 August 2024. Date of publication 14 August 2024; date of current version 5 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62301214 and in part by the Scientific Research Foundation for Doctoral Program of Hubei University of Technology under Grant XJ2022005901. *(Corresponding author: Xinguo Guo.)*

Yao Feng, Long Zhang, Yingwei Zhang, and Xinguo Guo are with the Shandong GEO-Surveying and Mapping Institute, Jinan 250013, China. (e-mail: zjxg15@foxmail.com).

Guangqi Xie is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China, and also with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China.

Chuang Liu is with the School of Computer Science, Hubei University of Technology, Wuhan 430068, China.

Shao Xiang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China.

Digital Object Identifier 10.1109/JSTARS.2024.3443400

contrast, the two-branch structure is capable of taking full advantage of the properties of the original image for different sources, thereby achieving high-fidelity sharpening results. However, both structures rely heavily on the stack of numerous residual or attention blocks, which results in a significant expansion of the model's parameters. In addition, due to the difference in scale between PAN and MS images, both traditional and DL-based methods need to extract and fuse features at different scales. In the fusion process, it is inevitable to use interpolation techniques to unify features at different scales. The most common upsampling methods include bicubic, bilinear, and deconvolution, all of which represent pixels as discrete points through explicit predefined functions. In the human visual system, images are presented in a continuous way. Consequently, the application of explicit interpolation techniques to fit data distributions from diverse modal data poses a challenge, potentially leading to the loss of critical texture details.

It can be inferred that a continuous function serves as a reliable representation of the actual state of the image. However, the precise form of this continuous function remains uncertain. Therefore, a number of researchers have attempted to approximate it through neural networks. Consequently, implicit neural representation (INR) has been developed, which can be employed to generate a continuous functional representation of an object by implicitly mapping the continuous coordinates to the signals within a specified domain. Considering that INR can establish a continuous representation between spatial and spectral information, we propose a progressive implicit neural feature fusion network (PINFNet). The PINFNet establishes a coordinate modal relationship between the spatial and spectral information through the guidance of the high-spatial features in PAN images, thereby enabling the neural network to learn and integrate both spatial and spectral information at different scales in a progressive manner. The main contributions are as follows.

1) We propose a novel PINFNet for pansharpening. Only two MLP layers are employed in the fusion stage, sparing the dramatic increment in parameters caused by stacking numerous convolutional or attention layers.
2) We develop a progressive implicit neural feature fusion (PINF) module inspired by INR, which adopts a progressive manner to fully utilize the scale information of different modalities in the original image pairs. The proposed PINF establishes a coordinate modal relationship between spatial information in the HR domain and spectral information in the low-resolution (LR) domain through continuous coordinate mapping. Subsequently, features at various scales are fused within the continuous domain.
3) By establishing implicit encoding representations of diverse modal features, high-fidelity fused images with richer edge texture information can be obtained. Extensive experiments show that the proposed method achieves optimal fusion performance while preserving promising effectiveness.

II. Related Works

A. Traditional Pansharpening Methods

In the CS-based method, the spatial component of the LRMS images is replaced by the PAN component. The sharpened results of CS can enhance spatial fidelity, but it is susceptible to spectral distortion in complex scenes. Methods, such as PRACS [11], GSA [12], and BDSD [15], are included in this category. The methods based on MRA extract spatial details from PAN by utilizing MRA tools. The MRA-based methods include additive wavelet luminance proportional (AWLP) [21] and morphological filters (MF) [22]. Furthermore, a technique known as BFLP has been proposed to enhance spatial fidelity [36]. Furthermore, Kalplan et al. [37] proposed an adaptive multiscale bilateral filtering method to improve the generalizability of BFLP. The spectral domain of the MRA-based methods permits them to retain sufficient spectral information, but it can also cause spatial distortion due to inadequate detail extraction. The VO-based method is divided into two components, including the construction of the energy function and the computation of the optimal solution. Compared with CS- and MRA-based methods, VO-based methods are able to achieve better sharpened results. There are numerous improved methods that have been proposed, such as LRFF [39] and CDIF [40]. These VO-based methods entail a significant amount of computation and numerous hyperparameters, which necessitate frequent modification for diverse sharpening scenarios.

B. DL-Based Pansharpening Methods

In recent years, DL has garnered significant attention, owing to its potent nonlinear fitting capabilities. The first pansharpening neural network (PNN) is proposed in [28]. Despite the fact that PNN only contains three convolutional layers, it demonstrated a notable sharpened effect. To preserve more spatial details, Yang et al. [29] extracted high-frequency feature from PAN and retained high-frequency details utilizing the residual structure. Zhang et al. [30] constructed a bidirectional pyramidal network (BDPN) using residual blocks, which is able to fuse the extracted spatial and spectral features through pairwise interactions. In contrast to PNN, BDPN exhibits the capability to attain a higher level of fidelity in sharpening quality. However, the stacking of numerous residual blocks results in a significant increase in the number of parameters of BDPN. Jia et al. [33] utilized an explicit bilinear interpolation technique to sample the PAN and LRMS images at the same scale. Subsequently, spatial and spectral transformers were designed to extract spatial and spectral features, respectively. In general, owing to the stacking of numerous transformers, this method is capable of preserving more spatial details. However, the predefined interpolation technique can result in misalignment of spatial and spectral features during the fusion stage, causing spatial distortion in the sharpened image. Based on the standard residual module, Wang et al. [35] proposed a residual module with multiscale receptive fields (CML-resblock), which was successfully applied to the field of pansharpening. Several methods employed

improved attention modules to construct the network, leading to satisfactory sharpened results, such as TANI [31], TRRNet [32], and RSANet [34]. However, the stacking of a large number of convolution blocks and attention blocks will significantly increase the amount of calculation and parameters of the model, thereby reducing the practicability of the model.

C. Implicit Neural Representation

Optical RS images are imaged on a continuous narrow band, so the correlation of images in different bands can be considered continuous. However, the traditional discrete signal representation method cannot accurately represent the true state of the image. INR can capture positional information of signals, embedded together with pixel encoding, thereby effectively capturing spatial correlations of images. Based on this property, INR has been widely used in many image reconstruction tasks, such as 3-D image reconstruction [41], [42], [43] and image super-resolution [44], [45], [46]. Zhu et al. [47] proposed an implicit neural sampling method that is capable of capturing diverse receptive fields by utilizing the concept of atrous convolution. Zhu et al. [48] reconstructed INR based on the structure of quadtree and successfully applied it to hyperspectral and MS fusion. Liu et al. [49] proposed an INR-based neural hybrid model that can fully simulate the imaging noise and multiple reflections of object spectra. Inspired by the above research, this article proposes a PINFNet. By accurately representing the relationships between modal data of different scales, it enhances the fusion capability without need for numerous convolutional or attention layers. This has the potential to reduce model parameters and computational complexity.

III. Proposed Method

A. Motivation and Overview

Due to the inherent scale differences between different modal images, existing DL-based methods need to sample different modal features at different scales to the same scale before fusion. In the human visual system, images are presented in a continuous way, but most sampling methods process images in a discrete expression. As a result, these techniques can lead to misalignment during the fusion process, which results in the loss of critical information. In addition, these methods can achieve good sharpening results by introducing advanced techniques. However, stacking numerous residual blocks and attention modules tends to bring about a significant increase in the number of model parameters and computation. Considering the continuous representation is more in line with the intuition of human eyes. We try to use the continuous property of INR to establish the coordinate modal relationship between different modal data. Then, the different modal data are mapped and fused continuously using implicit coding functions, such as MLP. In the meantime, the utilization of implicit coding functions can avoid the stacking of numerous residual blocks or self-attention modules, which has the potential to reduce the number of parameters and computation of the model.

Numerous studies have confirmed that directly using INR methods in image reconstruction tasks often leads to overfitting [44], [47], [48], [49]. In order to guarantee the stability of the model during training, we design an encoder–decoder architecture. As depicted in Fig. 1, the proposed PINFNet comprises two primary components, including an encoder and a PINF. The objective of the encoder is to extract the high-dimensional, stable semantic features of the original image pairs, which makes the training process less susceptible to noise, thereby ensuring the stability of model training. Specifically, the features of the LRMS image are first extracted by the encoder E_φ. The process can be represented as follows:

$$F_{\text{spe}} = E_\varphi(M), \tag{1}$$

where $F_{\text{spe}} \in R^{h\times w\times d}$ denotes the feature map that contains spectral information, φ is a learnable parameter in E_φ, and $M \in R^{h\times w\times c}$ is the LRMS image. In the meantime, we employ encoder E_δ to extract the spatial information from the PAN image. PAN and LRMS are captured by different sensors, and there exist radiometric differences between them. As a result, LRMS has spatial details that are complementary to PAN. To address this, LRMS is first upsampled to match the size of the PAN image, denoted as UPMS. And then the PAN and UPMS images are concatenated in the channel dimension and then fed to the encoder E_δ to extract comprehensive spatial details. The above process can be represented as follows:

$$F_{\text{spa}} = E_\delta\left(\left[P, \overset{\prime}{M}\right]\right), \tag{2}$$

where $F_{\text{spa}} \in R^{h\times w\times d}$ denotes the feature map that contains spatial information. [.] denotes the concatenation operation in the channel dimension. $P \in R^{H\times W\times 1}$ denotes the PAN image. $\overset{\prime}{M} \in R^{H\times W\times c}$ denotes the UPMS, which is upsampled by 23-tap polynomial interpolation. The height and width of $\overset{\prime}{M}$ are identical to those of P, while the number of channels is identical to that of M. δ is a learnable parameter in encoder E_δ. It is worth noting that E_φ and E_δ have the same structure, both consisting of convolutional residual blocks. As shown in Fig. 2, the detailed generation of F_{spe} can be represented as follows:

$$F_1^{hid} = Conv(M) \tag{3}$$

$$F_2^{hid} = ReLU\left(BN\left(Conv\left(F_1^{hid}\right)\right)\right), \tag{4}$$

$$F_3^{\text{hid}} = ReLU(BN\left(Conv\left(F_1^{hid} + F_2^{hid}\right)\right), \tag{5}$$

$$F_{spe} = F_3^{hid} + F_1^{hid}, \tag{6}$$

where F_1^{hid}, F_2^{hid}, and F_3^{hid} all denote the intermediate layer features. $Conv$, BN, and $ReLU$ denote the convolutional layer, batch normalization layer, and ReLU activation function, respectively. In the encoder, the first convolution is mainly used to enhance the dimensionality of the input image and the subsequent residual module is used to extract stable features.

Afterward, the extracted spectral and spatial feature maps are fed into PINF. The process can be represented as follows:

$$O = PINF_\theta\left(F_{spe}, F_{\text{spa}}, C\right), \tag{7}$$

where $O \in \mathrm{R}^{H\times W\times c}$ denotes the output of the PINF, and θ is a learnable parameter in PINF. Besides, and $C \in R^{H\times W\times 2}$ is a normalized 2-D coordinate map in the HR domain guided by P.

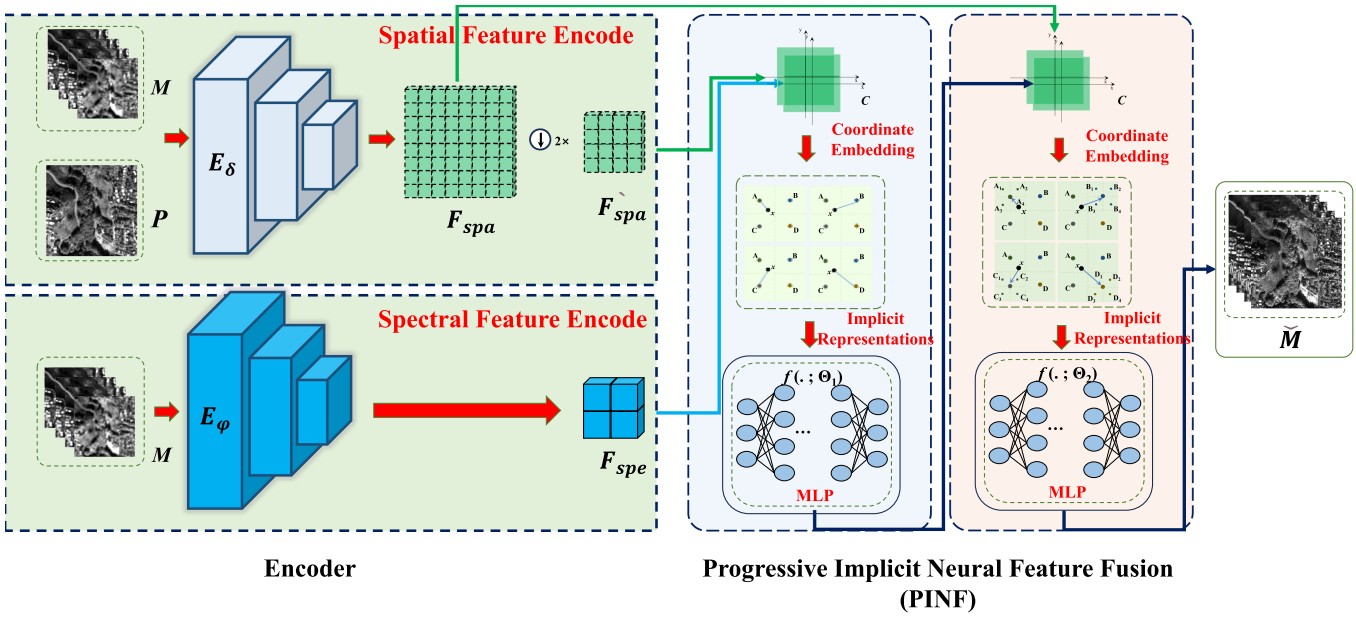

Fig. 1. Flowchart of the proposed PINFNet.

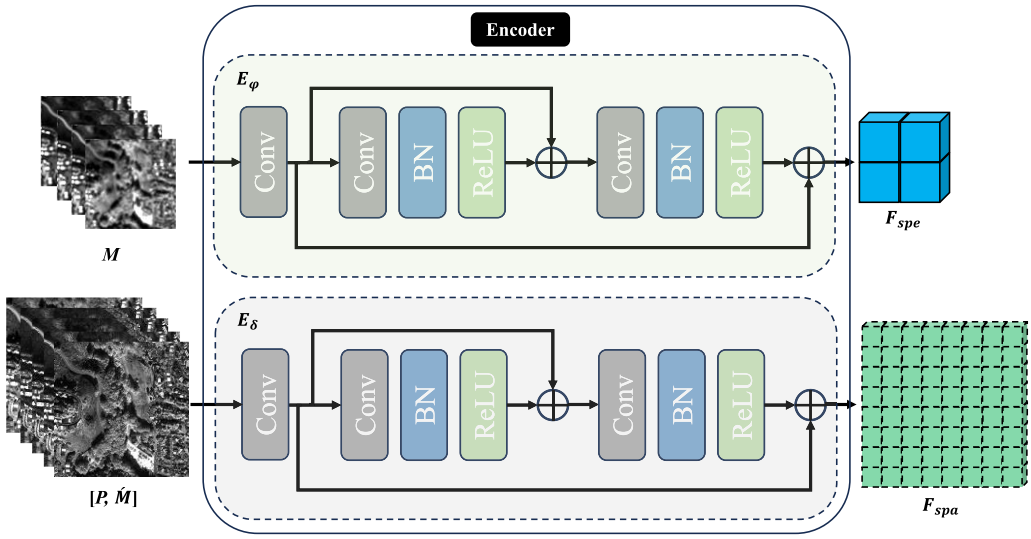

Fig. 2. Structure of encoder.

B. Progressive Implicit Neural Feature Fusion

The objective of pansharpening is to integrate P and M with diverse modal information at different spatial scales. Most of the existing pansharpening methods integrate original image pairs at various scales by frequently utilizing explicit upsampling and downsampling techniques during the fusion process. These explicit interpolation techniques represent pixels as discrete points through predefined functions, making it challenging to fit the distribution between diverse modal data. This results in the loss of crucial texture details. Inspired by INR, we develop a progressive implicit neural fusion scheme through HR guidance. As shown in Fig. 3, through the guidance of the HR feature maps extracted from P, PINF establishes a coordinate modality relationship between the LR and HR domains. Subsequently, MLP is employed to learn both spatial and spectral information in a progressive manner. Specifically, the fusion feature maps O_q located at position C_q can be represented as follows:

$$O_q = \sum_{k \in N_q} w_{q,k} F_{q,k}, \tag{8}$$

where N_q denotes the set of the four nearest query coordinates around C_q in the HR domain. $F_{q,k}$ and $w_{q,k}$ are the multimodal fusion information of the query coordinates C_q at the position C_k and corresponding weights, respectively. In the following, we describe in detail how to obtain weights and multimodal fusion information.

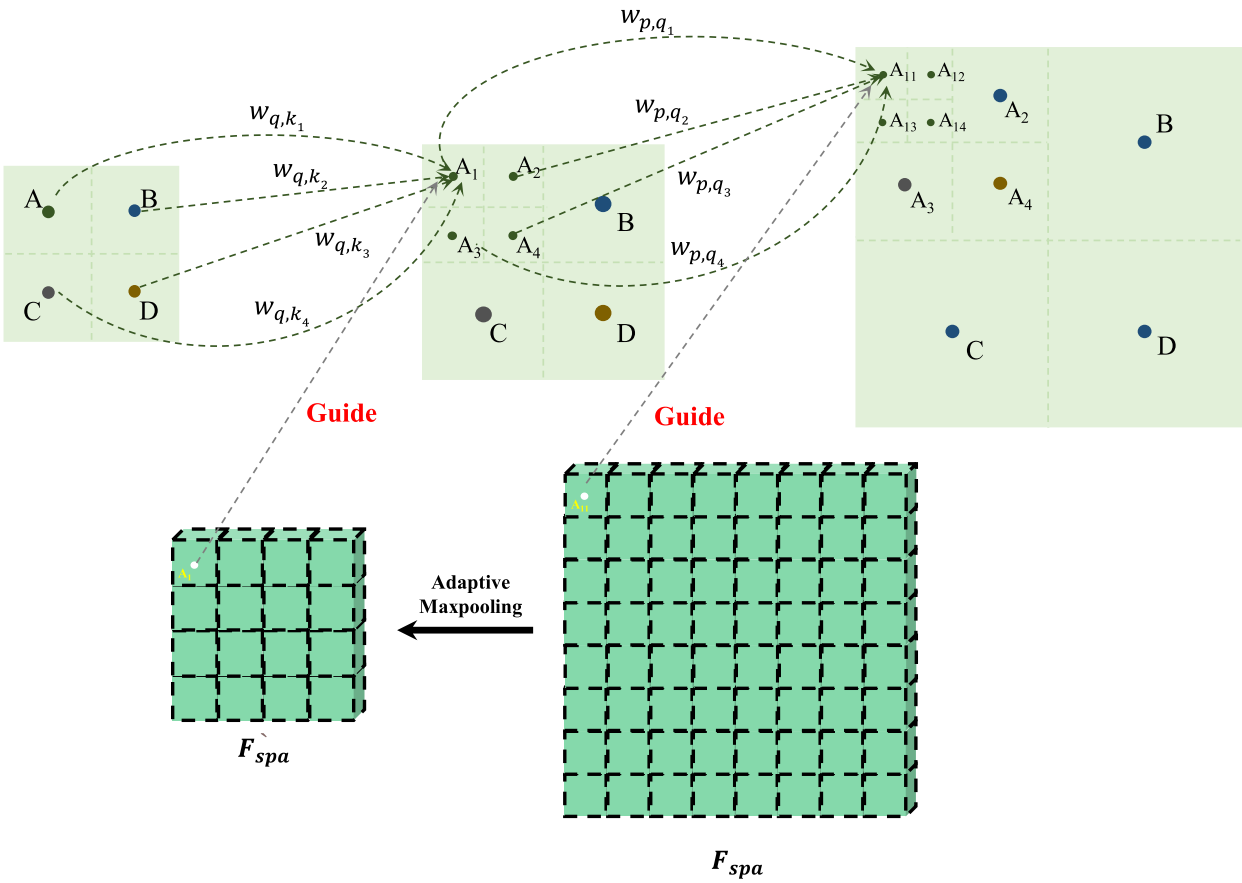

Fig. 3. Schematic of the HR guidance in PINF.

In order to ensure consistency in the coordinate scale between the LR domain where the M is located and the HR domain where the P is located, we represent pixels with the center position of the pixel and scale the coordinate map of $H \times W$ into a square grid of $[-1,\ 1] \times [-1,\ 1]$ to facilitate the sharing of the coordinates between the HR domain and the LR domain. The process of the normalization in the HR domain can be expressed as follows:

$$C\left(i,j\right) = \left[-1 + \frac{2i+1}{H}, -1 + \frac{2j+1}{W}\right], \tag{9}$$

where $i \in [0,\ H-1]$ and $j \in [0,\ W-1]$. In this way, each pixel feature in the original image pair can be continuously represented. Based on this, we obtain the weights by

$$w_{q,k} = \frac{E_{q,k}}{E}, \tag{10}$$

where $E_{q,k}$ represents the similarity value, which is computed from the final element of the vector $F_{q,k}$. Besides, $E = \sum_{i \in N_q} E_{q,k}$ represents the sum of $E_{q,k}$.

Due to the disparity in spatial resolution, it is imperative to unify the spatial scales of P and M within the same region. Pansharpening aims to improve the spatial resolution by using the texture information in P. Therefore, we upsample the M to achieve scale alignment with P in the proposed PINF. The spatial feature maps in the HR domain are utilized to gradually guide the spectral feature maps in the LR domain. In addition, the positional coordinates are utilized to guide the fusion of information from diverse domains. To establish a bridge between the LR and HR domains, we set a medium-resolution (MR) domain between the LR and HR domains. The sampled MR features are obtained after the fusion of the information of the LR pixels with the relative coordinates. The process can be expressed as follows:

$$O_{q,k}^{\text{mr}} = f\left(\grave{F}_{\text{spa}}\left(C_k\right), C_q - C_k, F_{\text{spe}}\left(C_k\right); \theta_1\right) \tag{11}$$

where $O_{q,k}^{\text{mr}} \in R^{1\times1\times d}$ denotes the multimodal fusion information of the query coordinate C_q at position C_k in MR domain, which is also a subset of $O^{\text{mr}} \in R^{\frac{H}{2}\times\frac{W}{2}\times d}$. $C_q - C_k$ denotes the coordinate modal information of C_q and C_k relative to each other. θ_1 denotes the learnable parameter of the current MLP $f(.)$. $\grave{F}_{\text{spa}}$ denotes the MR spatial feature map, which is the result of the spatial feature map in the HR domain after downsampling by a factor of two

$$\grave{F}_{spa} = MP\left(F_{spa}\right), \tag{12}$$

where MP denotes the adaptive maxpooling. Furthermore, the process of fusing the information from the MR pixel with the HR relative coordinates can be represented as follows:

$$O_{p,q}^{\text{hr}} = f\left(F_{\text{spa}}\left(C_q\right), C_p - C_q, O^{\text{mr}}\left(C_q\right); \theta_2\right), \tag{13}$$

where $O_{p,q}^{\text{hr}} \in R^{1\times1\times c}$ denotes the multimodal fusion information of the query coordinates C_p at position C_q, and $O_{p,q}^{\text{hr}}$is a subset of $O^{\text{hr}} \in \text{R}^{H\times W\times c}$.

TABLE I
QUANTITATIVE COMPARISON ON GF2 RSTESTING

Method	SSIM ↑	PSNR ↑	SAM ↓	sCC ↑	ERGAS ↓	Q2n ↑	RASE ↓
EXP$_{2002}$	0.9116±0.0291	38.0234±1.9236	1.8531±0.3459	0.9542±0.0203	2.4094±0.4647	0.7971±0.0430	8.5984±1.7342
C-BDSD$_{2015}$	0.9570±0.0132	39.7738±1.6343	1.8980±0.3134	0.9585±0.0125	1.9190±0.3652	0.8942±0.0255	7.0361±1.3643
BT-H$_{2017}$	0.9630±0.0119	41.9236±1.9184	1.6819±0.3084	0.9690±0.0125	1.5526±0.3548	0.9089±0.0284	5.5194±1.2519
AWLP$_{2005}$	0.9422±0.0229	40.5603±2.0507	1.9687±0.4955	0.9329±0.0311	1.7241±0.3574	0.8610±0.0328	6.8346±1.7655
MF$_{2016}$	0.9516±0.0149	40.5649±1.8525	1.6842±0.3115	0.9670±0.0104	1.7763±0.3009	0.8784±0.0240	6.4951±1.2576
FS$_{2018}$	0.9533±0.0160	41.4049±1.9117	1.6807±0.3394	0.9685±0.0118	1.6201±0.3526	0.8904±0.0250	5.8813±1.3128
PNN$_{2016}$	0.9585±0.0108	41.1461±1.5464	1.7958±0.2507	0.9525±0.0175	1.6230±0.2571	0.8950±0.0301	5.9599±0.9998
PanNet$_{2017}$	0.9720±0.0065	42.8281±1.3514	1.5439±0.2339	0.9751±0.0097	1.3438±0.1944	0.9154±0.0352	4.8840±0.7304
BDPN$_{2019}$	0.9625±0.0135	42.1568±2.2831	1.4359±0.2739	0.9765±0.0110	1.5190±0.3684	0.9234±0.0217	5.1054±1.3084
TANI$_{2022}$	0.9782±0.0076	44.3576±2.0788	1.1067±0.2118	0.9859±0.0064	1.1494±0.2541	0.9498±0.0204	4.1662±0.9110
TRRNet$_{2022}$	0.9896±0.0025	47.1221±1.2405	0.8858±0.1240	0.9927±0.0030	0.8246±0.1153	0.9685±0.0138	2.9839±0.4240
MSSTN$_{2023}$	0.9848±0.0044	46.1604±1.8511	0.9641±0.1796	0.9896±0.0046	0.9358±0.1764	0.9676±0.0088	3.3735±0.6521
RSANet$_{2023}$	0.9887±0.0026	47.4746±1.4732	0.8846±0.1514	0.9918±0.0034	0.8011±0.1248	0.9734±0.0103	2.8742±0.4696
CMLNet$_{2023}$	0.9884±0.0030	47.3663±1.6947	0.8780±0.1566	0.9920±0.0034	0.8134±0.1416	0.9737±0.0087	2.9231±0.5262
Ours	**0.9903±0.0026**	**48.1099±1.5703**	**0.8076±0.1351**	**0.9933±0.0032**	**0.7539±0.1272**	**0.9782±0.0087**	**2.7014±0.4807**

TABLE II
QUANTITATIVE COMPARISON ON GF2 FSTESTING

Method	D_λ↓	D_s↓	QNR↑
EXP$_{2002}$	0.0000±0.0000	0.0263±0.0176	0.9737±0.0176
C-BDSD$_{2015}$	0.0119±0.0066	0.0439±0.0157	0.9448±0.0177
BT-H$_{2017}$	0.0216±0.0114	0.0555±0.0179	0.9242±0.0262
AWLP$_{2005}$	0.0154±0.0147	0.0491±0.0188	0.9364±0.0300
MF$_{2016}$	0.0326±0.0188	0.0593±0.0208	0.9104±0.0361
FS$_{2018}$	0.0236±0.0134	0.0524±0.0173	0.9254±0.0273
PNN$_{2016}$	0.0206±0.0112	0.0433±0.0151	0.9370±0.0226
PanNet$_{2017}$	0.0121±0.0095	0.0356±0.0127	0.9528±0.0193
BDPN$_{2019}$	0.0117±0.0096	0.0364±0.0127	0.9524±0.0194
TANI$_{2022}$	0.0105±0.0083	0.0419±0.0157	0.9481±0.0205
TRRNet$_{2022}$	0.0074±0.0067	0.0310±0.0118	0.9619±0.0152
MSSTN$_{2023}$	0.0100±0.0081	0.0375±0.0132	0.9529±0.0184
RSANet$_{2023}$	0.0099±0.0083	0.0378±0.0133	0.9528±0.0164
CMLNet$_{2023}$	0.0096±0.0078	0.0384±0.0132	0.9524±0.0180
Ours	**0.0063±0.0045**	**0.0238±0.0110**	**0.9700±0.0135**

C. Training and Implementation Details

Since neural networks are better at predicting high-frequency texture edge details, we then use jump connections to combine $\acute{M}$ with the output of the PINF O, thereby enhancing the high-frequency component of the implicit representation. Accordingly, the reconstructed HRMS can be obtained by the following equation:

$$\check{M} = O + \acute{M}, \tag{14}$$

where $\check{M} \in R^{H\times W\times c}$ denotes the HRMS. Since the main contribution of the proposed PINFNet is PINF, we employ the commonly used $L1$ loss to constrain the entire model. This loss can be utilized to focus on learning edge pixel information, which further enriches the high-frequency component in the implicit representation

$$Loss = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{l=1}^{b}\|\check{M}(m,n,l) - G(m,n,l)\|_1}{HWb}, \tag{15}$$

where G denotes the ground truth (GT).

IV. EXPERIMENTS

A. Datasets

In order to fully validate the performance of the proposed method, we conduct extensive experiments on three publicly available datasets (Gaofen-2, QuickBird, and WorldView-3) [50]. Each dataset includes both reduced-scale testing (RSTesting) and full-scale testing (FSTesting). Due to the absence of real HRMS images as training labels, it is only feasible to obtain LR versions of the original image pairs and utilize the original MS as the corresponding labels. In other words, our method is trained on the reduced-scale dataset, so the RSTesting mainly evaluates the fitting ability of the model, while the FSTesting mainly evaluates the generalization ability of the model in real scenarios.

B. Evaluation Metrics

The RSTesting is capable of evaluating the fitting capability of each method. Therefore, we employ seven widely utilized reference metrics to evaluate the fusion performance of each method in RSTesting, namely SSIM [51], PSNR, SAM [52], sCC [53], ERGAS [54], Q2n [55], and RASE [56]. Additionally, the FSTesting is conducted to evaluate the generalization ability of each method. Consequently, we employ three nonreference metrics to verify the fusion performance of each method at full scale, namely D_λ, D_s, and QNR [57].

C. Comparative Methods

A total of 14 state-of-the-art methods were selected for comparison, including six traditional methods and eight DL-based methods. Among the traditional methods, EXP [58] is an upsampling method without the aid of PAN images. C-BDSD [15] and BT-H [13] are the CS-based methods, while AWLP [21], MF [22], and MTF-GLP-FS (FS) [23] are the MRA-based methods. The remaining eight DL-based methods comprise PNN [28], PanNet [29], BDPN [30], TANI [31], TRRNet [32], MSSTN [33], RSANet [34], and CMLNet [35]. For a fair comparison, we use the authors' officially released code and the setup described in the original article. All of the codes are executed on a computer that is equipped with an i5-11600 CPU

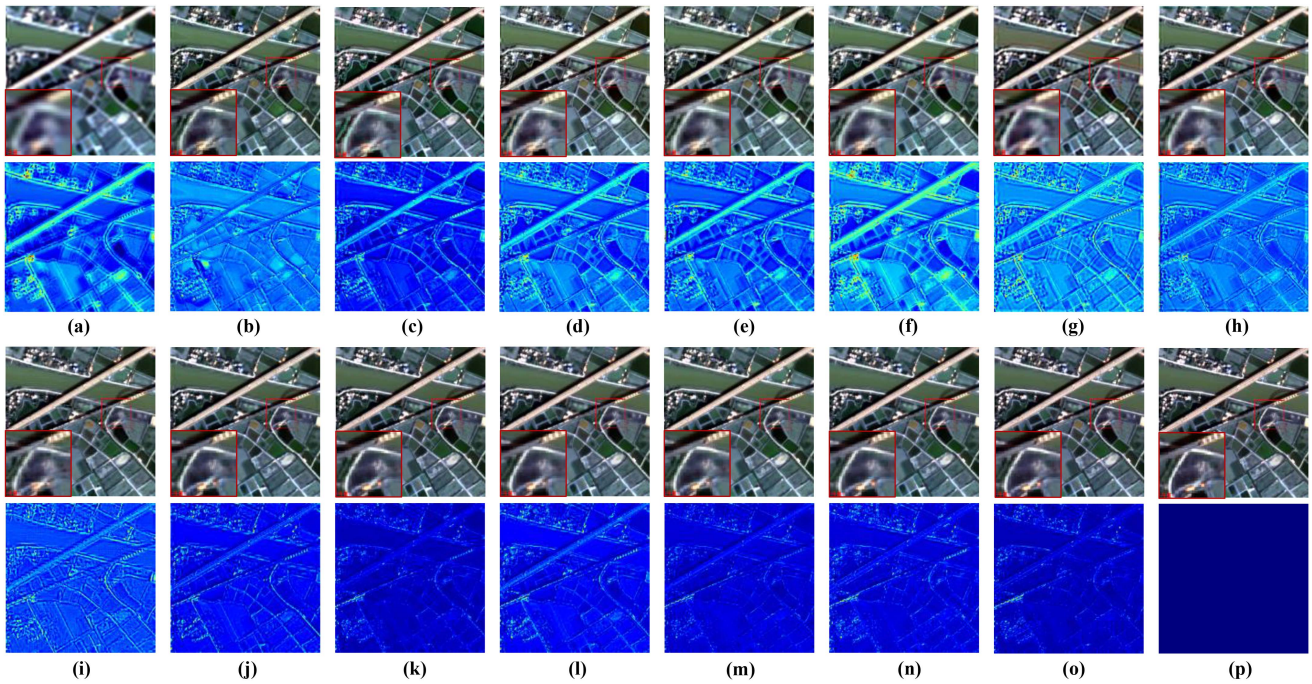

Fig. 4. Qualitative evaluation results on a reduced-scale sample acquired by GF2. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) GT. The fusion results are presented in the first and third rows, while the corresponding AEMs are listed in the second and fourth rows.

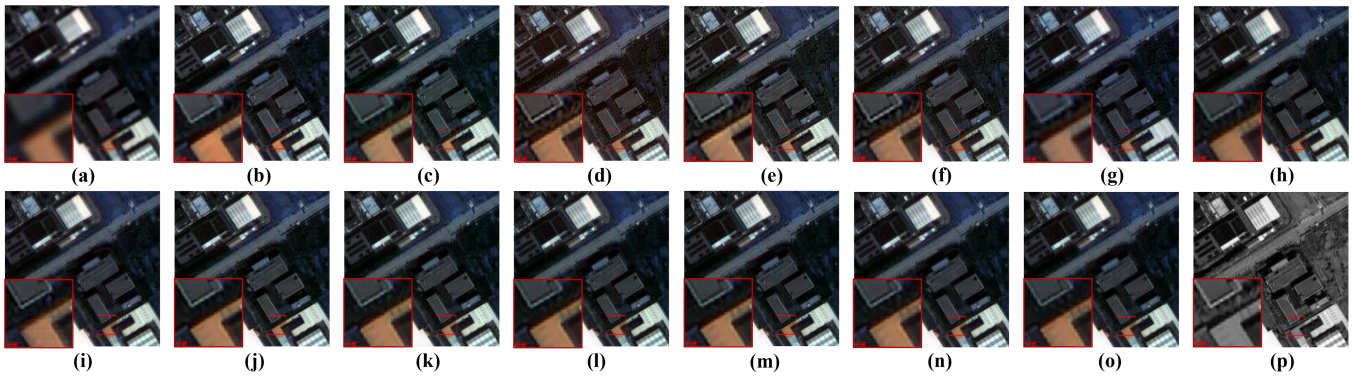

Fig. 5. Qualitative evaluation results on a full-scale sample acquired by GF2 dataset. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) PAN.

and two GTX-3060 GPUs. In our proposed methodology, the initial learning rate, epochs, and batch size are set to 1e-4, 1200, and 64, respectively. Besides, we choose Adam as the optimizer and decay the learning rate by one-quarter every 200 epochs.

D. Results on GF2 Dataset

Tables I and II present the quantitative results along with the mean and variance. The number labeled under the method name in the first column is the year of publication of the method. The top three scores are labeled red, green, and blue. It can be seen that, in general, DL-based methods perform better than the traditional methods on all metrics. Our proposed method achieves the best scores on all the metrics with significant advantages, which further validates the effectiveness of the proposed model.

Fig. 4 depicts the qualitative results of each method on the GF2 RSTesting. The first and third rows represent the fusion results for each method. The traditional methods appear to exhibit varying degrees of spatial and spectral distortion. The fusion results for PNN, PanNet, and BDPN in the red rectangular boxes show some details that do not exist in GT, and these implausible details lead to spectral distortion in local regions. The proposed method demonstrates the clearest edge details compared with other DL-based methods. Furthermore, in order to show the differences between the compared methods more visually, we provide the corresponding absolute error maps (AEMs) for each method, as shown in the second and fourth rows in Fig. 4. It is

TABLE III
QUANTITATIVE COMPARISON ON QB RSTESTING

Method	SSIM ↑	PSNR ↑	SAM ↓	sCC ↑	ERGAS ↓	Q2n ↑	RASE ↓
EXP$_{2002}$	0.6879±0.0840	28.3689±3.0528	8.5575±1.7025	0.8701±0.0264	12.0189±1.5432	0.5782±0.0774	46.6474±8.3320
C-BDSD$_{2015}$	0.8594±0.0318	31.0633±2.8605	8.2730±1.7657	0.9105±0.0284	8.7976±1.9225	0.8151±0.1053	34.1965±8.5797
BT-H$_{2017}$	0.8684±0.0288	32.3652±2.3924	7.2981±1.3674	0.9416±0.0104	4.4939±0.5895	0.8295±0.0963	29.1453±3.8422
AWLP$_{2005}$	0.8587±0.0319	32.1491±2.4621	8.3039±1.8186	0.9266±0.0204	7.7451±0.7005	0.8237±0.0982	29.9031±4.1927
MF$_{2016}$	0.8473±0.0353	31.7536±2.1175	8.0350±1.6997	0.9294±0.0135	8.9013±3.0752	0.8120±0.0950	33.1591±6.2940
FS$_{2018}$	0.8635±0.0281	32.4692±2.2643	7.8662±1.6282	0.9378±0.0124	7.4454±0.5512	0.8336±0.0932	29.0174±3.7189
PNN$_{2016}$	0.8607±0.0267	32.1623±2.1147	8.4605±1.5349	0.9217±0.0116	7.5907±0.5426	0.8222±0.1090	29.3501±3.5651
PanNet$_{2017}$	0.8804±0.0204	32.8161±2.0800	7.9971±1.4499	0.9320±0.0155	6.9987±0.5552	0.8413±0.1040	27.3428±3.4054
BDPN$_{2019}$	0.9270±0.0159	34.9920±2.1504	6.2399±1.1646	0.9609±0.0093	5.4729±0.6364	0.8965±0.0939	21.2544±3.1819
TANI$_{2022}$	0.9345±0.0134	35.5011±2.0416	5.3832±0.9007	0.9749±0.0046	5.2219±0.3829	0.9039±0.0961	20.1670±2.2983
TRRNet$_{2022}$	0.9445±0.0164	36.1991±2.2609	5.0151±0.9622	0.9786±0.0078	4.8495±0.9964	0.9146±0.0866	18.8421±4.3051
MSSTN$_{2023}$	0.9297±0.0163	35.0617±2.2083	5.3956±0.9889	0.9715±0.0079	5.5252±0.8150	0.8972±0.0959	21.3443±3.8320
RSANet$_{2023}$	0.9586±0.0052	37.9325±1.8095	4.7659±0.7969	0.9834±0.0037	3.8686±0.2722	0.9301±0.0916	15.0549±1.6040
CMLNet$_{2023}$	0.9482±0.0133	36.5725±2.1093	4.8879±0.8881	0.9799±0.0063	4.6149±0.8325	0.9176±0.0912	17.8909±3.6664
Ours	**0.9610±0.0054**	**38.0503±1.8677**	**4.5785±0.7738**	**0.9859±0.0034**	**3.8063±0.2920**	**0.9341±0.0857**	**14.8554**±1.6910

TABLE IV
QUANTITATIVE COMPARISON ON QB FSTESTING

Method	D_λ↓	D_s↓	QNR↑
EXP$_{2002}$	0.0001±0.0000	0.0529±0.0156	0.9470±0.0156
C-BDSD$_{2015}$	0.0372±0.0220	0.0586±0.0401	0.9066±0.0458
BT-H$_{2017}$	0.1457±0.0458	0.1244±0.0506	0.7489±0.0690
AWLP$_{2005}$	0.0467±0.0203	0.0975±0.0290	0.8605±0.0356
MF$_{2016}$	0.0350±0.0211	0.1105±0.0246	0.8589±0.0403
FS$_{2018}$	0.0342±0.0199	0.1146±0.0225	0.8554±0.0356
PNN$_{2016}$	0.1005±0.0952	0.0862±0.0435	0.8371±0.1745
PanNet$_{2017}$	0.0563±0.0560	0.1262±0.0369	0.8321±0.1465
BDPN$_{2019}$	0.0588±0.0546	0.0273±0.0147	0.9162±0.0638
TANI$_{2022}$	**0.0268±0.0251**	0.0920±0.0262	0.8841±0.0452
TRRNet$_{2022}$	0.0458±0.0389	0.0304±0.0103	0.9253±0.0406
MSSTN$_{2023}$	0.0268±0.0267	0.0937±0.0246	0.8826±0.0445
RSANet$_{2023}$	0.0348±0.0235	0.0250±0.0240	**0.9422±0.0355**
CMLNet$_{2023}$	0.0375±0.0348	0.0634±0.0264	0.9009±0.0275
Ours	0.0410±0.0380	**0.0193±0.0079**	0.9407±0.0419

clear that our method has the smallest residuals in addition to being the closest to GT in terms of fusion results. Fig. 5 shows the FSTesting results of each method on the GF2 dataset. The fusion results of the conventional methods demonstrate significant spatial and spectral distortions. The fusion results of PNN, PanNet, and BDPN are poor. In contrast, TRRNet, RSANet, and our proposed method demonstrate the most reasonable spectral distribution. By combined comparison, our method demonstrates the closest spatial details to PAN and the closest spectral details to EXP.

E. Results on QB Dataset

The quantitative results are shown in Tables III and IV, which indicate that our method still achieves the best scores on the QB dataset. Our method still demonstrates the strongest fitting ability, even when the features are more complex.

Figs. 6 and 7 depict the qualitative results on the QB dataset. The fusion results of traditional methods show obvious spectral distortion. The fusion results of PNN are even worse than those of traditional methods. The fusion results of PanNet and BDPN showed slight distortion. Furthermore, the fusion results of the remaining six DL-based methods demonstrate comparable fusion quality. As evident from the corresponding AEMs, RSANet and the proposed PINFNet show the closest visualization to GT. In contrast, the AEM corresponding to RSANet exhibits fewer residuals, whereas the AEM corresponding to PINFNet displays a darker hue. Similar to the results in RSTesting, the fusion results of the conventional method show significant spectral distortion in FSTesting. Most of the DL-based methods also demonstrate poor performance, e.g., PanNet, BDPN, TANI, TRRNet, RSANet, and CMLNet demonstrate significant spectral distortion. Visually, our method performs closest to the EXP in terms of spectral quality.

F. Results on WV3 Dataset

Our method achieves superior fusion results on two four-band datasets, and to further demonstrate the effectiveness of the proposed method, we conduct the corresponding qualitative and quantitative experiments on the eight-band WV3 dataset. The quantitative results are presented in Tables V and VI, where it is apparent that, as is customary, our method achieves the best scores on most metrics in RSTesting. Our proposed method does not perform the best on the full-resolution test metrics. Combined with the qualitative evaluation of FSTesting, we believe that the proposed method is still promising.

It is evident that the majority of the methods exhibit lower performance compared to the GF2 dataset. Similar outcomes can be observed in Figs. 8 and 9. In RSTesting, TRRNet, MSSTN, and PINFNet all show satisfactory fusion results, as shown in the corresponding AEMs. In contrast, the AEM corresponding to TRRNet exhibits the fewest residuals, while the AEMs corresponding to MSSTN and PINFNet are darker in color. In addition, the AEM corresponding to PINFNet exhibits fewer residuals compared with MSSTN, specifically in the flower bed portion of the zoomed-in region. In FSTesting, the fusion results of PanNet, TANI, and MSSTN demonstrate significant spectral distortion; whereas the fusion results of BDPN, TRRNet, RSANet, and CMLNet demonstrate significant spatial distortion as evidenced by severe distortion of the vehicle in the zoomed-in region. Additionally, the fusion result of PINFNet shows slight spectral distortion in the vehicle edge portion of the zoomed-in region, but it retains complete spatial details. In general, we believe that the proposed methodology remains promising due to its ability to strike a balance between the preservation of spatial and spectral information.

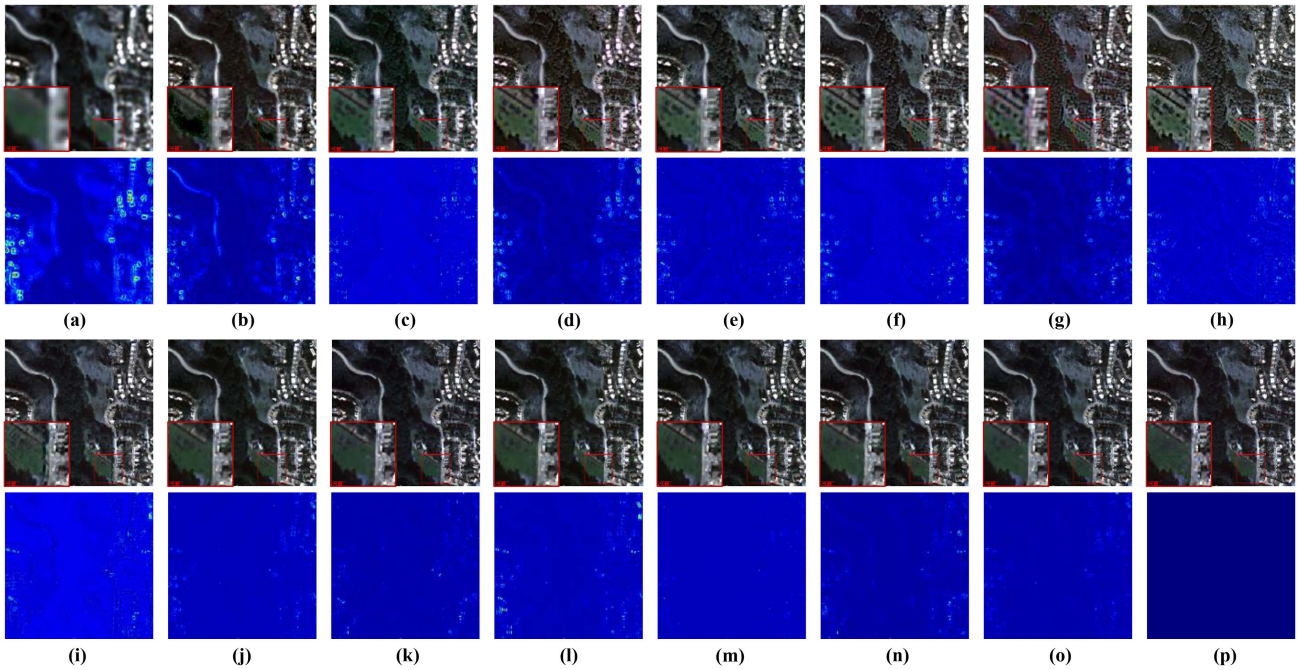

Fig. 6. Qualitative evaluation results on a reduced-scale sample acquired by QB. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) GT. The fusion results are presented in the first and third rows, while the corresponding AEMs are listed in the second and fourth rows.

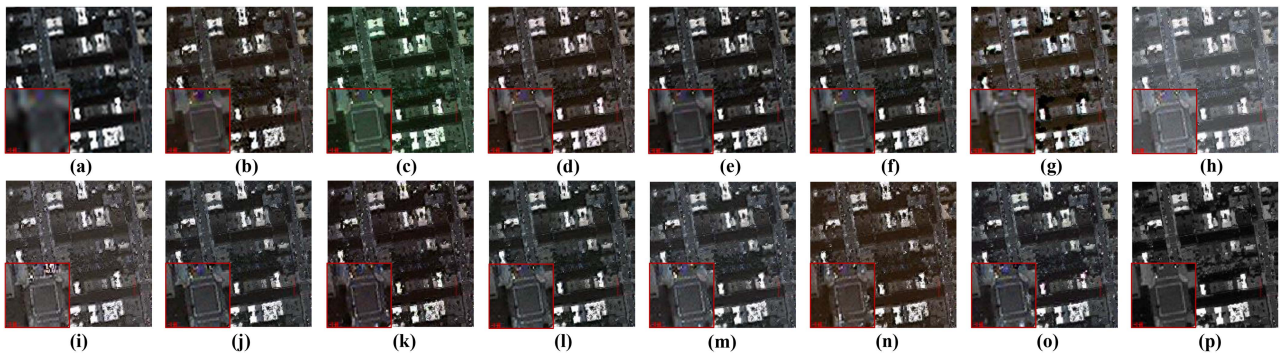

Fig. 7. Qualitative evaluation results on a full-scale sample acquired by QB dataset. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) PAN.

TABLE V
QUANTITATIVE COMPARISON ON WV3 RSTESTING

Method	SSIM ↑	PSNR ↑	SAM ↓	sCC ↑	ERGAS ↓	Q2n ↑	RASE ↓
EXP$_{2002}$	0.7246±0.0933	29.0499±3.3683	5.8351±1.6720	0.9226±0.0290	7.1354±1.5641	0.6027±0.0889	22.2931±6.3631
C-BDSD$_{2015}$	0.8717±0.0259	30.0994±2.7670	6.5466±1.5201	0.8998±0.0394	6.4582±2.5182	0.7530±0.1291	20.4397±8.7979
BT-H$_{2017}$	0.9018±0.0221	33.0920±2.5712	4.8984±1.2695	0.9586±0.0123	4.5107±1.2973	0.8182±0.0993	13.9308±4.7883
AWLP$_{2005}$	0.8952±0.0219	32.6717±2.6477	5.2762±1.3649	0.9508±0.0135	4.6971±1.3285	0.8066±0.1012	14.1126±4.7256
MF$_{2016}$	0.8846±0.0244	32.2962±2.7383	5.3162±1.4722	0.9527±0.0136	4.9191±1.2875	0.7957±0.1005	15.0103±4.7843
FS$_{2018}$	0.8905±0.0303	32.9542±2.4543	5.3228±1.6112	0.9560±0.0114	4.6450±1.4062	0.8177±0.0989	13.7713±4.8366
PNN$_{2016}$	0.8613±0.0281	31.3281±2.0218	6.5425±1.5494	0.9002±0.0422	5.2403±1.3255	0.8222±0.1090	29.3501±3.5651
PanNet$_{2017}$	0.8944±0.0183	32.0786±1.5508	6.9395±1.3863	0.9321±0.0355	5.2335±1.5127	0.7491±0.1464	15.5878±4.1562
BDPN$_{2019}$	0.9431±0.0128	35.3628±2.1251	4.4479±0.9630	0.9623±0.0160	3.3709±0.7864	0.8438±0.1075	9.9593±2.6070
TANI$_{2022}$	0.9639±0.0095	37.3004±2.6077	3.6054±0.6695	0.9781±0.0108	2.6869±0.5628	0.8912±0.0879	8.0661±2.3056
TRRNet$_{2022}$	0.9713±0.0078	38.2449±2.4122	3.2327±0.5808	0.9849±0.0059	2.3671±0.4995	0.9045±0.0858	7.3628±1.9141
MSSTN$_{2023}$	0.9622±0.0098	37.0005±2.5923	3.6973±0.7252	0.9734±0.0126	2.7775±0.6243	0.8965±0.0825	8.3879±2.4221
RSANet$_{2023}$	0.9717±0.0080	38.4846±2.5580	3.1252±0.5458	0.9864±0.0056	2.2904±0.4945	0.8995±0.0882	7.1681±1.8556
CMLNet$_{2023}$	0.9731±0.0073	**38.6805±2.4847**	3.0436±0.5397	0.9875±0.0050	**2.2344±0.4706**	0.9068±0.0850	**7.0343±1.8179**
Ours	**0.9734±0.0071**	38.6125±2.6158	**3.0288±0.5250**	**0.9875±0.0044**	2.2600±0.4542	**0.9117±0.0809**	7.1043±1.7623

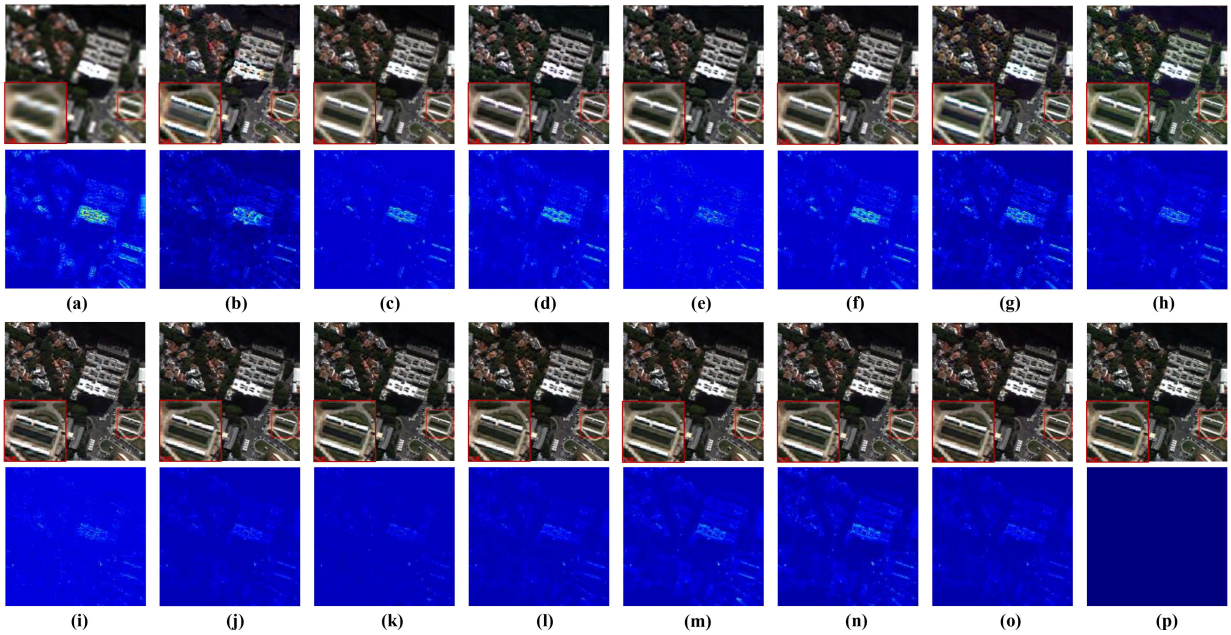

Fig. 8. Qualitative evaluation results on a reduced-scale sample acquired by WV3 dataset. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) GT. The fusion results are presented in the first and third rows, while the corresponding AEMs are listed in the second and fourth rows.

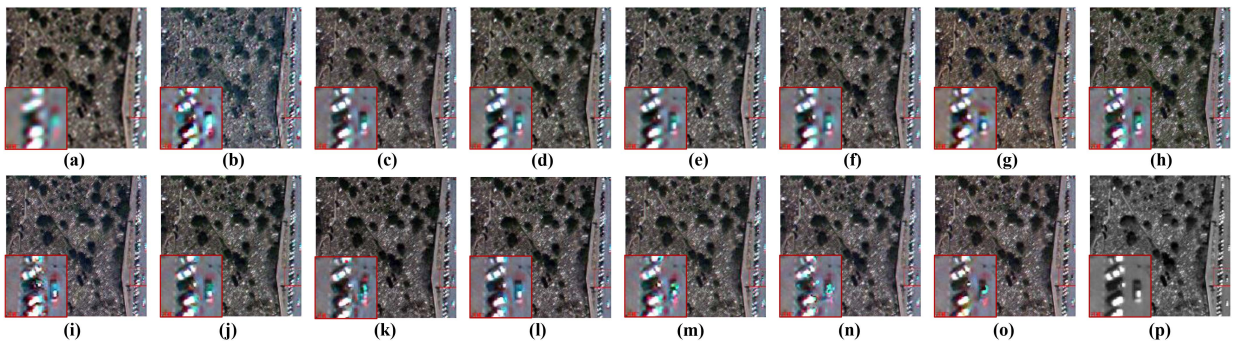

Fig. 9. Qualitative evaluation results on a full-scale sample acquired by WV3 dataset. (a) EXP. (b) C-BDSD. (c) BT-H. (d) AWLP. (e) MF. (f) FS. (g) PNN. (h) PanNet. (i) BDPN. (j) TANI. (k) TRRNet. (l) MSSTN. (m) RSANet. (n) CMLNet. (o) Ours. (p) PAN.

TABLE VI
QUANTITATIVE COMPARISON ON WV3 FSTESTING

Method	$D_\lambda\downarrow$	$D_s\downarrow$	QNR↑
EXP$_{2002}$	0.0000±0.0000	0.0340±0.0133	0.9660±0.0166
C-BDSD$_{2015}$	0.0624±0.0409	0.0691±0.0207	0.8735±0.0535
BT-H$_{2017}$	0.0268±0.0198	0.1001±0.0375	0.8764±0.0516
AWLP$_{2005}$	0.0219±0.0155	0.0761±0.0312	0.9040±0.0425
MF$_{2016}$	0.0266±0.0194	0.0853±0.0293	0.8909±0.0438
FS$_{2018}$	0.0197±0.0168	0.0851±0.0307	0.8973±0.0432
PNN$_{2016}$	0.0303±0.0139	0.0832±0.0304	0.8893±0.0405
PanNet$_{2017}$	0.0251±0.0192	0.0933±0.0348	0.8845±0.0485
BDPN$_{2019}$	0.0237±0.0160	0.0532±0.0167	0.9245±0.0281
TANI$_{2022}$	**0.0157±0.0151**	0.0849±0.0319	0.9012±0.0433
TRRNet$_{2022}$	0.0182±0.0141	0.0691±0.0213	0.9141±0.0318
MSSTN$_{2023}$	0.0200±0.0193	0.0860±0.0318	0.8963±0.0473
RSANet$_{2023}$	0.0205±0.0179	0.0498±0.0225	**0.9307±0.0275**
CMLNet$_{2023}$	0.0353±0.0182	**0.0436±0.0172**	0.9226±0.0260
Ours	0.0248±0.0158	0.0587±0.0222	0.9180±0.0271

G. Ablation Experiments

Our proposed method consists of two encoders and PINF, and in order to verify the effectiveness of the components, we performed the corresponding ablation experiments on each component. The results of the experiments are shown in Fig. 10 and Table VII, where it can be seen that our method performs the best in terms of fusion quality while demonstrating the fewest residuals in the AEMs. In addition, since the main innovation of the proposed method lies in the PINF, it can be seen that the fusion quality is significantly reduced when the PINF is removed, which can also be seen in the quantitative experimental results.

In addition, to verify the implicit representation capability of the proposed PINF, we compare it with four commonly used explicit interpolation techniques. We only change the upsampling method in the model with all other experimental settings identical. The experimental results are shown in Fig. 11 and Table VIII. It can be clearly seen that our method demonstrates the best edge details, while the AEMs can further demonstrate the superiority of the proposed method. Besides, the performance of our method is far ahead of the performance of

TABLE VII
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS CONDUCTED FOR EACH COMPONENT OF THE PROPOSED PINFNET

Method	SSIM ↑	PSNR ↑	SAM ↓	sCC ↑	ERGAS ↓	Q2n ↑	RASE ↓
w/o E_δ	0.9683±0.0080	37.8295±2.4098	3.2814±0.6431	0.9851±0.0066	2.4956±0.5437	0.8982±0.0884	7.8215±2.0825
w/o E_φ	0.9693±0.0110	38.1579±2.4570	3.2554±0.7308	0.9838±0.0085	2.4157±0.5246	0.9061±0.0820	7.4517±1.9431
w/o PINF	0.9711±0.0077	38.2403±2.4108	3.1644±0.5959	0.9857±0.0057	2.3722±0.5551	0.9057±0.0820	7.4405±2.0655
Ours	**0.9734±0.0071**	**38.6125±2.6158**	**3.0288±0.5250**	**0.9875±0.0044**	**2.2600±0.4542**	**0.9117±0.0809**	**7.1043±1.7623**

TABLE VIII
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS CONDUCTED FOR DIFFERENT INTERPOLATION METHODS

Method	SSIM ↑	PSNR ↑	SAM ↓	sCC ↑	ERGAS ↓	Q2n ↑	RASE ↓
Bicubic	0.9637±0.0112	37.4703±2.3036	3.4215±0.7092	0.9850±0.0066	2.5839±0.5567	0.8813±0.0894	8.0364±1.8862
Bilinear	0.9667±0.0095	37.7040±2.3720	3.3079±0.6646	0.9846±0.0079	2.4989±0.5288	0.8893±0.0862	7.8965±1.9389
Pixel Shuffle	0.9704±0.0095	38.1842±2.6198	3.1427±0.5874	0.9867±0.0060	2.3720±0.5278	0.9000±0.0822	7.5788±2.1088
Transpose	0.9682±0.0136	37.9389±2.7218	3.2217±0.6586	0.9855±0.0068	2.4841±0.7978	0.8941±0.0845	7.9491±2.9186
PINF	**0.9734±0.0071**	**38.6125±2.6158**	**3.0288±0.5250**	**0.9875±0.0044**	**2.2600±0.4542**	**0.9117±0.0809**	**7.1043±1.7623**

TABLE IX
QUANTITATIVE COMPARISON OF EFFICIENCY ON WV3 DATASET

	PNN_{2016}	**$PanNet_{2017}$**	**$BDPN_{2019}$**	**$TANI_{2022}$**	**$TRRNet_{2022}$**	**$MSSTN_{2023}$**	**$RSANet_{2023}$**	**$CMLNet_{2023}$**	**Ours**
Params (M)	0.18	0.31	5.92	**0.08**	5.23	25.35	0.20	0.33	0.23
FLOPs (G)	0.72	2.50	30.28	**0.34**	3.75	24.94	0.71	2.68	2.73
Time (s)	**0.002**	0.003	0.007	0.004	0.085	0.079	0.059	0.026	0.035

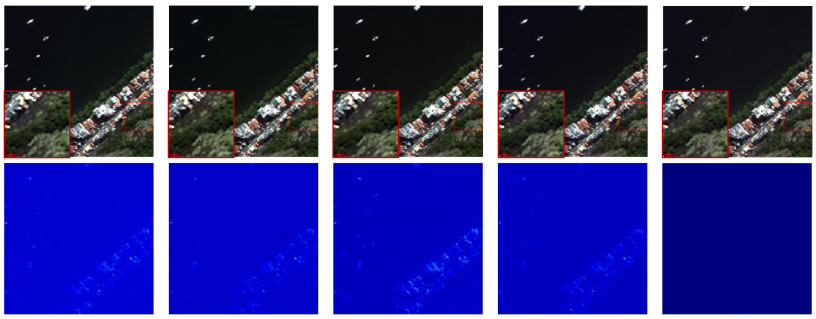

Fig. 10. Qualitative results of ablation experiments conducted for each component of the proposed PINFNet. (a) w/o E_δ. (b) w/o E_φ. (c) w/o PINF. (d) Ours. (e) GT.

other methods in all the metrics, which greatly indicates that PINF is more suitable for the upsampling requirement in the fusion process.

H. Efficiency Analysis

We validate the efficiency of each method in three aspects: model complexity (Params), computational complexity (FLOPs), and time complexity (testing time). As shown in Table IX, our method ranks medium on the three metrics. Specifically, our method is second only to PNN, TANI, and RSANet in terms of the number of parameters, and second only to PNN, PanNet, TANI, and CMLNet in terms of the amount of computation. In addition, in order to show the relationship between fusion performance and fusion efficiency more intuitively, we visualize the corresponding scatter plots. As illustrated in Fig. 12, our methodology achieves the best balance between fusion performance and efficiency.

V. DISCUSSION

Existing methods use discrete sampling techniques to sample images or features of different scales to the same scale, which tends to result in the loss of critical information. To alleviate this effect, we propose PINFNet inspired by INR. The proposed PINFNet establishes a coordinate modal relationship between spatial information in the HR domain and spectral information in the LR domain through continuous coordinate mapping. Subsequently, features at various scales are fused within the continuous domain, enabling high-fidelity fusion. The results of Section IV-D–F demonstrate that the proposed PINFNet is capable of achieving superior fusion outcomes with better spatial and spectral retention. Section IV-G validates the efficacy of the proposed methodology. Furthermore, the efficiency analysis section shows that our method effectively restricts the model parameters and computational complexity, which is attributed to the utilization of only two MLP layers in the fusion stage.

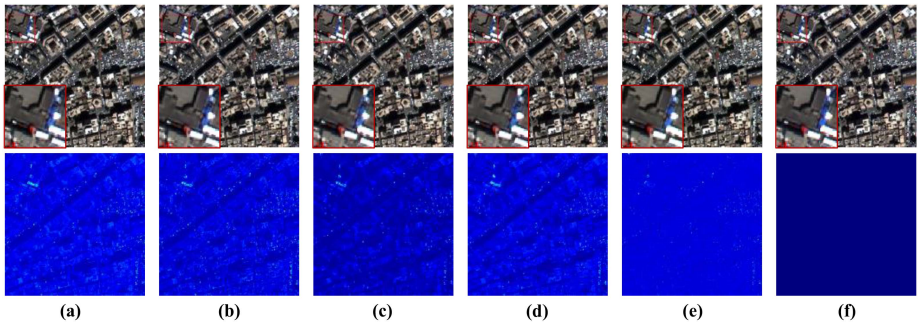

Fig. 11. Qualitative results of ablation experiments conducted for different interpolation methods. (a) Bicubic interpolation. (b) Bilinear interpolation. (c) PixelShuffle. (d) Transpose. (e) Proposed PINF. (f) GT.

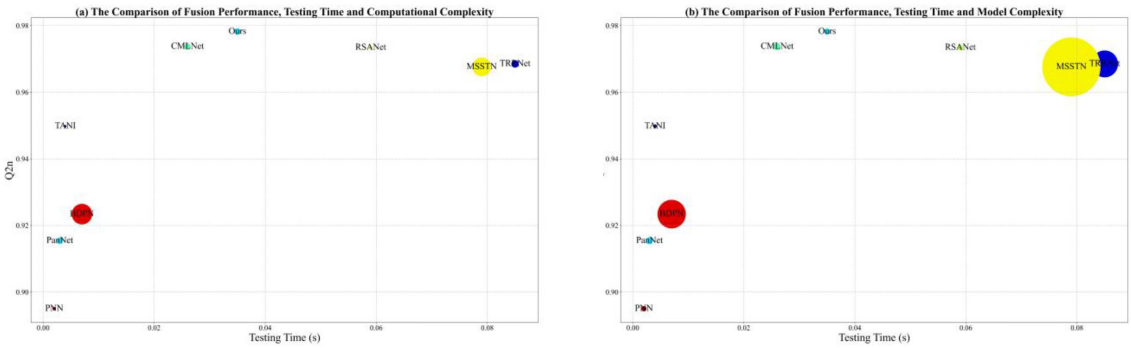

Fig. 12. Analysis of fusion performance and fusion efficiency.

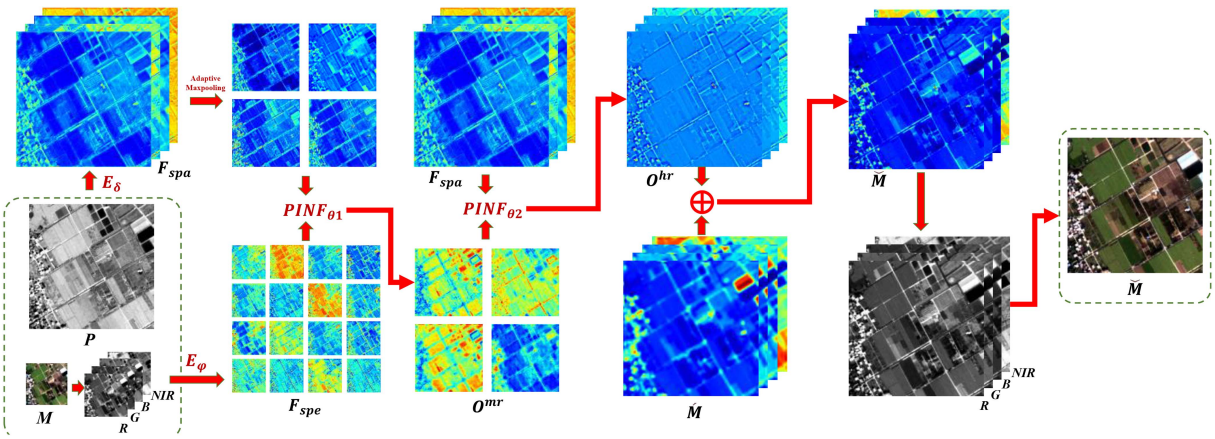

Fig. 13. Visualization of intermediate features in the proposed PINFNet.

However, the proposed PINF requires to compute the coordinates of different modal images to establish positional associations, an operation that increases the computational effort and runtime of the model to some extent. Overall, our method achieves a better balance between fusion performance and efficiency. In addition, we visualize the intermediate features of the model. An instance with images from the GF2 dataset is delivered, as shown in Fig. 13. It is apparent that there is no misalignment of different modal features during the fusion process, which further validates the effectiveness of the proposed PINFNet.

VI. Conclusion

In order to avoid the misalignment issue arising from the use of discrete sampling techniques, we propose a novel PINFNet. The proposed PINFNet establishes a coordinate modal relationship between spatial information in the HR domain

and spectral information in the LR domain through continuous coordinate mapping. Subsequently, features at various scales are fused within the continuous domain, enabling high-fidelity fusion. Numerous experiments have proved the effectiveness and superiority of the proposed methodology. However, the proposed method has much room for improvement in terms of timeliness. As PINFNet necessitates computing the coordinates of diverse modal data in the continuous domain during the fusion stage in order to establish positional correlation, this significantly elevates the computational complexity and runtime. In future work, we will optimize the proposed method to further improve the fusion performance and efficiency. We will attempt to perform local region-wide coordinate mapping to alleviate the computational burden of pixelwise computation.

REFERENCES

[1] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," *Med. Image Anal.*, vol. 76, Feb. 2022, Art. no. 102307.

[2] P. Duan, P. Ghamisi, X. Kang, B. Rasti, S. Li, and R. Gloaguen, "Fusion of dual spatial information for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7726–7738, Sep. 2021.

[3] Z. Xie, P. Duan, W. Liu, X. Kang, X. Wei, and S. Li, "Feature consistency-based prototype network for open-set hyperspectral image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9286–9296, Jul. 2024.

[4] Z. Xie, J. Hu, X. Kang, P. Duan, and S. Li, "Multilayer global spectral–spatial attention network for wetland hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518913.

[5] G. Vivone et al., "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.

[6] F. Dadrass 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 J. Photogramm. Remote Sens.*, vol. 171, pp. 101–117, Jan. 2021.

[7] F. J. Garcıá-Haro, M. A. Gilabert, and J. Meliá, "Extraction of end-members from spectral mixtures," *Remote Sens. Environ.*, vol. 68, no. 3, pp. 237–253, Jun. 1999.

[8] X. Kang, S. Li, and J. A. Benediktsson, "Pansharpening with matting model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5088–5099, Aug. 2014.

[9] M. Ghahremani and H. Ghassemian, "Nonlinear IHS: A promising method for pan-sharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 11, pp. 1606–1610, Nov. 2016.

[10] Q. Xu, B. Li, Y. Zhang, and L. Ding, "High-fidelity component substitution pansharpening by the fitting of substitution data," *IEEE Trans. Geosci Remote Sens.*, vol. 52, no. 11, pp. 7380–7392, Nov. 2014.

[11] 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. 2011.

[12] 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.

[13] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.

[14] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[15] A. Garzelli, "Pansharpening of multispectral images based on nonlocal parameter optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096–2107, Apr. 2015.

[16] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[17] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[18] J.-L. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, Jun. 2002.

[19] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.

[20] J. G. 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, Dec. 2000.

[21] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "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.

[22] R. Restaino, G. Vivone, M. Dalla Mura, and J. Chanussot, "Fusion of multispectral and panchromatic images based on morphological operators," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2882–2895, Jun. 2016.

[23] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[24] 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.

[25] Y. Zhang, S. De Backer, and P. Scheunders, "Noise-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3834–3843, Nov. 2009.

[26] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "Model-based fusion of multi- and hyperspectral images using PCA and wavelets," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2652–2663, May 2015.

[27] Y. Zhang, A. Duijster, and P. Scheunders, "A Bayesian restoration approach for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 9, pp. 3453–3462, Sep. 2012.

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

[29] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1753–1761.

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

[31] W. Diao, F. Zhang, H. Wang, J. Sun, and K. Zhang, "Pansharpening via triplet attention network with information interaction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3576–3588, Apr. 2022.

[32] K. Zhang, Z. Li, F. Zhang, W. Wan, and J. Sun, "Pan-sharpening based on transformer with redundancy reduction," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Jun. 2022, Art. no. 5513205.

[33] S. Jia, Z. Min, and X. Fu, "Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 96, pp. 117–129, Aug. 2023.

[34] C. Liu, W. Lü, Z. Zhang, X. Feng, and S. Xiang, "Recursive self-attention modules-based network for panchromatic and multispectral image fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 10067–10083, Oct. 2023.

[35] J.-D. Wang, L.-J. Deng, C.-Y. Zhao, X. Wu, H.-M. Chen, and G. Vivone, "Cascadic multireceptive learning for multispectral pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5408416.

[36] N. H. Kaplan and I. Erer, "Bilateral filtering-based enhanced pansharpening of multispectral satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 11, pp. 1941–1945, Nov. 2014.

[37] N. H. Kaplan, I. Erer, O. Ozcan, and N. Musaoglu, "MTF driven adaptive multiscale bilateral filtering for pansharpening," *Int. J. Remote Sens.*, vol. 40, no. 16, pp. 6262–6282, Aug. 2019.

[38] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.

[39] Y. Yang, C. Wan, S. Huang, H. Lu, and W. Wan, "Pansharpening based on low-rank fuzzy fusion and detail supplement," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5466–5479, Sep. 2020.
[40] J.-L. Xiao, T.-Z. Huang, L.-J. Deng, Z.-C. Wu, and G. Vivone, "A new context-aware details injection fidelity with adaptive coefficients estimation for variational pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Feb. 2022, Art. no. 5408015.
[41] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 1, pp. 3789–3799.
[42] C. Jiang, A. Sud, A. Makadia, J. Huang, M. NieBner, and T. Funkhouser, "Local implicit grid representations for 3D scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6000–6009.
[43] H. Li and H.-W. Shen, "Improving efficiency of iso-surface extraction on implicit neural representations using uncertainty propagation," *IEEE Trans. Visual. Comput. Graph.*, early access, Feb. 2024, doi: 10.1109/TVCG.2024.3365089.
[44] J. Tang, X. Chen, and G. Zeng, "Joint implicit image function for guided depth super-resolution," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4390–4399.
[45] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8624–8634.
[46] Z. Chen et al., "VideoINR: Learning video implicit neural representation for continuous space-time super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2037–2047.
[47] C. Zhu, R. Dai, L. Gong, L. Gao, N. Ta, and Q. Wu, "An adaptive multi-perceptual implicit sampling for hyperspectral and multispectral remote sensing image fusion," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 125, Dec. 2023, Art. no. 103560.
[48] C. Zhu, S. Deng, Y. Zhou, L.-J. Deng, and Q. Wu, "QIS-GAN: A lightweight adversarial network with quadtree implicit sampling for multispectral and hyperspectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Nov. 2023, Art. no. 5531115.
[49] L. Liu, Z. Zou, and Z. Shi, "Hyperspectral remote sensing image synthesis based on implicit neural spectral mixing models," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5500514.
[50] L.-J. Deng et al., "Machine learning in pansharpening: A benchmark, from shallow to deep networks," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 3, pp. 279–315, Sep. 2022.
[51] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[52] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, vol. 1, pp. 147–149.
[53] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, Jan. 1998.
[54] G. Vivone et al., "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.
[55] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.
[56] M. Choi, "A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1672–1682, Jun. 2006.
[57] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.
[58] 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.

Yao Feng received the bachelor's degree in surveying and mapping engineering from the Shandong University of Science and Technology, Qingdao, China, in 2012.

He is a Researcher with Shandong GEO-Surveying and Mapping Institute, Jinan, China. His research interests include UAV aerial survey, panchromatic sharpening, and deep learning.

Long Zhang received the master's degree in cartography and geographic information engineering from Yunnan Normal University, Kunming, China, in 2012.

He is a Researcher with Shandong GEO-Surveying and Mapping Institute, Jinan, China. His research interests include change detection, panchromatic sharpening, and deep learning.

Yingwei Zhang received the bachelor's degree in surveying and mapping engineering from Shandong Jiaotong University, Jinan, China, in 2007.

He is a Researcher with Shandong GEO-Surveying and Mapping Institute, Jinan, China. He specializes in unmanned aerial surveying, panchromatic sharpening, and deep learning.

Xinguo Guo received the bachelor's degree in civil engineering from the Shandong University of Science and Technology, Qingdao, China, in 2011.

He is a Researcher with Shandong GEO-Surveying and Mapping Institute, Jinan, China. His research interests include surveying and mapping, image processing, and data fusion.

Guangqi Xie received the Ph.D. degree photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2022.

His research interests include image matching and registration, panchromatic sharpening, and image super-resolution.

Chuang Liu is currently working toward the M.Eng. degree in computer technology with the School of Computer Science, Hubei University of Technology, Wuhan, China.

His research interests include remote sensing image processing, multimodal image fusion, low-level vision, machine learning, and deep learning.

Shao Xiang received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Surveying, Mapping and Remote Sensing Information Engineering, Wuhan University, Wuhan, China, in 2024.

His research interests include artificial intelligence, pattern recognition, and remote sensing image processing.