AWFLN: An Adaptive Weighted Feature Learning Network for Pansharpening

Hangyuan Lu, Yong Yang, *Senior Member, IEEE,* Shuying Huang, *Member, IEEE,* Xiaolong Chen, Biwei Chi, Aizhu Liu, Wei Tu

***Abstract*—Deep learning (DL)-based pansharpening methods have shown great advantages in extracting spectral-spatial features from multispectral (MS) and panchromatic (PAN) images compared with traditional methods. However, most DL-based methods ignore the local inner connection between the source images and high-resolution MS (HRMS) image, which cannot fully extract spectral-spatial information, and attempt to improve the quality of fusion by increasing the complexity of the network. To solve these problems, a lightweight network based on adaptive weighted feature learning (AWFLN) is proposed for pansharpening. Specifically, a novel detail extraction model is first built by exploring the local relationship between HRMS and source images, thereby improving the accuracy of details and the interpretability of the network. Guided by this model, we then design a residual multiple receptive-field structure to fully extract spectral-spatial features of source images. In this structure, an adaptive feature learning block based on spectral-spatial interleaving attention is proposed to adaptively learn the weights of features and improve the accuracy of the extracted details. Finally, the pansharpened result is obtained by a detail injection model in AWFLN. Numerous experiments are carried out to validate the effectiveness of the proposed method. Compared to traditional and state-of-the-art methods, AWFLN performs the best both subjectively and objectively, with high efficiency. The code is available at https://github.com/yotick/AWFLN.**

***Index Terms*—Pansharpening, lightweight network, adaptive feature learning, spectral-spatial fidelity priori.**

I. Introduction

With the widespread application of remote sensing images, the demand for high spatial-resolution multispectral (HRMS) images is increasing. Many tasks, such as disaster warning, geological exploration, and land object classification, require HRMS images with not only precise spectral information but also high spatial resolution [1]–[3]. However, owing to the limitations of satellite storage technology and image transmission technology, it is difficult to directly provide an HRMS image [4]. Only panchromatic (PAN) and multispectral (MS) images of the same scene are provided by satellites. In general, PAN images have high spatial resolution but little spectral information, whereas MS images have precise spectral information but low spatial resolution. To fulfill the applications' requirements, these two types of images should be fused into one image with both high spatial resolution and accurate spectral information [5], [6]. This fusion process is called remote sensing image fusion or pansharpening [7].

To date, many pansharpening techniques have been presented. Component substitution (CS)- and multiple resolution analysis (MRA)-based methods are two categories of traditional algorithms [8], [9]. To obtain the intensity (I) component in CS-based methods, the MS image is first projected into a new feature space. To increase the spatial resolution of the MS image, the I component is then substituted with a PAN image. The fused image is finally obtained by the inverse transformation. Representative algorithms include the intensity–hue–saturation transform (IHS) [10], Gram–Schmidt adaptive (GSA) [11], and band-dependent spatial detail (BDSD) [12]. To improve generality, the robust band-dependent spatial detail (RBDSD) approach was proposed [13], thus extending the method to the case of eight-channel images. Generally, CS-based methods can achieve high spatial resolution with high efficiency. However, owing to the variation in the low-frequency information in the MS image, it easily causes spectral distortion [14], [15].

In contrast to CS-based techniques, MRA-based techniques mainly extract the PAN image's details using multi-scale techniques and then inject those details into the original MS image to improve spatial resolution. Popular algorithms include wavelet transform [16], Laplacian pyramid [17], and contourlet transform [18]. Vivone *et al.* [19] revised the additive wavelet luminance proportional (AWLP-R) method to achieve fast and high performance, especially in the spectral aspect. MRA-based methods are more effective at addressing the issue of spectral distortion than CS-based methods. However, the effect of this kind of method is easily affected by the image registration results and is prone to producing edge artifacts [7], [20].

Another group of pansharpening methods is the variational optimization (VO). This kind of method mainly constructs the energy function through some assumptions and obtains the final fusion results by minimizing the energy function. Ballester *et al.* [21] proposed the first variational model, named P+XS, by assuming that the PAN image contains

This work is supported by the National Natural Science Foundation of China (No.62072218 and No.61862030), by the Natural Science Foundation of Zhejiang Province (No. LY22F020017), and by the Talent project of Jiangxi Thousand Talents Program (No. jxsq2019201056). (*Hangyuan Lu and Shuying Huang contributed equally to this work.*) (*Corresponding author: Yong Yang.*)

H. Lu, X. Chen, B. Chi, A. Liu are with the College of Information Engineering, and also with Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321007, China (e-mail: lhyhziee@163.com; xlycxl@126.com; chibiwei123@163.com; 20211003@jhc.edu.cn).

Y. Yang is with the School of Computer Science and Technology, Tiangong University, Tianjin 300387, China (e-mail: greatyangy@126.com).

S. Huang is with the School of Software, Tiangong University, Tianjin 300387, China (e-mail: shuyinghuang2010@ 126.com).

W. Tu is with the School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University, Nanchang 330038, China (e-mail: ncsytuwei@163.com).

the geometric information of the HRMS image. Since then, several VO models have been developed, such as Bayesian theory-based [22], sparse representing-based [23]–[25], and model-based methods [4], [26]. Yang *et al.* [3] proposed a pansharpening model based on conditional random fields (PCRF), which can well balance performance and efficiency. Xiao *et al.*. [27]further proposed a model based on context-aware detail injection fidelity (CDIF) and obtained impressive fusion results. These techniques can somewhat enhance the fusion quality. However, the settings of various parameters are uncertain, which may lead to inaccuracy of the constructed models [7], [28].

Many deep learning (DL)-based pansharpening techniques have emerged and have been widely used in recent years. Masi *et al.* [29] proposed a pansharpening method based on the convolutional neural network (PNN), which uses a simple super-resolution network with three layers. However, it regards the pansharpening as a black box, which lacks interpretability. Thus, Deng *et al.* [30] proposed FusionNet, which combines traditional CS- and MRA-based methods with the residue network. Further, Wu *et al.* [31] proposed VO+Net, which combines the PNN with a variational model. These methods can improve the interpretability and obtain better fusion results. However, they cannot fully extract the spatial features, which may lead to spatial distortion [15]. To improve the spatial quality according to the properties of the source images, some unsupervised pansharpening methods have been developed. For instance, Ma *et al.* [32] proposed unsupervised Pan-GAN by separately setting adversarial games with spectral and spatial decimators, Ozcelik *et al.* [33] considered the pansharpening process as the colorization of a PAN image and proposed PanColorGAN, and Qu *et al.* [34] proposed a self-attention mechanism based on a stick-breaking structure to extract the relevant details. These methods can achieve satisfactory spatial quality; however, they may easily cause spectral distortion due to the absence of a proper reference image. Some research has involved the construction of a multibranch network to fully utilize the spectral or spatial information. For example, Yang *et al.* [15] proposed a dual-stream convolutional neural network to enhance the spatial information, and Zhang *et al.* [35] proposed a triple-double network (TDNet) to progressively enhance the detail information. In general, the extraction of the enhanced information can achieve impressive spatial quality.

Existing DL-based pansharpening methods do not consider the inner relationship between the source images and the HRMS image, and usually rely on increasing the depth and complexity of the network, which may lead to an increased computational burden and redundant information [36]. Besides, there is still much room for the improvement of the interpretability and generalization of the network [27]. In addition, how to combine the prior knowledge with DL has always been a topic of interest. In fact, there is an intrinsic local relationship between the source images and HRMS images, and this relationship can be adaptively learned based on the spectral–spatial fidelity prior. For the spectral aspect, the HRMS image and MS image should be consistent in the channel dimension. For the spatial aspect, the details extracted from the source images should be consistent with those of HRMS image [4], [28]. According to the above analysis, this paper constructs a new detail injection model based on local spectral-spatial fidelity. Guided by this model, we propose an adaptive weighted feature learning network (AWFLN) with a lightweight structure, which obtains high-quality fusion results with good interpretability and generalization. The generalization in this work refers to the good performance on various datasets as well as on both reduced-scale (RS) and full-scale (FS) image fusion. The main contributions of this work are as follows.

(1) A novel detail injection model is designed by keeping local spectral-spatial fidelity between HRMS and source images, and a lightweight network called AWFLN is proposed based on the model, which has good interpretability and can obtain fusion results with high quality.

(2) Guided by the designed model, an adaptive feature learning block (AFLB) is proposed based on the spectral-spatial interleaving attention (SSIA) mechanism, which can better retain spectral information and obtain accurate details.

(3) A residual multiple receptive-field structure (RMRS) embedded with a multiscale convolution block (MCB) and AFLB is constructed to fully extract local features and obtain useful spectral–spatial information through the attention mechanism..

(4) The results of both the RS and FS experiments verify that the proposed AWFLN can achieve the best results in subjective and objective evaluations. The classification experiments further attest to the effectiveness of our methods in pansharpening application.

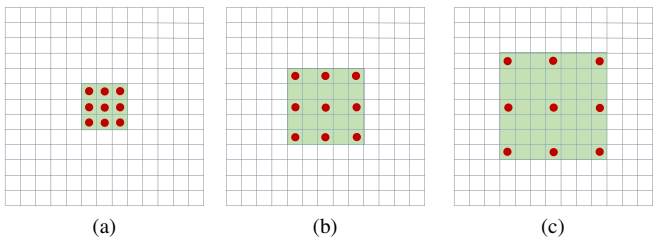
II. Related Work

A. Detail Injection Model

Both CS- and MRA-based methods can be considered as the detail injection approach [1], which can be expressed as follows:

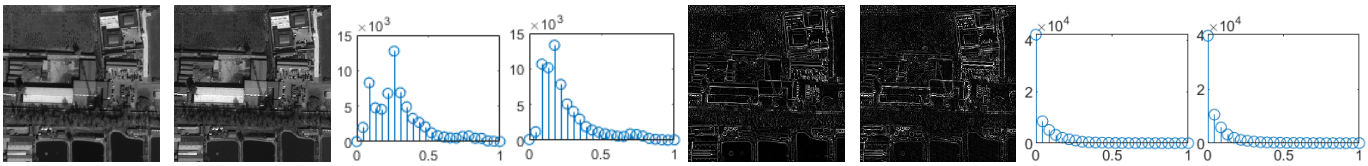
$$\widehat{M} = M_{\mathrm{up}} + GD \tag{1}$$

where $\widehat{M}$ and M_{up} represent the estimated HRMS and upsampled MS (UPMS) images, respectively. G represents the injection coefficient, and D denotes the extracted details. The difference between CS- and MRA-based approaches mainly lies in how to extract the details. For CS-based methods, D represents the difference between a PAN image (P) and the I component of an UPMS image (I), whereas for MRA-based method, D is expressed as the difference between P and the



(a) (b) (c)

Fig. 1. The dilation convolution with dilation rate of (a) 1, (b) 2, and (c) 3.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. Comparison of different bands in an HRMS image. (a) and (b) are R and NIR1bands, respectively, (c) and (d) are the corresponding histograms of (a) and (b), (e) and (f) are the corresponding detail information of (a) and (b), and (g) and (h) are the histograms of (e) and (f), respectively.

low-resolution version of P after applying a low-pass filter H on P. Thus, the detail extraction can be formulated as follows.

$$\begin{cases} D_{CS} = P - I \\ D_{MRA} = P - HP \end{cases} \tag{2}$$

where D_{CS} and D_{MRA} represent the details extracted by CS- and MRA-based methods, respectively.

The key to the detail injection model is how to obtain accurate details. Combing deep learning with detail injection model is a promising approach for extracting accurate details.

B. Dilation Convolution

In deep learning, down-sampling operation is usually conducted to increase the receptive field and reduce the computational complexity. Although the receptive field can be increased, the spatial resolution is reduced by this way. To expand the receptive field while maintaining spatial resolution, dilation convolution is carried out by injecting holes (0 value) into a standard convolution kernel [37]. Compared with ordinary convolution, dilation convolution has a hyper-parameter called dilation rate in addition to the size of convolution kernel. Dilation rate refers to the number of kernel intervals. Traditional convolution can be regarded as a dilation convolution with a dilation rate of 1. Fig. 1 displays the dilation convolutions when the dilation rate equals (a) 1, (b) 2, and (c) 3, respectively. The green rectangles are the dilated kernels and the red points represent nonzero weights. As dilation rate increases, so does the receptive field. In traditional convolution, the receptive field is linear with the number of layers, whereas in dilation convolution, the receptive field is exponential with the number of layers. Thus, the output of dilation convolution can obtain long-range information [38].

Although the dilation convolution can use a big dilation rate to obtain a much larger receptive field than traditional convolution, it may only be useful for detecting a large object and be poor at detecting a small object. Therefore, it is better to combine the traditional convolution with dilation convolution to obtain both local and long-range information.

III. Proposed Method

A. Model Construction

The detail injection model is popular because of its flexibility and interpretability, and the key to such a method is to extract accurate details so as to achieve a high level of spectral–spatial fidelity in an HRMS image. In this section, we build a detailed constraint model by exploring the spectral-spatial relationship between the source images and HRMS image. In a referenced HRMS image, there are diverse spatial structures in different channels, and there is an inner spatial relationship among the bands [3], [4]. To better illustrate the difference and connection of features between bands of HRMS image, we take an HRMS image in the WorldView-3 dataset as an example and compare the features of red (R) and near-infrared1 (NIR1) bands, as shown in Fig. 2. Fig. 2(a) and (b) show the R and NIR1 bands of the HRMS image, respectively; and Fig. 2(e) and (f) show the corresponding detail information of the R and NIR1 bands, respectively. The detail information is obtained by subtracting the UPMS image from the HRMS image. Fig. 2(c) and (d) display the histograms of (a) and (b), respectively; and Fig. 2(g) and (h) display the histograms of (e) and (f), respectively. From the histograms in the figure, we can see that the pixel distribution of NIR1 is different from that of R, but the detail distribution is similar. For the similarity of detail distribution, Lu *et al.* [4] further proposed the idea that the detail information of each band of an HRMS image is proportional to that of a PAN image. This proportional relationship can be expressed as an energy function, as follows:

$$E_D = \left\| CM_{\text{ref}}^b - w_b CP \right\| \tag{3}$$

where E_D represents a detail-related energy function, M_{ref} denotes the reference HRMS image, C represents a convolution operator for extracting details, b represents the b-th band in the HRMS image, and w_b represents the scale factor of the b-th band.

Because the PAN image and MS image are imaged by different sensors and there is a radiometric discrepancy between them, M_{ref} contains some spatial information that the PAN image lacks but is present in the MS image [20], [39]. Therefore, P in (3) does not contain all the detail features of the HRMS image. Besides, there are different features in local structures, and using a global scale factor is inappropriate for extracting accurate details. Furthermore, owing to the intrinsic correlation between the bands of the HRMS image, separating the bands as presented in model (3) may disrupt this connection. Taking these factors together, we mainly improve model (3) from three aspects. First, the part of the detail features not included in P can be compensated for by the detail features in M_{up}, and the compensated PAN image is expressed as P^*. Second, to improve the accuracy of detail extraction, we replace the global scale factor for each channel with multiple local factors. Third, to keep the structural integrity of the HRMS image, we treat all the bands of the HRMS image as a whole and construct its relationship with the PAN image. Thus, this paper defines a new energy function to express the

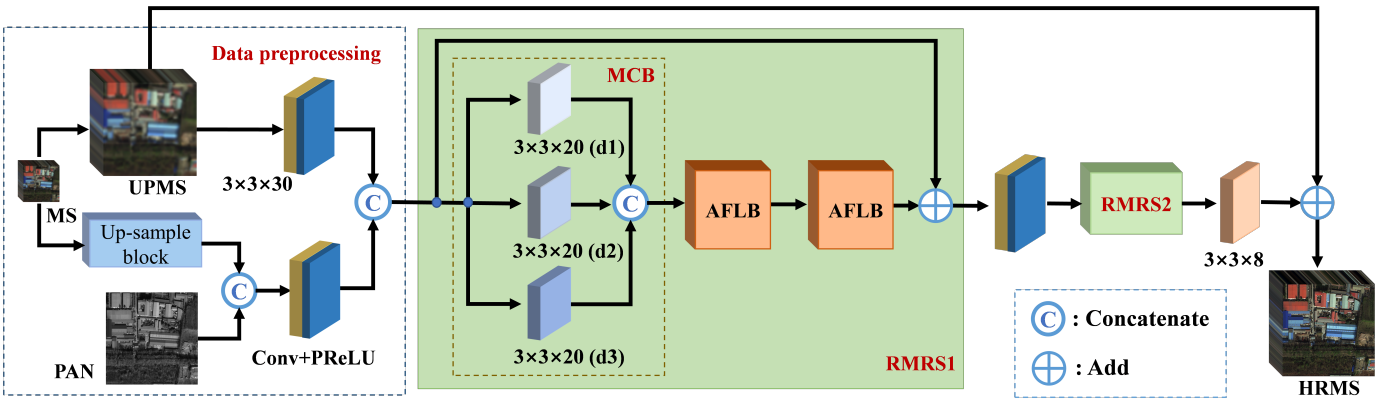


Fig. 3. Flowchart of the proposed AWFLN.

relationship between the local details of the HRMS and those of source images:

$$\widetilde{E}_D = \left\| CM_{\text{ref}}^{\Omega} - \omega_{\Omega} CP_{\Omega}^{*} \right\| \tag{4}$$

where Ω represents a local area, and ω_{Ω} represents a local weight.

In 4, ω_{Ω} plays an important role in extracting proper details, and two aspects need to be considered in the solution of ω_{Ω}. First, ω_{Ω} can be optimized by making the extracted details close to the reference details. Second, considering the requirement of dual fidelity of spectral–spatial features, the spectral information along the channel dimension and the local spatial structure should be considered when determining ω_{Ω}. Thus, ω_{Ω} is adaptively obtained by optimizing the formula defined as follows:

$$\omega_{\Omega} = \arg\min_{\omega_{\Omega}} \left\| D_{\Omega}^{ref} - \omega_{\Omega} CP_{\Omega}^{*} \right\| \quad s.t.\ \omega_{\Omega} = f(S_{\Omega}, \lambda_{\Omega}) \tag{5}$$

where $D_{\Omega}^{ref} = \nabla M_{\text{ref}}^{\Omega}$ represents the reference details; S_{Ω} and λ_{Ω} represent the local spectral-spatial information, respectively; and $f(\cdot)$ represents a function that integrates the extracted spectral–spatial information.

With ω_{Ω}, the estimated local details are obtained by $\widehat{D}_{\Omega} = \omega_{\Omega} CP_{\Omega}^{*}$, and the final HRMS image is obtained by the detail injection model, which is expressed as follows:

$$\widehat{M} = M_{\text{up}} + \widehat{D} \tag{6}$$

where $\widehat{D}$ represents the estimated details, which is a set of $\widehat{D}_{\Omega}$.

Based on the above analysis, we can infer that three issues need to be addressed for model (5) to obtain good fusion results. First, P^* needs to be determined for the precise extraction of details. Second, how to fully extract the spectral-spatial information from the source images needs to be solved. Third, how to integrate local spectral–spatial information when determining ω_{Ω} needs to be considered. By considering these three issues, we construct a deep network to achieve accurate extraction of detail information and obtain high-quality fusion results.

B. Overall Structure

For the three problems to be solved in the constructed model (5), an AFWLN based on SSIA is proposed, as shown in Fig. 3. For the solution of P^* in the above problems, a data preprocessing module is first constructed. Then, to fully extract spectral–spatial features, two RMRS modules combined with MCB and AFLBs are cascaded to adaptively learn the features and extract the details that are injected into the UPMS image. Next, the residual structure constructed by the jump link combines the extracted details with the UPMS image to output the fused image. The specific components of the AWFLN are elaborated on below.

C. Data Preprocessing

In the pansharpening task, extracting proper spectral–spatial features is critical for obtaining high-quality fusion results. Therefore, the first improvement of model (5) over model (3) is to construct a suitable P^* image that provides rich detail features for feature extraction in the later stage of the network. It can be known from the previous analysis that P^* is composed of the PAN image and some features from the MS image, so the MS image has two roles in the AFWLN: one is as a spectral benchmark [28], [40], and the other is to provide supplemental information for P^*. Therefore, the data preprocessing module is designed with two branches. First, the UPMS image up-sampled by the interpolation method passes through a 3×3 convolution layer and a PReLU activation function to form a spectral branch for further spectral feature extraction, which is expressed as follows:

$$f_S = PR(Conv(Inter(M))) \tag{7}$$

where M represents the original MS image, and f_S represents the extracted spectral features. $Inter(\cdot)$, $Conv(\cdot)$, and $PR(\cdot)$ represent an interpolation operation, a convolution layer, and a PReLU activation layer, respectively.

Then, another branch is constructed to supplement the spatial features for the PAN image with features from the MS image. The up-sampled MS image is obtained by constructing an up-sampled block to automatically extract spatial information rather than directly using the interpolation method, because the interpolation method may generate redundant

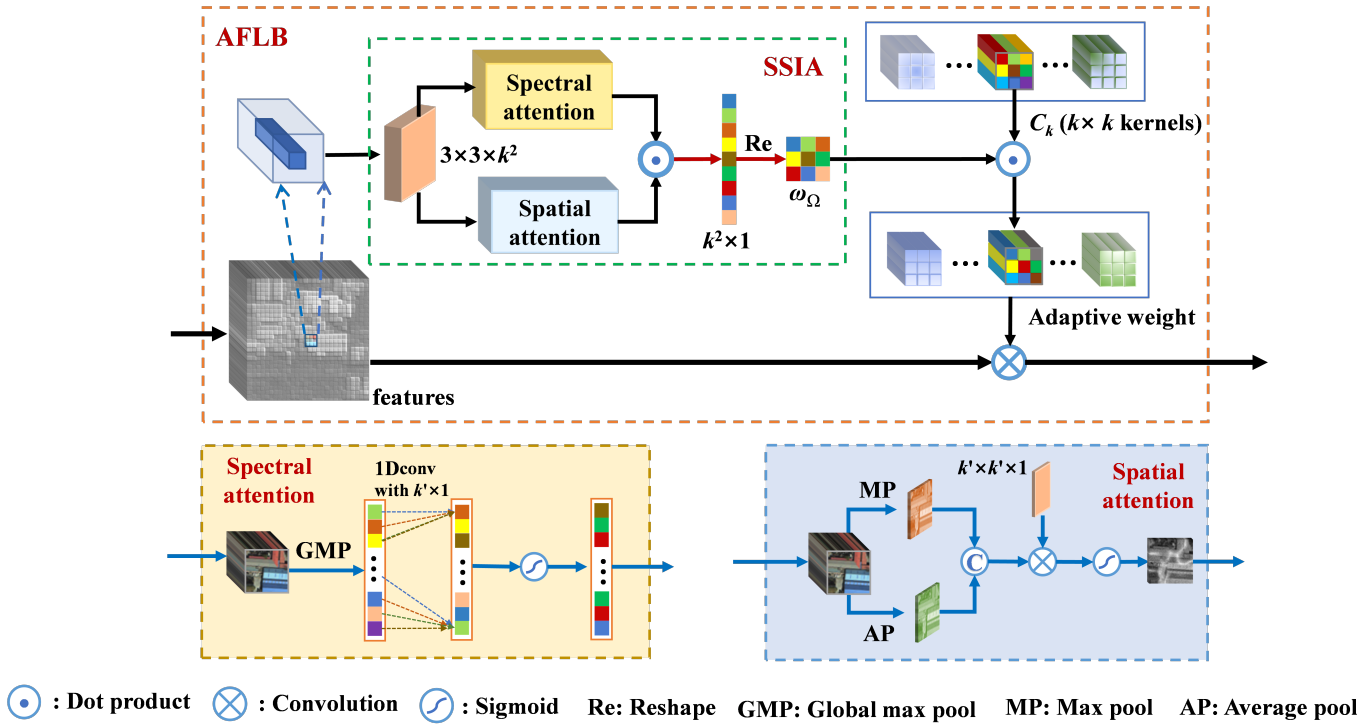


Fig. 4. Flowchart of the proposed AFLB.

features that affect the accuracy of detail extraction. The up-sample block contains a convolution layer, a PixelShuffle layer, and a PReLU activation function, which can be expressed as follows:

$$M_{\text{up}}^{*} = PR(PS(Conv(M))) \tag{8}$$

where M_{up}^{*} represents the up-sampled MS image through an up-sample block, and $PS(\cdot)$ represents a PixelShuffle layer.

Thus, the compensated PAN image P^{*} in the model (5) can be obtained by concatenating the learned M_{up}^{*} and the PAN image P. This process can be written as follows:

$$P^{*} = Cat(M_{\text{up}}^{*}, P) \tag{9}$$

where $Cat(\cdot)$ represents a concatenation operation. Next, P^{*} is sent to a convolution layer and a PReLU layer to form a spatial branch so as to obtain the initial spatial features, and the spatial branch is expressed as follows:

$$f_D = PR(Conv(P^{*})) \tag{10}$$

where f_D represents the spatial feature.

The feature maps from the spectral branch and the spatial branch are concatenated as the input of the first RMRS block (denoted as RMRS1), and two RMRS blocks are connected in series to gradually realize the refinement of spectral–spatial features.

D. RMRS Block

To address the second and third problems in model (5), an RMRS module is constructed. First, MCB is built to further extract features of different scales by using convolution kernels with different dilation rates. Specifically, d1, d2, and d3 in the MCB represent the dilation rates of 1, 2, and 3 for convolution operations, respectively. Then, an AFLB is designed to generate the adaptive local convolution kernels, and two AFLBs are concatenated to finely learn spectral–spatial features. The overall RMRS is designed as a residual structure to enhance the fitting ability.

In this paper, two cascaded RMRS blocks, RMRS1 and RMRS2, are set up to refine the spectral–spatial features. More RMRS blocks inevitably increase the computational burden of the network but may not necessarily improve the network performance. This is further verified in the ablation experiments. In addition, the RMRS1 and RMRS2 blocks are slightly different. We know that when the channel dimension of the input feature maps is much larger than that of the output feature maps, the direct dimensionality reduction operation leads to the loss of features in the feature integration process. To reduce the loss of features, the feature maps output by RMRS1 goes through a 3×3 convolutional layer before entering the RMRS2 block, and the channel dimension is reduced by half to balance the performance and efficiency of the network. Specifically, the number of channels of the feature maps received by RMRS1 is set to 60, and the number of channels of the feature maps received by RMRS2 is set to 30 by trial and error. As a result, the number of convolution kernels in each layer of the RMRS2 block is half that of the RMRS1 block.

In RMRS, AFLB is the core module for refined feature extraction, and its specific structure is described in detail below.

E. Structure of AFLB

The models (4) and (5) indicate that the estimated details should maintain the spectral–spatial fidelity of the reference details, and the task is to design a network representing $\omega_\Omega CP_\Omega^*$ and train the network to approximate the D_Ω^{ref} . Thus, AFLB is constructed to achieve this goal, and the flowchart is shown in Fig. 4. First, convolution kernels of size $k \times k$ (C_k) are designed to replace C in model (5) to extract features from different local regions, and their parameters are adaptively adjusted by the learned weights ω_Ω. Here, k is set to 3 to improve efficiency. Then, to ensure the integrity of the fused features, SSIA is constructed to generate the adaptive weights by interleaving the weight values learned by spectral attention and spatial attention. The specific structure of SSIA is as follows.

SSIA aims to learn a weight matrix ω_Ω that matches the convolution kernels of size $k \times k$. First, the feature maps received by AFLB go through a $3\times3\times k^2$ convolution layer to reduce the dimension of the feature maps and integrate channel features. Then, spectral attention and spatial attention are built to obtain channel and spatial weights, respectively. The spectral weights mainly reflect the importance of the spectral information of each band, and the spatial weights reflect the importance of the spatial information of all bands of the MS image. However, each band of an MS image has specific spectral–spatial characteristics, and there is a certain relationship between them. Therefore, in order to increase the correlation of spectral-spatial features, the spectral–spatial weights are finally interleaved, and an inner product operation is performed to obtain an interleaved attention weight matrix ω_Ω. The SSIA process for the local area can be expressed by the following equations:

$$F_\Omega^{'} = Conv(F_{in}^{\Omega}) \tag{11}$$

$$O_\Omega^{\lambda} = \sigma(Conv_{1D}(GMP(F_\Omega^{'}))) \tag{12}$$

$$O_\Omega^{S} = \sigma(Conv(Cat(MP(F_\Omega^{'}), AP(F_\Omega^{'})))) \tag{13}$$

$$\omega_\Omega = Re(O_\Omega^{S} \odot O_\Omega^{\lambda}) \tag{14}$$

where F_{in}^{Ω} represents the input local feature of AFLB, and O_Ω^{λ} and O_Ω^{S} represent the outputs of spectral attention and spatial attention, respectively. $GMP(\cdot)$ represents the global max pooling operation, $Conv1D(\cdot)$ represents one-dimensional convolution, and $\sigma(\cdot)$ represents a sigmoid activation function. $MP(\cdot)$ and $AP(\cdot)$ represent max pooling and average pooling, respectively. $\odot$ represents the inner product operation, and $Re(\cdot)$ represents a reshape operator.

Next, by weighting C_k with the weight matrix ω_Ω corresponding to each local area, the local convolution kernel can be obtained adaptively. The adaptive weighted kernels are used to perform the convolution operations on the feature maps received by the AFLB to extract the features from different local areas. The output of the AFLB (O_A) is defined as follows:

$$O_A = \omega_\Omega \odot C_k \otimes F_{in} \tag{15}$$

where F_{in} represents the input feature of the AFLB, and $\otimes$ represents the convolution operation.

TABLE I
SPECIFICATIONS OF THE DATASETS

Sensor	IKONOS	Pléiades	WorldView-3
MS/PAN resolutions	0.82/3.2(m)	0.5/2.0(m)	0.31/1.24(m)
MS sizes (RS/FS)	64×64×4 256×256×4	64×64×4 256×256×4	64×64×8 256×256×8
PAN sizes (RS/FS)	256×256 1024×1024	256×256 1024×1024	256×256 1024×1024
MS bands	red(R), green(G), blue(B),and near infrared(NIR)	R, G, B, and NIR	R, G, B, NIR1, coastal blue, yellow, red edge, and NIR2

Finally, with the constructed AFLB, ω_Ω and O_A can be iteratively optimized during the training process. Furthermore, L1-loss is employed to train the AWFLN to generate a fused image that has high spectral–spatial fidelity with the reference image.

IV. Experimental Result And Analysis

A. Experiment Setting

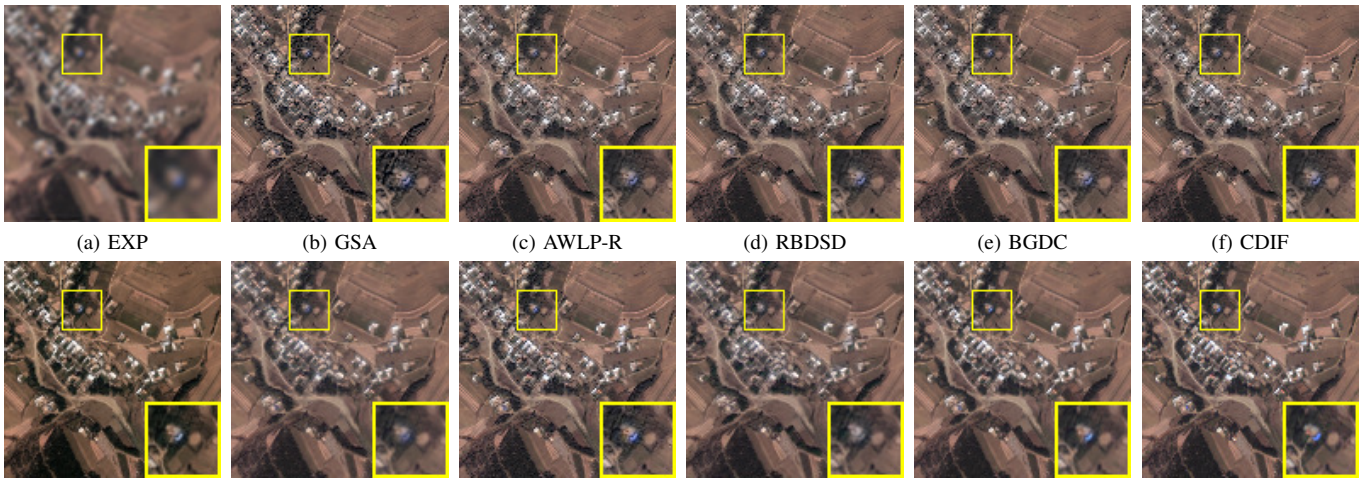
To verify the effectiveness of the proposed AWFLN, we collect three datasets: IKONOS, Pléiades, and WorldView-3. Table I presents detailed information about the three datasets. There are two types of experiments. The first is the RS experiment, in which the source images are first degraded by filters matched to the sensors' modulation transfer function (MTF) and then down-sampled by a factor of four. The original MS images are considered as the ground truth (GT), following Wald's protocol [41]. The second one is the FS experiment, where the images are fused at the original scale.

We collect 2000 RS image pairs from each dataset to participate in training, and the original MS image is used as the GT. During the training, the RS images are randomly cropped to half the original size in each epoch, that is, 128×128, which can greatly expand the training samples. The batch size is set to 16 according to the GPU capability. To stabilize the convergence, we set 300 epochs total during training, with the learning rate initially set at 0.0005 and halved every 100 epochs. During the testing process, another 100 image pairs from each dataset are collected for the RS and FS experiments.

Various state-of-the-art and conventional methods are collected and compared with the proposed method. For example, GSA (Aiazzi *et al.*, 2007) [11] and RBDSD (Vivone *et al.*, 2019) [13] are CS-based methods, AWLP-R (Vivone *et al.*, 2019) [19] is an MRA-based method, and band-adaptive gradient and detail correction (BGDC) (Lu *et al.*, 2022) [4] and CDIF (Xiao *et al.*, 2022) [27] are VO-based methods. DL-based methods include PanColorGAN (Ozcelik *et al.*, 2021) [33], FusionNet (Deng *et al.*, 2021) [30], VO+Net (Wu *et al.*, 2021) [31], and TDNet (Zhang *et al.*, 2022) [35]. The DL-based methods are retrained on our datasets for fair comparison. Besides, instead of being used as a comparison method, the up-sampling method with polynomial interpolation (EXP) [42] is used as the spectral benchmark. All the codes are provided by the authors, and the parameters are set according

TABLE II
ILLUSTRATION OF THE FUSION QUALITY METRICS

	Full name	Description	RS/FS	Optimal
PSNR↑	Peak Signal to Noise Ratio	Representing the quality of signal reconstruction	RS	-
UIQI↑	Universal Image Quality Index	Representing a global fusion quality	RS	1
SAM↓	Spectral Angle Mapper	Measuring the spectral distortion	RS	0
ERGAS↓	Erreur Relative Global Adimensionnelle De Synthese	Evaluating the spatial and spectral quality	RS	0
SCC↑	Spatial Correlation Coefficient	Indicating a spatial quality	RS	1
Q4/Q8↑	Q4 for a 4-band image, Q8 for an 8-band image	A vector extension of the Q-index to evaluating the global quality	RS	1
D_λ↓	-	Indicating spectral distortion	FS	0
D_s↓	-	Indicating spatial distortion	FS	0
QNR↑	Quality with No Reference	Comprehensive metric combining D_λ and D_s	FS	1



(a) EXP (b) GSA (c) AWLP-R (d) RBDSD (e) BGDC (f) CDIF

(g) PanColorGAN (h) FusionNet (i) VO+Net (j) TDNet (k) Proposed (l) GT

Fig. 5. Fusion results of the reduced-scale images from IKONOS dataset.

to the original papers. All comparison methods are run on a computer with i5-11400 CPU, RTX-3060 GPU, and 24 GB RAM.

To quantitatively evaluate the effect of various fusion methods, we employ some well-known metrics. They are PSNR↑ [27], UIQI↑ [43], SAM↓ [44], ERGAS↓ [40], SCC↑ [7], and Q2n↑ (Q4/Q8) [45] for RS experiments and D_λ↓, D_s↓, and QNR↑ [46] for FS experiments. The detailed information is presented in Table II. Here, the up arrow ↑ indicates that the higher the value, the better the fusion result, while the down arrow ↓ indicates the opposite.

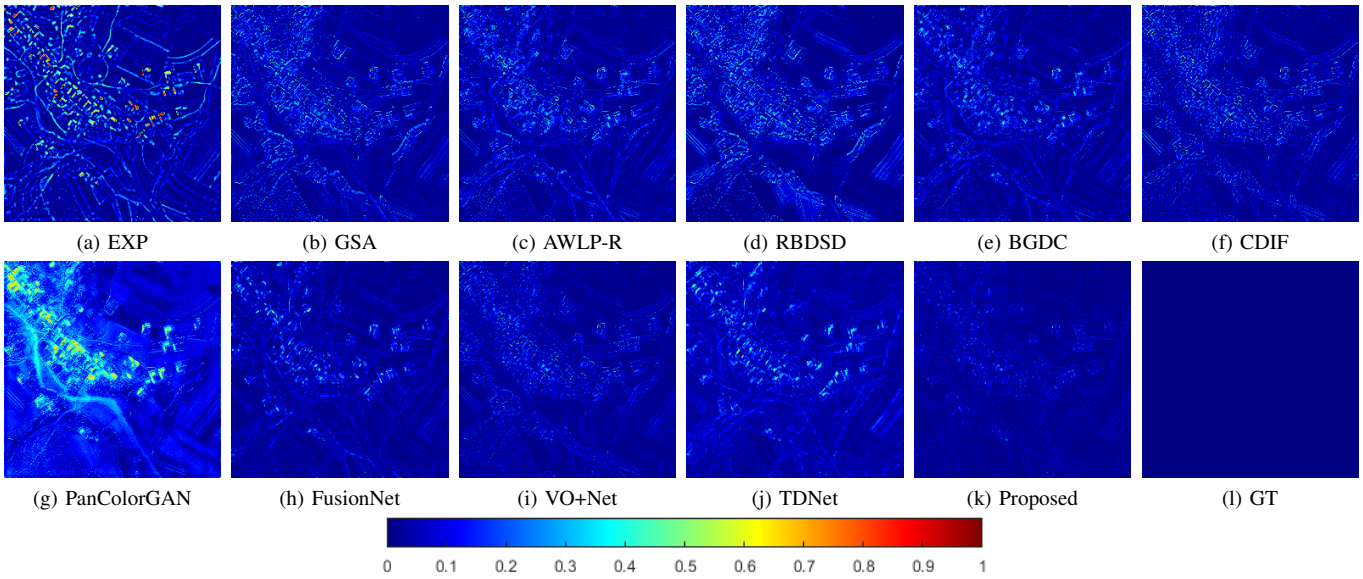
B. RS Experiments

Fig. 5 shows the fusion results for an image pair from the IKONOS dataset, with only RGB bands displayed for better visual effect. The yellow rectangles that have been enlarged are close-ups of the smaller ones. The figure shows that the PanColorGAN result has obvious spectral distortion, because the overall color is biased. The results of BGDC, FusionNet, and TDNet are blurred, as they have extracted insufficient spatial information. From the yellow rectangles, we can see that the results of GSA, AWLP-R, and RBDSD show a certain degree of spectral distortion, as the blue regions are changed in color. The results of CDIF, VO+Net, and the proposed method are closer to the GT; it is difficult to tell the difference between them.

To better visualize the differences in the results of the comparison methods, we use the absolute error maps (AEMs) [30], as shown in Fig. 6. From the figure, it is clear that the proposed method produces results that are similar to the GT and that contain noticeably fewer residues than the results produced by other methods.

Table III shows a quantitative evaluation of the fusion results on the IKONOS dataset. The left half of the table shows the results in Fig. 5, while the right half shows the average results for the 100 testing samples. The best results are highlighted in bold, while the second-best results are underlined. The table shows that the DL-based methods generally perform better than the traditional ones, and the proposed approach performs the best on all metrics and has obvious advantages. The objective evaluation further affirms the effectiveness of the proposed model.

A fusion example in the Pléiades dataset is shown in Fig. 7. We can see that the PanColorGAN result has significant spectral distortion. The enlarged yellow rectangles show that the results of the compared methods exhibit varying degrees of spectral distortion, particularly in the pink area. In contrast, the results of CDIF and the proposed method maintain good spectral-spatial fidelity. To further differentiate the fusion results visually, the AEMs of Fig. 7 are shown in Fig. 8, which demonstrates that the proposed method produces results with significantly fewer residues than those of other comparison methods. The objective assessment is presented in Table IV



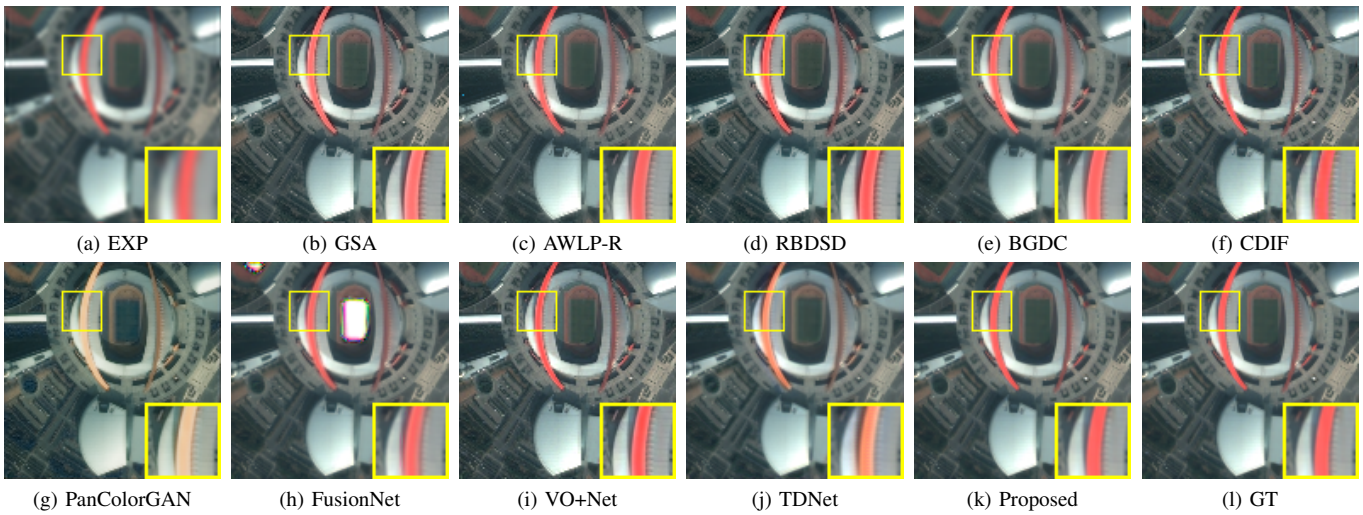
(a) EXP (b) GSA (c) AWLP-R (d) RBDSD (e) BGDC (f) CDIF

(g) PanColorGAN (h) FusionNet (i) VO+Net (j) TDNet (k) Proposed (l) GT

Fig. 6. The corresponding AEMs of the fusion results in Fig. 5.

TABLE III
QUANTITATIVE EVALUATION OF FUSION RESULTS IN FIG. 5, AND AVERAGE QUANTITATIVE EVALUATION ON IKONOS DATASET

	Result in Fig. 5						Average					
Methods	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [42]	23.7529	0.7585	3.6470	4.3244	0.6760	0.7612	26.6900	0.7627	3.3579	4.0487	0.6849	0.7574
GSA [11]	25.5501	0.9029	4.6613	3.3241	0.8613	0.8989	27.1903	0.8438	4.5543	3.9428	0.8284	0.8351
AWLP-R [19]	26.5230	0.9107	3.8261	3.0909	0.8765	0.9063	28.8912	0.8682	3.5731	3.3621	0.8540	0.8633
RBDSD [13]	25.1010	0.8950	4.9540	3.5094	0.8608	0.8841	27.4232	0.8529	4.4319	3.7717	0.8410	0.8427
BGDC [4]	27.3171	0.9209	3.6380	2.8181	0.8888	0.9180	29.5621	0.8833	3.4075	3.1049	0.8609	0.8798
CDIF [27]	26.7538	0.9278	3.6184	2.8852	0.8850	0.9229	29.7225	0.9068	3.3128	2.8697	0.8760	0.9033
PanColorGAN [33]	24.7995	0.9052	6.8689	4.6980	0.9536	0.9074	24.0976	0.8321	8.4360	6.0984	0.9463	0.8224
FusionNet [30]	29.0693	0.9457	3.1819	2.3258	0.9185	0.9449	31.1941	0.9118	3.1394	2.6173	0.8973	0.9088
VO+Net [31]	28.6561	0.9469	3.3867	2.3974	0.9074	0.9432	29.4322	0.8944	3.5710	3.1333	0.8685	0.8866
TDNet [35]	27.3888	0.9208	4.1004	2.8458	0.9046	0.9151	29.6391	0.8877	3.7885	3.0256	0.8896	0.8783
Proposed	**32.6707**	**0.9759**	**2.3307**	**1.5404**	**0.9600**	**0.9731**	**34.3431**	**0.9557**	**2.3082**	**1.6438**	**0.9531**	**0.9539**



(a) EXP (b) GSA (c) AWLP-R (d) RBDSD (e) BGDC (f) CDIF

(g) PanColorGAN (h) FusionNet (i) VO+Net (j) TDNet (k) Proposed (l) GT

Fig. 7. Fusion results of the reduced-scale images from Pléiades dataset.

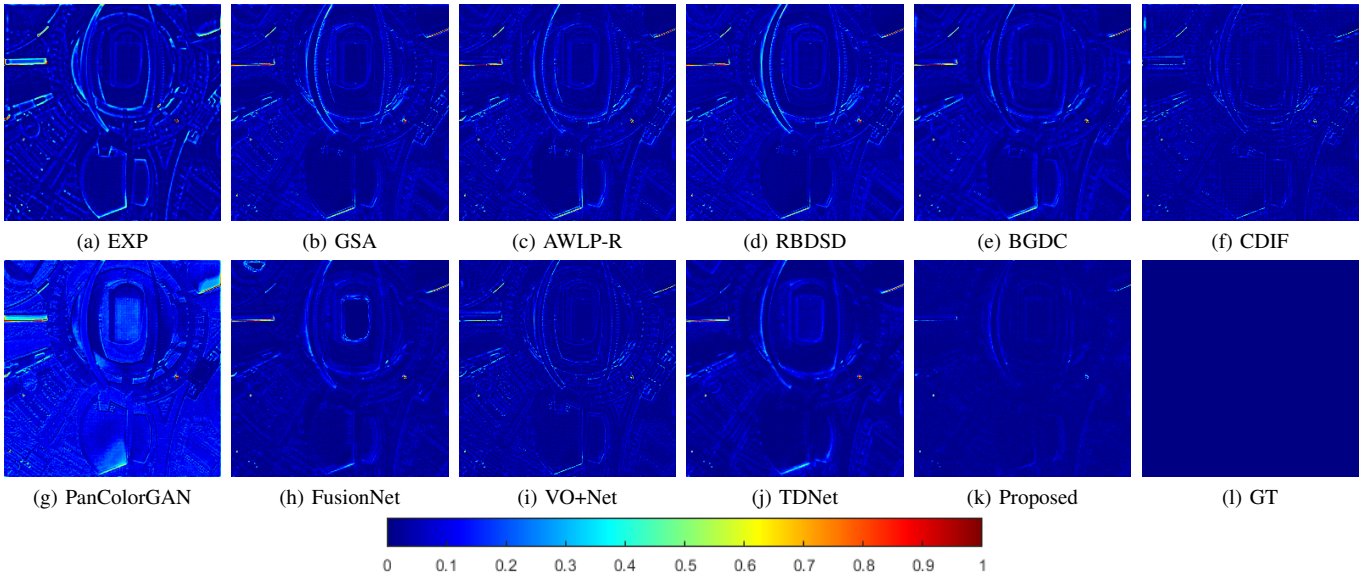


Fig. 8. The corresponding AEMs of the fusion results in Fig. 7.

TABLE IV
QUANTITATIVE EVALUATION OF FUSION RESULTS IN FIG. 7, AND AVERAGE QUANTITATIVE EVALUATION ON PLÉIADES DATASET

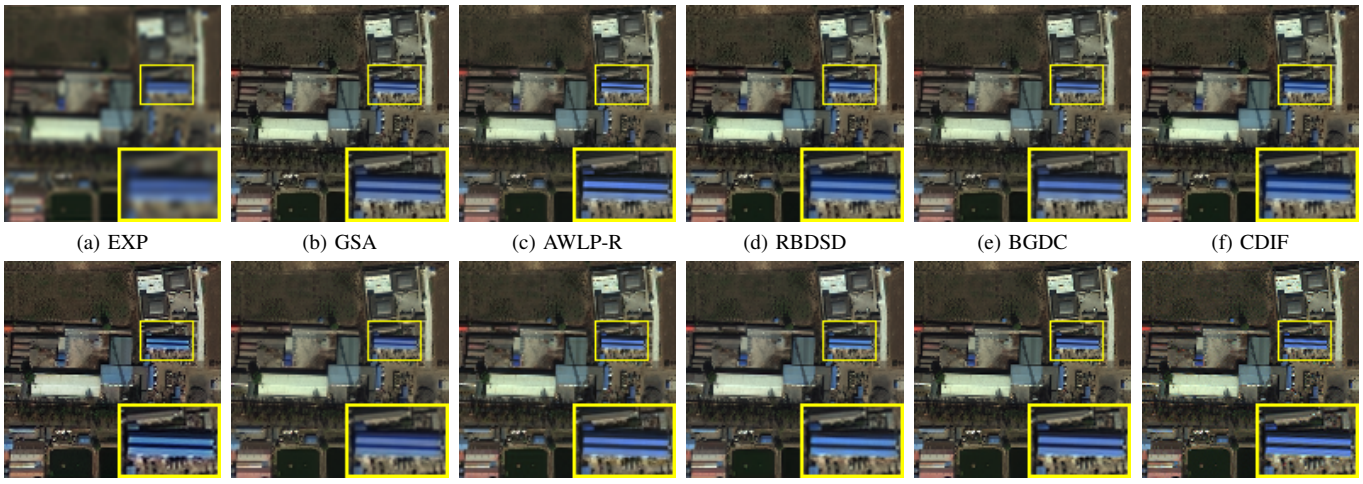
	Result in Fig. 7						Average					
Methods	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [42]	27.2385	0.8971	2.6933	2.8025	0.7842	0.8884	27.7443	0.8194	3.0838	3.9728	0.7522	0.8173
GSA [11]	27.5006	0.9252	2.9728	2.4843	0.8516	0.9138	28.5217	0.8751	3.3739	3.5754	0.8296	0.8709
AWLP-R [19]	28.6444	0.9345	2.6426	2.3830	0.8649	0.9268	29.2524	0.8863	3.0752	3.4263	0.8332	0.8846
RBDSD [13]	27.7510	0.9301	3.2575	2.5374	0.8620	0.9182	28.5048	0.8774	3.5755	3.6265	0.8354	0.8730
BGDC [4]	29.4487	0.9411	2.7422	2.1598	0.8829	0.9366	30.8217	0.9114	3.2388	2.8189	0.8696	0.9109
CDIF [27]	31.1904	0.9615	2.1998	1.7053	0.8952	0.9533	30.4693	0.9150	2.7529	2.8492	0.8620	0.9133
PanColorGAN [33]	24.8169	0.9081	5.4976	3.8929	0.8758	0.8540	27.0111	0.8653	7.7779	5.0045	0.8483	0.8052
FusionNet [30]	19.7971	0.8887	3.1617	5.7916	0.8753	0.9119	31.1747	0.9282	2.8778	2.9819	0.9039	0.9262
VO+Net [31]	30.0569	0.9530	2.1809	1.8801	0.8915	0.9430	30.3583	0.9122	2.7449	2.9123	0.8660	0.9090
TDNet [35]	28.4636	0.9354	3.3303	2.4320	0.9043	0.9317	31.3226	0.9215	3.5946	2.6642	0.9095	0.9164
Proposed	**34.3679**	**0.9743**	**1.7703**	**1.1140**	**0.9557**	**0.9705**	**35.1197**	**0.9565**	**2.2159**	**1.5428**	**0.9573**	**0.9566**

in order to quantify the results of the fusion. As shown in the table, the proposed method performs the best across all metrics, and the subjective evaluation is consistent with the objective evaluation.

Fig. 9 shows the fusion result for a pair of images from the WorldView-3 dataset. From the enlarged yellow rectangles, we can see that PanColorGAN exhibits impressive spatial quality but poor spectral quality, similar to the results on other datasets, because the color of the blue roofs is changed. The results of GSA, RBDSD, and FustionNet seem too dark compared to the GT. The results of BGDC and CDIF are somewhat blurry, and the result of TDNet has slight spectral distortion. The proposed method produces the result closest to the GT. Fig. 10 further shows the AEMs of the corresponding fusion results in Fig. 9, and it can be seen that the result produced by our method obviously contains fewer residues than those produced by other methods. Table V shows the objective evaluation of the result in Fig. 9 and the average evaluation of the test results on the WorldView-3 dataset, which confirm that our method outperforms other methods in all metrics.

C. FS Experiments

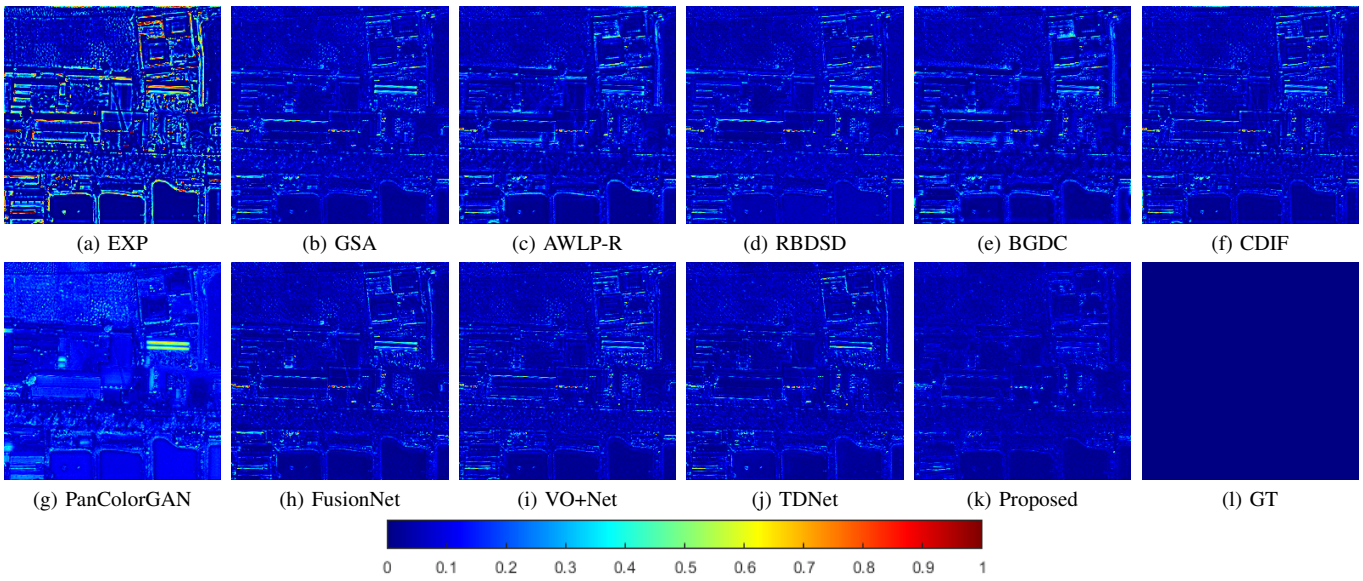
The FS experiment is illustrated using an image pair from the IKONOS dataset. The fusion results are shown in Fig. 11. Because the size of the FS image is $1024\times1024\times B$, where B represents the band number, the image is too large to display in the paper. We thus crop an area marked with a yellow rectangle and only show the close-up of the area. The EXP result is regarded as a spectral reference because of the lack of a reference image. As shown in the figure, the results of GSA, RBDSD, and PanColorGAN obviously exhibit spectral distortion, as the color of the trees turns brown when it should be green. The result of TDNet is blurred. Without the GT, it is difficult to distinguish between the results of other methods; so, we mainly refer to the quantitative evaluation shown in Table VI. The table shows that for the fusion results in Fig. 11, the proposed method yields the best results on D_λ and QNR and the second-best results on D_s. Furthermore, our method outperforms all other methods in the average evaluation with a clear advantage, demonstrating the efficacy of the proposed method.



(a) EXP (b) GSA (c) AWLP-R (d) RBDSD (e) BGDC (f) CDIF

(g) PanColorGAN (h) FusionNet (i) VO+Net (j) TDNet (k) Proposed (l) GT

Fig. 9. Fusion results of the reduced-scale images from WorldView-3 dataset.



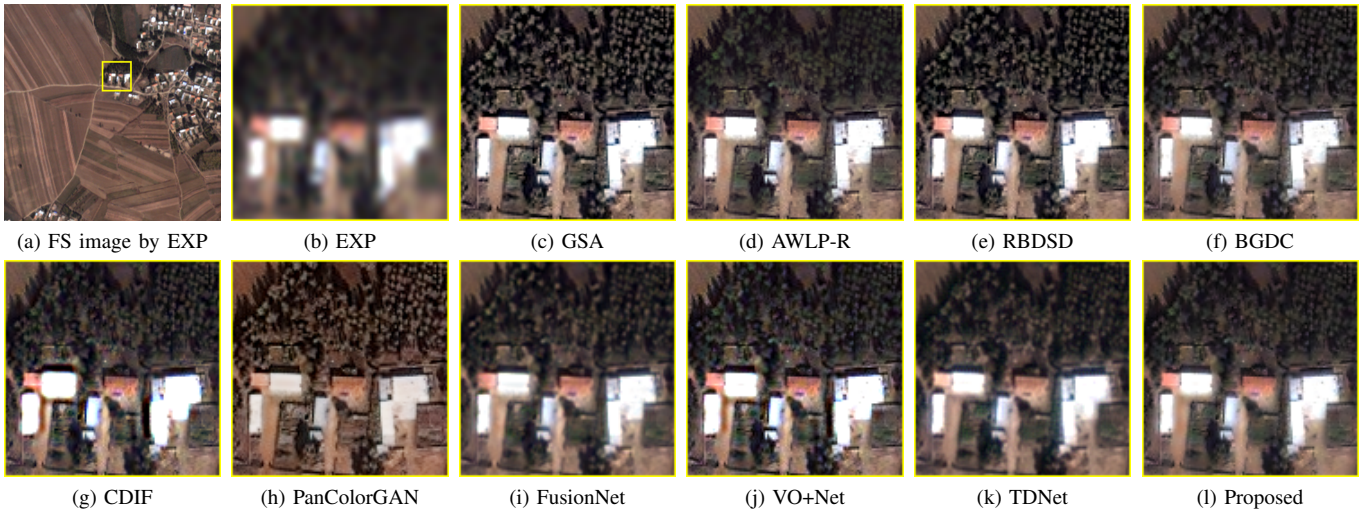
(a) EXP (b) GSA (c) AWLP-R (d) RBDSD (e) BGDC (f) CDIF

(g) PanColorGAN (h) FusionNet (i) VO+Net (j) TDNet (k) Proposed (l) GT

Fig. 10. The corresponding AEMs of the fusion results in Fig. 9.

TABLE V
QUANTITATIVE EVALUATION OF FUSION RESULTS IN FIG. 9, AND AVERAGE QUANTITATIVE EVALUATION ON WORLDVIEW-3 DATASET

	Result in Fig. 9						Average					
Methods	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑
EXP [42]	21.9986	0.6898	6.0953	7.3379	0.6211	0.7143	25.614	0.6567	5.443	6.1138	0.6284	0.6553
GSA [11]	28.6873	0.9152	6.7636	3.8956	0.9052	0.9305	29.0756	0.8589	6.4646	4.2272	0.8384	0.8720
AWLP-R [19]	27.9229	0.9171	6.1680	3.9334	0.9126	0.9237	29.5801	0.8769	5.4589	4.0417	0.8599	0.8805
RBDSD [13]	28.5178	0.9208	6.7489	3.8839	0.9115	0.9317	29.2933	0.8731	6.2859	4.1564	0.8539	0.8796
BGDC [4]	28.6515	0.9318	6.0100	3.7167	0.9198	0.9373	29.4501	0.8799	5.5510	4.1184	0.8613	0.8832
CDIF [27]	28.6807	0.9244	6.1032	3.7157	0.9124	0.9326	30.1201	0.8875	5.1814	3.7217	0.8641	0.8914
PanColorGAN [33]	27.4114	0.9252	7.9119	3.0019	0.9314	0.9235	25.3901	0.8401	8.3228	5.6929	0.8971	0.8302
FusionNet [30]	28.7356	0.9248	6.0122	3.7772	0.9150	0.9323	30.0092	0.8903	5.1457	3.7813	0.8736	0.8936
VO+Net [31]	29.2752	0.9281	6.2656	3.5994	0.9118	0.9359	29.8687	0.8842	5.7228	3.8555	0.8657	0.8846
TDNet [35]	29.1150	0.9322	6.1365	3.6603	0.9216	0.9372	30.1461	0.8912	5.3467	3.7361	0.8787	0.8936
Proposed	**31.2741**	**0.9565**	**5.3226**	**2.7279**	**0.9514**	**0.9603**	**33.7501**	**0.9408**	**4.1823**	**2.5070**	**0.9490**	**0.9431**



(a) FS image by EXP (b) EXP (c) GSA (d) AWLP-R (e) RBDSD (f) BGDC

(g) CDIF (h) PanColorGAN (i) FusionNet (j) VO+Net (k) TDNet (l) Proposed

Fig. 11. Fusion results of full-scale images from IKONOS dataset. (a) The full-scale image upsampled by EXP. (b) Close-up of the image upsampled by EXP. (c)-(l) Close-up of the pansharpened images by different methods.

TABLE VI
QUANTITATIVE EVALUATION OF FUSION RESULTS IN FIG. 11, AND AVERAGE QUANTITATIVE EVALUATION ON IKONOS DATASET

	Result in Fig. 11			Average		
Methods	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$	$D_\lambda\downarrow$	$D_s\downarrow$	QNR$\uparrow$
EXP [42]	0.0005	0.2445	0.7551	0.0006	0.2334	0.7662
GSA [11]	0.1385	0.1265	0.7538	0.1409	0.1207	0.7574
AWLP-R [19]	0.1378	0.1174	0.7610	0.1456	0.1597	0.7195
RBDSD [13]	0.1421	0.1232	0.7522	0.1293	0.1385	0.7530
BGDC [4]	0.0724	0.0351	0.8918	0.0868	0.0700	0.8500
CDIF [27]	0.0986	0.0712	0.8372	0.0866	0.0653	0.8549
PanColorGAN [33]	0.0524	0.1328	0.8217	0.1305	0.2374	0.6676
FusionNet [30]	0.0593	**0.0316**	0.9109	0.0772	0.0663	0.8618
VO+Net [31]	0.0972	0.0829	0.8280	0.1246	0.1219	0.7713
TDNet [35]	0.0936	0.0339	0.8756	0.1044	0.0772	0.8267
Proposed	**0.0508**	0.0337	**0.9172**	**0.0511**	**0.0581**	**0.8935**

D. Ablation Study

To verify the performance of the proposed components in the method, some ablation experiments are conducted in this section. Taking WorldView-3 dataset as an example, we retrain the ablated models on the training dataset and validate these models on 20 image pairs in the dataset.

1) Effect of Each Component in AWFLN : The proposed model mainly consists of a data preprocessing block (denoted as DPB) and an RMRS block including MCB and AFLB. Furthermore, AFLB contains the adaptive coefficient that is obtained by the proposed SSIA. Therefore, we train the models without AFLB, without MCB and SSIA, without SSIA, without MCB, and without DPB, successively. Note that in the model without DPB, we directly use the interpolated UPMS and PAN images as the source images. In the model without SSIA, we replace SSIA with a traditional convolution layer. Fig. 12 shows the fusion results of a WorldView-3 image pair, and Table VII shows the detailed ablation models as well as the average objective evaluation of the test results. The enlarged yellow rectangles in Fig. 12 show that the results in (a) and (b) are too dark, especially for the white roof, while others are close to each other. The AEMs show that the proposed model has fewer residues than other ablated models (see the orange arrow areas), which is further confirmed by the results in Table VII. Generally, we can see that the model without AFLB is inferior to the proposed method, which means that AFLB is critical to the performance of the network. Without other components such as MCB, SSIA, and DPB, the fusion results are inferior to some extent. Therefore, the proposed components are all necessary to constitute a high-performance network.

2) Convergence Experiment of Ablated Models: To further validate the convergence ability of the above ablated models, we show the loss curves of the ablated models in Fig. 13(a). Note that the training results are saved every 10 epochs. From the figure, we can see that the loss function of the proposed method can steadily converge to a smaller value compared to other models. To validate the trained models, taking the representative metric Q8 for evaluating the overall quality as an example, the results of the ablated models at different epochs are shown in Fig. 13(b). It can be seen that the proposed model can also achieve the best Q8 results and stability. Furthermore, the model without AFLB and that without MCB and SSIA obviously perform poorly both in the convergence of the loss function and the Q8 results, which proves that these components are essential for the proposed model. The models without SSIA, MCB, and DPB all perform poorly both in the fusion result and training stability. Therefore, the experimental results further verify that the proposed components can improve the robustness and effectiveness of the training process.

E. Effect of the Number of AFLBs and RMRSs

Setting the numbers of RMRSs and AFLBs inevitably affects the training result of the model. A shallow network may result in insufficient feature extraction, while overly deep networks may lead to large computational burden and overfitting phenomenon. As mentioned in Section III-D, this paper sets the number of RMRS to two for better feature extraction and progressive dimension reduction. The number

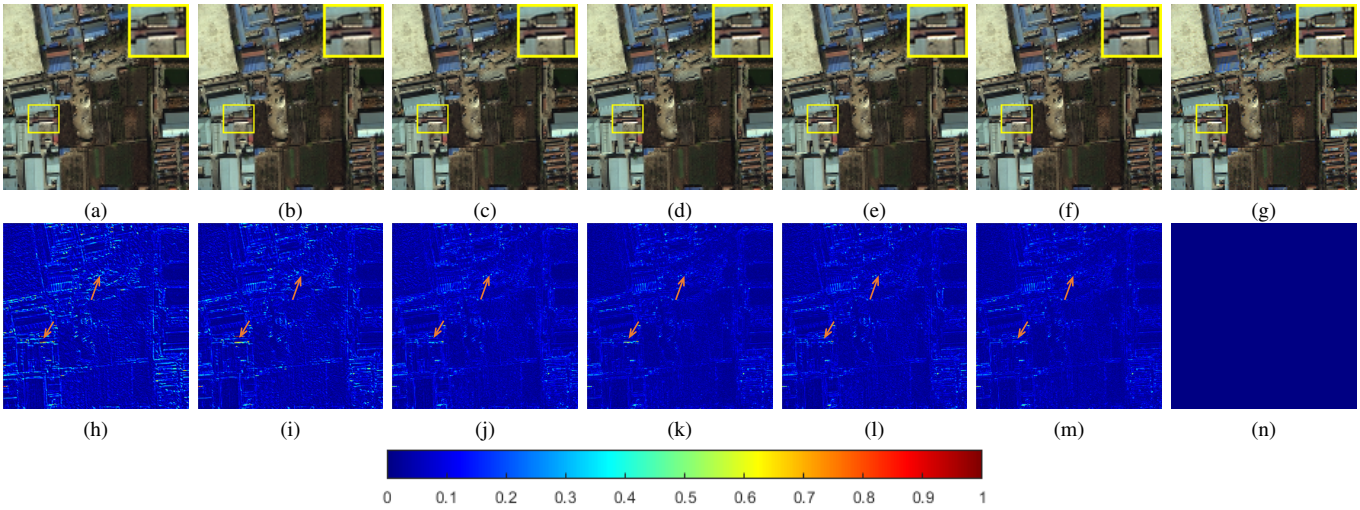


Fig. 12. Fusion results and AEMs of an image pair by different ablation models. (a) w/o AFLB. (b) w/o MCB & SSIA. (c) w/o SSIA. (d) w/o MCB. (e) w/o DPB. (f) Proposed. (g) GT. (h)-(n) are the corresponding AEMs of (a)-(g).

TABLE VII
AVERAGE OBJECTIVE EVALUATION OF DIFFERENT COMBINATIONS OF THE MODELS IN THE ABLATION STUDY

Models	AFLB	SSIA	MCB	DPB	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q8↑
w/o AFLB	×	×	✓	✓	31.5956	0.9468	5.1516	3.1987	0.9563	0.9085
w/o MCB & SSIA	✓	×	×	✓	32.1755	0.9527	5.0330	2.9834	0.9616	0.9167
w/o SSIA	✓	×	✓	✓	32.5699	0.9542	4.9318	2.7714	0.9673	0.9263
w/o MCB	✓	✓	×	✓	33.2628	0.9615	4.7170	2.6641	0.9697	0.9308
w/o DPB	✓	✓	✓	×	33.1042	0.9605	4.7496	2.7003	0.9688	0.9277
Proposed	✓	✓	✓	✓	**33.4553**	**0.9632**	**4.6126**	**2.6114**	**0.9711**	**0.9330**

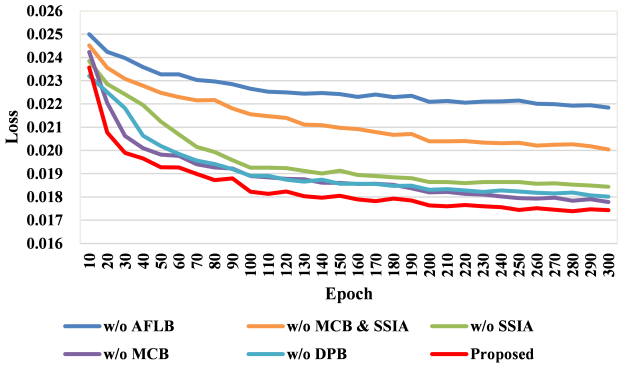
of AFLBs in each RMRS is also set to two to balance the computational efficiency and performance. To further verify the effect of the numbers of RMRSs and AFLBs, we train different combinations of the blocks when the number of RMRS is n and that of AFLB is m in each RMRS (abbreviated as R$\{n\}\times$A$\{m\}$), where $n, m \in 1, 2, 3$. Then, we test the models on the WorldView-3 image pairs. Note that when n = 3, the new RMRS is added at the end of RMRS2 in Fig. 3. Because the fusion results are similar and it is difficult to tell the difference visually, we mainly refer to the objective evaluation shown in Table VIII. Notably, the best, second best, and third best results are displayed in red, green, and blue, respectively. The last column of the table shows the average training time per epoch, which represents the computational complexity. We can see that as the number of blocks increases, the fusion results generally improve. In particular, the models R2×A2, R2×A3, and R3×A3 outperform other combinations. Furthermore, we can see that deepening the network from R2×A3 to R3×A3 does not lead to better fusion results but increases the computational complexity. Therefore, overfitting may occur for the model R3×A3. Besides, on deepening the network from R2×A2 to R2×A3, we can see that the fusion results are slightly improved, but the computational complexity is significantly increased. Thus, if we only desire high performance, we can choose to use R2×A3 or our default setting R2×A2 to balance the performance and efficiency.

TABLE VIII
AVERAGE OBJECTIVE EVALUATION OF DIFFERENT NUMBER OF RMRSS AND AFLBS IN THE ABLATION STUDY

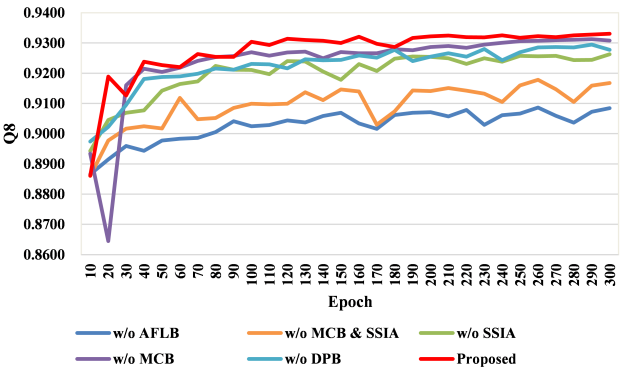
Models	PSNR↑	UIQI↑	SAM↓	ERGAS↓	SCC↑	Q4↑	Time(s)
R1×A1	32.9998	0.9599	4.8588	2.7336	0.9683	0.9290	42.5
R1×A2	33.1110	0.9608	4.7878	2.7092	0.9691	0.9296	52.8
R1×A3	33.4132	0.9629	4.6146	2.6222	0.9709	0.9325	61.7
R2×A1	33.2603	0.9620	4.7421	2.6578	0.9699	0.9313	51.5
R2×A2	33.4553	0.9632	4.6126	2.6114	0.9711	0.9330	69.4
R2×A3	33.6310	0.9643	4.5448	2.5593	0.9720	0.9345	84.1
R3×A1	33.3721	0.9626	4.6390	2.6348	0.9706	0.9325	63.7
R3×A2	33.4259	0.9628	4.6543	2.6161	0.9710	0.9323	84.1
R3×A3	33.6195	0.9642	4.4995	2.5642	0.9721	0.9341	113.5

TABLE IX
AVERAGE RUNNING TIME (S) AND NUMBER OF PARAMETERS

Method	Time (s)	Parameters
EXP [42]	-	-
GSA [11]	0.03(CPU)	-
AWLP-R [19]	0.04(CPU)	-
RBDSD [13]	0.03(CPU)	-
BGDC [4]	0.24(CPU)	-
CDIF [27]	22.76(CPU)	-
PanColorGAN [33]	0.004(GPU)	3.3×10^7
FusionNet [30]	0.0002(GPU)	2.3×10^5
VO+Net [31]	11.65(CPU)	3.1×10^5
TDNet [35]	0.0003(GPU)	4.9×10^5
Proposed	0.0003(GPU)	1.6×10^5



(a) Loss functions of ablated models at different epochs



(b) Average Q8 of testing results at different epochs

Fig. 13. The convergence experiment for ablated models.

F. Efficiency Analysis

To evaluate the efficiency of the proposed model, we compare the testing time of our method with that of other approaches and determine the total number of parameters used in DL-based approaches. When we take IKONOS dataset as an example, the size of UPMS images is 256×256×4, and the test results are as shown in Table IX. Note that the traditional methods run on CPU, while DL-based methods are tested on GPU, except for VO+Net, which combines PNN with a VO-based model and runs on CPU. From the table, we can see that our method contains fewer parameters than other DL-based methods, which is a significant advantage for constructing a lightweight network. Because our method involves matrix calculation in obtaining adaptive convolution kernels, the testing time is slightly longer than that of FusionNet. Specifically, our method has a much shorter running time and fewer parameters than PanColorGAN. Overall, the proposed AWFLN is a lightweight network with high performance.

G. Classification Experiments

To verify the effectiveness of the pansharpened images in applications such as object classification, using the fused images in Fig. 5 as an example, we classify the objects with the ENVI classification tool, which can be found at https://www.l3harrisgeospatial.com/Software-Technology/ENVI. Fig. 14 shows the results of the classification. The results of DL-based methods are typically superior to those of the non-DL-based methods, as shown by the black and crimson rectangles. Obviously, the proposed method produces results that are closer to the GT than other tested methods.

TABLE X
OBJECTIVE EVALUATION OF CLASSIFICATION RESULT

Method	OA↑	K↑	CE↓	OE↓	PA↑	UA↑
EXP [42]	0.7220	0.6061	0.2709	0.2878	0.7182	0.7546
GSA [11]	0.8014	0.7189	0.1868	0.2112	0.8005	0.8050
AWLP-R [19]	0.8014	0.7189	0.1868	0.2112	0.8005	0.8050
RBDSD [13]	0.7732	0.6769	0.2071	0.2439	0.7650	0.7745
BGDC [4]	0.8160	0.7381	0.1721	0.1950	0.8146	0.8254
CDIF [27]	0.8171	0.7406	0.1692	0.1895	0.8101	0.8164
PanColorGAN [33]	0.8329	0.7625	0.1529	0.1759	0.8245	0.8396
FusionNet [30]	0.8564	0.7974	0.1392	0.1518	0.8602	0.8627
VO+Net [31]	0.8477	0.7843	0.1416	0.1617	0.8440	0.8471
TDNet [35]	0.8251	0.7530	0.1687	0.1859	0.8289	0.8350
Proposed	**0.9010**	**0.8603**	**0.0950**	**0.1037**	**0.9024**	**0.9010**

To further quantify the classification quality, some metrics embedded in ENVI are employed. They are producer accuracy (PA) and user accuracy (UA), representing accuracy from the perspectives of the producer and user, respectively. The overall accuracy (OA) and kappa coefficient (K) are used for consistency testing, and the commission error (CE) and omission error (OE) represent the classification error. The results are shown in Table X, from which we can see that the classification result of the image fused by the proposed method is the best in all metrics, with a significant advantage. This demonstrates once more that the proposed pansharpening method works well when it comes to applying classification.

V. CONCLUSION

Aiming at the problems of insufficient feature extraction, lack of interpretability, and high computational complexity, in this paper, a lightweight network called AWFLN, which has high performance and efficiency, is proposed. To improve interpretability, we first construct a detail extraction model based on spectral–spatial fidelity. Guided by the model, data preprocessing is first performed on the source image to prepare for further feature extraction. Then, MCB is designed to fully extract spectral–spatial information in multiple receptive fields. Next, AFLB is designed based on SSIA to adaptively learn the weights of features and improve the accuracy of the extracted details. Combining MCB and AFLB, RMRS is constructed using a residue structure to improve the robustness of the training process. We use two RMRSs and two AFLBs in each RMRS to balance efficiency and performance. Our method outperforms other conventional and state-of-the-art methods in both RS and FS experiments on several datasets, indicating the good generalization of the proposed method. The results of additional ablation studies support the effectiveness of the proposed components. Moreover, the classification experimental results demonstrate that the proposed method outperforms all others in the pansharpening application.

REFERENCES

[1] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of

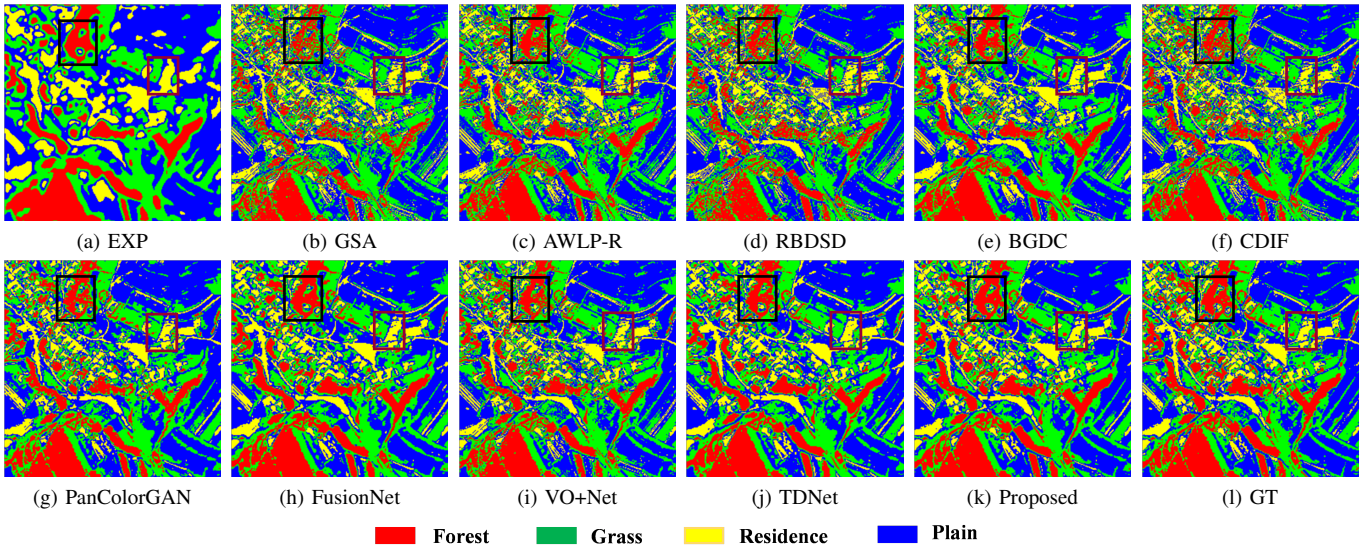


Fig. 14. Classification results of the images in Fig. 5

meta-analysis: Practical discussion and challenges," *Inf. Fusion*, vol. 46, pp. 102–113, Mar. 2019.

[2] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "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.

[3] Y. Yang, H. Lu, S. Huang, Y. Fang, and W. Tu, "An efficient and high-quality pansharpening model based on conditional random fields," *Inf. Sci.*, vol. 553, pp. 1–18, Apr. 2021.

[4] H. Lu, Y. Yang, S. Huang, W. Tu, and W. Wan, "A Unified Pansharpening Model Based on Band-Adaptive Gradient and Detail Correction," *IEEE Trans. on Image Process.*, vol. 31, pp. 918–933, 2022.

[5] 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. Photogram. Remote Sens.*, vol. 171, pp. 101–117, Jan. 2021.

[6] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogram. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[7] H. Lu, Y. Yang, S. Huang, and W. Tu, "An Efficient Pansharpening Approach Based on Texture Correction and Detail Refinement," *IEEE Geosci. Remote Sensing Lett.*, vol. 19, pp. 1–5, 2022.

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

[9] B. Meher, S. Agrawal, R. Panda, and A. Abraham, "A survey on region based image fusion methods," *Inf. Fusion*, vol. 48, pp. 119–132, Aug. 2019.

[10] X. Zhou, J. Liu, S. Liu, L. Cao, Q. Zhou, and H. Huang, "A GIHS-based spectral preservation fusion method for remote sensing images using edge restored spectral modulation," *ISPRS J. Photogram. Remote Sens.*, vol. 88, pp. 16–27, Feb. 2014.

[11] B. Aiazzi, S. Baronti, and M. Selva, "Improving Component Substitution Pansharpening Through Multivariate Regression of MS +Pan Data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Sep. 2007.

[12] 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.

[13] 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.

[14] Y. Yang, H. Lu, S. Huang, and W. Tu, "Pansharpening Based on Joint-Guided Detail Extraction," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 14, pp. 389–401, 2021.

[15] Y. Yang, W. Tu, S. Huang, H. Lu, W. Wan, and L. Gan, "Dual-Stream Convolutional Neural Network With Residual Information Enhancement for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–16, 2022.

[16] L. Zhang and J. Zhang, "A New Saliency-Driven Fusion Method Based on Complex Wavelet Transform for Remote Sensing Images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2433–2437, Dec. 2017.

[17] G. Vivone, S. Marano, and J. Chanussot, "Pansharpening: Context-Based Generalized Laplacian Pyramids by Robust Regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6152–6167, 2020.

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

[19] G. Vivone, L. Alparone, A. Garzelli, and S. Lolli, "Fast Reproducible Pansharpening Based on Instrument and Acquisition Modeling: AWLP Revisited," *Remote Sensing*, vol. 11, no. 19, p. 2315, 2019.

[20] Y. Yang, H. Lu, S. Huang, and W. Tu, "Remote Sensing Image Fusion Based on Fuzzy Logic and Salience Measure," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1943–1947, Dec. 2019.

[21] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A Variational Model for P+XS Image Fusion," *Int J Comput Vision*, vol. 69, no. 1, pp. 43–58, Aug. 2006.

[22] T. Wang, F. Fang, F. Li, and G. Zhang, "High-Quality Bayesian Pansharpening," *IEEE Trans. on Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[23] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior," *Information Fusion*, vol. 52, pp. 76–89, Dec. 2019.

[24] M. Ghahremani, Y. Liu, P. Yuen, and A. Behera, "Remote sensing image fusion via compressive sensing," *ISPRS J. Photogram. Remote Sens.*, vol. 152, pp. 34–48, Jun. 2019.

[25] Y. Peng, W. Li, X. Luo, J. Du, Y. Gan, and X. Gao, "Integrated fusion framework based on semicoupled sparse tensor factorization for spatio-temporal–spectral fusion of remote sensing images," *Inf. Fusion*, vol. 65, pp. 21–36, Jan. 2021.

[26] F. Palsson, M. O. Ulfarsson, and J. R. Sveinsson, "Model-Based Reduced-Rank Pansharpening," *IEEE Geosci. Remote Sensing Lett.*, vol. 17, no. 4, pp. 656–660, Apr. 2020.

[27] 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, pp. 1–15, 2022.

[28] Y. Yang, H. Lu, S. Huang, W. Wan, and L. Li, "Pansharpening Based on Variational Fractional-Order Geometry Model and Optimized Injection Gains," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 15, pp. 2128–2141, 2022.

[29] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by Convolutional Neural Networks." *Remote Sensing*, vol. 8, no. 7, p. 594, Jul. 2016.

[30] L.-J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 8, pp. 6995–7010, 2021.

[31] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J.-F. Hu, and G. Vivone, "VO+Net: An adaptive approach using variational optimization and deep learning for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

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

[33] F. Ozcelik, U. Alganci, E. Sertel, and G. Unal, "Rethinking CNN-Based Pansharpening: Guided Colorization of Panchromatic Images via GANs," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 4, pp. 3486–3501, Apr. 2021.

[34] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised Pansharpening Based on Self-Attention Mechanism," *IEEE Trans. Geosci. Remote Sensing*, vol. 59, no. 4, pp. 3192–3208, Apr. 2021.

[35] T.-J. Zhang, L.-J. Deng, T.-Z. Huang, J. Chanussot, and G. Vivone, "A Triple-Double Convolutional Neural Network for Panchromatic Sharpening," *IEEE Trans. Neural Netw. Learning Syst.*, pp. 1–14, 2022.

[36] L.-j. Deng, G. Vivone, M. E. Paoletti, G. Scarpa, J. He, Y. Zhang, J. Chanussot, and A. Plaza, "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.

[37] T. Yamashita, H. Furukawa, and H. Fujiyoshi, "Multiple Skip Connections of Dilated Convolution Network for Semantic Segmentation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1593–1597.

[38] T. Kim, J. Kim, and D. Kim, "SpaceMeshLab: Spatial Context Memoization And Meshgrid Atrous Convolution Consensus For Semantic Segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2259–2263.

[39] 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.

[40] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A Critical Comparison Among Pansharpening Algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

[41] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Engineering & Remote Sensing*, vol. 63, pp. 691–699, 1997.

[42] 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, Jan. 2002.

[43] Zhou W and A. C. Bovik, "A universal image quality index," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[44] R. H. Yuhas, A. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm," in *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop.*, vol. 1, 1992, pp. 147–149.

[45] G. Vivone, M. Dalla Mura, A. Garzelli, and F. Pacifici, "A Benchmarking Protocol for Pansharpening: Dataset, Preprocessing, and Quality Assessment," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing*, vol. 14, pp. 6102–6118, 2021.

[46] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and Panchromatic Data Fusion Assessment Without Reference," *photogramm eng remote sensing*, vol. 74, no. 2, pp. 193–200, Feb. 2008.