D2TNet: A ConvLSTM Network With Dual-Direction Transfer for Pan-Sharpening

Meiqi Gong, Jiayi Ma, Senior Member, IEEE, Han Xu, Graduate Student Member, IEEE, Xin Tian, Member, IEEE, and Xiao-Ping Zhang, Fellow, IEEE

Abstract—In this article, we propose an efficient convolutional long short-term memory (ConvLSTM) network with dual-direction transfer for pan-sharpening, termed D2TNet. We design a specially structured ConvLSTM network that allows for dual-directional communication, including multiscale information and multilevel information. On the one hand, due to the sensitivity of spatial information to scales and the sensitivity of spectral information to levels, multiscale and multilevel information is extracted to facilitate the fuller use of source images. On the other hand, ConvLSTM is employed to capture the strong dependencies between multiscale information and multilevel information. Besides, we introduce a multiscale loss to enable different scales contributing to each other to generate high-resolution multispectral images that are closer to the ground truth. Extensive experiments, including qualitative evaluation, quantitative evaluation, and efficiency comparison, are implemented to verify that our D2TNet outperforms state-of-the-art methods indeed.

Index Terms—Convolutional long short-term memory (ConvLSTM), dual-direction transfer, multiscale loss, pan-sharpening.

I. INTRODUCTION

Due to the powerful ground survey capabilities of satellites, remote sensing images captured by their sensors contain a wealth of ground information. Low-resolution multispectral (LRMS) images and panchromatic (PAN) images are two common captured modalities. The former possesses high spectral resolution but low spatial resolution [1], while the latter presents the opposite characteristics [2]. To meet the needs of some practical applications (e.g., land survey [3], environmental monitoring [4], and object detection [5]), pan-sharpening methods fuse the captured LRMS image and PAN image to produce desirable high-resolution multispectral (HRMS) image. Because of the excellent characteristics of the generated HRMS image, pan-sharpening has become a research hotspot in the field of remote sensing image processing.

Over the past few decades, the field of pan-sharpening has caught more and more attention. Various traditional methods have been proposed to solve the pan-sharpening problem. In general, traditional pan-sharpening methods can be largely classified into four categories: component substitution (CS)-based methods [6], multiresolution analysis (MRA)-based methods [7], CS/MRA hybrid-based methods [8], and model-based methods [9]. Due to the complexity of ground objects and the diversity of spectral features captured by different sensors, handcrafted designs of traditional methods make it difficult to establish a link between original images and the targeted HRMS image.

Fortunately, in the past few years, as the powerful feature extraction capabilities and nonlinearity of neural networks, deep learning has come into the limelight and has been introduced to a wide variety of tasks [10]–[12], including image fusion. Deep-learning-based pan-sharpening methods can be classified into a convolutional neural network (CNN)-based methods and generative adversarial network (GAN)-based methods. Most CNN-based methods construct networks to extract features, fuse them, and reconstruct HRMS. Encoder–decoder networks, densely convolutional networks, and residual convolutional networks are commonly used network structures. On this basis, GAN-based methods introduce a generator and a discriminator to implement the fusion process through the min–max game between them without ground truth [13]. Training with large amounts of data, both CNN- and GAN-based methods have the ability to build more robust nonlinear mappings from source images to target images, so as to escape the limitations of traditional methods and achieve state-of-the-art performance.

Although current deep learning-based pan-sharpening algorithms have achieved impressive results, there are still some pressing issues that need to be addressed. On the one hand, most previous works typically input the LRMS and PAN images with the original size to the network directly. However, features of different ground objects captured from different sensors exhibit large-scale differences [14]. Therefore, images at different scales can contain partially nonoverlapping information. Considering multiscale information and enhancing the interaction between them facilitate better use of multiscale information from the source images, so as to force the fusion
results to maintain richer features. On the other hand, although a few pan-sharpening methods consider multiscale information, they tend to relate information of different scales and different levels through dense blocks or Resblocks. However, information between different scales and different levels exists strong dependencies. Indiscriminate delivery of information leads to an increase of invalid or redundant information while reducing the status of valid information. How to transfer information correctly is an issue worth considering.

To inherit the advantages of deep learning and cope with the issues aforementioned, we propose an efficient pan-sharpening method with a dual-direction transfer, termed D2TNet. Specifically, the dual-direction transfer includes multiscale and multi-level information interaction. Combining the advantages of convolutional long short-term memory (ConvLSTM) in dealing with long-term information dependencies, an eight-shaped (as shown in Fig. 1) ConvLSTM network is designed to better solve the issues of dual-direction information interaction. This specific structure utilizes three gates in ConvLSTM to implement long-term information interaction between different scales and levels. It enables fuller use of original information, so as to achieve richer spatial details and more realistic spectral features. In addition to the eight-shaped ConvLSTM structure, we introduce three-scale information loss in the total loss functions, which promotes the generation of HRMS images with spatial and spectral distributions closer to the ground truth.

Our contributions can be concluded as follows.

1) We propose an effective-information dual-direction transfer pan-sharpening method by a specific ConvLSTM structure, which enables long-term information interaction between different scales and levels, thus making fuller use of original information and achieving richer spatial details and more realistic spectral features.

2) We introduce a novel loss function, including three-scale loss, which enhances the consistency of the fusion result and the ground truth.

3) Extensive experiments are conducted to verify that our D2TNet stands out from the state-of-the-art methods while possessing high efficiency.

II. RELATED WORK

A. Deep-Learning-Based Pan-Sharpening Methods

Recently, with the development of deep learning in the image processing field, deep learning-based pan-sharpening methods are gaining popularity. These methods can be broadly classified into CNN-based methods and GAN-based methods.

Inspired by SRCNN [15], a CNN-based method processing image super-resolution, Masi et al. [16] introduced PNN to solve pan-sharpening problems, which is the first CNN-based pan-sharpening method. It stacks the interpolated LRMS and original PAN images, and uses them as the input to generate HRMS images. PNN possesses a simple network, thus achieving high efficiency. In addition, Liu et al. [17] proposed TFNet, which exploits the feature extraction ability of CNN. It constructs an encoder–decoder network to implement feature extraction, feature fusion, and reconstruction process. Xu et al. [18] proposed SDPNet focusing on spatial information and spectral information. Specifically, it designs a spatial encoder–decoder and a spectral encoder–decoder to select the unique feature maps between the two original images. In addition, Wang et al. [19] introduced MPNet based on ConvLSTM. It exploits the original ConvLSTM to fuse features of LRMS and PAN images at different levels while not taking full advantage of ConvLSTM to drive the fused images to contain more effective information. Besides the methods mentioned above, there are also some methods based on multiscale features. Wang et al. [20] proposed MSDRN, which is a multiscale deep residual network. It downsamples the concatenated original images to different scales and links them through upconvolution and concatenation. Xu et al. [21] proposed a multiscale network named CPNet. It first downsamples PAN images two and four times, and upsamples LRMS for corresponding times to get three groups of different scales as inputs. In our method, we follow the way of obtaining multiscale images in CPNet. However, the way they relate different scale images is “pixel shuffle,” which is a subjective human decision, leading to a risk of losing information. After that, Jin et al. [22] proposed a novel pan-sharpening method using the Laplacian pyramid to separate images to different scales. For each scale, it designs a fusion CNN to get the fusion results. However, it related multiscale features only by sharing parameters, a link that appears to be weak and insufficient to fully exploit multiscale features. In addition, multilevel information transfer in the abovementioned methods is realized by dense blocks or Resblocks, which ignores the relationships between shallow levels and deep levels.

Different from CNN-based methods, GAN-based methods fulfill fusion through an adversarial process between the generator and the discriminator. Liu et al. [23] proposed PSGAN, which is the first time introducing GAN to pan-sharpening. It designs a generator to fuse PAN and MS images,
and leverages a discriminator to reduce the gap between the fused image and the ground truth. After that, RED-cGAN was proposed by Shao et al. [24] using a residual encoder–decoder network. The design of the conditional discriminator can further supplement the spatial information in the final results. Furthermore, Ma et al. [25] presented Pan-GAN using double discriminators, which is an unsupervised method without ground truth. Double discriminators force the result to look like both the PAN and the LRMS images, thus possessing the spatial information of the PAN image and the spectral information of the LRMS image.

In the methods aforementioned, multiscale and multilevel information between the two original images is not exploited or is not appropriately related, either of them may result in spectral distortion or spatial distortion. In this article, we introduce a novel method considering effective multiscale and multilevel information communication to make fuller use of the original information.

B. Convolutional Long Short-Term Memory

Long short-term memory (LSTM) [26] is a network that excels in processing long sequences of memory problems. Compared to the normal network structure, LSTM changes the internal network structure by adding three gates, i.e., the input gate, the output gate, and the forget gate. The input gate does a nonlinear transformation of two elements, including the output of the previous timestamp and the input of the current timestamp, to obtain the new input. The forget gate selectively updates the state vector based on the state of the previous timestamp and the current timestamp. The output gate controls the output of the current timestamp based on the forget gate.

When the temporal data are 3-D images, the ordinary LSTM is difficult to describe the complex spatial characteristics between points. ConvLSTM is introduced to better describe the spatiotemporal relationship between images. It was first proposed by Xingjian et al. [27], and the authors experimentally verified that ConvLSTM is better than LSTM in obtaining spatiotemporal relations.

Owing to the success of ConvLSTM in transferring image information, it has been broadly involved in image processing fields, including image classification [28], image segmentation [29], [30], and so on. Only Wang et al. [19] introduced MPNet based on ConvLSTM to address the pan-sharpening problem. Nevertheless, they exploited the original ConvLSTM to fuse features of LRMS and PAN images at different levels, without taking full advantage of ConvLSTM to drive the fused images to contain more effective information. Since ConvLSTM can reasonably screen for useful information and pass it on to the next timestamp, we utilize it to enhance information communication between multiscales and multilevels.

III. PROPOSED METHOD

In this section, we introduce our D2TNet in detail. We first give the problem formulation of our method and then present the network architecture and loss functions.

A. Problem Formulation

On the one hand, it is necessary to extract hierarchical features at different levels as they contribute to more comprehensive representations of the original information. Besides, the deep low-frequency features extracted by the CNN can be seen as a further extraction of the shallow high-frequency features; the deep level has a strong dependency on the shallow level. As such, we design multilevel ConvLSTM to capture their discrepancy, so as to learn more accurate hierarchical spectral features. On the other hand, since spatial details and spectral features at different scales ponder divergences, relating multiscale information is conducive to maintaining richer spatial details and more realistic spectral features. Besides, low-scale and high-scale information relies on each other for the same reason. Therefore, we also design multiscale ConvLSTM to relate multiscale features so as to maintain richer spatial details and more realistic spectral features.

Thereby, in order to make better use of original information and interact with multiscale and multilevel information effectively, we take advantage of the outstanding performance of ConvLSTM in transmitting information and propose a pan-sharpening method with dual-direction (multiscale and multilevel) transfer via a ConvLSTM network, termed D2TNet.

The whole framework is shown in Fig. 2. Multiscale images are produced to obtain hierarchical information. Specifically, LRMS images are upsampled to get LRMS↑2 and LRMS↑4. Similarly, PAN images are downsampled to get PAN↓2 and PAN↓4. Three sets of images of the same scale are concatenated and fed into the three-stream (top, middle, and bottom) network, respectively, as shown in Fig. 2.

To achieve our goal, the eight-shaped ConvLSTM network is designed to link information at different scales and different levels. In order to feed the same type of features to the ConvLSTM network, we let the convolutional layers right before ConvLSTM share parameters. In addition, since our loss functions utilize all products of three-stream networks, the last convolutional layers also share parameters to guarantee that the middle and the bottom stream networks contribute to the generation of HRMS.

B. Network Architectures

Combining considerations mentioned in Section III-A, the final network structures are determined as Fig. 2. Network parameters of the top stream network are given in Fig. 3. In fact, in the top, middle, and bottom stream networks, the corresponding convolutional layers have the same number of input or output channels and differ only in their scale sizes. Briefly, we just present the network parameters of the top stream network.

Three parameters in Conv(·) represent the kernel size, numbers of input channels, and output channels, respectively. The activation function for all the convolutional layers is leaky rectified linear unit (ReLU) (lrelu) except for the last layer, which employs tanh. Three parameters in ConvLSTM(·) represent numbers of units, input channels for the first unit, and output channels for the last unit, respectively. More specifically, each unit has the same input channels 32 and the same output channels 32, which makes it prone to transmit states between multiscale and multilevel. In addition, the residual
network is exploited during implement process due to the advantage of learning efficiency.

For each unit of ConvLSTM, its internal network architecture is shown in Fig. 4. The calculation process can be formulated as follows:

\[ i_t = \sigma_i(W_{xi} \ast X_t + W_{hi} \ast H_{t-1} + W_{ci} \ast C_{t-1} + b_i) \]  
\[ f_t = \sigma_f(W_{xf} \ast X_t + W_{hf} \ast H_{t-1} + W_{cf} \ast C_{t-1} + b_f) \]  
\[ C_t = f_t \cdot C_{t-1} + i_t \cdot \tanh(W_{xc} \ast X_t + W_{hc} \ast H_{t-1} + b_c) \]  
\[ o_t = \sigma_o(W_{xo} \ast X_t + W_{ho} \ast H_{t-1} + W_{co} \ast C_t + b_o) \]  
\[ H_t = o_t \cdot \tanh(C_t) \]

where \( \cdot \) represents the multiplication and \( \ast \) represents the convolution.

In our method, \( X_t \) denotes the input of this unit, and \( C_{t-1}, H_{t-1}, C_t, \) and \( H_t \) are symbols representing the state of the last unit, the output of the last unit, the state of this unit, and the output of this unit, respectively. When this unit is the first one, we set \( C_{t-1} \) and \( H_{t-1} \) as all-zero, which are also called the initial state. From Fig. 2, we find that a unit has the possibility to have two input states, e.g., unit5 of ConvLSTM1; it receives not only the state passed from unit2 but also the state passed from unit4. In this case, unit4 is first upsampled to the size of unit2; then, we add up all the input states to get the final input state. The specific operation of each unit follows (1)–(5).

C. Loss Functions

Our loss functions contain three parts corresponding to three-stream networks. Compared with the traditional constraint only on the fused image, this constraint is stronger and allows the final fused image to be closer to the ground truth. The whole loss function can be expressed as

\[ L_{all} = L_{top} + \lambda_1 L_{middle} + \lambda_2 L_{bottom} \]  

Fig. 2. Framework of our D2TNet. The PAN and LRMS images in the pink shadows are the original images. The HRMS image in the blue shadow is the fused image. \( C_0 \) and \( H_0 \) are the initial state and the initial input. In this article, all-zero matrices are employed for them.

Fig. 3. Parameters of the top stream network of D2TNet. Three parameters in Conv\((\cdot)\) represent the kernel size, channels of the input, and channels of the output, respectively. Three parameters in ConvLSTM\((\cdot)\) represent numbers of units, channels of the input for the first unit, and channels of the output for the last unit, respectively.

Fig. 4. Internal network architecture of the ConvLSTM units. \( \sigma_f, \sigma_i, \) and \( \sigma_o \) represent the sigmoid functions corresponding to the forget gate, the input gate, and the output gate, respectively. \( \otimes \) means the elementwise multiplication, and \( \oplus \) means the elementwise summation.
where $L_{\text{top}}$, $L_{\text{middle}}$, and $L_{\text{bottom}}$ represent the loss functions of those three stream networks, respectively. $\lambda_1$ and $\lambda_2$ are employed to access a tradeoff among the three parts in (6).

1) **Loss Function of Top Stream Network:** For the top stream network, we expect the generated HRMS to approach the ground truth as much as possible. We constrain the generation of HRMS from both spectral and spatial perspectives. Specifically, we use structural similarity (SSIM) index measurement and the Frobenius norm to constrain the similarity of spectral information between HRMS and the ground truth, and the gradient loss to constrain the similarity of spatial details. In addition, to further constrain the features, we downsample the obtained HRMS to LRMS size and force their feature information to converge. Therefore, $L_{\text{top}}$ is determined as

$$L_{\text{top}} = \frac{1}{\text{HWC}} \sum_{H} \sum_{W} \sum_{C} (\|\text{HRMS} - G\|_F + \xi_1\|\text{HRMS} \downarrow 4 - \text{LRMS}\|_F + \xi_2\|\nabla \text{HRMS} - \nabla G\|_F + \xi_3(1 - \text{SSIM}(\text{HRMS}, G))) \quad (7)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. HRMS represents the generated image of the top stream network, which is also the final result. $G$ represents the ground truth, which is obtained according to Wald’s protocol introduced in [31]. $H$, $W$, and $C$ represent the height, the width, and the number of channels of the HRMS image, respectively. SSIM(·) stands for SSIM between two elements. $\xi_1$ and $\xi_2$ are employed to access a tradeoff among the four parts in (7).

2) **Loss Functions of Middle and Bottom Stream Networks:** For the middle and bottom stream networks, we constrain them in the same way as we process the top stream network. Their loss functions are displayed as follows:

$$L_{\text{middle}} = \frac{1}{\text{HWC}} \sum_{H} \sum_{W} \sum_{C} (\|\text{HRMS}_2 - G \downarrow 2\|_F + \xi_1\|\text{HRMS}_2 \downarrow 2 - \text{LRMS}\|_F + \xi_2\|\nabla \text{HRMS}_2 - \nabla G \downarrow 2\|_F + \xi_3(1 - \text{SSIM}(\text{HRMS}_2, G \downarrow 2))) \quad (8)$$

$$L_{\text{bottom}} = \frac{1}{\text{HWC}} \sum_{H} \sum_{W} \sum_{C} (1 + \xi_3)\|\text{HRMS}_4 - G \downarrow 4\|_F + \xi_2\|\nabla \text{HRMS}_4 - \nabla G \downarrow 4\|_F + \xi_3(1 - \text{SSIM}(\text{HRMS}_4, G \downarrow 4))) \quad (9)$$

where HRMS$_2$ and HRMS$_4$ represent the products of the middle and bottom stream networks, respectively. $G \downarrow 2$ and $G \downarrow 4$ are products downsampling ground truth to one-half or one-quarter of the original size.

IV. EXPERIMENTS AND EVALUATIONS

In this section, we first introduce some experimental details, including training data, hyperparameters, and experimental devices. After that, we exhibit some evaluation metrics preparing for the analysis of later experimental results. Then, we conduct extensive experiments to explain the superiority and stability of our D2TNet in terms of effectiveness and efficiency. In the end, we implement some ablation experiments to validate the necessity of some architecture designs, some loss components, and the settings of some parameters.

A. Experimental Details

1) **Training Details:** The QuickBird and GF-2 datasets are employed for evaluation. According to Wald’s protocol introduced in [31], PAN images and MS images are first downsampled, and the original MS images are used as the ground truth. In order to obtain more training data, we crop the downsampled PAN images and downsampled MS (LRMS) images into $264 \times 264 \times 1$ and $66 \times 66 \times 4$ size of patches. Correspondingly, the original MS images are cropped to patches with size $264 \times 264 \times 4$ to get the ground truth. All image patches are randomly cropped in addition to being rotated and color enhanced with a certain probability of 0.25. Totally 20000 pairs of images are gained for each training dataset.

In Section III-C, five hyperparameters ($\lambda_1$, $\lambda_2$, $\xi_1$, $\xi_2$, and $\xi_3$) are used to enable better training of the network. $\lambda_1$ and $\lambda_2$ are exploited to measure the importance of the three stream networks. The weights of the top and the bottom stream networks should be larger than that of the middle stream network; the reason is that the inputs of the top and the bottom stream networks contain original image information, i.e., information from the PAN image and information from the LRMS image. Based on the above considerations, $\lambda_1$ and $\lambda_2$ are assigned as 0.2 and 1, respectively. $\xi_1$, $\xi_2$, and $\xi_3$ are exploited to make a tradeoff among the different loss components, and they are eventually decided as 5, 40, and 40, respectively. The slope value setting of the lrelu activation function is 0.02, and the method used for upsampling is nearest-neighbor interpolation while for downsampling is max pooling.

The experiments are conducted on a 3.4-GHz Intel Core i5-7500 CPU and NVIDIA GeForce GTX Titan X GPU. We set the number of epochs as 100 and the batch size as 16. Tensorflow is chosen as the experimental platform.

2) **Evaluation Metrics:** In order to objectively evaluate the results, we employ seven evaluation metrics for reduced-resolution images and three evaluation metrics for full-resolution images. The seven metrics mentioned in the former include the relative dimensionless global error in synthesis (ERGAS) [32], the root mean square error (RMSE), the structural similarity coefficient (SSIM) [33], the Pearson correlation coefficient (CC), the relative average spectral error (RASE) [34], the spectral angle mapper (SAM) [35], and the visual information fidelity (VIF) [36]. The three metrics mentioned in the latter include the quality-with-no-reference (QNR) [37], $D_2$, and $D_4$. They are introduced in detail as follows.

1) **ERGAS** is a metric measuring dynamic range change between the fused image and the ground truth. A smaller ERGAS value means less spectral distortion. The calculation of ERGAS is mathematically expressed as

$$\text{ERGAS}_{b,f} = 100 \frac{\text{size}_p}{\text{size}_l} \left( \frac{1}{B} \sum_{b=1}^{B} \left( \frac{\text{RMSE}_{gb,bf}}{\mu(b)} \right)^2 \right)^{1/2} \quad (10)$$
where \( \text{size}_p/\text{size}_f \) represents the ratio between the spatial size of the PAN image and the LRMS image. The subscript of ERGAS represents the inputs required for the calculation. \( B \) stands for the number of LRMS bands, and \( b \) represents the \( b \)th band of LRMS images. \( \text{RMSE}_{g,b,f_0} \) represents the RMSE value between the \( b \)th ground truth band and the \( b \)th generated HRMS band, while \( \mu(b) \) represents the mean value of the \( b \)th LRMS band. The calculation of RMSE is described in the immediately following part.

2) RMSE describes the degree of the pixel difference between the fused image and the ground truth. A smaller RMSE value means that the fused image is closer to the ground truth in terms of the pixel value. It is calculated as

\[
\text{RMSE}_{g,f} = 100 \cdot \sqrt{\frac{1}{HWC} \sum_{c=1}^{C} \sum_{w=1}^{W} \sum_{h=1}^{H} [f(h, w, c) - g(h, w, c)]^2}
\]

where \( H, W, \) and \( C \) represent the height, the width, and the number of channels of the fused image, respectively.

3) SSIM calculates the structural similarity between the fused image and the ground truth. A higher SSIM value means that the fused image shows higher correspondence with the ground truth in terms of luminance, contrast, and structure. It can be mathematically formalized as

\[
\text{SSIM}_{g,f} = \sum_{g,f} \frac{2\mu_g \mu_f + c_1}{\mu_g^2 + \mu_f^2 + c_1} \cdot \frac{2\sigma_g \sigma_f + c_2}{\sigma_g^2 + \sigma_f^2 + c_2} \cdot \frac{\sigma_{g,f} + c_3}{\sigma_g \sigma_f + c_3}
\]

where \( \mu \) and \( \sigma \) represent the mean value and standard deviation, \( \sigma_{g,f} \) stands for the covariance value between \( g \) and \( f \), \( c_1, c_2, \) and \( c_3 \) are parameters stabilizing equations as the denominator approaches zero.

4) CC evaluates the spectral correlation between the fused image and the ground truth. A higher CC value represents stronger correlation. It is calculated as

\[
\text{CC}_{g,f} = \sum_{g,f} \frac{\sigma_{g,f}}{\sigma_g \sigma_f}.
\]

5) RASE measures spectral quality through calculating the relative spectral error between the fused HRMS image and the ground truth. A smaller RASE value means higher spectral quality. It is calculated as follows:

\[
\text{RASE}_{g,f} = \frac{100}{M} \sqrt{\frac{1}{B} \sum_{b=1}^{B} \text{RMSE}_{g,b,f_0}^2}
\]

where \( M \) is the mean radiance of the \( B \) spectral bands of the LRMS image.

6) SAM calculates the angle change between the ground truth and the fused HRMS image, so as to evaluate its spectral distortion. A smaller SAM value means higher spectral quality. It is calculated as follows:

\[
\text{SAM}_{g,f} = \cos^{-1} \left( \frac{g^T f}{\|g\| \|f\|} \right).
\]

7) VIF is proposed based on statistical models of natural images, models of image distortion, and models of the human visual system. Compared with other metrics, it maintains a higher degree of consistency with subjective visual effects. A higher VIF value means better visual. It is calculated as follows:

\[
\text{VIF}_{g,f} = \frac{\sum_b \text{MI}(r_N, f_b)}{\sum_b \text{MI}(r_N, g_b)}
\]

where \( r_N \) is \( N \) random elements and MI represents mutual information.

8) QNR is a comprehensive metric measuring both spatial distortion between the PAN image and the fused HRMS image, and spectral distortion between the LRMS image and the HRMS image. A higher QNR represents less overall distortion, so as to achieve higher quality. It is calculated as

\[
\text{QNR}_{l,p,f} = (1 - D_s)(1 - D_l)
\]

where \( l \) and \( p \) represent the original LRMS and PAN images, respectively. \( D_s \) and \( D_l \) are introduced in the following.

9) \( D_s \) is the metric measuring spectral distortion between the LRMS image and the fused HRMS image. A smaller \( D_s \) value means higher spectral consistency. It can be mathematically formalized as

\[
D_s(l, f) = \sqrt{\frac{1}{B(B - 1)} \sum_{i=1}^{B} \sum_{j=1}^{B} |Q(f_i, f_j) - Q(l_i, l_j)|}
\]

where \( f_i \) represents the \( i \)th band of the generated HRMS image and \( l_i \) represents the \( i \)th band of the LRMS image.

10) \( D_l \) is the metric measuring spatial distortion between the PAN image and the fused HRMS image. A smaller \( D_l \) value means higher spatial consistency. It can be mathematically formalized as

\[
D_l(l, f) = \sqrt{\frac{1}{B} \sum_{i=1}^{B} |Q(f_i, p) - Q(l_i, \bar{p})|}
\]

where \( \bar{p} \) represents the downsampled PAN image.

B. Results on QuickBird Dataset

In this section, we conduct experiments on the QuickBird dataset and compare our D2TNet with eight state-of-the-art methods, e.g., LGC [38], AVWP [39], CDIF [40], PNN [16], PanNet [41], SDPNet [18], PSGAN [23], and CPNet [21]. We first implement qualitative experiments to demonstrate the superiority of our D2TNet. Then, the first eight metrics and the last three metrics introduced in Section IV-A2 are used to further evaluate on reduced-resolution images and full-resolution images, respectively.
Fig. 5. Comparison of results for a group of reduced-resolution images from the QuickBird dataset, i.e., LRMS images with the size of $66 \times 66 \times 4$, PAN images with the size of $264 \times 264$, and the size of the results and the ground truth is $264 \times 264$. From (Left) to (Right), the first row presents the LRMS image, the PAN image, and the results of LGC, AVWP, CDIF, and PNN. The second row shows the results of PanNet, SDPNet, PSGAN, CPNet, our D2TNet, and the ground truth. The last line shows the residuals between different results and the ground truth. (a) LRMS. (b) PAN. (c) LGC. (d) AVWP. (e) CDIF. (f) PNN. (g) PanNet. (h) SDPNet. (i) PSGAN. (j) CPNet. (k) Ours. (l) Ground truth. (m) LGC. (n) AVWP. (o) CDIF. (p) PNN. (q) PanNet. (r) SDPNet. (s) PSGAN. (t) CPNet. (u) Ours.

TABLE I

<table>
<thead>
<tr>
<th>metrics</th>
<th>LGC</th>
<th>AVWP</th>
<th>PNN</th>
<th>PanNet</th>
<th>SDPNet</th>
<th>PSGAN</th>
<th>CPNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>reduced-resolution images</td>
<td>ERGAS$^*$</td>
<td>1.3601±0.3266</td>
<td>1.7185±0.4005</td>
<td>2.4799±0.6530</td>
<td>1.4052±0.3276</td>
<td>1.8271±0.3857</td>
<td>1.3947±0.2994</td>
<td>1.2840±0.3233</td>
</tr>
<tr>
<td>RMSE</td>
<td>3.0724±0.9411</td>
<td>3.8885±1.2380</td>
<td>5.5578±1.7818</td>
<td>3.1472±0.9229</td>
<td>4.0853±1.0964</td>
<td>3.1543±0.9735</td>
<td>3.1720±0.9735</td>
<td>3.0650±0.8873</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.9432±0.0256</td>
<td>0.9175±0.0332</td>
<td>0.8932±0.0432</td>
<td>0.9441±0.0252</td>
<td>0.9035±0.0375</td>
<td>0.9465±0.0246</td>
<td>0.9502±0.0230</td>
<td>0.9465±0.0236</td>
</tr>
<tr>
<td>CC$^*$</td>
<td>0.9625±0.0644</td>
<td>0.9385±0.0760</td>
<td>0.9270±0.0768</td>
<td>0.9618±0.0743</td>
<td>0.9395±0.0707</td>
<td>0.9613±0.0687</td>
<td>0.9561±0.0736</td>
<td>0.9624±0.0716</td>
</tr>
<tr>
<td>RASE</td>
<td>5.6288±1.8280</td>
<td>6.8899±1.7021</td>
<td>9.8377±2.7149</td>
<td>5.5956±1.3395</td>
<td>7.2737±1.6403</td>
<td>5.7064±1.2748</td>
<td>5.1467±1.3562</td>
<td>5.4321±1.2763</td>
</tr>
<tr>
<td>SAM$^*$</td>
<td>1.7651±0.5399</td>
<td>2.2167±0.6404</td>
<td>2.3171±0.6270</td>
<td>1.8866±0.3866</td>
<td>2.4736±0.4087</td>
<td>1.8352±0.5227</td>
<td>1.4203±0.4343</td>
<td>1.7809±0.4777</td>
</tr>
<tr>
<td>VIF</td>
<td>0.5970±0.1093</td>
<td>0.4850±0.0956</td>
<td>0.4833±0.0782</td>
<td>0.5560±0.1164</td>
<td>0.4308±0.0804</td>
<td>0.5375±0.1070</td>
<td>0.5834±0.1159</td>
<td>0.5338±0.1080</td>
</tr>
<tr>
<td>full-resolution images</td>
<td>QNR$^*$</td>
<td>0.5216±0.0835</td>
<td>0.3788±0.0912</td>
<td>0.4770±0.0973</td>
<td>0.4863±0.0932</td>
<td>0.4913±0.0800</td>
<td>0.4552±0.0830</td>
<td>0.4913±0.0806</td>
</tr>
<tr>
<td>$D_1^*$</td>
<td>0.0773±0.0243</td>
<td>0.0957±0.0453</td>
<td>0.1045±0.0235</td>
<td>0.0550±0.0249</td>
<td>0.0750±0.0211</td>
<td>0.0750±0.0273</td>
<td>0.0644±0.0269</td>
<td>0.0757±0.0233</td>
</tr>
<tr>
<td>$D_2^*$</td>
<td>0.4330±0.1093</td>
<td>0.5833±0.0954</td>
<td>0.4871±0.1081</td>
<td>0.4858±0.1021</td>
<td>0.4885±0.0922</td>
<td>0.5078±0.0888</td>
<td>0.4705±0.0973</td>
<td>0.4924±0.0988</td>
</tr>
</tbody>
</table>

1) Qualitative Analysis: We display two groups of reduced-resolution results for intuitive illustration, as shown in Figs. 5 and 6. In addition, Fig. 7 displays the results of full-resolution images. The residual results of different methods and the ground truth, as shown in the latter part of each figure, are calculated by absolute differences.

For reduced-resolution images, the results of our D2TNet demonstrate high spectral consistency with the ground truth, which is clearly shown in Fig. 5. The enlargements in red boxes demonstrate that the results of LGC and our method are closer to the ground truth in terms of spectral retention. Other methods either suffer from severe spectral distortion, such as AVWP, CDIF, PanNet, SDPNet, and CPNet, or have undesirable artifacts or poor details, such as PNN and PSGAN. AVWP also presents severe block artifacts. In addition, our D2TNet also displays the preponderance of detail retaining, which is shown in Fig. 6. It is clear that only the result of our D2TNet preserves comprehensive details relatively well.

Conversely, the result of LGC suffers from severe noise, and other methods show either poor details or blurred edges. The results of residuals in both Figs. 5 and 6 illustrate the higher consistency between our D2TNet and the ground truth than other methods.

Fig. 7 shows a group of comparative results for full-resolution images. Our D2TNet presents both the clearest details and the sharpest edges. Some other methods are blurred in detail and appear with abnormal spectral information, e.g., the enlargements of LGC, CDIF, PanNet, and PSGAN. The results of SDPNet and CPNet are failed to preserve sharp edges, and the result of AVWP suffers from block artifacts. It indicates that our D2TNet achieves the best performance in preserving both spatial details and spectral features. Overall, our method achieves outstanding performance on full-resolution images from the QuickBird dataset.

2) Quantitative Analysis: This part gives the quantitative analysis of the QuickBird dataset considering seven metrics for

TABLE I


GONG et al. : D2TNet: ConvLSTM NETWORK WITH DUAL-DIRECTION TRANSFER FOR PAN-SHARPENING
reduced-resolution images, including ERGAS, RMSE, SSIM, CC, RASE, SAM, and VIF. In addition, metrics with no reference, such as QNR, $D_\lambda$, and $D_s$, are employed for full-resolution images. All metrics mentioned above are introduced in detail in Section IV-A2. Table I shows the comparative results of eight state-of-the-art methods and our D2TNet. The first seven lines display the results on reduced-resolution images, and the last three lines show the comparative results on full-resolution images.

Considering reduced-resolution images, as presented in the table, our D2TNet ranks first on the first six metrics and ranks second on $VIF$. The smallest ERGAS value and the highest CC value indicate that our method has the superiority of correlation between the results of D2TNet and the ground truth. The highest SSIM and the smallest RMSE indicate the strongest SSIM and pixel consistency between the results of D2TNet and the ground truth, respectively. The smallest RASE and SAM illustrate that our D2TNet achieves the best spectral effects. The competitive VIF value illustrates the relatively good visual impact of our approach. Overall, our D2TNet possesses the best performance considering both spatial details and spectral consistency.

For full-resolution images, our D2TNet ranks second on QNR and $D_s$ and fourth on $D_\lambda$. It is worth noting that $D_\lambda$ is computed by comparing the fused image with the original LRMS image, which has a different resolution from the HRMS image. Due to scale differences, it is rational for the spectral distribution of low-resolution images to be not exactly
consistent with that of high-resolution images. Similarly, the spatial distribution after downsampling may also be different from the real one. Both of them illustrate that, although the objective metrics of our method are not the highest, the subjective results are the best.

C. Results on GF-2 Dataset

In this part, we conduct experiments on the GF-2 dataset. We implement both qualitative and quantitative experiments to justify the superiority of our D2TNet.

1) Qualitative Analysis: Fig. 8 shows a group of reduced-resolution results on the GF-2 dataset. The enlargements in the figure demonstrate the advantages of our D2TNet in terms of detail retaining. Meanwhile, our result also presents high consistency with the spectral information from the ground truth, while the results of AVWP, PNN, and PanNet show severe spectral distortion in overall images. In addition, residual images also demonstrate the highest consistency between the result of our D2TNet and the ground truth.

Fig. 9 shows a group of full-resolution results on the GF-2 dataset. In this figure, the results of AVWP and D2TNet show the clearest details. However, the result of AVWP not only suffers from spectral distortion but also exists block artifacts. All other methods suffer from varying degrees of blurred edges and loss of details.

2) Quantitative Analysis: Table II displays the quantitative results of our D2TNet and eight comparison methods. The table shows that our D2TNet achieves the best performance when testing reduced-resolution images and ranks second when testing full-resolution images. The highest values of metrics demonstrate that our D2TNet performs best in terms of both detail fidelity and spectral fidelity. Although D2TNet ranks second on full-resolution images, we attribute it to inconsistencies in scale, which has been detailed illustrated in Section IV-B2. As shown in Fig. 9, the enlargements clarify that the result of LGC is not as effective as ours. This part indicates the generalization ability of our model as the similarity performance on both the QuickBird dataset and the GF-2 dataset.

D. Reliability and Stability Validation

In this part, the t-test between training data and testing data is introduced to evaluate the reliability and stability of our model. Specifically, we randomly select 72 pairs of training data and 70 pairs of testing data, which comes from reduced-resolution images from the QuickBird dataset. Experiments are implemented based on the two-tailed test, and the degree is set as 140. Experimental results are reported in Table III.

The original hypothesis claims that there are no significant differences between the two sample groups, while the alternative hypothesis is the opposite. Generally, we support the original hypothesis when \( p > 0.05 \). According to Table III, all \( p \) values are much larger than 0.05. It illustrates that there are no significant divergences between the training data and the testing data, thus corroborating the reliability and stability of our model.

E. Efficiency Comparison

To illustrate the efficiency of our D2TNet, we exhibit the mean times for eight comparison methods and D2TNet when testing images from the QuickBird and GF-2 datasets. The traditional methods, LGC, AVWP, and CDIF, are implemented on CPU, and other methods are implemented on GPU. Comparative results for time are presented in Table IV. In addition,
Fig. 9. Comparison of results for a group of full-resolution images from the GF-2 dataset, i.e., LRMS images with the size of 400 × 400 × 4, PAN images with the size of 1600 × 1600, and the size of the results is 1600 × 1600 × 4. (a) LRMS. (b) PAN. (c) LGD. (d) AVWP. (e) CDIF. (f) PNN. (g) PanNet. (h) SDPNet. (i) PSGAN. (j) CPNet. (k) Ours.

<table>
<thead>
<tr>
<th>TABLE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUANTITATIVE COMPARISON ON 90 PAIRS OF REDUCED-RESOLUTION IMAGES AND 80 PAIRS OF FULL-RESOLUTION IMAGES FROM THE GF-2 DATASET.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>metric</th>
<th>LGC</th>
<th>AFWP</th>
<th>CDIF</th>
<th>PNN</th>
<th>PanNet</th>
<th>SDPNet</th>
<th>PSGAN</th>
<th>CPNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERGAS</td>
<td>1.9643 ± 0.3352</td>
<td>2.9915 ± 0.4887</td>
<td>3.3151 ± 0.6092</td>
<td>3.8070 ± 0.3688</td>
<td>1.6239 ± 0.2176</td>
<td>1.5642 ± 0.3194</td>
<td>1.3099 ± 0.3242</td>
<td>1.4833 ± 0.2240</td>
<td>1.4415 ± 0.1659</td>
</tr>
<tr>
<td>RMSE</td>
<td>7.6939 ± 1.5179</td>
<td>11.8962 ± 2.5297</td>
<td>12.9112 ± 2.4909</td>
<td>7.4865 ± 0.6906</td>
<td>6.4095 ± 1.0013</td>
<td>6.1717 ± 0.9907</td>
<td>5.9475 ± 1.0019</td>
<td>5.6063 ± 0.9553</td>
<td>4.4824 ± 0.7274</td>
</tr>
<tr>
<td>SSIM</td>
<td>0.2668 ± 0.0366</td>
<td>0.8565 ± 0.0566</td>
<td>0.7191 ± 0.0336</td>
<td>0.8370 ± 0.0256</td>
<td>0.8586 ± 0.0208</td>
<td>0.8166 ± 0.0124</td>
<td>0.8846 ± 0.0169</td>
<td>0.8534 ± 0.0207</td>
<td>0.9111 ± 0.0225</td>
</tr>
<tr>
<td>CCI</td>
<td>0.9322 ± 0.0195</td>
<td>0.8411 ± 0.0202</td>
<td>0.8765 ± 0.0320</td>
<td>0.8982 ± 0.0287</td>
<td>0.9512 ± 0.0138</td>
<td>0.9815 ± 0.0117</td>
<td>0.9609 ± 0.0110</td>
<td>0.9622 ± 0.0118</td>
<td>0.9570 ± 0.0074</td>
</tr>
<tr>
<td>RASE</td>
<td>7.4419 ± 1.7271</td>
<td>11.5577 ± 1.5903</td>
<td>12.5314 ± 2.2845</td>
<td>7.3157 ± 1.4573</td>
<td>6.2112 ± 0.8425</td>
<td>5.9785 ± 0.8425</td>
<td>5.7825 ± 0.8496</td>
<td>5.9807 ± 0.8560</td>
<td>4.3476 ± 0.6397</td>
</tr>
<tr>
<td>SAM</td>
<td>1.8400 ± 0.2800</td>
<td>1.9475 ± 0.2801</td>
<td>2.3309 ± 0.4424</td>
<td>3.6569 ± 0.5512</td>
<td>2.3045 ± 0.3042</td>
<td>1.7112 ± 0.5307</td>
<td>1.9776 ± 0.2800</td>
<td>1.6144 ± 0.2946</td>
<td>1.3299 ± 0.2019</td>
</tr>
<tr>
<td>VIF</td>
<td>0.4722 ± 0.0008</td>
<td>0.5283 ± 0.1316</td>
<td>0.3834 ± 0.0217</td>
<td>0.4596 ± 0.0464</td>
<td>0.4612 ± 0.0204</td>
<td>0.4966 ± 0.0610</td>
<td>0.5259 ± 0.1790</td>
<td>0.5972 ± 0.0593</td>
<td>0.592 ± 0.0078</td>
</tr>
</tbody>
</table>

| TABLE III |
| EXPERIMENTAL RESULTS OF T-TEST ON SOME REDUCED-RESOLUTION IMAGES FROM THE QUICKBIRD DATASET |

<table>
<thead>
<tr>
<th>metric</th>
<th>ERGAS</th>
<th>RMSE</th>
<th>SSIM</th>
<th>CCI</th>
<th>RASE</th>
<th>SAM</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>training data</td>
<td>1.1615</td>
<td>4.4837</td>
<td>1.4480</td>
<td>0.9534</td>
<td>2.6112</td>
<td>0.9751</td>
<td>0.6381</td>
</tr>
<tr>
<td>testing data</td>
<td>1.1673</td>
<td>4.4623</td>
<td>1.4927</td>
<td>0.9593</td>
<td>2.7626</td>
<td>0.9715</td>
<td>0.6285</td>
</tr>
<tr>
<td>calculated p-value</td>
<td>0.2168</td>
<td>0.6248</td>
<td>0.4603</td>
<td>0.1512</td>
<td>0.2163</td>
<td>0.6102</td>
<td>0.2147</td>
</tr>
<tr>
<td>critical p-value</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

we also compare FLOPs and parameters with five deep-learning-based methods and report the results in Table V.

Owing to the simple architecture, our D2TNet achieves the highest efficiency when testing images on the QuickBird dataset, as shown in Table IV. As PSGAN only employs the generator during testing, it achieves the highest efficiency on the GF-2 dataset. Overall, D2TNet shows competitive running efficiency on both the QuickBird dataset and the GF-2 dataset.

Table V reports the comparison of FLOPs and parameters. Our D2TNet shows relatively complex results due to the internal operations in ConvLSTM. Owing to the simple overall structure, PNN and PanNet achieve the best performance on FLOPs and parameters, respectively.

F. Ablation Study

In order to validate the necessity of some designs exploited in our D2TNet, we conduct ablation experiments for both structures and loss functions. The former part includes D2TNet without ConvLSTM, D2TNet without multiscale ConvLSTM, and D2TNet without multiscale ConvLSTM. The latter part involves some loss components and parameters. We describe each of these ablation experiments in detail and also exhibit some qualitative results and quantitative values for illustration. They are described in the following parts.

1) D2TNet Without ConvLSTM: The introduction of ConvLSTM to pan-sharpening is one of our contributions. To validate the effectiveness of ConvLSTM, we substitute ConvLSTM with common convolution. Specifically, a unit is replaced by a single convolutional layer, the kernel size is set as 3 × 3, and the inputs and the outputs are the same as the designs of ConvLSTM. The remaining conditions keep the same.

2) D2TNet Without Multiscale or Multilevel ConvLSTM: The specific eight-shaped ConvLSTM guarantees multiscale and multilevel information communication. Next, we separately verify the effectiveness of multiscale ConvLSTM and multilevel ConvLSTM. We use structures in Fig. 10 to replace the original ConvLSTM structure; other conditions remain the same. When implementing D2TNet without multiscale ConvLSTM (the structure shown in the left blue box in Fig. 10), it is obvious that three stream networks only communicate information by parameters sharing, while, in experiments abandoning multilevel ConvLSTM, the structure shown in the right red box in Fig. 10 is employed as a replacement. In this case, multilevel information communication only relies on traditional inputs and outputs.
TABLE IV
EFFICIENCY DEMONSTRATION ON EIGHT COMPARISON METHODS AND D2TNet. TESTING IMAGES ARE FROM THE QUICKBIRD DATASET AND THE GF-2 DATASET. THE PRESENTED VALUES REPRESENT MEAN VALUE ± STANDARD DEVIATION (UNIT: SECOND) (RED MARKED VALUES: THE BEST)

<table>
<thead>
<tr>
<th>DataSet</th>
<th>LGC</th>
<th>AVWP</th>
<th>CDEF</th>
<th>PNN</th>
<th>PanNet</th>
<th>SDNet</th>
<th>PSGAN</th>
<th>CPNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>QuickBird</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reduced-resolution</td>
<td>14.86±1.09</td>
<td>20.43±1.45</td>
<td>32.33±2.55</td>
<td>0.03±0.04</td>
<td>0.29±1.09</td>
<td>0.15±0.05</td>
<td>0.04±0.05</td>
<td>0.04±0.40</td>
<td>0.02±0.00</td>
</tr>
<tr>
<td>full-resolution</td>
<td>552.47±26.39</td>
<td>634.32±27.77</td>
<td>699.93±25.40</td>
<td>0.62±0.01</td>
<td>0.84±1.15</td>
<td>0.36±0.03</td>
<td>0.28±0.07</td>
<td>0.74±0.02</td>
<td>0.23±0.02</td>
</tr>
<tr>
<td>GF-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reduced-resolution</td>
<td>41.33±2.22</td>
<td>76.98±2.10</td>
<td>92.53±4.94</td>
<td>0.08±0.01</td>
<td>0.30±0.01</td>
<td>0.18±0.23</td>
<td>0.03±0.02</td>
<td>0.05±0.00</td>
<td>0.04±0.00</td>
</tr>
<tr>
<td>full-resolution</td>
<td>750.64±19.05</td>
<td>929.69±38.23</td>
<td>1694.43±87.95</td>
<td>0.72±0.01</td>
<td>0.80±0.20</td>
<td>1.03±0.72</td>
<td>0.50±0.03</td>
<td>0.82±0.09</td>
<td>0.55±0.00</td>
</tr>
</tbody>
</table>

Fig. 11. Qualitative results of some ablation experiments. The first three rows represent seven ablation results of three pairs of reduced-resolution images, and the last two rows are those of full-resolution images, where the column of the ground truth is absent. For the first three rows, images from (Left) to (Right) represent results of removing multiscale ConvLSTM, removing multilevel ConvLSTM, removing low-scale loss, setting $\lambda$ as 0.02, setting $\lambda$ as 2, setting $\xi_2$ and $\xi_3$ as 4, and setting $\xi_2$ and $\xi_3$ as 400, our D2TNet, and the ground truth.

TABLE VI
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON THE QUICKBIRD DATASET. THE TEST DATA FOR THE FIRST EIGHT METRICS ARE REDUCED-RESOLUTION IMAGES, AND THE TEST DATA FOR THE LAST THREE METRICS ARE FULL-RESOLUTION IMAGES. THE PRESENTED VALUES REPRESENT MEAN VALUE ± STANDARD DEVIATION (RED MARKED VALUES: THE BEST)

<table>
<thead>
<tr>
<th>DataSet</th>
<th>reduced-resolution images</th>
<th>full-resolution images</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o MS ConvLSTM</td>
<td>w/o MS ConvLSTM</td>
<td>w/o MS ConvLSTM</td>
</tr>
<tr>
<td>w/o ML ConvLSTM</td>
<td>w/o ML ConvLSTM</td>
<td>w/o MS ConvLSTM</td>
</tr>
<tr>
<td>w/o DC loss ($\xi_3 = 0$)</td>
<td>w/o SSIM loss ($\xi_2 = 0$)</td>
<td>w/o LS loss ($\lambda_1 = 0$ &amp; $\lambda_3 = 0$)</td>
</tr>
<tr>
<td>w/o SSIM loss ($\xi_2 = 0$)</td>
<td>w/o LS loss ($\lambda_1 = 0$ &amp; $\lambda_3 = 0$)</td>
<td>w/o SSIM loss ($\xi_2 = 0$)</td>
</tr>
<tr>
<td>$\lambda_1 = 0.02$</td>
<td>$\lambda_2 = 2$</td>
<td>$\lambda_3 = 4$ &amp; $\xi_2 = 4$ &amp; $\xi_3 = 400$</td>
</tr>
<tr>
<td>$\lambda_3 = 4$ &amp; $\xi_2 = 4$ &amp; $\xi_3 = 400$</td>
<td>$\lambda_3 = 4$ &amp; $\xi_2 = 4$ &amp; $\xi_3 = 400$</td>
<td>$\lambda_3 = 4$ &amp; $\xi_2 = 4$ &amp; $\xi_3 = 400$</td>
</tr>
</tbody>
</table>

3) D2TNet Without Some Loss Components: In order to substantiate the effectiveness of some loss components, this part implements the ablation experiments of D2TNet without SSIM loss (components with weight $\xi_3$), downsampling consistency loss (components with weight $\xi_1$), and low-scale loss (components with weights $\lambda_1$ and $\lambda_2$). Specifically, $L_{top}$ is illustrated as an example for the former two situations. As the mse loss and the gradient loss are commonly used to constrain spectral consistency and spatial consistency, they are not verified.
TABLE VII
QUANTITATIVE RESULTS OF ABLATION EXPERIMENTS ON THE GF-2 DATASET

<table>
<thead>
<tr>
<th>Dataset</th>
<th>PRQA</th>
<th>MSE</th>
<th>SSIM</th>
<th>CC1</th>
<th>RMSE</th>
<th>SAM</th>
<th>VIF*</th>
<th>QNR</th>
<th>DR</th>
<th>DRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o ConvLSTM</td>
<td>1.849 x 10^{-5}</td>
<td>6.330 x 10^{-1}</td>
<td>1.372</td>
<td>0.758 x 10^{-1}</td>
<td>0.976 x 10^{-1}</td>
<td>0.979 x 10^{-1}</td>
<td>0.978 x 10^{-1}</td>
<td>0.975 x 10^{-1}</td>
<td>0.971 x 10^{-1}</td>
<td>0.971 x 10^{-1}</td>
</tr>
<tr>
<td>w/o MS ConvLSTM</td>
<td>1.620 x 10^{-5}</td>
<td>5.617 x 10^{-1}</td>
<td>1.460</td>
<td>0.810 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
</tr>
<tr>
<td>w/o ML ConvLSTM</td>
<td>1.507 x 10^{-5}</td>
<td>5.691 x 10^{-1}</td>
<td>1.465</td>
<td>0.810 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
<td>0.944 x 10^{-1}</td>
</tr>
<tr>
<td>w/o DC loss (ξ2 = 0)</td>
<td>2.863 x 10^{-5}</td>
<td>1.032 x 10^{-1}</td>
<td>0.920</td>
<td>0.930 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
</tr>
<tr>
<td>w/o SSIM loss (ξ1 = 0)</td>
<td>3.224 x 10^{-5}</td>
<td>2.512 x 10^{-1}</td>
<td>0.918</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
</tr>
<tr>
<td>w/o LS loss (λ1 = 0 &amp; λ2 = 0)</td>
<td>3.200 x 10^{-5}</td>
<td>4.695 x 10^{-1}</td>
<td>0.916</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
</tr>
<tr>
<td>λ1 = 0.02</td>
<td>1.251 x 10^{-5}</td>
<td>4.977 x 10^{-1}</td>
<td>0.920</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
<td>0.920 x 10^{-1}</td>
</tr>
<tr>
<td>λ1 = 2</td>
<td>1.099 x 10^{-5}</td>
<td>5.859 x 10^{-1}</td>
<td>0.906</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
<td>0.906 x 10^{-1}</td>
</tr>
<tr>
<td>ξ1 = 4 &amp; ξ2 = 4</td>
<td>3.430 x 10^{-5}</td>
<td>1.790 x 10^{-1}</td>
<td>0.935</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
<td>0.935 x 10^{-1}</td>
</tr>
<tr>
<td>ξ1 = 400 &amp; ξ2 = 400</td>
<td>3.145 x 10^{-5}</td>
<td>4.482 x 10^{-1}</td>
<td>0.931</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
<td>0.931 x 10^{-1}</td>
</tr>
</tbody>
</table>

As for another experiment of D2TNet without low-scale loss, the total \( L_{\text{all}} \) is formalized as

\[
L_{\text{all}} = L_{\text{top}}
\]

where \( L_{\text{top}} \) remains the same as original settings.

4) D2TNet With Some Parameter Settings: In this part, \( \lambda_1, \xi_2, \) and \( \xi_3 \) are involved to implement parameter ablation experiments. First, based on the considerations in Section IV-A1, \( \lambda_1 \) is determined as 0.2. This part gives another two settings 0.02 and 2 to verify its rationality. Second, \( \xi_2 \) and \( \xi_3 \) are considered equally important as they constrain spatial consistency and spectral consistency, respectively. They are set as 40, and two other settings 4 and 400 are also given to verify the rationality.

5) Ablation Result Analysis: Metrics mentioned in Section IV-A2 are employed for all above ablation experiments. The comparative results on the QuickBird dataset and the GF-2 dataset are presented in Tables VI and VII. They demonstrate the superiority of our method on both datasets in spatial consistency, spectral consistency, structure similarity, and so on, thus verifying the necessity of the specific designs of the eight-shaped ConvLSTM and some loss components. To simplify, “multiscale,” “multilevel,” “downsampling consistency,” and “low-scale” are abbreviated as “MS,” “ML,” “DC,” and “LS,” respectively.

In addition, we also display three pairs of reduced-resolution images and two pairs of full-resolution images collected from the QuickBird and GF-2 datasets. We select seven ablation experiments that perform relatively well on metrics as examples to show their subjective results, including experiments w/o MS ConvLSTM, w/o ML ConvLSTM, w/o LS loss, \( \lambda_1 = 0.02, \lambda_1 = 2, \xi_2 = 4 \) & \( \xi_3 = 4, \) and \( \xi_2 = 400 \) & \( \xi_3 = 400, \) as shown in Fig. 11.

The first three rows display three groups of results on reduced-resolution images. In the first row, the results of ablation experiments all express the shortcoming of blurred edges, and they also suffer from detail loss in the third row. In addition, ablation experiments introduce block artifacts in the second row. The reason is that the absence of multiscale or multilevel ConvLSTM leads to weaker information communication, thus losing some important details as the network deepens. Besides, experiments of \( \lambda_1 = 2 \) put more emphasis on upsampled or downsampled images, when they should have been given less attention than the original image.

Nevertheless, underappreciated multiscale information also

Fig. 10. Structure in the left blue box is exploited to replace ConvLSTM in the experiment without multiscale ConvLSTM. The right one in the red box is the structure used to replace ConvLSTM in the experiment without multilevel ConvLSTM.

TABLE V
FLOPS AND PARAMETERS ON FIVE DEEP-LEARNING-BASED METHODS AND D2TNet. TESTING IMAGES ARE REDUCED IMAGES FROM THE GF-2 DATASET; THE SIZE OF RESULTS IS 400 x 400 x 4 (RED MARKED VALUES: THE BEST)

<table>
<thead>
<tr>
<th>FLOPs/parameters</th>
<th>PNN</th>
<th>PanNet</th>
<th>SDPNet</th>
<th>PSCGAN</th>
<th>CPNet</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRQA x 10^{-6}</td>
<td>12.87</td>
<td>24.50</td>
<td>180.06</td>
<td>46.30</td>
<td>179.01</td>
<td>133.07</td>
</tr>
<tr>
<td>PanNet x 10^{-6}</td>
<td>8.04</td>
<td>7.09</td>
<td>57.64</td>
<td>146.97</td>
<td>64.16</td>
<td>92.59</td>
</tr>
</tbody>
</table>
results in information distortion, e.g., when setting $\lambda_1$ to 0.02. The experiment of $\xi_2 = 4$ & $\xi_3 = 4$ weakens constraints on spectrum and details. Albeit the experiment of $\xi_2 = 400$ & $\xi_3 = 400$ performs well, the too heavily weighted compared with mse loss causes that it still falls somewhat short of our approach, whereas, in the ablation experiments removing low-scale loss, they reduce the constraints on the fusion result, leading to a greater gap between the fused image and the ground truth. The last two rows display comparative results on full-resolution images; they indicate that our D2TNet achieves the most normal and clearest spectral information, while results of all ablation experiments introduce varying degrees of artifacts.

V. CONCLUSION

In this article, we present a novel efficient pan-sharpening method, called D2TNet. Owing to the advantages of ConvLSTM in dealing with long-term information dependencies, a specially structured ConvLSTM network is designed to allow for dual-directional communication. The dual-direction transfer includes multiscale and multilevel information communication, which is beneficial for maintaining richer spatial details and more consistent spectral features with source images, respectively. Besides, we constrain the generated HRMS from three different scales. It not only strengthens the constraints on the network but also increases the contribution of middle and bottom stream networks to the top stream network. Experimental results indicate that our D2TNet obtains impressive results both qualitatively and quantitatively on different images compared to the state-of-the-art methods. Moreover, our method has higher testing efficiency and stable performance on different images compared to the state-of-the-art methods.

REFERENCES


Meiqi Gong received the B.S. degree from the School of Science, Jiangnan University, Wuxi, China, in 2021. She is currently pursuing the master’s degree with the Electronic Information School, Wuhan University, Wuhan, China. Her research interests include computer vision and image processing.

Jiayi Ma (Senior Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan. He has authored or coauthored more than 200 refereed journal articles and conference papers, including IEEE TPAMI, IEEE TIP, IJCV, CVPR, ICCV, and ECCV. His research interests include computer vision, machine learning, and robotics.

Han Xu (Graduate Student Member, IEEE) received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree with the Multi-Spectral Vision Processing Laboratory, Electronic Information School.

She has first-authored over ten refereed journal articles and conference papers, including IEEE TPAMI, IEEE TIP, IEEE TGRS, IEEE TCI, IEEE TIM, AAAI, and IJCAI. Her research interests include computer vision and image processing.

Xin Tian (Member, IEEE) received the B.S. degree from the Department of Electronic Science and Technology, Huazhong University of Science and Technology, Wuhan, China, in 2004, and the Ph.D. degree from the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, in 2010. He is currently an Associate Professor with the School of Electronic Information, Wuhan University, Wuhan. He has authored or coauthored over 50 refereed journal articles and conference papers. He has been granted more than 30 invention patents. His research interests include computational imaging, image fusion, and sparse coding.

Dr. Tian’s awards include the First Prize for Scientific and Technological Progress in Surveying and Mapping in China, the First Prize for Excellent Achievements of Information Technology in Electric Power Industry of China, and so on.

Xiao-Ping Zhang (Fellow, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1992 and 1996, respectively, and the M.B.A. degree (Hons.) in finance, economics, and entrepreneurship from The University of Chicago Booth School of Business, Chicago, IL, USA.

Since Fall 2000, he has been with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of the Communication and Signal Processing Applications Laboratory. He has served as the program director for graduate studies. He is cross-appointed to the Finance Department with the Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent consultant for biotech companies and investment firms. His research interests include image and multimedia content analysis, sensor networks and IoT, machine learning, statistical signal processing, and applications in big data, finance, and marketing.

Dr. Zhang is also a fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada, a Registered Professional Engineer in Ontario, Canada, and a member of the Beta Gamma Sigma Honor Society. He was an Elected Member of the ICME Steering Committee. He received the 2020 Sarwan Sahota Ryerson Distinguished Scholar Award—the Ryerson University Highest Honor for scholarly, research and creative achievements. He is the General Co-Chair for the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2021, the 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He is the General Chair for the IEEE International Workshop on Multimedia Signal Processing, 2015. He is Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is a Senior Area Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING. He served as a Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS. He is selected as a IEEE Distinguished Lecturer by the IEEE Signal Processing Society for the term 2020 to 2021, and by the IEEE Circuits and Systems Society for the term 2021 to 2022.