MFMIF-Net: Multi-Scale Feature Memory Interactive Fusion Net For Pansharpening

Yu Xin, Wu Zheng

***Abstract*—Pansharpening is the fusion of panchromatic (PAN) image with multispectral (MS) image to obtain high spatial resolution multispectral (HRMS) image. Due to the limitations of convolution operations and the diversity of remote sensing image features, multi-scale remote sensing pan-sharpening methods cannot effectively establish the connection between features at different scales. In order to establish the connection between different features, we use the "memory" mechanism of GRU and introduce it into the task of Pansharpening of remote sensing image, Establish connections between features of different scales and features at different levels, eliminate unnecessary noise and information redundancy in the process of feature extraction, while retaining important information. Specifically, we proposed a progressive fusion block, in which we proposed a Multi-Scale Memory Interaction Fusion Block and an Adaptive Feature Fusion Block. The former fully extracts features of different scales and establishes the connection between features of different scales, and then The shallow features are fused in a progressive manner to extract features of different depths, while establishing connections between features of different depths. Finally, the Adaptive Feature Fusion Block is used to adaptively fuse the shallow features and deep features to generate a sharpened HRMS. Extensive experiments prove that our proposed method is superior to existing state-of-the-art Pansharpening methods.**

***Index Terms*—Pansharpening, remote sensing image fusion,ConvGRU.**

INTRODUCTION

With the rapid increase in the amount of satellite data, remote sensing technology has advanced by leaps and bounds, promoting progress in many fields such as agriculture and environmental protection [1,2]. Due to the satellite's powerful ground measurement capabilitiy, the remote sensing images captured by its sensors contain rich ground information. Low-resolution multispectral (LRMS) image and panchromatic (PAN) image are two common satellite data. The former has high spectral resolution but low spatial resolution [3], while the latter has low spectral resolution but high spatial resolution [4]. Unfortunately, the lack of high-resolution multispectral (HRMS) images affects the application of remote sensing image in the above fields. In this context, pan-sharpening technology emerged [5-7], which aims to fuse high-resolution panchromatic (HRPAN) and low-resolution multispectral (LRMS) image to produce HRMS image. Based on real-life requirements. In recent years, leveraging Pansharpening techniques for the fusion of remote sensing image has become a prominent focus of research.

A large number of remote sensing image Pansharpening methods have been proposed. They can be divided into four groups: (1) Methods based on component substitution (CS), This kind of method employs the tailored transformation to transfer MS and PAN image into a new domain, and then replaces the specific component of MS image with the spatial information of PAN image to achieve texture enhancement. Representative methods include the hue-intensity-saturation(HIS) [8] and principal component analysis(PCA) [9][10].(2) Methods based on multiresolution analysis (MRA), Including wavelet transforms (WT)s) [11], smoothing filter intensity modulation (SFIM) [12], modulation transfer function with generalized Laplacian pyramid, MTF-GLP) [13], and MTF-GLP with high-pass modulation (MTF-GLP-HPM) [14].(3) Methods based on variational optimization (VO), These methods use known prior information to construct regular terms to reasonably constrain the model, and obtains the final panchromatic sharpening result through efficient solving algorithms. Representative methods include pan-sharpening models based on sparse prior construction regular terms [15], non-local similarity based on image [16] and pan-sharpening methods (TV) based on total variation models [17]. (4) Methods based on deep learning (DL). In the past few years, deep learning has become the focus of attention due to the powerful nonlinear modeling and feature extraction capabilities of neural networks. Deep neural networks have been successfully used in target detection [18] and behavior recognition [19] , super-resolution [20] and other fields. When dealing with the issue of pan-sharpening in remote sensing image, deep learning techniques exhibit distinct advantages. This approach does not require prior manual knowledge and allows for end-to-end training using existing low-resolution multispectral (LRMS) and panchromatic (PAN) image, learning their mapping relationships in high-resolution multispectral (HRMS) image.

Inspired by image super-resolution network (SRCNN) [21] · Masi et al. [22] proposed a simple three-layer convolutional network structure (PCNN), which achieved greater performance improvement compared with traditional methods. Subsequently, Yang et al. [23] proposed a general pan-sharpening framework (PanNet), which leverages domain knowledge to propose two main goals of pan-sharpening: preserving spectral and

Manuscript received April 17, 2025
Yu Xin, School of Computer Science and Technology, Tiangong University, Tianjin, China
Wu Zheng, School of Computer Science and Technology, Tiangong University, Tianjin, China

spatial information. It introduces high-pass filtering and spectrum mapping based on residual networks [24], which makes PanNet better transferable between different satellites by training on the high-pass domain. In addition, PSGAN [25] applies Generative Adversarial Network (GAN) [26] to Pansharpening of remote sensing image for the first time. This method enables the two networks to form a dynamic "game process", which can greatly improve the quality of the fused image, making PSGAN one of the most effective neural networks. Xu et al. [27] proposed a model-based deep learning pan-sharpening method (GPPNN), which applies prior knowledge to optimize the generation model of PAN and LRMS images,This is the first model-driven neural network pan-sharpening method in remote sensing. Fu et al. [28] proposed a new convolutional neural network-based method using dilated convolution (DMDNet) for panchromatic sharpening by combining deep learning techniques with domain-specific knowledge. Deng et al. [29] proposed a new deep convolutional fusion network (FusionNet) based on the traditional CS and MRA framework, absorbing the advantages of simple structure and fewer network parameters of traditional methods. Inspired by the back-projection (BP) mechanism, Zhang et al. [30] proposed a BP-driven model, spatial-spectral double back-projection network (S2DBPN), to fuse low spatial resolution multispectral (LRMS) image and high spatial resolution Resolution Panchromatic(PAN) Image, which utilizes reflection projection in the spatial and spectral domains to generate image.

In remote sensing images, different objects usually present diverse colors, textures, and shapes, and their distribution in the image is also different. Some of the above methods [23], [24], [27] and other methods use fixed convolution kernel designs, which limits the ability to extract different features and cannot effectively extract features of objects of different sizes. Some methods use dilated convolution [28] and multi-scale convolution [31] to effectively extract features of objects of different sizes, but they fail to effectively establish the connection between features of different scales and lack attention to global information. In recent years, vision transformers (ViTs) [32] have achieved excellent performance in machine vision tasks. It uses a self-attention mechanism to capture global and local features in image, and can better handle long-range dependencies in image, but Vit requires a large amount of computing resources and training data to train the model, so it requires longer training time and higher cost. In order to solve the above problems, we introduce the convolution gate recurrent unit (ConvGRU), which is a variant of GRU (the specific structure is introduced in Ⅱ.B). GRU is a special type of RNN. It has a reset gate and an update gate, which can better capture and propagate important information in the sequence. It is often used for time series data prediction and recognition tasks [33, 34]. Specifically, we designed a multi-scale memory interactive fusion block, which uses ConvGRU to establish connections between features at different scales, eliminate unnecessary noise and redundant information in the process of selecting local features, and at the same time establish long-range dependencies in image.

Deep features focus on semantic information, while shallow features focus on detail information [35]. Semantic information is beneficial to the recovery of edge information of objects in image [36], and shallow features are beneficial to the recovery of texture information of objects in image. As the depth of deep neural network increases, the network will pay more attention to the semantic information in the image, but this will bring about the problem of Gradient Vanishing. The difference from [30] is that we do not use residual density to solve the Gradient Vanishing problem, and use the "memory" mechanism of GRU to selectively select available information from different depth features and update it in the hidden state features to pass it to the next One layer, while filtering useless information and forgetting it. Compared with [30], we use this method to effectively reduce the transmission of useless information, which is beneficial to the training and optimization of the network. In order to make full use of the detail information in shallow features and the semantic information in deep features, we designed an adaptive feature fusion block that can adaptively associate shallow features and deep features. By aggregating image features at different depths, we can model the correlation between shallow and deep features, thereby enhancing the network's ability to reconstruct the target image.

In general, the main contributions of this article can be divided into the following three points:

(1)A Pansharpening network (MFMIF-Net) based on multi-scale feature memory interactive fusion is proposed. It can establish the connection between different scale features and different depth features, fully interact between different features, and can achieve excellent results to complete the Pansharpening task. Extensive experiments on three datasets demonstrate the superior performance of the proposed MFMIF-Net.

(2)We introduced ConvGRU into the pansharpening task, and proposed Multi-Scale Memory Interaction Fusion Block (MMIFB) and Progressive Feature Fusion Block (PFFB) with the help of GRU's "memory" mechanism. The former establishes the connection between features at different scales. The latter establishes connections between features of different depths.

(3)We proposed an Adaptive Feature Fusion Block that can adaptively associate shallow features and deep features, and fully utilize the detail information of shallow features and the semantic information of deep features to reconstruct the target image.

RELATED WORK

A.Deep-Learning-Based Pansharpening Methods

With the development of deep learning technology, various multispectral panchromatic sharpening methods based on deep learning have been proposed. CNN-based pan-sharpening (PNN) [22] is the first work to apply CNN to remote sensing pan-sharpening, which uses three convolutional layers to fuse PAN and LR-MS image. In [23], [43], and [44], global residual connections were

utilized to simplify HRMS regression by solely learning the residual parts. The Multiscale and Multidepth Convolutional Neural Network (MSDCNN) [45] attempted to aggregate multiscale features by employing multiple convolutional layers with different kernel configurations. Due to the characteristics of remote sensing images, different objects typically exhibit diverse colors, textures, and shapes, and they also vary in distribution within the images [46]. The aforementioned methods failed to establish connections between different features, inadequately considering the similarity between different features and the differences among identical features. Therefore, we consider establishing connections between different features to facilitate interactions among them, adequately considering both the similarities between different features and the differences among identical features, thereby enhancing the performance of pansharpening tasks.

B. CONVGRU

GRU [39] is a variant similar to LSTM. It also handles long-term dependencies by introducing a gating mechanism. It is often used in time series data prediction and identification tasks [33, 34]. Compared with LSTM, GRU has a more structured structure. For simplicity, GRU has fewer parameters and can achieve the same performance as LSTM. The original LSTM and GRU are mainly used to process one-dimensional sequence data. Due to their superior performance in processing time series, data prediction and other tasks, many researchers consider introducing them into computer vision tasks. In [40], shi et al. proposed a method for precipitation nowcasting method that introduced LSTM into computational vision tasks for the first time. In this method, shi et al. proposed convLSTM to process radar wave signals to achieve accurate precipitation nowcasting. Thanks to the excellent performance of convLSTM, many researchers have proposed various variants of GRU to handle specific computer vision tasks. In [41] Yuan et al. proposed an attention convolution GRU module (AttConvGRU) to learn the spatial correlation and long-range context dependence information of smoke, and achieved satisfactory results on the smoke segmentation task. In order to enable GRU to handle 2D features, Yuan et al. in [41] replaced the fully connected layer of the original GRU [39] with two-dimensional convolution, while using an attention mechanism to guide the current state. In order to make full use of shallow features, Wang et al. proposed a multi-scale convGRU module in [42] to combine feature details between different levels, and can also implement feature selection to retain more useful information. Inspired by the above work, we use the "memory" mechanism of GRU and introduce it into the task of Pansharpening of remote sensing image, establishing connections between features of different scales and features at different levels, and eliminating unnecessary noise in the process of feature
extraction and information redundancy while retaining important information. The convGRU structure we use is shown in Fig 1. Referring to the GRU formula in [39], similar to [41], we use 3*3 convolution to replace the original fully connected layer. Xt-1 is the input of the current ConvGRU, Ht−1 is the hidden layer state and the output of the previous ConvGRU, Ht is the hidden layer state and the output of the current ConvGRU. The yellow sigmoid function simulates the reset gate in the original GRU, and the blue sigmoid function simulates the update gate in the original GRU.

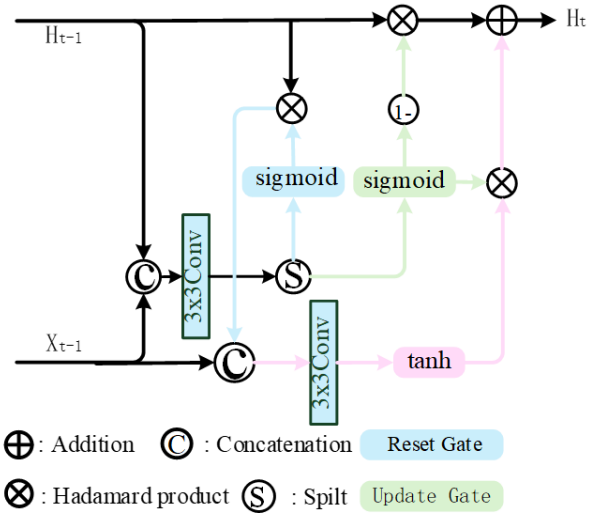

Fig.1 The structure of ConvGRU

PROPOSED METHOD

A. Overall Framework

The overall framework of our proposed MFMIF-Net is illustrated in Fig. 2(a). As we can see, MFMIF-Net completes the pan-sharpening process of remote sensing images in two stages. The processes of the two stages are similar, and we will illustrate using the first stage as an example. Initially, the original PAN image undergoes downsampling through the Down-Block, obtaining PAN↓2 with half the spatial resolution of the original PAN image. Similarly, the original LRMS image undergoes upsampling through the UP-Block, obtaining LRMS↑2 with twice the spatial resolution of the original LRMS image. Subsequently, PAN↓2 and LRMS↑2 are fed into the Progressive Feature Fusion Block (PFFB) to generate HRMS↑2 with twice the spatial resolution of the original LRMS. In the PFFB, we utilize the Multi-Scale Memory Interaction Fusion Block (MMIFB) to extract and interact different scale features to achieve more refined feature selection. We employ the [38] Dual Attention Unit (DAU) to suppress the transmission of irrelevant information in various features and enhance the further transmission of important features. Simultaneously, to preserve and propagate features of different depths, we utilize ConvGRU to retain features of different depths and propagate them to deeper levels. To avoid the issues of gradient vanishing and feature loss with the increase in network depth, we divide the process into multiple stages (subsequent experiments prove to set it as 3, detailed structure will be introduced in III.B) and progressively fuse features from the DUA branch, MMIFB branch, and ConvGRU branch. Finally, shallow features and progressively fused features are input into the Adaptive Feature Fusion Block (AFFB) to generate HRMS. The

aforementioned process can be represented by the following formula:

$$HRMS\uparrow 2 = F_{PFFB}(PAN\downarrow 2 + LRMS\uparrow 2)$$

$$HRMS = F_{PFFB}(PAN + (HRMS\uparrow 2)\uparrow 2)$$

F_{PFFB} representing the operation of the Progressive Feature Fusion Block. HRMS stands for the image after pan-sharpening.

B. Progressive Feature Fusion Block

In deep neural networks, shallow features pay more attention to the detailed information in the image, while deep features focus more on the semantic information

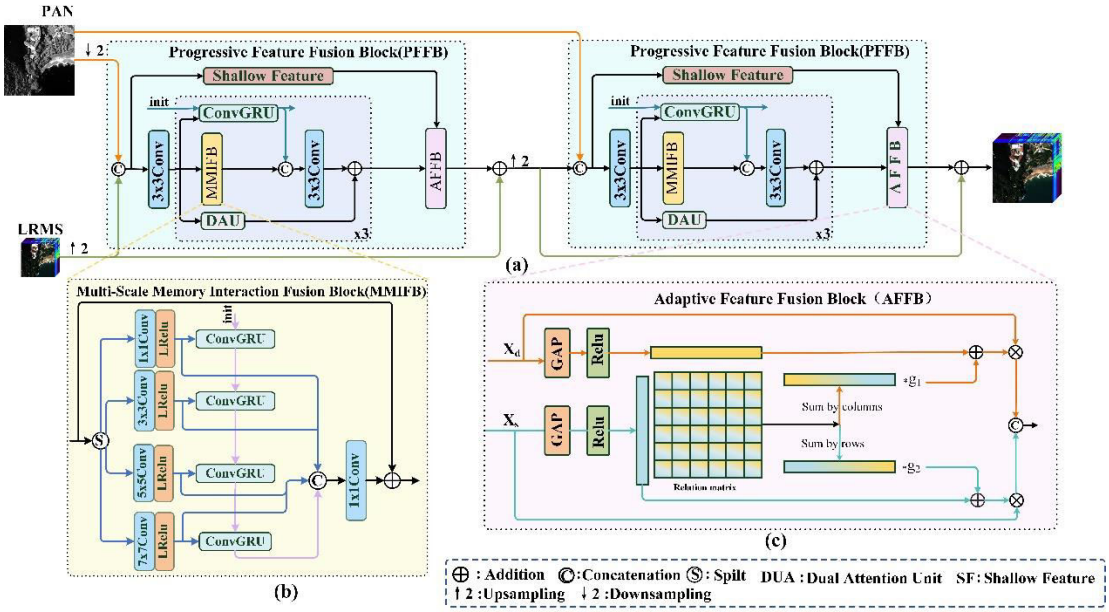

Fig2 The architecture of MFMIF-NET

[35]. Semantic information is advantageous for more accurate image restoration and edge recovery [36,37]. However, as the depth of neural networks increases, although performance can be improved, shallow information is also lost. Although networks using dense connection structures can solve this problem, they also come with a significant increase in computational complexity, which adversely affects training. To address these issues, we propose a novel Progressive Feature Fusion Block (PFFB). This module consists of a Multi-Scale Memory Interaction Fusion Block (MMIFB), an Adaptive Feature Fusion Block (AFFB), a Dual Attention Unit (DAU) [38], and a series of concatenated ConvGRUs.

To fully utilize shallow and deep features in the Adaptive Feature Fusion Block, we employ a dual-branch structure to extract these shallow and deep features. To suppress the transmission of irrelevant information and enhance the further propagation of important features, we use the Dual Attention Unit (DAU) from [38], which simultaneously employs channel attention [47] and spatial attention [48] to enhance features. Inspired by [49], we believe that channel attention focuses more on spectral information, while spatial attention focuses more on spatial information within the features. To fully exploit features from different layers, we propose a novel progressive fusion structure, as shown in Fig. 2(a). We slightly modify the shallow feature extraction structure from [35] as our first branch to extract shallow features. In the second branch, we use three cascaded ConvGRUs to achieve interaction of different deep features, selectively choosing usable information from different depths and updating it in the hidden state feature to pass to the next layer. This way, we can retain shallow information while effectively extracting deep semantic information. To avoid the problems of gradient vanishing and feature loss as network depth increases, we progressively fuse features from the DUA branch, MMIFB branch, and ConvGRU branch in three stages (as will be explained in subsequent experiments). To keep the model complexity wit except for the initial input image, is set to 32.hin a reasonable level, the channel dimension of all features,

$$P\_LMS = cat(pan\uparrow, LMS)$$

$$Conv_1 = conv(P\_LMS)$$

$$H_1 = Convgru(\text{in}it, Conv1)$$

$$Out_1 = DUA(Conv_1) + conv(cat(MMIFB(Conv_1), H_1))$$

Pan $\downarrow$ represents either the down-sampled by a factor of two or the original PAN image. LMS represents the LRMS image up-sampled by a factor of two $\cdot$ H_1 represents the output result of the first ConvGRU and the subsequent hidden state. Out_1 represents the fusion result of the first stage. Repeated the above steps three times, we obtained the second-stage fusion result Out_2 and the third-stage fusion result Out_3 progressively. Finally, the shallow-layer feature XS from the first branch and the deep-layer feature Out_3 from the second branch are fed into the Adaptive Feature Fusion Block to generate HRMS.

C. Multi-Scale Memory Interaction Fusion Block

Deep learning methods have achieved good performance in image processing, but often use fixed convolution kernel designs, which limits the ability to capture different feature representations. In remote sensing images, different objects usually present diverse colors, textures, and shapes, and their distribution in the image is also different [46]. By using small-size convolution kernels, we can obtain more detailed object features, while large-size convolution kernels help obtain more global information [35]. However, the current multi-scale remote sensing pan-sharpening method has a problem, that is, it fails to fully establish the connection between features at different scales and ignores the differences between features of the same type and the similarities between different features. In order to solve the above problems, we propose an innovative multi-scale memory interactive fusion module. This module simultaneously captures the features of two remote sensing images under different receptive fields and uses convolution gated recurrent unit (GRU) convolution to establish the connection between features at different scales. In this way, we can effectively eliminate unnecessary noise and redundant information while extracting features of different scales. This design aims to better mine the details and global information of objects in remote sensing images, making the model more accurate when processing different scale features of the same object.

Our proposed Multi-Scale Memory Interaction Fusion Block is shown in Fig. 2(b). Inspired by [40, 41, 42], we utilize the "memory" mechanism of GRU and introduce ConvGRU to establish connections between features of different scales, eliminating unnecessary noise and information redundancy in the process of feature extraction, while retaining important Information. As shown in Figure 2(b), four convolution operations with different convolution kernel sizes are used to extract the features of the input features at different scales. This operation fully considers the response of different scale features to convolution kernels of different sizes. Some previous methods did not fully consider the connection between features at different scales and ignored the differences between the same objects and the similarities between different objects.For this reason, we use convGRU to establish connections between features at different scales and fully extract the same features.Difference information between objects and similar information between different objects to achieve more refined feature selection, while making up for the lack of global information modeling capabilities of convolution operations.The difference from [41, 42] is that we use feature sequences extracted at different scales to replace the temporal data sequence of the original GRU as the input before the convGRU module, and pass the information processed by ConvGRU as a hidden state to the next layer(The initial hidden state is assigned all 0 values).

D.Adaptive Feature Fusion Block

In deep neural networks, [35] deep features contain rich semantic information, and shallow features contain rich detailed information.Deep features are extracted from shallow features, and there is a certain relationship between them. How to adaptively associate shallow features and deep features and make full use of both features to reconstruct the target image.The adaptive feature fusion module adaptively selects important features based on the characteristics of different features themselves, retaining both detailed information in shallow features and semantic information in deep features. As shown in Fig 2(c), in the adaptive feature fusion module, the use of global average

pooling (GAP) allows features of different depths to be calculated in the same dimension.The correlation matrix is then calculated through matrix multiplication, which captures the relationship between deep and shallow features. We use X_d and X_s to represent deep features containing semantic information and shallow features containing rich detailed information respectively.The feature maps of Xd and Xs can be expressed as v_d and $v_s \in RC$ respectively.The calculation process of feature mapping is as follows:

$$v_d = \mathrm{Re}lu(GAP(X_d))$$

$$v_s = \mathrm{Re}lu(GAP(X_s))$$

The correlation matrix M between Xd and Xs can be expressed as:

$$M = v_d v_s^T$$

Then add the correlation matrices by rows and columns to get two one-dimensional correlation vectors $v_d^{'}$ and $v_s^{'}$. The element values in $v_d^{'}$ represent the correlation between deep features and all shallow features in each channel dimension. Similarly, the element values in $v_s^{'}$ represent the correlation between shallow features and all deep features in each channel dimension. In order to retain the original features while adaptively enhancing the features, we use two learnable weight coefficients g_1 and g_2 add the correlation vectors $v_d^{'}$ and $v_s^{'}$ to v_d and v_s to obtain the enhanced feature map. Then the enhanced deep features and shallow features are obtained through the following operations:

$$R_d = (v_d + g_1 * v_d^{'}) \odot X_d$$

$$R_s = (v_s + g_2 * v_s^{'}) \odot X_s$$

R_d and R_s are enhanced deep features and shallow features respectively. $\odot$ represents the Hadamard product. Finally, the calculation process of fusing features to obtain the target image is as follows:

$$HRMS = conv(\mathrm{Re}lu(conv(\mathrm{Re}lu(conv(cat(R_d, R_s))))))$$

E.Loss Function

Because our network completes the remote sensing image pan-sharpening task in two stages, the total loss function is defined as follows:

$$L = \sum_{k=1}^{K} || HRMS - GT ||_2 + \lambda || HRMS\uparrow 2 - GT\downarrow 2 ||_2$$

where GT represents the ground truth reference image, and represents the down-sampled reference image by a factor of two. and the total number of training images is K,λ is 4。we consider the L2 loss for training, which is implemented on the PyTorch framework and NVIDIA GeForce RTX A6000 GPU. The number of epochs is set to 1500, and the batch size is 32. Moreover, we use the Adam optimizer with a learning rate of 0.0002. The parameters in the optimizer, beta1 and beta2, are set as values of 0.9 and 0.999, respectively.

IV EXPERIMENTAL RESULTS

In this section, the performance of the proposed MFMIF-Net is assessed on datasets from the GaoFen-2 ,WorldView-3 and Quickbird satellites[57]. To compare, we employ state-of-the-art methods, including both traditional approaches and （Deep Neural Network） DNN-based methods.The former methods are variational optimization-based VO-TV[50], the generalized Laplacian pyramid with MTF-matched filters with an FS regression-based injectionmodel (MTF-GLP-FS)[51], the generalized Laplacian pyramid with MTF-matched filters and a high-pass modulation injection model with a preliminary regression-based spectral matching phase (MTF-GLP-HPM-R) [52], the band-dependent spatial detail with physical constraints (BDSD-PC) [53].The five DNN-based methods are CNN-based PanNet[23], FusionNet[29], ADKNet[54], BiMPAN [55] and PSCF-NET[56]. To ensure the fairness of comparisons, we train all the DNN-based methods using the same PyTorch framework and NVIDIA GeForce RTX A6000 GPU.According to their performance on the validation dataset, the best model is selected for testing.

A.Experimental Settings

Datasets:To demonstrate the effectiveness of our MFMIF-Net, we conduct experiments using datasets from the GaoFen-2[57], the spatial resolutions of PAN images in GaoFen-2 is 0.8m. its corresponding MS images is 3.2. Moreover, the MS image from the GaoFen-2 satellite consists of four spectral bands, including red, green, blue, and near infrared.

Evaluation Metrics:To quantitatively assess the rationality and superiority of the proposed method, we introduce several reduced-resolution evaluation metrics and full-resolution evaluation metrics in our experiment. For the reduced resolution, the evaluation metrics the spatial correlation coefficient (SCC)[58], the spectral angle mapper (SAM)[59], the relative dimension less global error in synthesis (ERGAS)[60], relative average spectral error (RASE)[61], the structural similarity index(SSIM)[62], the peak signal-to-noise ratio (PSNR).SSIM and SCC are more effective in measuring the spatial similarity of the results while SAM focuses on distinguishing the spectral differences, PSNR and ERGAS refer to spectral and spatial errors to evaluate the model performance. RMSE is a pixel level reflection of the difference between the fusion results and the reference image.

For the full-resolution experiments, we use three of the more popular metrics to evaluate the performance of the model on real images, including the no-reference evaluation metric quality without reference(QNR), the spectral distortion($D\lambda$), the spatial distortion index(Ds). QNR it combines spectral distortion and spatial distortion. QNR includes evaluation metrics Dλ and DS for spectral and spatial distortion, respectively. Past studies have shown that $D\lambda$, Ds and QNR can roughly reflect the panchromatic sharpening performance at full resolution. A large QNR indicates better quality, while a small $D\lambda$ and Ds indicate less distortion.

B.Experiments on Reduced-Resolution Datasets

In this section, we perform Reduced-Resolution experiments on the GaoFen-2 ,WorldView-3 and Quickbird datasets. HRMS and GT images are covered in this section with a size of 256 × 256 to facilitate visual presentation. All metrics are calculated from the average of the 20 samples in the test set.

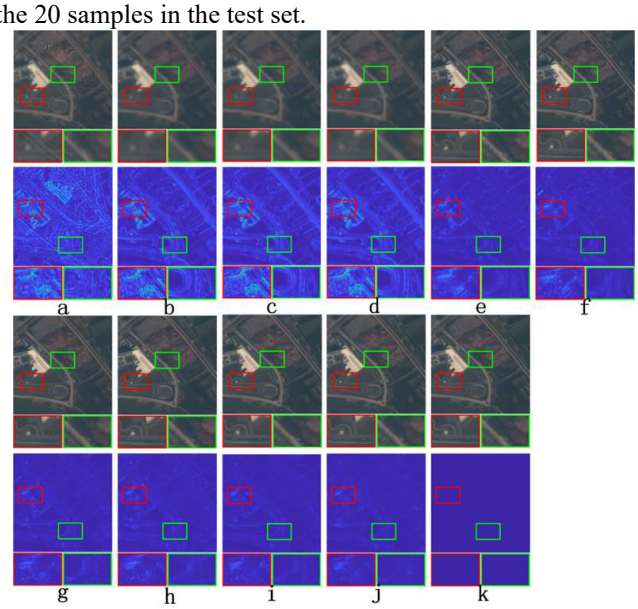

Fig 3. Visual comparisons on a reduced-resolution GaoFen-2 dataset.(a)VO-TV.(b) MTF-GLP-FS.(c) MTF-GLP-HPM-R.(d) BDSD-PC.(e)PanNet.[f]FusionNet.(g) ADKNet.(h)BiMPAN.(i)PSCF-NET[56].(j)OURS.(k)GT

As shown in Fig. 3, all these methods show excellent fusion performance on the GaoFen-2 dataset, in order to better compare the different methods, we use the error map to measure the difference between each image and GT, the brighter the brightness of the error map indicates that the fused image is more different from GT, and the artifacts in VO-TV, and the subjective visual maps of MTF-GLP-FS and MTF-GLP-HPM-R are blurred.darker the brightness of the error map indicates that the fused image is less different from GT. In terms of subjective visual maps,

traditional methods do not perform very well compared to DNN methods, e.g., there are obvious. It is clear from the error map that the DNN method has less error compared to the traditional method, indicating that the DNN method has better fusion accuracy. In DNN methods, it can be seen from the partial zoomed-in map of error map (red and green boxes in the error map) that the brightness of the error map of our proposed method is low, and it can also be seen from the partial zoomed-in map of the fused image that our proposed method has the closest spatial and spectral information to GT, which indicates that our proposed method achieves the best performance in terms of the subjective effect.

In Table I all the evaluation metrics of our proposed method outperform the other comparative methods, which is consistent with the performance of the subjective results, especially the PSNR is far more than the other methods. Our proposed method has the largest SSIM and SCC values and the smallest SAM value, which is indicated that our method performs well in spatial and spectral information reconstruction. In addition ADKNet, BiMPAN and SCF-NET also have better evaluation indexes.

TABLE I
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE REDUCED-RESOLUTION GAOFEN-2 DATASET

Methods	SCC ↑	SAM ↓	ERGAS ↓	RASE ↓	SSIM ↑	PSNR ↑
VO-TV[50]	0.8486	1.9106	1.7371	0.0334	0.7596	29.9405
MTF_GLP_FS[51]	0.8520	1.6578	1.5994	0.0263	0.8144	31.8641
MTF_GLP_HPM_R[52]	0.8522	1.6526	1.5986	0.0264	0.8139	31.8395
BDSD_PC[53]	0.8526	1.6763	1.6505	0.0259	0.8241	31.9999
PanNet[23]	0.9699	1.1217	1.2001	0.0129	0.9534	38.0035
FusionNET[29]	0.9797	0.9687	0.9558	0.0104	0.9656	39.9475
ADKNet[54]	0.9860	0.8506	0.7643	0.0084	0.9751	41.7222
BiMPAN[55]	0.9853	0.9504	0.8530	0.0092	0.9724	40.8555
PSCF-NET[56]	0.9900	0.7421	0.6705	0.0072	0.9813	43.0210
OURS	**0.9913**	**0.7039**	**0.6214**	**0.0067**	**0.9833**	**43.6364**

C.Experiments on Full-Resolution Datasets

The reduced resolution experiments mainly proved the effectiveness of our method on simulated data, and in order to verify the effectiveness of our method on real data, in this section, we will compare the various methods on the real dataset of GaoFen-2, where the size of the PAN image is 512×512 and the size of the LRMS is 128×128. Table II and Fig 4 give the objective evaluation results and visual results. In Table IV, the evaluation metrics Dλ and Ds clearly show that the DNN-based methods are superior to the traditional methods, which indicates that the DNN-based methods are able to recover the spatial and spectral information of the images better. Our proposed method is better than others in both spatial and spectral evaluation metrics. In Fig 4, we can clearly see that the visual results of the traditional method are worse in the recovery of spatial details and spectral information. the DNN-based methods outperform the traditional method in the recovery of both spatial detail and spectral information. Due to the lack of GT in real images, we compare our method with recent methods, and we can see that the visual results of our proposed method are very close to those of BiMPAN and PSCF-NET [56], and even in the recovery of spatial detail information our method is better than that of PSCF-NET [56].From the partial enlargement section of the visual results (green box), we can see that in the Figure square part, our method contains more texture information than PSCF-NET [56]. Overall, our method still performs well on the real dataset.

TABLE II
QUANTITATIVE METRICS FOR ALL THE COMPARISON METHODS ON THE FULL-RESOLUTION GaoFen-2 DATASET

Methods	QNR↑	D_λ↓	D_s↓
VO-TV[50]	0.8052	0.0144	0.1831
MTF_GLP_FS[51]	0.8173	0.0139	0.1711
MTF_GLP_HPM_R[52]	0.8145	0.0142	0.1738
BDSD_PC[53]	0.8378	0.0154	0.1491
PanNet[23]	0.9681	0.0187	0.0134
FusionNET[29]	0.9574	0.0289	0.0141
ADKNet[54]	0.9688	0.0180	0.0135
BiMPAN[55]	0.9721	0.0138	0.0142
PSCF-NET[56]	0.9722	0.0147	0.0132
OURS	**0.9751**	**0.0123**	**0.0127**

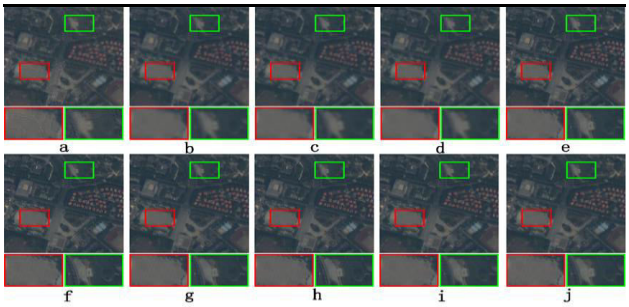

Fig 4. Visual comparisons on a full-resolution GaoFen-2 case. (a)VO-TV.(b) MTF-GLP-FS.(c) MTF-GLP-HPM-R.(d) BDSD-PC.(e)PanNet.[f]FusionNet.(g) ADKNet. (h)BiMPAN.(i)PSCF-NET[56].(j)OURS

Ⅴ.CONCLUSION

In this study, we proposed a MFMIF-Net for Pansharpening, hich achieves state-of-the-art performance, both quantitatively and qualitatively. Our proposed method is divided into two stages, first to achieve Pansharpening with 2 times the original resolution, and then to achieve Pansharpening with 4 times the original resolution. Each stage consists a Progressive Feature Fusion Block In this network we propose a Multi-Scale Memory Interaction Fusion Block, with the help of ConvGRU to interact between different scales features, to make full use of different scales features, and then fuse the shallow features to extract the different depth features in an progressive way, and at the same time to establish the connection between the different depth features. finally to use the Adaptive Feature Fusion Block to adaptively fuse the shallow features with the deep features to generate the sharpened HRMS. After a large number of experiments, it is proved that our method has advanced performance.Although the proposed method brings promising results, some notable issues remain and call for further research. We only consider the relationship between different scales and different depth features, but not the correlation between the two image-specific features, LRMS and PAN. In later work, we will investigate the correlation between the two image-specific features to better reconstruct HRMS. and improve the possibility of practical application of MFMIF-Net.

REFERENCES

[1] B. Zhang, D. Wu, L. Zhang, Q. Jiao, Q. Li, Application of hyperspectral remote sensing for environment monitoring in mining areas, Environ. Earth Sci. 65 (3) (2012) 649–658.

[2] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (2021) 323–336.

[3] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, “Image fusion meets deep learning: A survey and perspective,” Inf. Fusion, vol. 76, pp. 323–336, Dec. 2021.

[4] H. Ghassemian, “A review of remote sensing image fusion methods,” Inf. Fusion, vol. 32, pp. 75–89, Nov. 2016.

[5] C. Thomas, T. Ranchin, L. Wald, J. Chanussot, Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics, IEEE Trans. Geosci. Remote Sens. 46 (5) (2008) 1301–1312.

[6] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion32 (2016) 75–89.

[7] C.S. Yilmaz, V. Yilmaz, O. Gungor, A theoretical and practical survey of image fusion methods for multispectral pansharpening, Inf. Fusion 79 (2022) 1–43.

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

[9] L. Wenzhi et al., “Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS data fusion contest,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 8,no. 6, pp. 2984–2996, Jun. 2015.

[10] V. K. Shettigara, “A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolutiondata set,” Photogram. Eng. Remote Sens., vol. 58, no. 5, pp. 561–567,May 1992.

[11] 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.

[12] 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, doi:10.1080/014311600750037499.

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

[14] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, “MTF-tailored multiscale fusion of high-resolution MS and PAN imagery,”Photogramm. Eng. Remote Sens., vol. 72, no. 5, pp. 591–596, May 2015.

[15] ang F M, Li F, Shen C M and Zhang G X. 2013. A variational approach for pan-sharpening. IEEE Transactions on Image Processing, 22(7): 2822-2834,doi: 10.1109/TIP.2013.2258355

[16] Buades A, Coll B, Duran J and Sbert C. 2014. Implementation of nonlocal pansharpening image fusion. Image Processing On Line, 4: 1-15,doi: 10.5201/ipol.2014.98

[17] Palsson F, Sveinsson J R and Ulfarsson M O. 2014. A new pansharpening algorithm based on total variation. IEEE Geoscience and Remote Sensing Letters, 11(1): 318-322,doi: 10.1109/LGRS.2013.2257669

[18] M. Zhang, K. Pang, C. Gao, and M. Xin, “Multi-scale aerial target detection based on densely connected inception ResNet,” IEEE Access,vol. 8, pp. 84867-84878, 2020.

[19] M. Z. Uddin, M. M. Hassan, A. Alsanad, and C. Savaglio, “A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare,” Inf. Fusion, vol. 55,pp. 105–115, Mar. 2020.

[20] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, “Learning texture transformer network for image super-resolution,” in Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 5791–5800.

[21] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. Eur. Conf. Comput. Vis.Berlin, Germany: Springer, 2014, pp. 184–199.

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

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

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

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

[26] Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., vol. 27, 2014, pp. 2672–2680.

[27] S. Xu, J. Zhang, Z. Zhao, K. Sun, J. Liu, and C. Zhang, “Deep gradient projection networks for pan-sharpening,” in Proc. IEEE/CVF Conf.Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1366–1375.

[28] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, “Deep multiscale detail networks for multiband spectral image sharpening,” IEEE Trans.Neural Netw. Learn. Syst., vol. 32, no. 5, pp. 2090–2104, May 2021.

[29] L. J. Deng, F. Vivone, C. Jin, and J. Chanussot, “Detail injection-based deep convolutional neural networks for pansharpening,” IEEE Trans. Geosci. Remote Sens., 2020, doi: 10.1109/TGRS.2020.3031366.

[30] Zhang K, Wang A, Zhang F, et al. Spatial-Spectral Dual Back-Projection Network for Pansharpening[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023.

[31] Lei, Dajiang, et al. “MHANet: A Multiscale Hierarchical Pansharpening Method With Adaptive Optimization.” *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-15.

[32] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprintarXiv:2010.11929, 2020. 1, 3

[33] S.T. Rajamani, K.T. Rajamani, A. Mallol-Ragolta,S. Liu, and B. Schuller, “A novel attention-based gated recurrent unit and

its efficacy in speech emotion recognition," in Proceedings ofICASSP, 2021, pp. 6294–6298.
[34] Y. Li, Y. Yang, K. Zhu, and J. Zhang, "Clothing sale forecasting by a composite gru–prophet model with anattention mechanism," IEEE Transactions on Industrial Informatics, vol. 17, no. 12, pp. 8335–8344, 2021.
[35] Zhang, H., Wang, H., Tian, X., & Ma, J. (2023). P2Sharpen: A progressive pansharpening network with deep spectral transformation. *Information Fusion*, *91*, 103-122.
[36] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.
[37] Nazeri, Kamyar, et al. "Edgeconnect: Generative image inpainting with adversarial edge learning." *arxiv preprint arxiv:1901.00212* (2019).
[38] Zamir, Syed Waqas, et al. "Learning enriched features for real image restoration and enhancement." *Computer Vision–ECCV 2020: 16th European Conference.*
[39] Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).
[40] [Shi X, Chen Z, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[J]. Advances in neural information processing systems, 2015, 28.
[41] F. Yuan, L. Zhang, X. Xia, Q. Huang and X. Li, "A Gated Recurrent Network With Dual Classification Assistance for Smoke Semantic Segmentation," in IEEE Transactions on Image Processing, vol. 30, pp. 4409-4422, 2021, doi: 10.1109/TIP.2021.3069318.
[42] Wang, Qidong, et al. "SFEMGN: Image Denoising with Shallow Feature Enhancement Network and Multi-Scale ConvGRU." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.
[43] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 9,pp. 5443–5457, Sep. 2018.
[44] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," IEEE Trans.Geosci. Remote Sens., vol. 59, no. 8, pp. 6995–7010, Aug. 2021.
[45] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.,vol. 11, no. 3, pp. 978–989, Mar. 2018.
[46] Fan, Junyu, et al. "Frequency-aware robust multidimensional information fusion framework for remote sensing image segmentation." *Engineering Applications of Artificial Intelligence* 129 (2024): 107638.
[47] Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[48] Woo, Sanghyun, et al. "Cbam: Convolutional block attention module." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
[49] Gong, Meiqi, et al. "D2TNet: A ConvLSTM network with dual-direction transfer for pan-sharpening." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-14.
[50] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation, " IEEE Geosci. Remote Sens. Lett., vol. 11, no. 1, pp. 318±322, Jan. 2014.
[51] 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.
[52] G. Vivone, R. Restaino, and J. Chanussot, "A regression-based high-pass modulation pansharpening approach," IEEE Trans. Geosci. Remote Sens., vol. 56, no. 2, pp. 984–996, Feb. 2018.
[53] 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.
[54] Peng, Siran, et al. "Source-adaptive discriminative kernels based network for remote sensing pansharpening." *Proc. 31st Int. Joint Conf. Artif. Intell.*. 2022.
[55] J. Hou, Q. Cao, R. Ran, C. Liu, J. Li, and L.-J. Deng, "Bidomain modeling paradigm for pansharpening," in Proc. 31st ACM Int. Conf. Multimedia, Oct. 2023, pp. 347–357.
[56] Peng, Siyuan, et al. "PSCF-Net: Deeply coupled feedback network for pansharpening."*IEEE Transactions on Geoscience and Remote Sensing* 61 (2023): 1-12.
[57] L. 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.
[58] 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.
[59] 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, vol. 1, 1992, pp. 147–149.
[60] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?" in Proc. 3rd Conf. Fusion Earth Data, Merging Point Meas., Raster Maps Remotely Sensed Images, 2000, pp. 99–103.
[61] M. Choi, A new intensity-hue-saturation fusion approach to image fusion with a tradeoff parameter, IEEE Trans. Geosci. Remote Sens. 44 (6) (2006) 1672–1682.
[62] 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.
[63] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, M. Selva, Multispectral and panchromatic data fusion assessment without reference, Photogramm. Eng. Remote Sens. 74 (2) (2008) 193–200.

Yu Xin Postgraduate student. His research interests include computer vision and Pansharpening.

Wu Zheng Postgraduate student. His research interests include computer vision.