IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 September 2024, accepted 8 October 2024, date of publication 16 October 2024, date of current version 12 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3481466

RESEARCH ARTICLE

Remote Sensing Image Pansharpening Using Deep Internal Learning With Residual Double-Attention Network

RIKA SUSTIKA[1,2], ANDRIYAN B. SUKSMONO[1,3,4], (Senior Member, IEEE), DONNY DANUDIRDJO[1], (Member, IEEE), AND KETUT WIKANTIKA[5]

[1]School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung 40132, Indonesia
[2]Research Center for Artificial Intelligence and Cybersecurity, National Research and Innovation Agency (BRIN), Bandung 40135, Indonesia
[3]ITB Research Center on ICT (PPTIK-ITB), Bandung 40132, Indonesia
[4]Research Collaboration Center for Quantum Technology 2.0, STEI ITB, Bandung 40132, Indonesia
[5]Faculty of Earth Sciences and Technology, Bandung Institute of Technology, Bandung 40132, Indonesia

Corresponding author: Rika Sustika (rika004@brin.go.id)

This work was supported in part by Bandung Institute of Technology (ITB) Grant of Research, and in part by the National Research and Innovation Agency (BRIN).

ABSTRACT In recent years, deep convolutional neural networks (CNNs) have significantly improved pansharpening performance compared to traditional methods. However, existing CNN-based methods for pansharpening still lack spatial detail and suffer from spectral distortion. To address this problem, this study proposed a deep learning network based on channel and spatial attention mechanisms to enhance the spatial resolution and decrease the spectral distortion of a pansharpened image. The proposed network consists of a shallow feature extraction (SFE) unit to exploit the spatial and spectral features of the panchromatic (PAN) and multispectral (MS) input images. Furthermore, a double-attention feature fusion (DAFF) module, which consists of residual double-attention modules (RDAMs) with long and short skip connections, was designed to improve the spatial resolution and alleviate the spectral distortion of the output image. In the experiments, we utilized a deep internal learning strategy in which training data were extracted from a large scene of the observed image itself. We evaluated the effectiveness of the proposed method using WorldView-3, Spot-7, Pleiades, and Geoeye datasets. The performance of the proposed method was compared with some existing deep learning-based pansharpening techniques: deep residual pansharpening neural network (DRPNN), residual network (ResNet), residual dense model for pansharpening network (RDMPSnet), symmetric skipped connection convolutional neural network (SSC-CNN), and triplet attention network with information interaction (TANI). The experimental results revealed that the proposed method outperformed all the other methods in terms of quality evaluation metrics and visualization.

INDEX TERMS Channel attention, deep internal learning, multispectral, pansharpening, residual, spatial attention.

I. INTRODUCTION

Most satellite systems acquire two types of images: single-channel panchromatic (PAN) images with high spatial resolution and multispectral (MS) images with low spatial resolution. Many remote sensing implementations, such as environmental monitoring, target detection, classification, scene interpretation, urban planning, and surveillance [1], [2], [3], [4], require high-resolution multispectral (HRMS) images. Pansharpening is an effective method for obtaining these types of images. This technique is used to improve the spatial resolution of an MS image by fusing it with a high-resolution PAN image, as shown in Fig. 1.

Pansharpening is an active research topic, and several studies have proposed pansharpening algorithms [5], [6], [7], [8], [9]. These algorithms can be divided into two main classes: classical and deep learning-based methods. Among the various classical approaches, component substitution and multiresolution analysis are two widely representative categories [6]. Examples of component substitution-based

The associate editor coordinating the review of this manuscript and approving it for publication was Yi Zhang.

methods include intensity hue saturation (IHS) [10], Brovey transform [11], Gram-Schmidt (GS) [12], and principal component analysis (PCA) [13]. Component substitution-based methods estimated spatial details using information from PAN image. The MS channels were upsampled to PAN resolution and transformed into an alternate color space by spectral transformation (such as IHS, GS, PCA, and Brovey transform). The component in the transformed space was then substituted with a PAN image to enhance the spatial resolution. A sharpened image was produced after applying the reverse transformation. These methods were fast to execute, simple to implement, and resistant to aliasing and misregistration issues; however, they introduced more pronounced spectral distortions [14]. Meanwhile, multiresolution analysis-based methods extracted spatial details from PAN image and injected them into the MS bands. Several approaches were used to perform the extraction, such as using a wavelet transform [15], Laplacian pyramid [16], or a smoothing filter [17]. Multiresolution analysis-based methods were effective for reproducing spectral content; however, they typically provided poor fusion results regarding spatial detail [5].

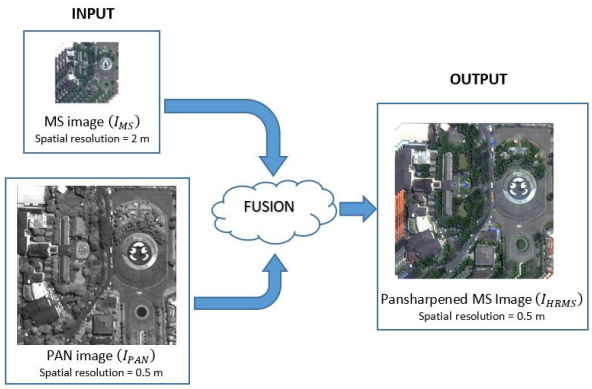

FIGURE 1. Pansharpening concept. The PAN and MS images are combined to produce an MS image with higher spatial resolution.

In the last few years, deep learning-based approaches have been proposed in many studies for remote sensing image enhancement tasks, such as single-image super-resolution [18], [19], [20], [21], [22] and multi-image super-resolution or pansharpening [23], [24], [25], [26], [27], [28], [29]. Deep learning has gained popularity owing to its robust capabilities and comprehensive learning approaches [14]. The first deep learning approach for pansharpening, known as pansharpening using convolutional neural network (PNN), was introduced by Masi et al. [23]. The PNN comprised three convolutional layers, with different feature maps for each layer. The structure was adapted from a super-resolution convolutional neural network (SRCNN), designed for single-image super-resolution [30]. Following PNN, deeper and wider convolutional neural networks (CNN) with different architectures and learning methodologies were explored to increase the model's performance and robustness.

Some researchers proposed pansharpening architectures with residual learning and specific adaptations to overcome the limitations of deep networks. Wei et al. proposed deep residual convolutional neural network (DRPNN) and obtained better results than PNN and classical methods [31], [32]. Palsson et al. used a deep CNN with a residual network (ResNet) to fuse fine and coarse spatial resolution bands in Sentinel-2 images. The study revealed that the residual architecture accelerated the convergence of deeper networks by freeing the network from learning the coarse spatial resolution part of the inputs, thereby allowing it to concentrate on building missing fine spatial details [29]. Vionthini et al. proposed RDMPSNet, a deep residual dense model for pansharpening satellite data. In this method, a densely connected layer in the residual network was proposed to preserve spectral information from a low-resolution MS image and spatial information from a high-resolution PAN image [33]. Nguyen et al. used a symmetric skipped connection convolutional neural network (SSC-CNN), inspired by the U-Net architecture to improve the spatial resolution of remote sensing images. The skipped connection in this method improved convergence without the use of too many layers [28].

In another study, generative adversarial network (GAN) strategies were used for pansharpening. In 2018, Liu et al. proposed PSGAN (pansharpening using a generative adversarial network) to effectively preserve the latent information of features using a two-branch architecture as the generator and a three-convolutional layer network as the discriminator. The two-branch architecture in the generator network was better than the one-branch architecture at improving the spatial details of the fused image [34]. Recently, a multiscale unsupervised network based on generative adversarial networks (Mun-GAN) consisting of a generator and two discriminators was proposed for pansharpening and competitive results were obtained [35].

Previous studies with deeper and wider convolutional neural networks have improved the fusion quality. Nevertheless, a few issues associated with previous CNN-based methods were introduced during fusion. MS and PAN images contain different types of information. PAN images provide high-frequency and spatial information, whereas MS images are rich in spectral information and low-frequency components. With considerable redundancy, it is challenging to determine the relationship between them [36]. Each convolutional operator had only one local receptive field throughout the convolutional process. Because context information could not be adequately exploited, the obtained features were likewise devoid of context information [36]. Furthermore, most current CNN-based methods handled all channel-wise features equally. This process hindered representational ability, and pansharpened images lacked spatial details and suffered from spectral distortion [37]. This condition affected remote sensing image analysis tasks that rely on spectral features and spatial information, such as target segmentation and object identification.

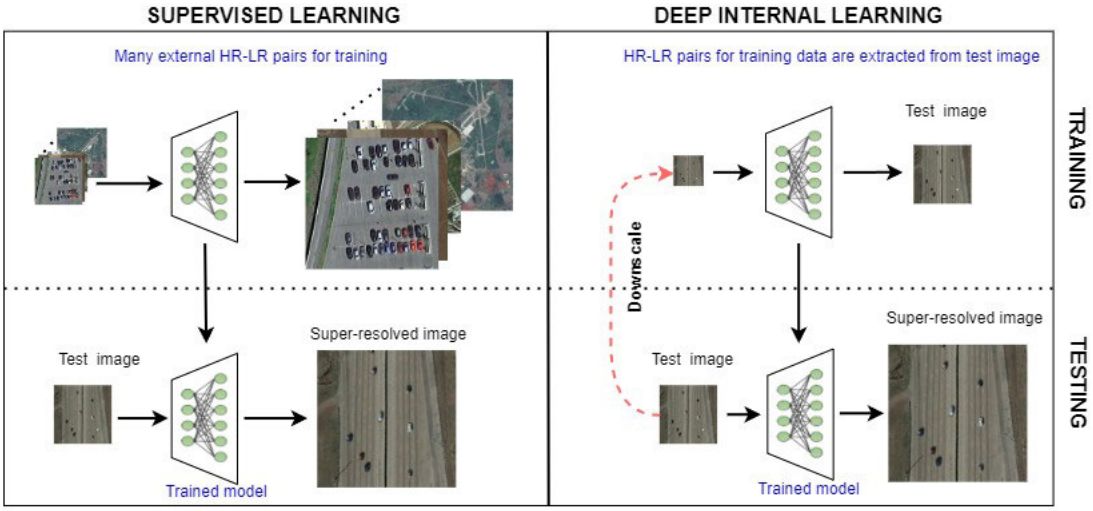

FIGURE 2. The difference between supervised learning and deep internal learning for super-resolution. In supervised learning, training was done using many external datasets of both low- and high-resolution image pairs. Then, the trained model was used to increase the test images that were not used in the training process. In deep internal learning, the model learns to recover a test image from the downgrade resolution of the test image itself [38].

To solve this problem, this study proposed a pansharpening method that combined channel and spatial attention mechanism networks, namely the residual double-attention network (RDAN). The attention mechanism, which can learn correlations between channels, has resulted in significant improvements in image classification [39], object detection [40], and super-resolution tasks [41]. Recently, it has been proven to be effective in improving pansharpening performance. For example, Zhang et al. added an attention mechanism module to the Tri-UNet structure to capture multi-level features and minimize the loss caused by downsampling [37]. Salvetti et al. used a residual channel attention mechanism in a two-stream network to learn the interdependence between channels. Based on this dependency, the correlation features among the channels were adapted such that spatial and spectral information were extracted exclusively [42]. Li et al. proposed a cross-attention-based depth unfolding iteration (CADUI) for pansharpening, which optimized deep prior regularization and combined it with a cross-attention mechanism. This method was proven superior to the other evaluated methods [43]. Diao et al. adopted attention mechanisms with information interaction to learn the spatial and spectral components in source images more efficiently. This method produced competitive results compared with state-of-the-art methods [44].

The contributions of this study are summarized as follows. First, we introduced a network with spatial and channel-wise attention with residuals in the residual architecture. Channel and spatial attention caused the network to focus on key information, and the residual structure preserved the spectral resolution of the fused image. Second, a feature extraction was performed using shallow convolution layers before fusing the MS and PAN feature images. The feature extraction module takes advantage of the spectral and spatial properties of MS and PAN images. Third, a deep internal learning strategy was used to address the limitations of the training data. In the deep internal learning method, pansharpening was performed without extensive external data for training, relying solely on a large number of small image patches of the observed image.

The remainder of this paper is organized as follows. The details of the proposed method are presented in Section II. The experimental setup is described in Section III. Section IV summarizes the experimental results and discussion, and Section V concludes the paper.

II. PROPOSED METHOD

A. DEEP INTERNAL LEARNING

Pansharpening is a technique for improving the spatial resolution of an MS remote sensing image by fusing it with spatial information from a PAN image, as illustrated in Fig. 1. This technique is known as multi-image super-resolution in remote sensing implementations, in which two types of images are used as inputs.

In this study, an observed MS image is denoted by $I_{MS} \in \mathbb{R}^{w \times h \times c}$, the PAN image is represented by $I_{PAN} \in \mathbb{R}^{rw \times rh}$, and the pansharpened MS image is represented by $I_{HRMS} \in \mathbb{R}^{rw \times rh \times c}$. w and h are the width and height of the MS image, respectively, r is the ratio of the spatial resolution between I_{PAN} and I_{MS}, and c is the number of channels in the MS image. The pansharpening problem can be formulated in (1).

$$I_{HRMS} = f([\tilde{I_{MS}}, I_{PAN}]; \theta) \tag{1}$$

where $f(.)$ is a pansharpening model to produce I_{HRMS} from the interpolated MS image ($\tilde{I_{MS}}$) and I_{PAN} as inputs, and θ is a network parameter consisting of weights and biases.

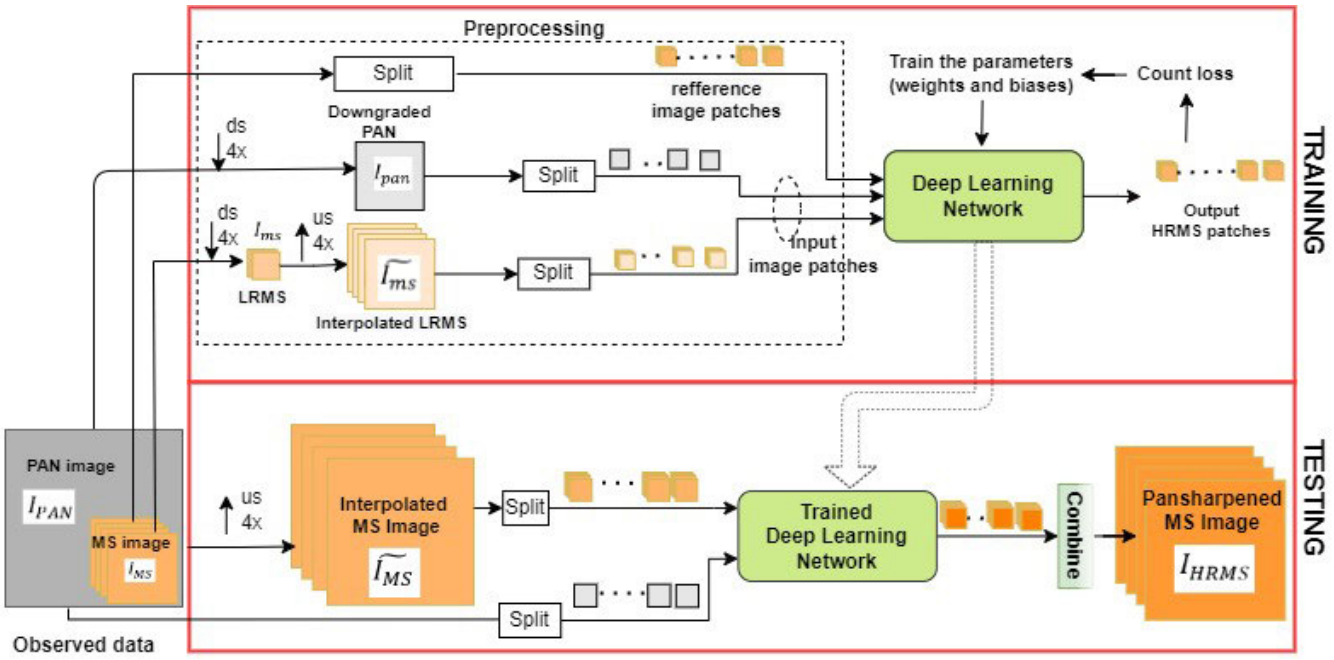

FIGURE 3. Schematic diagram of pansharpening using a deep internal learning strategy. In the training stage, training data were extracted from the observed data (a large scene of PAN and MS images, denoted as I_{PAN} and I_{MS}, respectively). In the upper channel of the training phase, the MS image was split, and the image patches were treated as references. In the two lower channels, the PAN and MS images were reduced in resolution (denoted as I_{pan} and I_{ms}, respectively) by their respective ratio, and then the MS component was interpolated. The results were split and fed to the network. The network learned to reconstruct the MS image patches from their reduced-resolution. In the testing stage, the PAN and interpolated MS image were split and fed to the trained model to produce the panshapened MS image patches. In the last step, the patches were tiled to obtain the whole pansharpened MS image.

Although deep learning technologies have significantly improved the pansharpening performance, most implementations employ supervised learning, which requires many low- and high-resolution image pairs for training. In practice, pansharpening implementations often face unavailability of such image pairs. To overcome this problem, a deep internal learning strategy was utilized. This strategy was inspired by zero-shot super-resolution using deep internal learning, as described in [38], for single-image super-resolution. Fig. 2 shows the differences between the deep internal and supervised learning methods for single-image super-resolution. In the supervised learning method, training was performed using an extensive external database of low- and high-resolution image pairs. The trained model was then used to increase the resolution of the test images that were not used in the training process. Unlike supervised learning, in the deep internal learning approach, the model was trained to recover the test image from the downgraded resolution of the test image itself. During the testing stage, the trained model was used to increase the resolution of the test image. The deep internal learning approach did not require extensive training data, and the training was faster than supervised learning [38].

The implementation of the deep internal learning approach for pansharpening is illustrated in Fig. 3. Unlike single-image super-resolution, pansharpening uses two types of inputs for training: PAN and MS images. The details of pansharpening using deep internal learning, which consists of training and testing stages, are described below.

1) TRAINING STAGE

The training process requires a pair of low-resolution multispectral (LRMS) and high-resolution multispectral (HRMS) image patches. The HRMS image patches were used as references to calculate the loss function. Because the HRMS images were not available, training data were created using the Wald protocol [45]. According to the Wald protocol, the observed MS image (I_{MS}) was treated as a reference, and then this image was downgraded to produce the LRMS image (I_{ms}). In Fig. 3, in the first channel of the training phase, the observed MS image (I_{MS}) was divided into N patches of size $n \times n$ pixels, and the MS image patches were used as a reference for training the network. In the two lower channels, the observed MS and PAN images were downgraded, and the MS component was interpolated ($\tilde{I_{ms}}$). The resulting images ($\tilde{I_{ms}}$ and I_{pan}) were split into the same size as the reference image patch. The N-stacked patches from $\tilde{I_{ms}}$ and I_{pan} were used as the inputs for the network.

The training stage aims to obtain optimal weights and biases that minimize a loss function. To achieve this objective, the mean square error (MSE) between the patches of the pansharpened MS image and the patches of the reference image (I_{MS}) was used as the loss function. Equation (2) used to calculate the MSE.

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|f([\tilde{I_{ms_i}}, I_{pan_i}]; \theta) - I_{MS_i}\|^2. \quad (2)$$

In (2), $f([\tilde{I_{ms_i}}, I_{pan_i}]; \theta)$ is i^{th} pansharpened image patch, I_{MS_i} is i^{th} reference image patch, θ is a network parameter, and N is the number of image patches. The result of the training stage was a trained model with network parameter θ. This model was used in the testing stage to improve the spatial resolution of the observed image (I_{MS}).

2) TESTING STAGE

In the testing/prediction stage, the observed MS image (I_{MS}) was upsampled by a ratio of r, denoted by $I_{\tilde{MS}}$. The observed PAN (I_{PAN}) and upsampled MS ($I_{\tilde{MS}}$) images were divided into smaller patches of $m \times m$ pixels. These image patches ($I_{\tilde{MS}_i}$ and I_{PAN_i}) were fed to the trained model to obtain the patches of the pansharpened MS image. In the last stage, the reconstructed image patches were tiled to obtain the entire image ($I_{HRMS} \in \mathbb{R}^{rw \times rh \times c}$).

B. NETWORK ARCHITECTURE

This study proposed a pansharpening network called residual double-attention network (RDAN). The network was inspired by the residual dense network (RDN) [46], residual channel attention network (RCAN) [41], and convolutional block attention module (CBAM) [40] for single-image super-resolution. Fig. 4 shows the proposed architecture.

In the preprocessing step, the observed MS image patch (I_{MS}) was upsampled to the PAN size using a polynomial interpolator. The upsampled MS image ($I_{\tilde{MS}}$) and PAN image (I_{PAN}) were used as the inputs for the RDAN network.

The proposed network consists of three main parts: shallow feature extraction (SFE) unit, double-attention feature fusion (DAFF) module, and image reconstruction (IR) part. The SFE unit was responsible for extracting the spatial and spectral features from the PAN and MS images. Spatial and spectral information were combined using the DAFF module. Finally, the pansharpened MS image were reconstructed using the IR part. The details of each RDAN component are explained as follows:

1) SHALLOW FEATURE EXTRACTION (SFE)

Pansharpening techniques based on deep learning can be classified into two main categories: single-branch and dual-branch neural networks [34], [47]. In a single-branch neural network, the MS image was concatenated with the PAN and the composite image was then sent to the deep learning model as one input. In the dual-branch architecture, the MS and PAN images were processed using a feature extraction module consisting of two feature extractors. The extracted features were then concatenated and merged using a fusion network.

In this study, a dual-branch architecture with an SFE unit was used. The SFE unit consisted of a two-branch neural network for extracting the features of the MS and PAN images. A four-channel MS image was used as the input for the first branch, and a single-band PAN image was used as the input for the other branch. The two branches had similar structures with different weights and biases. The structure contained two consecutive convolution layers with 64 channels and a 3×3 kernel size, followed by a rectified linear unit (ReLU) activation function. Equation (3) describes the shallow feature extraction process.

$$\begin{aligned} F_{pan} &= H_{SFE}(I_{PAN}) \\ F_{ms} &= H_{SFE}(\tilde{I_{MS}}), \end{aligned} \tag{3}$$

where $H_{SFE}(.)$ denotes a convolution operation. Output from the SFE unit contained two feature maps that represent the spatial and spectral information of the PAN and MS images.

2) DOUBLE-ATTENTION FEATURE FUSION (DAFF)

To fully use the feature maps from the SFE part, we fused them using a DAFF module. The DAFF contained a set of residual double-attention modules (RDAMs) that consisted of convolution operations, multiplication, channel attention, and spatial attention mechanisms. In the DAFF, the feature maps were concatenated to form a compact feature map, as shown in (4).

$$F_{in} = F_{ms} \parallel F_{pan} \tag{4}$$

where $\parallel$ denotes a concatenation operation. F_{in} is then utilized for further shallow feature extraction as expressed in (5):

$$F_1 = H_{SFE}(F_{in}). \tag{5}$$

F_1 in (5) is then used as the input to the first RDAM.

RDAM combined channel and spatial attention to recalibrate feature weights. RDAM contained convolution layers with a rectified linear unit (ReLU) activation function and a channel and spatial attention block with a short skip connection (SSC), as shown in Fig. 4. An SSC allowed shallow information to propagate in a straightforward manner through identity mapping, which was advantageous for information flow.

In this network, let F_n be the input and F_{n+1} the output of the n-th RDAM. First, we performed two convolution operations, as formulated in (6).

$$F'_n = H(F_n) \tag{6}$$

where H was the convolution network consisting of two convolutional layers. RDAM inferred attention maps sequentially along two distinct dimensions: channel and spatial. The attention maps were then multiplied by the input feature map to improve the adaptive features. Equation (7) describes the double-attention mechanism in RDAM [40]:

$$\begin{aligned} F_{ca} &= M_c(F'_n) \otimes F'_n \\ F_{sa} &= M_s(F_{ca}) \otimes F_{ca}. \end{aligned} \tag{7}$$

In (7), $\otimes$ denotes element-wise multiplication, F'_n is the input feature map, M_c is the 1D channel attention map, M_s is the 2D spatial attention map, F_{ca} is the channel-refined feature and F_{sa} is the final refined output.

An SSC was introduced to obtain the output of the n-th RDAM. It was acquired by element-wise addition between

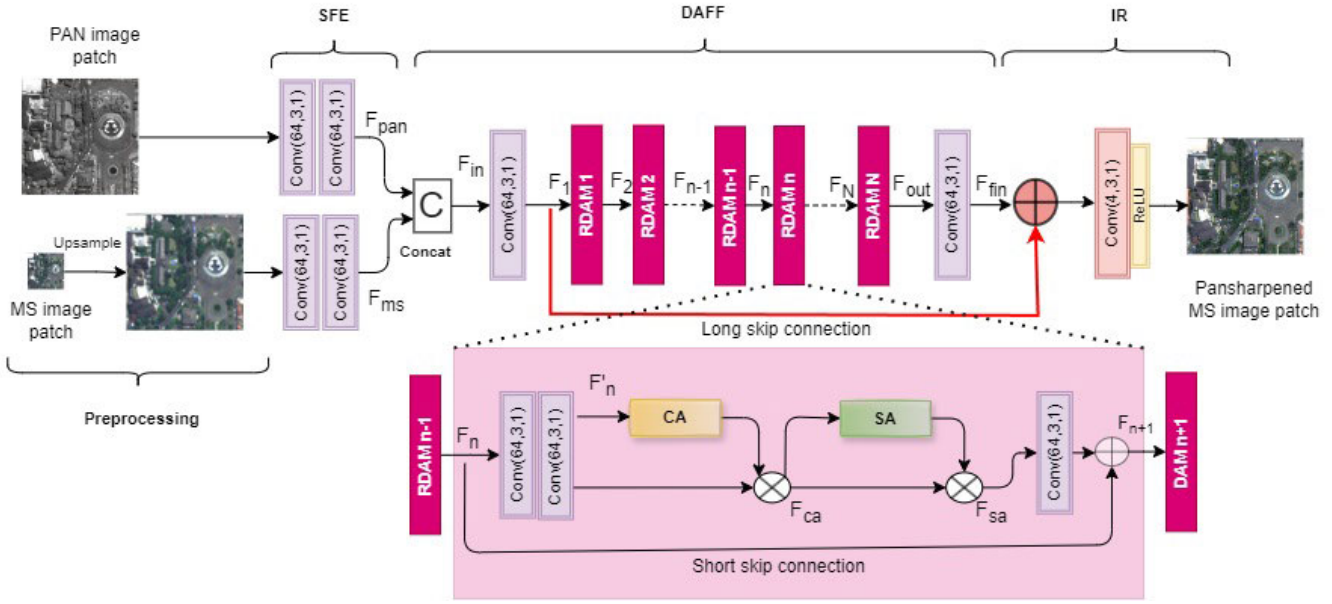

FIGURE 4. The proposed network architecture. The network consisted of an SFE unit, a DAFF module, and IR parts. The DAFF contained a set of RDAMs with channel and spatial attention mechanisms.

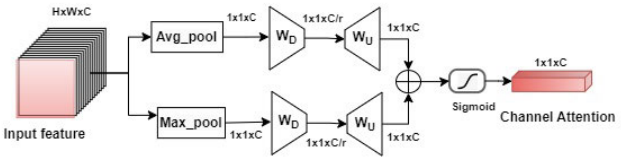

FIGURE 5. Channel attention (CA) module. The spatial information of the input feature was aggregated by using average and max pooling operations. Then, it was forwarded to a shared network to create a channel attention map.

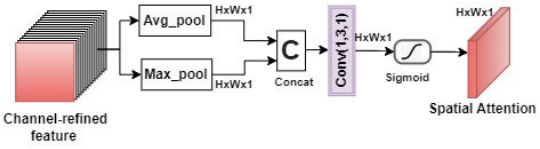

FIGURE 6. Spatial attention (SA) module. The spatial attention map was generated by applying average and max pooling operations along the channel axis, concatenating them, and then forwarding them to a convolution layer.

the output of the channel and the spatial attention block with the input feature map, as formulated in (8).

$$F_{n+1} = H(F_{sa}) \oplus F_n \tag{8}$$

where $\oplus$ is element-wise addition and F_{n+1} denotes the output of the n-th RDAM.

Suppose we have N RDAMs. Equation (9) expresses the output of the N-th RDAM.

$$\begin{aligned} F_{out} &= H_{RDAM,N}(F_N) \\ &= H_{RDAM,N}(H_{RDAM,N-1}(\ldots(H_{RDAM,1}(F_1)))) \end{aligned} \tag{9}$$

where F_{out} is the output of the N-th RDAM and $H_{RDAM,N}$ is the composite function of N-th RDAM operation.

More details regarding the channel and spatial attention are provided below:

- Channel attention (CA)
  In contrast to the current CNN which treats all channel features equally, the channel attention mechanism assigns weights to distinct channels to focus on more useful information and suppress useless information in an input feature [40]. Channel attention was computed by compressing the spatial information of the input tensor to generate a weight for each channel through a pooling operation. Fig. 5 describes the process of the channel attention mechanism used in this study [40].
  Let $F'_n \in \mathbb{R}^{H\times W\times C}$ be the input of the attention module, where C is the number of features of size $H \times W$. Global average pooling and max pooling functions were used to convert channel-wise global spatial information into the channel descriptor. The output was then sent to a shared network to create a channel attention map M_c. The shared network was composed of channel-downscaling (W_D) and channel-upscaling (W_U) with a ratio of r. Channel-wise dependencies were extracted from the aggregate information using a gating mechanism with sigmoid functions. Equation (10) expresses the channel attention mechanism [40].

$$\begin{aligned} M_c(F'_n) = \sigma((W_U ReLU(W_D F_{avg}(F'_n))) \\ + W_U ReLU(W_D F_{max}(F'_n))) \end{aligned} \tag{10}$$

In (10), $F'_n \in \mathbb{R}^{C\times H\times W}$ is the CA module input, F_{avg} is the average pooling feature, and F_{max} is the max-pooling feature obtained by shrinking F'_n through

TABLE 1. The datasets used in the experiments.

Satellite	Spatial resolution		Image size		Number of image patches
	MS image	PAN image	MS image	PAN image	
WorldView-3	0.5 m	2 m	996×996	$3,984 \times 3,984$	13,689
Spot-7	1.5 m	6 m	$2048 \times 2,048$	$8,192 \times 8,192$	63,504
Pleiades	0.5 m	2 m	$2048 \times 2,048$	$8,192 \times 8,192$	63,504
Geoeye	0.41 m	1.64 m	984×984	3936×3936	13,225

spatial dimension $H \times W$ by the reduction ratio of r. σ is the sigmoid function, W_U and W_D represent weight matrices, which act as channel-upscaling and channel-downscaling, respectively, with a ratio of r.

- Spatial attention (SA)
  Channel attention methods only considered inter-channel dependencies and neglected spatial information. The spatial attention (SA) module has often been used in CNNs to increase their ability to model spatial information and has shown great success [40]. The spatial attention module created a spatial attention map using the inter-spatial relationships of features to show the location of important information in the feature maps. Channel attention focused on refining the feature maps, whereas spatial attention focused on the importance of learning inside the feature map. Combining these two methods significantly enhanced the feature maps and improved the model performance [40]. Fig. 6 shows the spatial attention module diagram.
  In the SA module, the average and max pooling operations were used along the channel axis to generate two 2D maps, $F_{avg}(F'_n) \in \mathbb{R}^{1\times H\times W}$ and $F_{max}(F'_n) \in \mathbb{R}^{1\times H\times W}$, which were then concatenated to generate an efficient feature descriptor and convolved using a convolutional operation to produce a 2D spatial attention map $M_s(F'_n) \in \mathbb{R}^{H\times W}$. Equation (11) describes the spatial attention mechanism [40].

$$M_s(F'_n) = \sigma(H^{(7\times7)}([F_{avg}(F'_n) \parallel F_{max}(F'_n)])) \quad (11)$$

  where σ denotes the sigmoid function, and $H^{7\times7}$ represents the convolution operation with 7×7 filter size.

3) IMAGE RECONSTRUCTION (IR)

In the reconstruction stage, a convolution layer with four filters was employed after extracting and fusing the spectral and spatial features with a set of RDAMs. The number of filters used was the same as the MS image channels used in the experiments. To compensate for the potential loss of spectral information because of feature extraction or feature fusion operations, a long skip connection (LSC) was used, in which the input MS image passed directly into the reconstructed MS image. The LSC allowed the input image's spectral information to complement the pansharpened image's spectral information and prevented spatial information lost as the network deepened. The final pansharpened output I_{HRMS} is described in (12).

$$\begin{aligned} F_{fin} &= H(F_N) \\ I_{HRMS} &= H(F_{con} + F_{fin}) \end{aligned} \quad (12)$$

where F_{con} represents the concatenation of the PAN and MS feature maps and H represents the function of the entire pansharpening model.

III. EXPERIMENTS

A. DATASET

We performed several experiments using remote sensing images from Bandung, Indonesia, collected from the WorldView-3, Spot-7, and Pleiades satellites. In addition, we used data from Geoeye satellite for computational cost analysis. Table 1 summarizes the datasets used in the experiments. The data consisted of PAN and MS images with a spatial resolution ratio of four. All MS images have four bands, and the PAN images have a single channel. The images were divided into 64×64-pixel patches with 8-pixel overlap. The training data were downscaled four times and downgraded using a Gaussian kernel of size 9×9.

In the testing phase, we split the PAN and upsampled MS images into 100×100 pixels, added 28 zero paddings around the image patches and fed them into the network. In image reconstruction, the output image edges were cropped to restore the 100×100 image patches, and then all reconstructed patches were combined to form a whole pansharpened MS image.

B. EXPERIMENTAL CONFIGURATIONS

The experiments were implemented using Python 3.7, TensorFlow GPU 2.2.0, and PyTorch for CUDA 12.2, on an NVIDIA DGX cluster with 8x NVIDIA Tesla V100. We used the mean square error (MSE) as the loss function and Adam (adaptive moment estimation) as the optimizer. In this study, we did not evaluate the effects of hyper-parameter settings such as the learning rate, filter size, filter number, batch size, and number of training epochs. We set most hyper-parameters as in previous studies [23], [29], [41]. The learning rate was set to 5×10^{-4} with a momentum of 0.9. Training was completed in 30 epochs with a batch size of 32. We used 64 filters with a kernel size 3×3 in all the convolutional layers. In the proposed network, we set the number of RDAMs to 5.

Two types of experiments were conducted to assess the performance of the proposed method. The first was an experiment with reduced-resolution images, and the second

was with full-resolution images. In the first experiment, the reference quality metrics were used to evaluate the performance of the proposed method. The properties of the reconstructed image were compared with those of the reference image by measuring the deviation between the images. Because the HRMS image as a reference was unavailable, we evaluated the performance of our proposed method at reduced-resolution based on Wald's protocol [45]. The observed MS image was treated as a reference and the testing image was a downgraded version of the observed MS image. We used the peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [48], spectral angle mapper (SAM) [49], relative dimensionless global error synthesis (ERGAS) [50], spatial correlation coefficient (SCC) [51], and Q index [52] as the reference quality evaluation metrics. Based on PSNR, the best performance was achieved with the highest PSNR. One is the ideal value of the SSIM, SCC, and Q index, whereas zero is the best value for SAM and ERGAS.

The second experiment was performed at full-resolution. This experiment followed the diagram shown in Fig. 3. The network was trained using downgraded data, and all methods were tested using the original MS image. This experiment had no ground-truth images to measure the performance using the reference quality evaluation metrics. Therefore, we measured the quality of the pansharpened images using the no-reference quantitative metrics (QNR) [53]. The QNR consists of spectral distortion (D_λ) and spatial distortion (D_s) components. The D_λ measured the interband distortion between the original MS image and the pansharpened image, while the D_s focused on the spatial distortion by calculating the Q-index value between the MS image and the PAN image. The best D_λ and D_s were indicated by a zero value, whereas the best QNR was one [53].

IV. RESULTS AND DISCUSSION

The experimental results are presented in this section. We compared the performance of the proposed method with several existing methods in both the reduced- and full-resolution versions. At full-resolution, we also assessed the effects of channel and spatial attention, skip connections, and feature extraction module on the proposed network.

A. COMPARISON WITH EXISTING METHODS

In this section, we evaluated the performance of the proposed method and compared it with several existing deep learning-based methods, including DRPNN [31], ResNet [29], RDMP-SNet [33], SSC-CNN [28], and TANI [44]. In addition, we assessed the performance of two classical methods: IHS [10] and SFIM [17]. The IHS was based on the component substitution method, and the SFIM was based on multiresolution analysis method. The performance was compared at reduced and full-resolution.

The quantitative metric values of Pleiades, Spot-7, and WorldView-3 at reduced and full-resolutions are summarized in Table 2, Table 3, and Table 4, respectively. In the reduced-resolution experiments, reference metrics were used for the quantitative evaluation. For example, SAM was used to examine spectral distortion, ERGAS was used for spatial distortion, and SSIM was used as a comprehensive metric. The best outcomes were marked in bold, and the second-best results were marked in italics. According to Table 2 - Table 4, in reduced-resolution experiments all deep learning methods had better values than the classical methods for almost all metrics. The deep learning methods with attention mechanisms (TANI and RDAN proposed in this study)

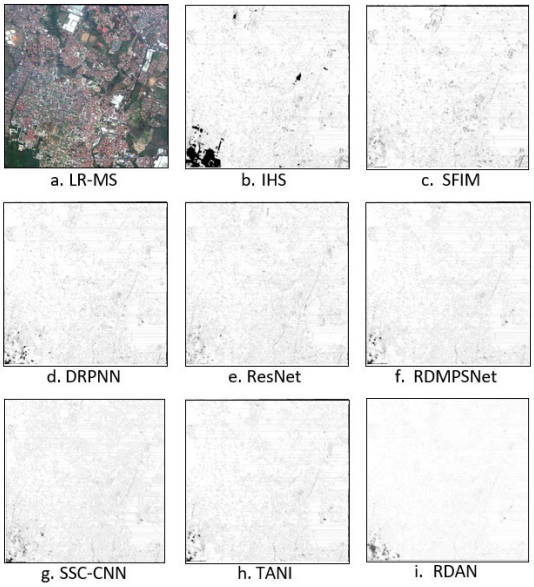

FIGURE 7. (a) Observed MS image; (b)-(i) Residual plot of IHS, SFIM, DRPNN, ResNet, RDMPSNet, SSC-CNN, TANI, and the proposed method (RDAN), respectively, in the reduced-resolution WorldView-3 image.

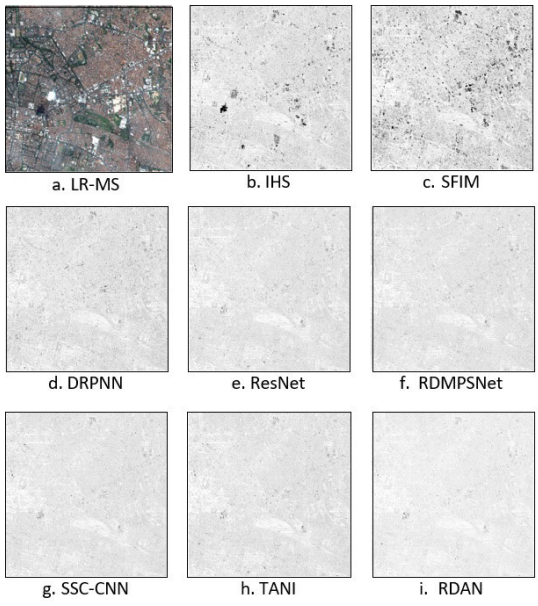

FIGURE 8. (a) Observed MS image; (b)-(i) Residual plot of IHS, SFIM, DRPNN, ResNet, RDMPSNet, SSC-CNN, TANI, and the proposed method (RDAN), respectively, in the reduced-resolution Pleiades image.

TABLE 2. Performance comparison of the proposed method with the other pansharpening methods in reduced and full resolution on the WorldView-3 data.

Methods	reduced-resolution						full-resolution		
	PSNR (dB)(↑)	SSIM (↑)	SAM (↓)	ERGAS (↓)	SCC (↑)	Q (↑)	D_λ (↓)	D_s (↓)	QNR (↑)
IHS	31.2097	0.7611	0.0761	3.5600	0.7862	0.4205	0.0590	0.2123	0.7413
SFIM	31.0012	0.7488	0.0745	3.8206	0.7653	0.3991	0.0908	0.1334	0.7879
DRPNN	31.2704	0.7639	0.0809	3.5467	0.7987	0.3845	0.0605	0.1371	0.8107
ResNet	31.5841	0.7925	0.0895	3.4115	0.8125	0.4924	0.0575	0.0872	0.8603
RDMPSNet	32.4177	0.8282	0.0703	3.0699	0.8457	0.5996	*0.0247*	0.0433	0.9330
SSC-CNN	32.8544	0.8396	0.0638	2.8800	0.8704	0.6170	0.0280	*0.0380*	*0.9350*
TANI	*34.8019*	*0.9007*	*0.0526*	*2.4149*	*0.8791*	*0.6703*	0.0739	0.0492	0.8805
RDAN (proposed method)	**36.8346**	**0.9442**	**0.0509**	**1.8015**	**0.9492**	**0.8326**	**0.0225**	**0.0277**	**0.9504**

TABLE 3. Performance comparison of the proposed method with the other pansharpening methods in reduced and full-resolution on the Spot-7 data.

Methods	reduced-resolution						full-resolution		
	PSNR (dB)(↑)	SSIM (↑)	SAM (↓)	ERGAS (↓)	SCC (↑)	Q (↑)	D_λ (↓)	D_s (↓)	QNR (↑)
IHS	31.3470	0.8609	0.0477	3.5751	0.8707	0.4934	0.1062	0.1454	0.7638
SFIM	33.8629	0.8549	0.0394	2.8198	0.9050	0.5284	0.0396	0.1748	0.7925
DRPNN	36.2600	0.9312	0.0306	1.6450	0.9645	0.7184	0.0644	*0.0461*	0.8925
ResNet	35.1581	0.9088	0.0410	2.3209	0.9305	0.5662	0.0393	0.0657	0.8976
RDMPSNet	35.9711	0.9210	0.0325	2.0523	0.9503	0.6752	0.0402	0.0609	0.9013
SSC-CNN	35.8892	0.9195	0.0300	1.6215	0.9666	0.7352	0.0397	0.0683	0.8947
TANI	*37.8048*	*0.9527*	*0.0213*	*1.2184*	*0.9730*	*0.7543*	*0.0276*	0.0707	*0.9036*
RDAN (proposed method)	**38.2319**	**0.9555**	**0.0225**	**1.0148**	**0.9848**	**0.8239**	**0.0246**	**0.0314**	**0.9448**

TABLE 4. Performance comparison of the proposed method with the other pansharpening methods in reduced and full-resolution on the Pleiades data.

Methods	reduced-resolution						full-resolution		
	PSNR (dB)(↑)	SSIM (↑)	SAM (↓)	ERGAS (↓)	SCC (↑)	Q (↑)	D_λ (↓)	D_s (↓)	QNR (↑)
IHS	35.1732	0.8869	0.0720	3.0780	0.9105	0.6991	0.0744	0.0498	0.8795
SFIM	34.2205	0.8591	0.0685	3.6849	0.8873	0.6544	**0.0408**	0.1516	0.8137
DRPNN	34.7930	0.8593	0.0792	3.1673	0.8885	0.6185	0.0693	0.0475	0.8865
ResNet	36.8940	0.9130	0.0567	2.3550	0.8993	0.609	0.0576	0.0708	0.8757
RDMPSNet	37.3931	0.9250	0.0501	2.2278	0.9544	0.8216	0.0646	*0.0459*	*0.8925*
SSC-CNN	37.3394	0.9234	0.0509	2.2349	0.9529	0.8152	0.0631	0.0571	0.8834
TANI	*40.1944*	*0.9589*	*0.0369*	*1.5264*	*0.9753*	*0.8571*	*0.0482*	0.0719	0.8834
RDAN (proposed method)	**41.8599**	**0.9725**	**0.0374**	**1.2274**	**0.9851**	**0.9159**	0.0634	**0.0426**	**0.8967**

performed better than the other methods. The tables also show that the proposed RDAN algorithm performed best for all indicators.

Non-reference metrics were used to examine the pansharpening performance of the full-resolution experiments. D_λ was used to evaluate spectral distortions, D_s was used to assess the spatial distortions, and *QNR* was used as a comprehensive metric. Table 2, Table 3, and Table 4 show that in the full-resolution condition, the deep learning-based methods performed better than the classical methods. In reduced-resolution experiments, TANI achieved the second-best performance, however, in the full-resolution experiments, the second-best QNR on the three datasets was achieved using different methods. SSC-CNN had the second-best *QNR* on the WorldView-3 dataset, TANI had the second-best result on the Spot-7 data, and RDMPSNet achieved the second-best performance on the Pleiades data. TANI performs well on down-scaled images, however, this method performed poorly on the full-resolution images. For the Worldview-3 and Spot-7 data, our proposed network outperformed the other methods in terms of D_s, D_λ, and *QNR*. On the Pleiades image, our method obtained the fourth rank for D_λ. However, the D_s and the *QNR* achieved the best results. These results verified that the proposed approach improved the spectral and spatial information in the MS images better than the other methods.

From the experiments with reduced-resolution, the differences (residuals) between the fused image and reference were plotted in Fig. 7, Fig. 8, and Fig. 9 for the WorldView-3, Pleiades, and Spot-7 images, respectively. The images show the residuals for the entire test image. A darker image indicates more significant errors between the reference and the enhanced images. Fig. 7 shows that the proposed method produced the brightest residual image compared with the other fused images in the experiment using the WorldView-3 dataset. These results indicated that the fused image obtained using the proposed method was the closest to the reference image. Similar results are shown in Fig. 8 and Fig. 9. Using the proposed method, the residual images between the pansharpened and the reference on the WorldView-3, Spot-7, and Pleiades data are whiter than those obtained using the other methods. These results demonstrated that the proposed method effectively improved the resolution of the MS images.

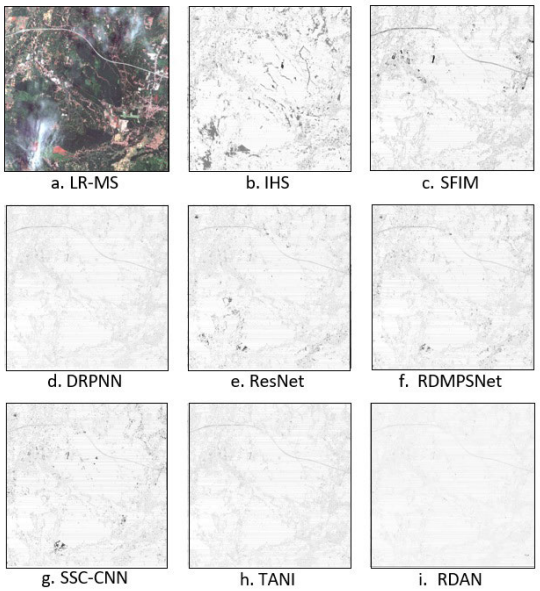

FIGURE 9. (a) Observed MS image; (b)-(i) Residual plot of IHS, SFIM, DRPNN, ResNet, RDMPSNet, SSC-CNN, TANI, and the proposed method (RDAN), respectively, in the reduced-resolution Spot-7 image.

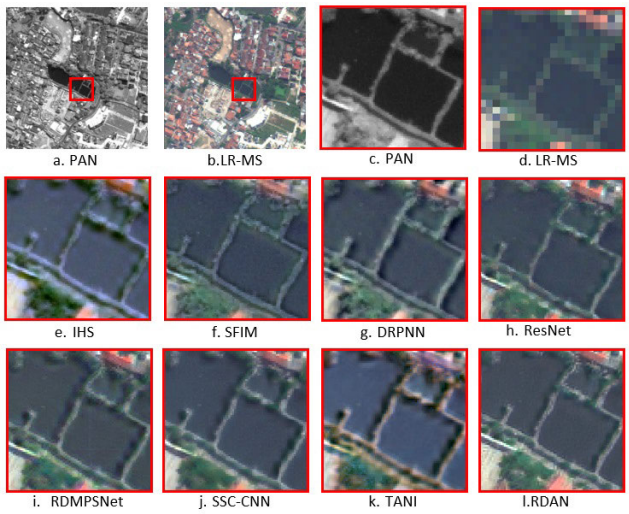

FIGURE 10. Pansharpening results in a full-resolution WorldView-3 image. (a) PAN image; (b) Observed MS image; (c) enlargement of the red area in the PAN image; (d) enlargement of the red area in the MS image; (e-k) enlargement of the pansharpening results of IHS, SFIM, DRPNN, ResNet, RDMPSNet, SSC-CNN, TANI, and the proposed method (RDAN), respectively.

The visualization results of the experiments at full-resolution are shown in Fig. 10, Fig. 11, and Fig. 12. For a better visual comparison, we enlarged the red areas in the figures. From the WorldView-3 and Pleiades data, as shown in Fig. 10 and Fig. 12, respectively, the pansharpened images from the proposed method are clearer and provide better spatial resolution than those from the other methods. From the experiments on the Spot-7 data, Fig. 11c and Fig. 11d show that the classical methods based on component substitution (IHS) provided better spatial resolution than the multiresolution analysis-based methods (SFIM). However, artifacts and blurs were still generated near the edges. Fig. 11c shows that classical methods based on component substitution (IHS) suffered from spectral and color distortions. In contrast, Fig. 11d to Fig. 11j show that classical method based on multiresolution analysis (SFIM) and deep learning-based methods performed well in terms of spectral aspects. They could preserve the color and spectral information in the pansharpened image. Fig. 10 to Fig. 12 also show that deep learning-based methods performed better than non-deep learning approaches, and our proposed method (RDAN) provided the best results with appropriate spectral and spatial resolutions.

B. EVALUATION OF CHANNEL ATTENTION AND SPATIAL ATTENTION MODULES

The experiments were conducted under three conditions to demonstrate the effects of the CA and SA modules. First, we conducted an experiment using a complete network, as illustrated in Fig. 4. The second step was to remove the CA module from the network, and the last step was to remove the SA module. The experimental results for the full-resolution data are presented in Table 5. The table shows that the network with channel attention without spatial attention performed better for all datasets than the network with only spatial attention. The experimental results also showed that the combination of channel and spatial attention outperformed a network with only channel or spatial attention. These comparisons demonstrated that utilizing both types of attention was effective for MS image pansharpening.

TABLE 5. Effect of channel and spatial attention modules.

Satellite	Methods	D_λ (↓)	D_s (↓)	QNR (↑))
WorldView-3	only SA	0.0282	0.0361	0.9367
	only CA	0.0334	0.0354	0.9324
	Proposed network	**0.0225**	**0.0277**	**0.9504**
Spot-7	only SA	0.0410	0.0454	0.9155
	only CA	0.0365	0.0323	0.9324
	Proposed network	**0.0246**	**0.0314**	**0.9448**
Pleiades	only SA	0.0638	0.0457	0.8934
	only CA	0.0634	**0.0426**	0.8967
	Proposed network	**0.0620**	0.0436	**0.8971**

C. EVALUATION OF LONG AND SHORT SKIP CONNECTIONS

To evaluate the effect of skip connections, we first removed the SSC in all RDAM modules and then removed the LSC from the proposed network. Table 6 presents the results. The table shows that the performance decreased significantly when the SSCs or LSC were removed from the network. For example, when the LSC was removed, the QNR decreased from 0.9504 to 0.8880 in the WorldView-3 image and from 0.9448 to 0.7218 in the Spot-7 image. The best D_λ values achieved by the proposed method indicated that the long

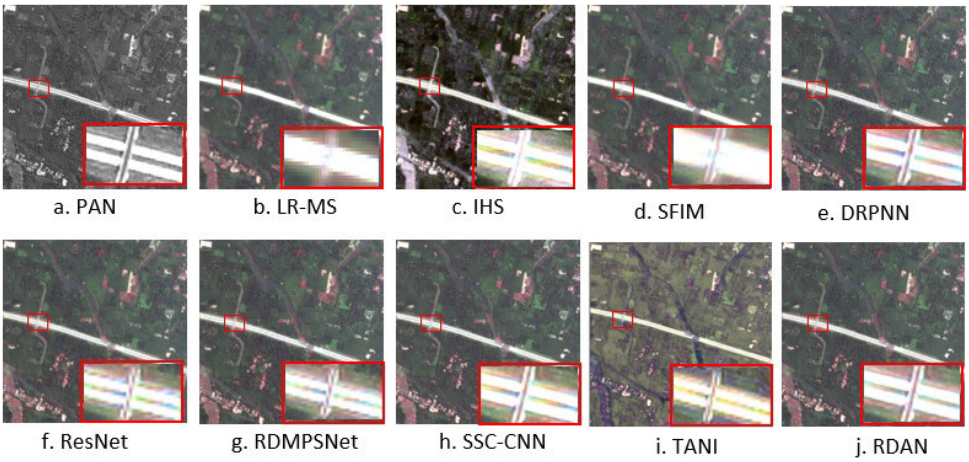

FIGURE 11. Pansharpening results in a full-resolution Spot-7 image. (a) PAN image; (b) Observed MS image; (c) IHS; (d) SFIM; (e) DRPNN; (f) ResNet; (g) RDMPSNet; (h) SSC-CNN; (i) TANI; (j) The proposed method (RDAN).

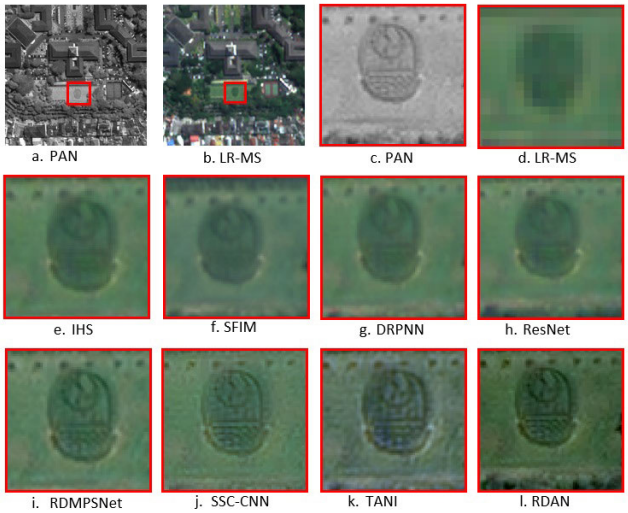

FIGURE 12. Pansharpening results in a full-resolution Pleiades image. (a) PAN image; (b) Observed MS image; (c) enlargement of the red area in the PAN image; (d) enlargement of the red area in the MS image; (e-k) enlargement of the pansharpening results of IHS, SFIM, DRPNN, ResNet, RDMPSNet, SSC-CNN, TANI, and the proposed method (RDAN), respectively.

and short skip connections are advantageous for improving spectral information. The LSC made the spectral information of the MS input image was directly used to supplement the spectral information of the pansharpened image. The D_s indicator of the proposed method shows that the use of skip connections could address the problem of spatial information loss. The experimental results verified that the combination of short and long skip connections enhanced the spectral and spatial details of the fused image.

D. EVALUATION OF FEATURE EXTRACTION MODULE

In this study, we evaluated the effect of using the SFE module in a dual-branch neural network and compared it with that of a single-branch neural network (without the SFE module). The results are presented in Table 7. Based on the quantitative metrics, the fused image generated using the feature extraction module performed better than those generated using a single-branch neural network. Compared to the single-branch feature extraction network, this dual-branch network exhibited enhanced performance in reducing redundant information across different MS image channels and better utilized the spatial information in the MS and PAN images.

TABLE 6. Effects of long and short skip connections.

Satellite	Methods	D_λ (↓)	D_s (↓)	QNR (↑))
WorldView-3	Without LSC	0.0830	0.0317	0.8880
	Without SSC	0.0775	0.0707	0.8573
	Proposed network	**0.0225**	**0.0277**	**0.9504**
Spot-7	Without LSC	0.1558	0.1449	0.7218
	Without SSC	0.0466	0.0351	0.9200
	Proposed network	**0.0246**	**0.0314**	**0.9448**
Pleiades	Without LSC	0.0711	0.0582	0.8748
	Without SSC	0.0682	**0.0418**	0.8929
	Proposed network	**0.0634**	0.0426	**0.8967**

TABLE 7. Effect of the feature extraction module.

Satellite	Methods	D_λ (↓)	D_s (↓)	QNR (↑))
WorldView-3	Without SFE	0.0254	0.0567	0.9193
	Proposed network	**0.0225**	**0.0277**	**0.9504**
Spot-7	Without SFE	0.1558	0.1449	0.7218
	Proposed network	**0.0246**	**0.0314**	**0.9448**
Pleiades	Without SFE	0.0711	0.0582	0.8748
	Proposed network	**0.0638**	**0.0457**	**0.8934**

E. EVALUATION OF COMPUTATIONAL COST

Generally, there is a trade-off between the quality and computational burden. A more extensive network, indicated by a more significant number of parameters, provides better

performance but increases computational cost [9]. In this study, we evaluated the effectiveness of the proposed method by investigating the number of parameters or model size, computation time, and QNR of a pansharpened image at full-resolution. We assessed the computational cost using a remote sensing image acquired by a Geoeye sensor. The results are summarized in Table 8.

TABLE 8. The comparison of parameter numbers, QNR, and computation time of four deep learning-based methods in Geoeye image.

Methods	Number of parameters (M)	Computation time (s)	QNR(↑)
DRPNN	*0.30*	**384**	0.9001
Resnet	0.90	3,376	0.8910
RDMPSNet	1.08	1,167	*0.9295*
SSC-CNN	1.34	997	0.912
TANI	**0.17**	2,797	0.8934
RDAN	0.78	*725*	**0.9393**

Table 8 shows that there is no direct correlation between number of parameters with computational cost and quality of pansharpened image in our experiments using Geoeye image. The experimental results show that the model size is not the only factor affecting the computation burden and model performance. Many factors affected the deep learning computation time and model performance, including the model framework, size, and optimization process. The DRPNN had the quickest computing time in our experiments, but the performance could have been better. TANI had the most minor parameters; however, this method performed more poorly and required significant computation time. The proposed method (RDAN) offered the largest *QNR* with the second-lowest computation time. These results demonstrated that our proposed method can perform better with an acceptable computational cost.

V. CONCLUSION

This study proposed a pansharpening method to learn the spectral and spatial information in source images more efficiently by combining channel and spatial attention in residual connections. The combination of channel and spatial attention empirically verifies that exploiting both is superior to using only a single type of attention. The residual connections within the low and high layers allow lower-level information to propagate to the higher layer, helping the network to preserve information that would have been lost through the training process with many layers. Using a feature extraction module before fusing multispectral and panchromatic images also improves pansharpening performance. A deep internal learning strategy is used in this experiment. This approach can enhance the spatial resolution of multispectral images without requiring a large amount of external training data. However, it only uses many small patches extracted from a large scene of the observed image itself. This strategy overcomes the problem of the unavailability of extensive training data.

To evaluate the performance of our proposed method, we conducted experiments on downgraded and full-resolution remote sensing images of Bandung, Indonesia, taken from the WorldView-3, Spot-7, Pleiades, and Geoeye satellites. The experimental results show that the proposed method performs better in terms of spatial and spectral evaluation metrics with an acceptable computational cost. In the future, we will explore the potential of implementing the proposed method for pansharpening with other sensors such as Ikonos, Gaofen, Quickbird, or Landsat, including pansharpening hyperspectral images.

Applying deep learning techniques has proven effective in improving the spatial resolution of multispectral remote sensing images. However, some specific issues still need to be solved. For example, the generalization ability of the most recent deep learning approaches for pansharpening still need to be improved. To solve this problem, we will explore a fine-tuning technique to enhance the generalization ability of the network in the following study. Furthermore, we will explore unsupervised learning methods and generative adversarial networks, combined with attention mechanisms for better performance to overcome the situation in which reference images are unavailable.

ACKNOWLEDGMENT

The computation in this study was performed using the Mahameru high-performance computing facilities of the National Research and Innovation Agency (BRIN), Indonesia.

REFERENCES

[1] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of pansharpened urban satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 1, pp. 281–297, Feb. 2012.

[2] A. Khalel, O. Tasar, G. Charpiat, and Y. Tarabalka, "Multi-task deep learning for satellite image pansharpening and segmentation," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 4869–4872.

[3] V. Tarverdiyev, I. Erer, N. H. Kaplan, and N. Musaoglu, "Target detection in multispectral images via detail enhanced pansharpening," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 1544–1547.

[4] Z. Cui, J. Leng, Y. Liu, T. Zhang, P. Quan, and W. Zhao, "SKNet: Detecting rotated ships as keypoints in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8826–8840, Oct. 2021.

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

[6] G. Vivone, L. Alparone, J. Chanussot, M. D. 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.

[7] Z. Wang, Y. Ma, and Y. Zhang, "Review of pixel-level remote sensing image fusion based on deep learning," *Inf. Fusion*, vol. 90, pp. 36–58, Feb. 2023.

[8] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926.

[9] 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.

[10] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman, "An adaptive IHS pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746–750, Oct. 2010.

[11] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. channel ratio and 'chromaticity' transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, Aug. 1987.

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

[13] P. S. C. Jr and A. Y. Kwarteng, "Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis," *Remote Photogrammetric Eng. Remote Sens.*, vol. 55, no. 3, pp. 1–20, 1989.

[14] H. Hallabia, A. Kallel, and A. B. Hamida, "Image pansharpening: Comparison of methods based on multiresolution analysis and component substitution," in *Proc. 1st Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Mar. 2014, pp. 25–30.

[15] 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.

[16] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

[17] 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, Jan. 2000.

[18] R. Fernandez-Beltran, P. Latorre-Carmona, and F. Pla, "Single-frame super-resolution in remote sensing: A practical overview," *Int. J. Remote Sens.*, vol. 38, no. 1, pp. 314–354, Jan. 2017.

[19] J. Fu, Y. Liu, and F. Li, "Single frame super resolution with convolutional neural network for remote sensing imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2018, pp. 8014–8017.

[20] L. Liebel and M. Körner, "Single-image super resolution for multispectral remote sensing data using convolutional neural networks," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 883–890, Jun. 2016.

[21] S. Nakazawa and A. Iwasaki, "Super-resolution imaging using remote sensing platform," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 1987–1990.

[22] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-resolution of remote sensing images via a dense residual generative adversarial network," *Remote Sens.*, vol. 11, no. 21, p. 2578, Nov. 2019.

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

[24] G. He, S. Xing, Z. Xia, Q. Huang, and J. Fan, "Panchromatic and multi-spectral image fusion for new satellites based on multi-channel deep model," *Mach. Vis. Appl.*, vol. 29, no. 6, pp. 933–946, Aug. 2018.

[25] J. Hu, Z. He, and J. Wu, "Deep self-learning network for adaptive pansharpening," *Remote Sens.*, vol. 11, no. 20, p. 2395, Oct. 2019.

[26] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.

[27] 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," *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020.

[28] H. V. Nguyen, M. O. Ulfarsson, J. R. Sveinsson, and J. Sigurdsson, "Zero-shot Sentinel-2 sharpening using a symmetric skipped connection convolutional neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Sep. 2020, pp. 613–616.

[29] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sens.*, vol. 10, no. 8, p. 1290, Aug. 2018.

[30] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision—ECCV*, vol. 8689. Cham, Switzerland: Springer, 2014, pp. 184–199.

[31] Y. Wei and Q. Yuan, "Deep residual learning for remote sensed imagery pansharpening," in *Proc. Int. Workshop Remote Sens. Intell. Process. (RSIP)*, May 2017, pp. 1–4.

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

[33] D. S. Vinothini and B. S. Bama, "Residual dense network for pan-sharpening satellite data," *IEEE Sensors J.*, vol. 19, no. 24, pp. 12279–12285, Dec. 2019.

[34] 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.

[35] X. Liu, X. Liu, H. Dai, X. Kang, A. Plaza, and W. Zu, "Mun-GAN: A multiscale unsupervised network for remote sensing image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402418.

[36] M. Huang, S. Liu, Z. Li, S. Feng, D. Wu, Y. Wu, and F. Shu, "Remote sensing image fusion algorithm based on two-stream fusion network and residual channel attention mechanism," *Wireless Commun. Mobile Comput.*, vol. 2022, pp. 1–14, Jan. 2022.

[37] W. Zhang, J. Li, and Z. Hua, "Attention-based tri-UNet for remote sensing image pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3719–3732, 2021.

[38] A. Shocher, N. Cohen, and M. Irani, "Zero-shot super-resolution using deep internal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3118–3126.

[39] T. Liu, R. Luo, L. Xu, D. Feng, L. Cao, S. Liu, and J. Guo, "Spatial channel attention for deep convolutional neural networks," *Mathematics*, vol. 10, no. 10, p. 1750, May 2022.

[40] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2018, pp. 3–19.

[41] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Computer Vision—ECCV*. Cham, Switzerland: Cham, Switzerland: Springer, 2018, pp. 294–310.

[42] F. Salvetti, V. Mazzia, A. Khaliq, and M. Chiaberge, "Multi-image super resolution of remotely sensed images using residual attention deep neural networks," *Remote Sens.*, vol. 12, no. 14, p. 2207, Jul. 2020.

[43] Z. Li, J. Li, F. Zhang, and L. Fan, "CADUI: Cross-Attention-Based depth unfolding iteration network for pansharpening remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5402420.

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

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

[46] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.

[47] Q. Liu, L. Han, R. Tan, H. Fan, W. Li, H. Zhu, B. Du, and S. Liu, "Hybrid attention based residual network for pansharpening," *Remote Sens.*, vol. 13, no. 10, p. 1962, May 2021.

[48] 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.

[49] F. A. Kruse, A. B. Lefkoff, J. W. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. Goetz, "The spectral image processing system (sips) interactive visualization and analysis of imaging spectrometer data," *Remote Sens. Environ.*, vol. 44, no. 2, pp. 145–163, 1993.

[50] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 245–253, May 1989.

[51] K. Wang, G. Qi, Z. Zhu, and Y. Chai, "A novel geometric dictionary construction approach for sparse representation based image fusion," *Entropy*, vol. 19, no. 7, p. 306, Jun. 2017.

[52] K. Nikolakopoulos and D. Oikonomidis, "Quality assessment of ten fusion techniques applied on Worldview-2," *Eur. J. Remote Sens.*, vol. 48, no. 1, pp. 141–167, Jan. 2015.

[53] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.

RIKA SUSTIKA received the bachelor's and master's degrees in electrical engineering from Bandung Institute of Technology (ITB), Indonesia, where she is currently pursuing the Ph.D. degree with the School of Electrical and Informatics Engineering (STEI), specializing in research on the utilization of deep learning techniques for remote sensing image super-resolution. She is a Researcher with the Research Center for Artificial Intelligence and Cybersecurity, National Research and Innovation Agency (BRIN), Indonesia. Her research interests include signal and image processing, machine learning, and artificial intelligence.

ANDRIYAN B. SUKSMONO (Senior Member, IEEE) received the B.Sc. degree in physics and the M.T. degree in electrical engineering from Bandung Institute of Technology (ITB), Bandung, Indonesia, in 1990 and 1996, respectively, and the Ph.D. degree in engineering from The University of Tokyo, Japan, in 2022. From 1996 to 2005, he was with ITB, as an Instructor; from 2005 to 2009, he was an Associate Professor; and since 2009, he has been a Professor. He joined the School of Electrical Engineering and Informatics, ITB. His research interests include signal processing, imaging, and quantum computing. He is currently a Professional Member of ACM. He was a recipient of several international research funds, grants, scholarships, and fellowships from RCAST-Tokyo University, Monbukagakusho, JSPS, AIGRP, Asahi Glass, and the Hitachi Foundation. In 2004, he received the Best Paper Award in Theoretical Development from APNNA, for his work on spatiotemporal ultra-wideband neuro-beamforming. He was also a recipient of the Outstanding Faculty Award of ITB, in June 2007; the Outstanding Faculty Award of the Republic of Indonesia, in August 2007; the ITB Innovation Award, in 2017; the ITB Research Award, in 2020; the 114 Indonesia Innovation Award, in 2022; and the 115 Indonesia Innovation Award, in 2023. His work on quantum algorithms for finding the Hadamard matrix was included in Nature's Scientific Report 100 in Physics 2020 (eighth place) and 2023 (second place), Editor's Choice (2023), and is listed in the Quantum Algorithm Zoo.

DONNY DANUDIRDJO (Member, IEEE) received the bachelor's and master's degrees in electrical engineering from Bandung Institute of Technology (ITB), Indonesia, and the Ph.D. degree in electrical engineering from the University of Tokyo, Japan, in 2013. He is currently a Lecturer with the Research Group of Biomedical Engineering, School of Electrical and Informatics Engineering, ITB. His research interests include medical imaging and signal processing.

KETUT WIKANTIKA was the Head of the Remote Sensing and GIS Research Group, Institut Teknologi Bandung (ITB), Indonesia, from 2012 to 2023, and the Director of the Center for Remote Sensing (CRS), ITB, from 2005 to 2020. He has cooperated with numerous institutions, such as Chiba University, Tottori University, Nagoya University, Kochi University, Japan International Research Center for Agricultural Sciences, University of Oklahoma, Asian Institute of Technology, University of Salzburg, Universiti Teknologi Malaysia, and The Pennsylvania State University. He was the Founder of ForMIND (Indonesian Young Researcher Forum) and the ForMIND Insitute, in 2013 and 2016, respectively. He is currently a Professor of environmental remote sensing with the Faculty of Earth Science and Technology, ITB. His research interests include geospatial approaches, including artificial intelligence (AI) application for demography, land use-land cover and its changes, biogeography and biodiversity, and disaster. In the past, he also served on Indonesian Society for Remote Sensing, as the Secretary General, from 2003 to 2006, and the General Chairperson, from 2006 to 2009.