Convolutional LSTM-Based Hierarchical Feature Fusion for Multispectral Pan-Sharpening

Wang, Dong; Bai, Yunpeng; Wu, Chanyue; Li, Ying; Shang, Changjing; Shen, Qiang

Published in:
IEEE Transactions on Geoscience and Remote Sensing

DOI:
10.1109/TGRS.2021.3104221

Publication date:
2022

Citation for published version (APA):
Wang, D., Bai, Y., Wu, C., Li, Y., Shang, C., & Shen, Q. (2022). Convolutional LSTM-Based Hierarchical Feature Fusion for Multispectral Pan-Sharpening. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, Article 5404016. https://doi.org/10.1109/TGRS.2021.3104221

Convolutional LSTM-Based Hierarchical Feature Fusion for Multispectral Pan-Sharpening

Dong Wang, Yunpeng Bai, Chanyue Wu, Ying Li, Changjing Shang, and Qiang Shen

***Abstract*—Multispectral (MS) pan-sharpening aims at producing High Resolution (HR) MS images in both spatial and spectral domains, by merging single-band panchromatic (PAN) images and corresponding MS images with low spatial resolution. The intuitive way to accomplish such MS pan-sharpening tasks, or to reconstruct ideal HR-MS images, is to extract feature pairs from the given PAN and MS images and to fuse the results. Therefore, feature extraction and feature fusion are two key components for MS pan-sharpening. This paper presents a novel MS Pan-sharpening Network (MPNet), including a heterogeneous pair of Feature Extraction Pathways (FEPs) and a Convolutional LSTM (ConvLSTM)-based Hierarchical Feature Fusion Module (HFFM). Specifically, we design a PAN FEP to extract 2D feature maps via 2D convolutions and dual attention, while an MS FEP is introduced in an effort to obtain 3D representations of MS image by 3D convolutions and triple attention. To merge the resulting hierarchical features, the ConvLSTM-based HFFM is developed, leveraging intra-level fusion, inter-level fusion, and information exchange within one single framework. Here, the inter-level fusion is implemented with the ConvLSTM to capture the dependencies amongst hierarchical features, reduce redundant information, and effectively integrate them via its recurrent architecture. The information exchange between different FEPs helps enhance the representations for subsequent processing. Systematic comparative experiments have been conducted on three publicly available data sets at both reduced-resolution and full-resolution, demonstrating that the proposed MPNet outperforms state-of-the-art methods in the literature.**

***Index Terms*—Hierarchical Feature Fusion, Convolutional LSTM, Multispectral Pan-sharpening, Information Fusion, Triple Attention.**

I. Introduction

HIGH Resolution (HR) remote sensing images in both spatial and spectral domains are desirable for many practical applications, e.g., environmental monitoring [1], object detection [2], land cover classification [3], [4], remote sensing scene classification [5]–[7], etc. However, due to hardware limitations, it is often difficult to obtain such ideal images. Instead, only panchromatic (PAN) images with high spatial resolution and low spatial resolution MS images may be captured by some sensors, e.g., IKONOS, QuickBird, and WorldView-4. From this regard, techniques for MS pan-sharpening are desirable that are developed to obtain HR-MS images by fusing the PAN images and the corresponding MS images [8], [9].

Over the last decades, many and various MS pan-sharpening approaches have been proposed, e.g., Component Substitution (CS)-based methods [10], Multi-Resolution Analysis (MRA)-based methods [11], model-based methods [12]–[14], CS/MRA hybrid methods [15], CNN-based methods [16]–[20], etc. The advantages of CS-based methods are fast, easy to implement, and robustness to misregistration errors and aliasing. MRA-based approaches typically characterize temporal coherence, spectral consistency, and robustness to aliasing under certain appropriate conditions. CS/MRA hybrid methods combine both of them and inherit their advantages. Generally speaking, model-based methods can obtain fused images of relatively high quality. However, these approaches bear their own limitations. For CS-based methods and MRA-based approaches, there is a conflict between retaining the spectral information in MS images and improving the spatial resolution, especially when the spectrum range of the MS images and that of the PAN images are only partially overlapping. Model-based methods rely heavily on priori knowledge and hyper-parameters, while requiring high computational resources.

Recently, Convolutional Neural Network (CNN)-based techniques have shown great potential in the field of MS pan-sharpening, thanks to the high non-linearity of deep CNNs that facilitates sophisticated modeling tasks. Such fusion methods can be coarsely classified into two groups: single-level feature fusion (Fig.1a and Fig.1b) and multi-level feature fusion (Fig.1c), with most of which [16], [21]–[24] adopting the single-level feature fusion approach, involving early fusion (Fig.1a) or late fusion (Fig.1b), through fusing the representations from different sources regarding a given position. Single-level feature fusion can only merge partial information to perform MS pan-sharpening however, which may hinder the full achievement of fusion potential. In contrast, multilevel features are capable of representing different characteristics of PAN and MS images, thereby significantly improving the fusion performance.

Despite the promising performance obtained by the above methods, three key issues have not been solved yet. One is

This work is supported in part by the National Natural Science Foundation of China(61871460), in part by the Shaanxi Provincial Key Research and Development Program of China (2020KW-003), and in part by the Strategic Partner Acceleration Award (80761-AU201) of United Kingdom.(Corresponding author: Ying Li.)

D. Wang, C. Wu and Y. Li are with School of Computer Science, National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Shaanxi Provincial Key Laboratory of Speech & Image Information Processing, Northwestern Polytechnical University, Xi'an 710000, China (e-mail: dongwang@mail.nwpu.edu.cn; wuchanyuec@163.com; lybyp@nwpu.edu.cn)

Y. Bai is with the Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K., and also with the School of Communication and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710000, China (e-mail: cloudbai@nwpu.edu.cn)

C. Shang and Q. Shen are with the Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth SY23 3DB, U.K. (e-mail: cns@aber.ac.uk; qqs@aber.ac.uk).

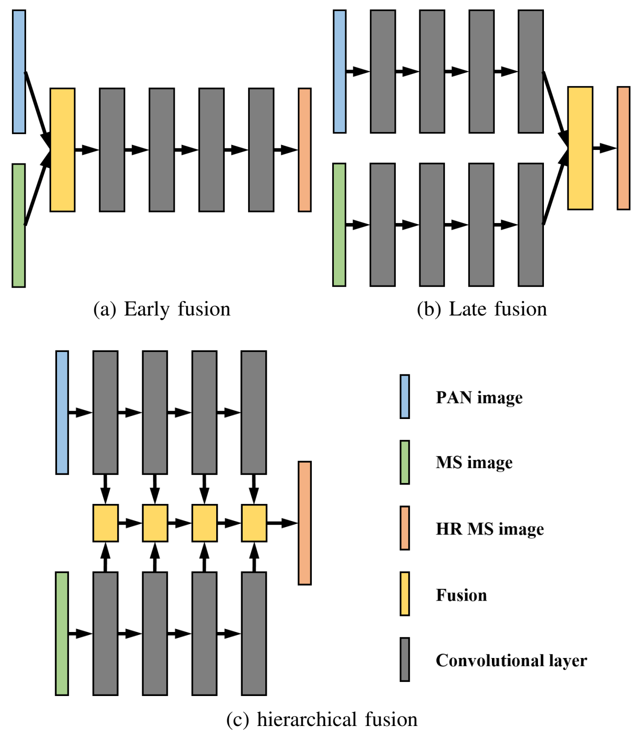


Fig. 1. Example of early fusion, late fusion, and hierarchical fusion.

that most of the existing multilevel feature fusion methods [20], [25] only use simple fusion manners, such as summation, concatenation, etc., to fuse features of several levels. The hierarchical features may provide redundant but complementary information since different-level features contain specific information. Only fusing a portion of hierarchical features can limit the pan-sharpening performance. In addition, such naive fusion methods hardly reduces the redundancy of the hierarchical features without learning the dependency of the hierarchical features. Another issue is that most existing methods treat diverse features at each spatial-spectral position equally, which lacks flexibility in dealing with different types of information. For instance, edges containing high-frequency information are much more difficult to reconstruct than flat regions and should gain more attention during pan-sharpening. The third issue has to do with the fact that most of these approaches employ 2D convolutions for both spatial and spectral information processing. Unfortunately, 2D convolutions typically cause the extracted features in the spectral dimension of a layer to be averaged and collapsed to a scalar [26], resulting in low spectral resolution.

Having taken notice of the aforementioned issues, a novel network for MS Pan-sharpening (MPNet) is proposed in this work, where attention-based Feature Extraction Pathways (FEPs) and ConvLSTM-based Hierarchical Feature Fusion Module (HFFM) are exploited. First, A heterogeneous pair of FEPs is employed to extract hierarchical spatial and spectral features. In PAN FEP, 2D convolutions and dual attention, i.e., Channel-Spatial Attention (CSA), are utilized to obtain more informative feature maps. In contrast, the MS FPS extracts the 3D representations by 3D convolutions and triple attention, i.e., Channel-Spectral-Spatial Attention (CSSA). We then employ the hierarchical fusion (shown in Fig.1c), one kind of multi-level feature fusion, to merge all hierarchical features in a shallow-to-deep manner. Specifically, all generated hierarchical features are merged with the ConvLSTM-based HFFM, which captures the dependencies amongst hierarchical features, reduce redundant information, and integrate them effectively via its recurrent architecture. The main contributions of this paper are outlined as follows:

1) We develop a heterogeneous pair of FEPs for MS pan-sharpening. The PAN FEP equipped with 2D convolutions and CSA obtains the feature maps from PAN images, while the MS FEP acquires 3D representations via 3D convolutions and CSSA from MS images. This provides a novel approach that integrates the CSSA mechanism to address the issue of MS pan-sharpening, which can adaptively learn further informative channel-wise, spectral-wise, and spatial-wise features simultaneously.
2) Different from most existing pan-sharpening methods that adopt single-level feature fusion, the multi-level representations are investigated for fusion in this study. To fully exploit the potential representation capacity of multi-level features, hierarchical fusion is investigated that merges all hierarchical features in a shallow-to-deep manner. Taking advantage of the representations of a wider range of levels, information on different sources can be better fused.
3) We present a ConvLSTM-based HFFM to merge the hierarchical spatial and spectral features of different levels. The HFFM leverages intra-level fusion, inter-level fusion, and information exchange within one single framework. The inter-level fusion is herein implemented originally with the ConvLSTM to capture the dependencies amongst hierarchical features, reducing redundant information, and to integrate them effectively via its recurrent architecture.
4) Extensive comparative experiments on three publicly available benchmark data sets (namely, IKONOS, QuickBird, and WorldView-4) are conducted at both the full-resolution and the reduced-resolution level. In reduced-resolution experiments, the proposed MPNet substantially outperforms the State-Of-The-Art (SOTA) methods. Results of full-resolution experiments also demonstrate that MPNet achieves competitive performance.

The remainder of this paper is organized as follows. Section II briefly introduces the background knowledge of the ConvLSTM and the existing CNN-based pan-sharpening methods. The proposed approach is detailed in Section III. The experimental results are presented and discussed in Section IV. Finally, the conclusion of this paper is given in Section V.

II. Related Work

For academic completeness, this section presents an overview of the relevant background, regarding CNN-based pan-sharpening approaches and ConvLSTM.

A. *CNN-Based Pan-Sharpening*

Recently, CNNs have become increasingly popular in the implementation of systems for MS pan-sharpening. The following part describes CNN-based methods according to the classification of early fusion, late fusion, and multilevel fusion.

Early fusion-based methods concatenate the up-sampled MS image and the PAN image and then reconstruct the HR-MS image. Inspired by the significant work of SRCNN [27], PNN [21] was proposed, being the first utilising CNN for pan-sharpening. In the architecture of PNN, each MS image is up-sampled and concatenated with the PAN image, thereby implementing early fusion. PNN has been further extended with residual learning [16], leading to a significant performance gain over the original PNN. A target-adaptive tuning phase is introduced in the PNN+ [16] to solve the problem of insufficient data. As with PNN, a Multi-Scale and multi-Depth Convolutional Neural Network (MSDCNN) is proposed [24], which also concatenates the PAN band and the MS bands together, feeding the concatenated into the network. Recently, a novel unsupervised framework has been introduced for pan-sharpening based on GAN and PNN, termed as Pan-GAN, which does not rely on the availability of information on ground-truth during the phase of network training [19].

Late fusion is also employed in some pansharening methods. Unlike the methods mentioned above, a remote sensing image fusion mechanism, named RSIFNN, is considered in [28] that can adequately extract spectral and spatial features from the source images. In RSIFNN, the spatial and spectral features are only integrated at the late stage, without leveraging the Hierarchical features of the PAN and MS streams. Liu et al. [18] proposed a Two-stream Fusion Network (TFNet) that extracts CNN features from PAN and MS images with two 2D CNN and subsequently fuses the deepest features with concatenation operation. Subsequently, they proposed a generative adversarial network for remote sensing image pan-sharpening (PSGAN) [29], consisting of a generative network (i.e., TFNet) and a discriminative network.

Another popular family is based on the multilevel feature fusion. Zhang et al. [25] introduced a new end-to-end bidirectional pyramid network (BDPN) for pan-sharpening. Two bidirectional pyramid branches process MS and PAN images separately, and merge them at only two levels with the summation operation. Cai et al. [20] propose and develop a novel pan-sharpening algorithm that is guided by a deep super-resolution convolutional neural network, where the progressive pan-sharpening with two-level fusion is used to achieve a gradual and stable pan-sharpening process.

B. ConvLSTM

LSTM has achieved great success for sequence modeling in performing various natural language processing tasks, including speech recognition [30] and visual question answering [31]. With the gates, LSTMs can remove or add information to the cell states and can model the long-term dependencies.

Note however, that LSTMs only take as input 1D vectors and thus, cannot be applied for 2D feature maps. The 2D convolution operation is therefore introduced to LSTM, resulting in ConvLSTM [32], which can process 2D feature maps, automatically capturing temporal dependencies between states. ConvLSTMs can also be exploited for 3D data processing. For instance, a fast video salient object detection model is proposed in [33], based on Pyramid Dilated Bidirectional ConvLSTM (PDB-ConvLSTM). In [34], a powerful tree-structure based traversal method is presented to model 3D skeletons, with LSTM employed to handle the noise and occlusions in the 3D data. Also, an Object-to-Motion convolutional neural network (OM-CNN) has been reported [35], in which a two-layer ConvLSTM (2C-LSTM) network is utilised to predict video saliency.

Although aforementioned CNN-based methods have achieved great advances in the field of pan-sharpening, there are three issues still existing. One is that the existing multilevel feature fusion methods only merge part levels with a simple fusion manner, e.g., summation or concatenation. As demonstrated by many visualization works, shallow level features contain more details. With the increase of layers, the features will become more abstract. The hierarchical features can provide redundant but complementary information. Only fusing a portion of hierarchical features with such naive fusion methods may limit the pan-sharpening performance. Another issue goes that most existing treat diverse features at each spatial-spectral position equally, which lack flexibility in dealing with different types of information. The last issue is that most of them employ 2D convolutions for both spatial and spectral information processing. Unfortunately, 2D convolutions cause the extracted features in the spectral dimension of a layer to be averaged and collapsed to a scalar [26], which leads to low spectral resolution.

III. Proposed Approach

The purpose of MS pan-sharpening is to obtain HR-MS images by fusing single-band PAN images and the corresponding MS images with B bands (e.g., $B = 4$ for IKONOS, QuickBird, and WorldView-4 satellites). In this paper, the observed PAN images are denoted as $X_P \in R^{H\times W}$, where H and W are the height and width, respectively. Also, $X_M \in R^{\frac{H}{4}\times\frac{W}{4}\times B}$ represents the MS images, with 4 being the spatial resolution ratio between the PAN images and the corresponding MS images. The ideal HR-MS images are denoted as $Y_M \in R^{H\times W\times B}$. A detailed illustration of the proposed MPNet is shown in Fig. 2. Particularly highlighted in the yellow and blue background, the PAN and MS FEPs extract hierarchical features from the PAN and MS images, respectively. The HFFM fuses the resulting hierarchical features level by level. The reconstruction module is devised to recover the ideal HR-MS images. More details about the four main parts of our newly proposed MPNet and the object function are given in Section III-A-III-E.

A. PAN FEP

PAN FEP is designed to extract 2D hierarchical features from the PAN image. Without loss of generality, the idea of the ResNet [36] is applied to implement this FEP, where the DenseNet [37] can be readily constructed by replacing ResBlocks with DenseBlocks [37]. In addition, the 2D features which contain distinct information across channels or spatial positions, contribute differently to the pan-sharpening process. The channel attention and the spatial attention should highlight

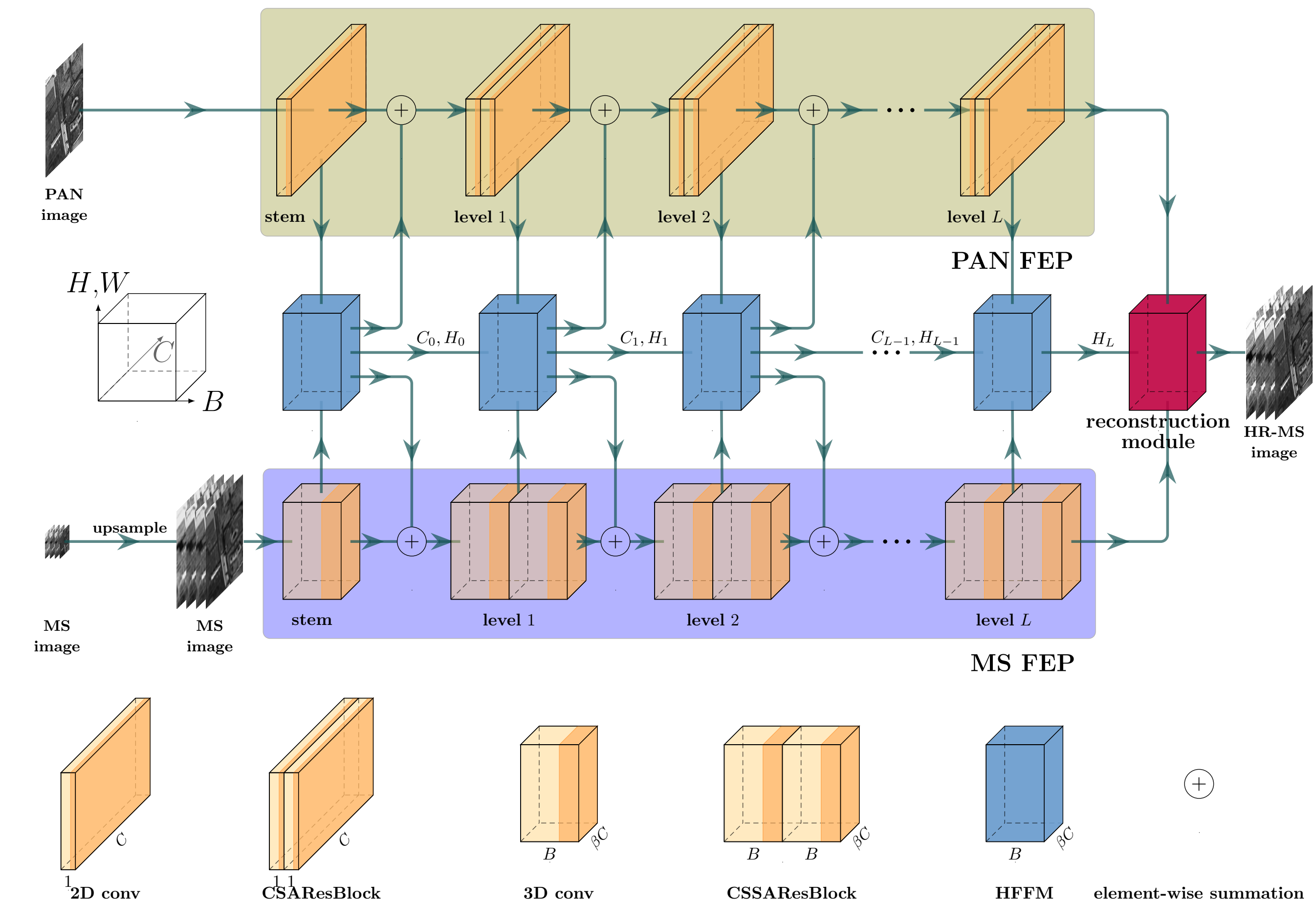


Fig. 2. Architecture of the proposed MPNet. The left coordinate system indicates dimensionalities of width, height, bands, and channels. H and W represent image height and width. B indicates the number of bands. C denotes the number of channels. FEP denotes the feature extraction pathway. HFFM represents the hierarchical feature fusion module. CSA and CSSA indicate channel-spatial attention and channel-spatial-spectral attention, respectively. β is the filter number ratio between the FEPs. H_l and C_l denote hidden and cell states of ConvLSTM, respectively.

the most informative feature maps and regions, respectively. Thus, the PAN FEP contains a stem layer and L stacked CSAResBlocks with 2D convolutions and these attentions. Batch normalization and sampling operation are removed to reduce the brought noise and to preserve details, respectively. In short, the PAN FEP can be formulated as follows:

$$F^0_{2D} = f_0(X_P) \tag{1}$$

$$F^L_{2D} = f^l_{2D}(F^{l-1}_{2D} + F^{l-1}_{H,2D}) \tag{2}$$

where f_0 denotes the stem layer consisting of a 2D convolution and a Parametric Rectified Linear Unit (PReLU); f^l_{2D} represents the CSAResBlocks; l indexes the level number, ranging from 1 to L. CSAResBlock utilizes CSA to improve the representation ability of the FEP. We construct a channel-attention module and a spatial-attention module to exploit the interchannel relationships and the interspatial dependencies, respectively. The channel-attention module and the spatial-attention module are shown in Fig. 3, where r is the reduction ratio (which is herein set to 16 according to the practice in the literature [38]).

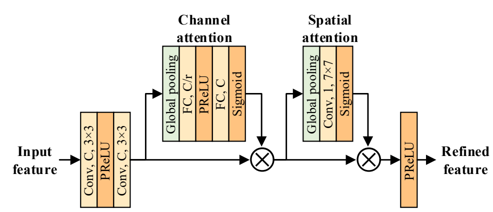


Fig. 3. Structure of CSAResBlock.

B. MS FEP

MS FEP is devised to extract 3D representations with 3D convolutions. Note that 3D CNNs have been employed in 3D data-related tasks despite their high computation complexity to improve system performance [39]–[41]. This is due to the recognition that the 3D convolutions are more consistent with the underlying characteristics of 3D data. Although CNNs has proven to be effective in the field of pan-sharpening, they may be hindered by their modelling of all spectral bands with the same weight, as generally not all bands are equally informative and predictive [42]. Therefore, the CSSA mechanism is integrated to address the issue of MS pan-sharpening, which can

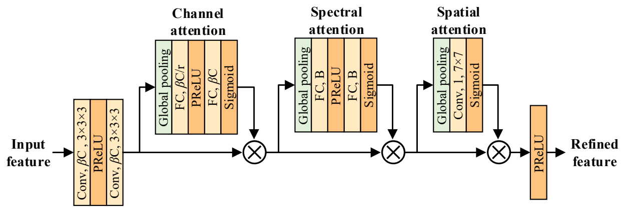


Fig. 4. Structure of CSSAResBlock.

adaptively learn more informative channel-wise, spectral-wise, and spatial-domain features simultaneously.

The overall structure of the MS FEP is similar to the PAN FEP. Using the bicubic interpolation algorithm [43], the input MS image is up-sampled to align with the PAN image. In particular, the first layer is a stem layer, in which a 3D convolution layer [44] replaces the 2D counterpart, and the others are CSSAResBlocks with 3D convolutions. The number of channels is reduced to βC to relax the computational burden (where β is empirically set to 0.5 in this paper to balance the computational cost of these two FEPs).

Apart from channel attention and spatial attention, CSSAResBlock learns how to pay attention to spectral domain. Although the general structures of the channel-attention module and the spatial-attention module are the same as that of CSAResBlock, certain components have been customized for 3D representation, e.g., the global pooling used in these modules is replaced with spatial-spectral average-pooling and channel-spectral average pooling, respectively. The structure of CSSAResBlocks is outlined in Fig. 4.

The **spectral attention** is comprised of a global pooling layer and MLP. The channel-spatial information of the input features is aggregated by channel-spatial average-pooling. The spectral attention is computed such that

$$\begin{aligned} M_{spe}(F) &= \sigma(MLP(AvgPool(F))) \\ &= \sigma(W_1\delta(W_2F_{avg}^{c,spe}))) \end{aligned} \tag{3}$$

where σ represents the sigmoid function; δ denotes the PReLU; $W_0 \in R^{B\times B}$; $W_1 \in R^{B\times B}$; and $F_{avg}^{c,spe}$ is the generated channel-spetial context descriptor.

C. ConvLSTM-Based HFFM

Once the hierarchical representations are obtained, they will be fused in the next step. However, the normal multilevel feature methods only merge the multi-level features at several levels. In addition, most existing multi-level feature-based approaches employ simple fusion operations, e.g. concatenation, summation, and multiplication, etc. The hierarchical features representing the input image at different positions have complementary and redundant information. With more levels involved, it may be more difficult for the feature fusion methods to decide what information needs to throw away or pick up with more levels involved. Such naive fusion manners ignore dependencies between features at different levels, which may hinder the fusion performance. In this paper, the ConvLSTM-based HFFM is utilized to integrate these features. ConvLSTM has the ability to retain long-term information

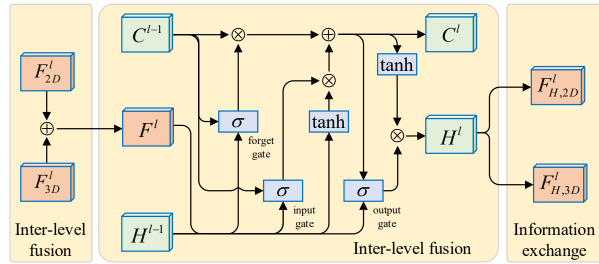


Fig. 5. Architecture of HFFM, where ⊗ denotes element-wise multiplication, ⊕ represents element-wise summation, and σ indicates sigmoid function.

with the cell states, which facilitates mining their dependencies in a learnable way. With the learned dependencies, HFFM can reduce redundant information, and integrate effectively via its recurrent architecture.

The architecture of HFFM is shown in Fig. 5. At each layer, it has two inputs (F_{2D}^l and F_{3D}^l) and produces two outputs ($F_{H,2D}^l$ and $F_{H,3D}^l$ for PAN and MS FEPs, respectively). Intra-level fusion integrates the spatial and spectral information from the FEPs, which is shown in Fig. 5. Inter-level fusion is implemented with the ConvLSTM, where the gates learn the dependencies of different levels from data. The resulting gates can reduce the redundant information in hierarchical features, which may boost the fusion performance. The exchanged information is demonstrated in the right part of Fig. 5.

1) Intra-Level Fusion: The responsibility of intra-level fusion is to integrate the spatial feature and spectral information of the same level. The key motivation behind this is that feature representations at different levels may differ significantly, e.g., high-level features have abstract information, while their low-level counterparts are of minor details concerning edges and curves. Of course, the fusion of the same type of features is more accessible than different types. Instead of integrating a mass of low-level and high-level features in one step, individual representations of the same level are fused before merging those of different levels. The intra-level fusion can be formulated as:

$$F^l = W_{C,1} *^T F_{2D}^l + W_{C,2} * F_{3D}^l \tag{4}$$

where $*^T$ and $*$ denote a transpose convolution and a normal convolution, respectively; $W_{C,1}$ and $W_{C,2}$ are weights of the two filters. The $*^T$ increase the spectral dimension of the of F_{2D}.

2) Inter-Level Fusion via ConvLSTM: Whilst Conv-LSTM is often applied for time sequence-related tasks, especially for handling sequential inputs like video frames, LSTMs are also widely used to model sequential relationships between different image bands [45], [46]. For example, the issue of spectral feature extraction has been considered as a sequence learning problem [45] and as shown in [46], for each pixel, spectral values from different channels are fed into spectral LSTM one by one to learn the spectral features. Indeed, the inputs of ConvLSTM in this work are closely related to the features extracted from different levels. Yet, the inputs of ConvLSTM in video processing are not necessarily independent features obtained simply at the same level because the frames in videos may also be closely correlated.

In this work, ConvLSTM manages to mine the dependencies amongst hierarchical representations and to integrate them. As the cell states contains long-term information of previous levels, they act as the bridge to connect the current situation to prior levels. The input gate and the forget gate learn the dependencies between different levels and decide what information is to be removed from the cell states of the previous level and what new information to be selected for fusion. Based on C^{l-1}, H^{l-1}, and F^l_C, the gate outputs a number between 0 and 1 for each spatial-spectral element, where 0 and 1 indicates completely forgetting and keeping the corresponding element, respectively. The input gate, then, decides on which part of the fused spatial-spectral feature will flow into the cell with a sigmoid layer. ConvLSTM automatically extracts hidden states with the output gate .

Following the standard method for developing a ConvLSTM, apart from the current inputs, previous hidden states are integrated. Note that the group convolution is employed to release the restriction over the input image size produced by the Hadamard product of the original ConvLSTM. The elimination of this restriction enables the block effect in the fused image to be addressed. In addition, the ConvLSTM in this paper is constructed with 3D convolutions instead of 2D convolutions in the original ConvLSTM. This procedure is summarised in the following equations:

$$i^l = \sigma(W_i * F^l + W_{hi} * H^{l-1} + W_{ci} * C^{l-1} + b_i) \tag{5}$$

$$f^l = \sigma(W_f * F^l + W_{hf} * H^{l-1} + W_{cf} * C^{l-1} + b_f) \tag{6}$$

$$C^l = f^l \circ C^{l-1} + i^l \circ \tanh(W_c * F^l + W_{hc} * H^{l-1} + b_c) \tag{7}$$

$$o^l = \sigma(W_o *^T F^l + W_{h,o} * H^l + W_{co} * C^l + b_o) \tag{8}$$

$$H^l = o^l \circ \tanh(C^l) \tag{9}$$

where H^{l-1} indicates the hidden state of previous level; W and b are learnable parameters; $*$ stands for the convolution operation; $tanh$ represents the tanh activation function; i^l denotes the input gate at level l; $\circ$ represents the element-wise multiplication; o^l is the output gate; the notation $*$ in $W_{ci} * C^{l-1}$, $W_{hf} * H^{l-1}$, and $W_{h,o} * H^l$ is a grouped convolution, where the number of groups is the same as the channel dimensionality.

3) Information Exchange: Inspired by [47] in which information exchange between different channels can enhance overall information content, the information exchange is also incorporated in our method. The hidden states, as the output information, are obtained by the output gate, and are fed back to the FEPs.

The information $F_{H,2D}$ and $F_{H,3D}$ fed back the PAN and MS FEPs is shown as at the right part of Fig. 5, respectively. Since the ConvLSTM operates on 3D data, the output features $F^l_{C,2D}$ for the PAN FEP need to be transformed to 2D format. These are obtained through the following computation:

$$F^l_{H,3D} = H^l \tag{10}$$

$$F^l_{H,2D}(i,j) = f_{B\times1\times1}(H^l) \tag{11}$$

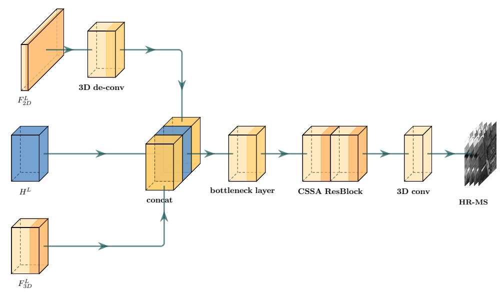


Fig. 6. Architecture of reconstruction module.

where $F^l_{C,3D}$ denotes the information fed back to the spectral FEP; $f_{B\times1\times1}(H^l)$ is a $B \times 1 \times 1$ convolution layer.

D. Reconstruction Module

The responsibility of the reconstruction module is to recover the desired HR-MS images from the fused feature H^L. Inspired by the work of [18], F^L_{2D} and F^L_{3D} are also fed to the reconstruction module. The redundant information of these three items is reduced with a bottleneck layer. CSSAResBlock is also employed to obtain further informative representations, improving the non-linearity and alleviating the gradient vanishing problem. Finally, the HR-MS image is recovered by a convolutional layer from the feature space.

Fig. 6 illustrates the reconstruction module, which can be divided into four components: a 3D de-convolution layer, a bottleneck layer, a CSSAResBlock, and a 3D convolution layer without activation. First, F^L_{2D} is projected into $R^{\beta C\times B\times H\times W}$ by a de-convolution layer. Next, it is concatenated with F^L_M and H^L. Then, the bottleneck layer is added to weight the three 3D features by βC filters of size $1\times1\times1$. After that, the output of this layer is fed into the 3D residual block R_{3D}, in an effort to transform the weighted features into the recovery domain. Finally, a filter of size $3\times3\times3$ recovers the ideal HR-MS image.

E. Objective Function

In the training phase, given the MPNet denoted as $\Phi(\cdot)$, which is parameterized by θ, the objective is to determine the optimum θ. Accordingly, the object function can be formulated as:

$$\theta = \arg\min_{\theta} \frac{1}{N} \sum_{y_{M,n}\in X_{tain}} L\left(\hat{y}_{M,n}, y_{M,n} \middle| \theta\right) \tag{12}$$

$$L\left(\hat{y}_M, y_M \,|\, \theta\right) = \|\Phi\left(x_P, x_M|\theta\right) - y_M\|_1 + \lambda\|\theta\|_2^2 \tag{13}$$

where $L(\cdot)$ is a loss function; X_{train} indicates the training data set, which has N pairs of x_P (PAN image), x_M (MS image), and y_M (ground truth); $\hat{y}_M$ represents a fused MS image. The first part of this loss function is the L1 norm, which is computationally efficient and can obtain relatively sharp edges [16], [18], [48]. To prevent over-fitting, the loss function is regularized with the L2 penalty $\|\theta\|_2^2$ [49]; λ is a balancing parameter that balances the importance of the L1-loss and the regularization term, which is empirically set to 10^{-5} after trial and error in the present implementation. Upon convergence, the parameter θ is fixed and used for tests with both full-resolution and reduced-resolution.

IV. Experimental Investigation

This section presents systematic performance evaluation of the proposed approach. The open-available large-scale data sets and experimental setup are first outlined, followed by discussions about the results.

A. Data Sets

Three publicly available large-scale data sets [50] are adopted to compare the performance of MPNet with SOTA methods. The original date are acquired by three satellites: IKONOS, QuickBird, and WorldView-4. Each satellite carries a PAN sensor and an MS sensor. As the electromagnetic spectrum of IKONOS, QuickBird, and WorldView-4 differs from each other, it does not make practical sense attempting to merge these three data sets, be it for training or testing. This is why there are separately treated. The remote sensing images are cut into patches. The PAN and MS images have a dimension of 1024×1024 and that of 256×256×4, respectively. The images are gathered in 11-bit radiometric resolution.

Following the common practice in the literature [51], original images with 4 bands are used as the ground truth, and they are down-sampled to obtain the simulated MS and PAN images with low spatial resolution according to Wald protocol [52]. The parameters about Modulation Transfer Function (MTF) are set the same as [53]. The specification of patch numbers used during these experimental evaluations, per data set, is given in the newly added Table I.

TABLE I
Distribution of patches for training, validation, and testing.

Data set	Train set	Validation set	Test set
IKONOS	120 (256×256)	20 (256×256)	60 (256×256)
QuickBird	300 (256×256)	50 (256×256)	150 (256×256)
WorldView-4	300 (256×256)	50 (256×256)	150 (256×256)

B. Experimental Setting

MPNet is implemented within the PyTorch framework [54]. For each data set, Adam [55] is selected to train the proposed network, and the number of epochs for loss converge is set to 600. The experiments are carried out on a GPU server. Two NVIDIA GeForce TITAN Xp GPUs (12GB memory per GPU) are used for training. The batch size is set to 10 with limited size of GPU memory. The learning rate is initially set to 0.001 and reduced 20% per 150 epochs. The other hyper-parameters of MPNet in this paper are shown in Table II. The kernel dimensionalities of the PAN FEP are denoted by $W^2 \times C/S$ for width, channel, and stride sizes. In the MS FEP, the kernels and strides are represented as $W^3 \times B \times C/S$, where B indicates the number of bands. The representation of features takes the form of $W^2 \times C$ and $W^2 \times B \times C$, respectively.

C. Reduced-Resolution Experiments

For both qualitative and quantitative performance evaluation, the proposed MPNet is compared with seven SOTA methods including: PNN+ [16], Pan-GAN [19], ResTFNet [18], SRPPNN [20], BDSD-PC [10], MTFGLPFS [11], and FE-HPM [14]. All these compared methods are implemented with the publicly available codes, where the parameters of these methods are set according to their original specifications in the corresponding references.

TABLE II
Hyper-parameters of FEPs.

FEP		Stem	Level 1, 2, and 3	Level 4
PAN FEP	Kernel/Stride	$3^2 \times 64/1$	$3^2 \times 64/1$	$3^2 \times 64/1$
	Input	256^2	$256^2 \times 64$	$256^2 \times 64$
	F_{2D}	$256^2 \times 64$	$256^2 \times 64$	$256^2 \times 64$
MS FEP	Kernel/Stride	$3^3 \times 32/1$	$3^3 \times 32/1$	$3^3 \times 32/1$
	Input	$256^2 \times 4$	$256^2 \times 4 \times 32$	$256^2 \times 4 \times 32$
	F_{3D}	$256^2 \times 4 \times 32$	$256^2 \times 4 \times 32$	$256^2 \times 4 \times 32$

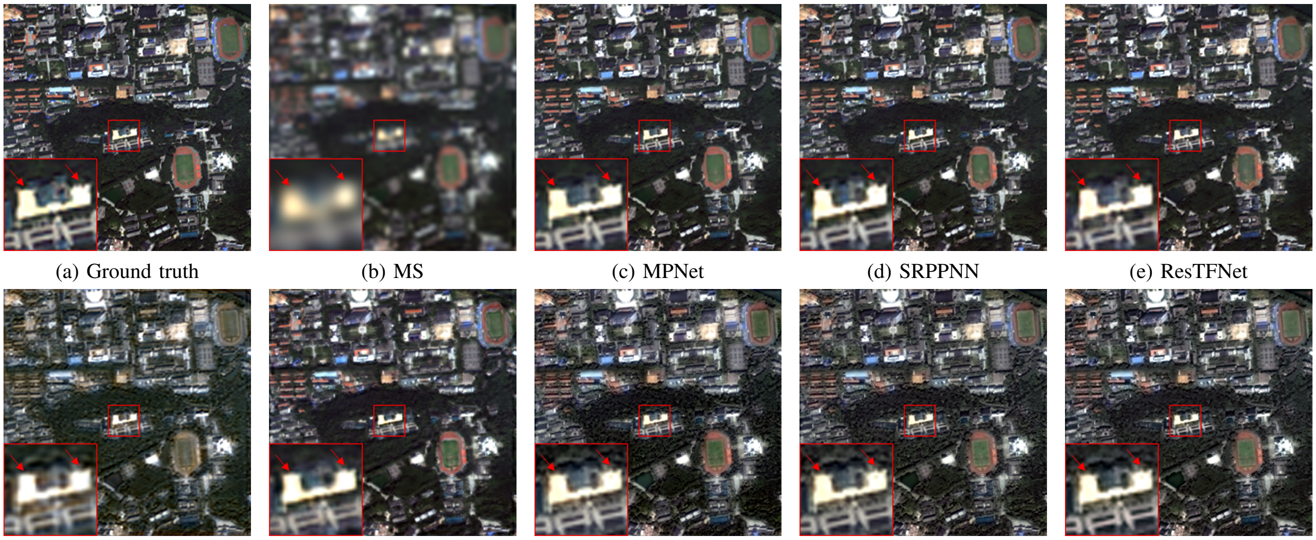
For **qualitative evaluation**, the fused images are visualized to check spatial and spectral distortions. First, consider the IKONOS data set. Fig. 7 shows an example of the experimental results performed on an IKONOS image. Since the MS images have more than three bands, only red, green, and blue bands are extracted to synthesize the TrueColor images in this illustration. The ground truth is shown in Fig. 7a, with Fig. 7c-(h) displaying the pan-sharpened images by different methods. The proposed MPNet produces the pan-sharpened image with the best visual quality in terms of spatial preservation, e.g., the shape of the white building of MPNet is the closest to the ground truth. The residual maps in Fig. 8 also show that the MPNet produces the least distortion.

Fig. 9 shows the visualized results of an experiment performed on the QuickBird data set. BDSD-PC, MTFGLPFS, and FE-HPM generate more details than the ground truth, which indicates over-sharpening, one kind of spatial distortion. PNN+, Pan-GAN, ResTFNet, and SRPPNN produce blurred results. MPNet can obtain the most similar pan-sharpened image compared with other methods. Besides, MPNet produces the lest error according to the residual maps in Fig. 10.

Fig. 11 shows the results on the WorldView-4 data set. Although it fails to identify the apparent distortion, the quality of fused images can be identified in some details. For example, the left boundary of the white building in the enlarged area is recovered by MPNet, which shows it is better than other methods. In addition, the residual maps in Fig. 12 demonstrate the superior performance of MPNet over the rest.

For **quantitative evaluation**, MPNet and SOTA methods are compared using five popular performance indices, namely: Q4 [56], Universal Image Quality Index (UIQI) [57], Spectral Angle Mapper (SAM) [58], relative dimensionless global error in synthesis (ERGAS) [59] and Spatial Correlation Coefficient (SCC) [60]. The indices Q4, UIQI, and ERGAS are exploited to comprehensively assess the spectral and spatial quality of fused images. SCC is another widely used index to measure the spatial quality of a fused image. In addition, SAM is employed to effectively measure any spectral distortion in a fused image in comparison with the ground truth.

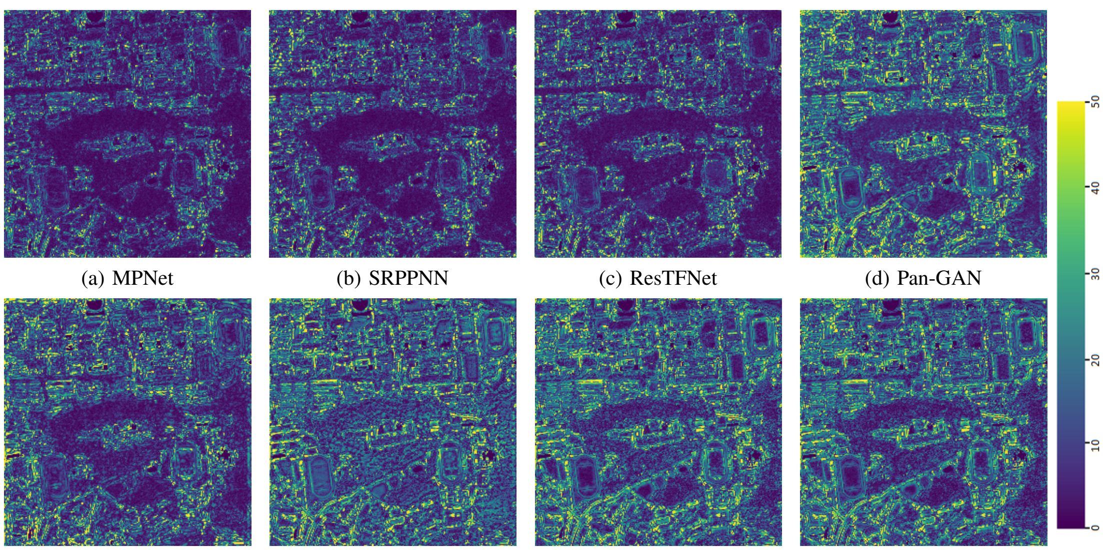
The quantitative evaluation results are shown in Tables III-V. The optimal results are highlighted in bold font. For the spectral metric SAM, the spatial metric SCC, and indeed for other global metrics, MPNet significantly outperforms the



(a) Ground truth (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet

(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 7. Pan-sharpened images by different methods on the IKONOS data set.



(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN

(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 8. Residual maps between pan-sharpened images of different methods and the reference image (IKONOS images).

TABLE III

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON IKONOS DATA SET. OPTIMAL RESULTS ARE INDICATED IN BOLD FONT.

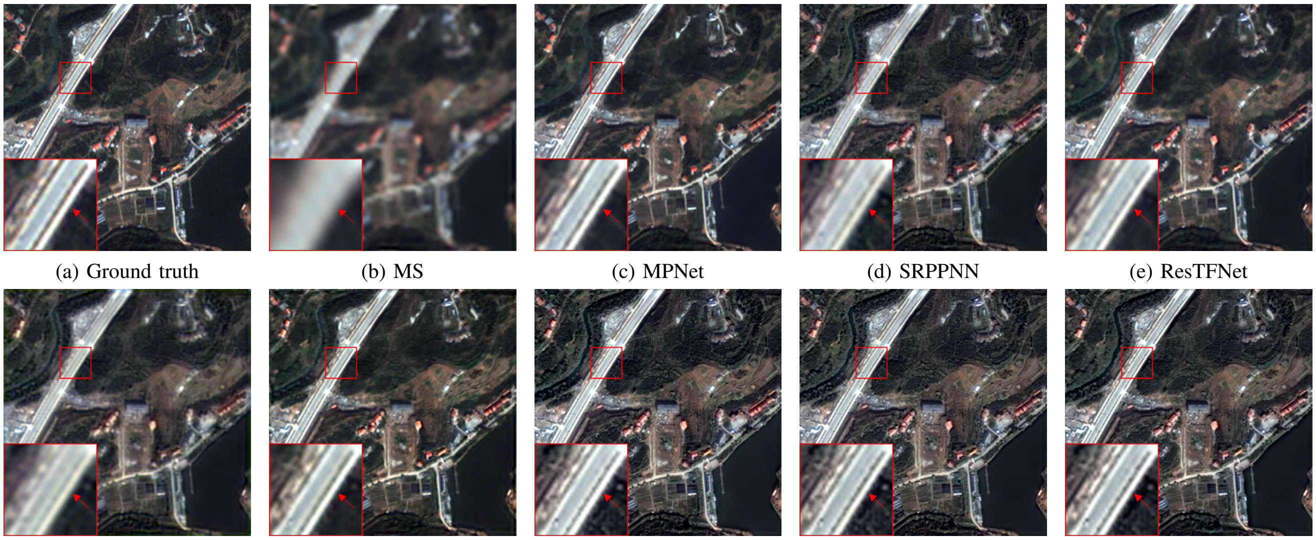
Method	Q4	UIQI	SAM	ERGAS	SCC	D_λ	D_S	QNR
FE-HPM [14]	.7324	.7319	2.4053	1.8557	.9336	.0553	.0627	.8853
MTFGLPFS [11]	.7290	.7246	2.5450	1.9798	.9168	.0533	.0584	.8909
BDSD-PC [10]	.6973	.7213	2.6005	1.9368	.9272	.0391	.0502	.9132
PNN+ [16]	.7472	.7932	2.0571	1.8185	.9459	.0639	.1219	.8219
Pan-GAN [19]	.7452	.7951	2.0450	2.1358	.9276	.1191	.0709	.8186
ResTFNet [18]	.8852	.8895	1.7461	1.3612	.9722	.0858	.0492	.8692
SRPPNN [20]	.9017	.9011	1.6187	1.2644	.9749	.0647	.0516	.8873
MPNet	**.9071**	**.9078**	**1.5359**	**1.2237**	**.9763**	**.0329**	**.0473**	**.9223**
Ideal value	1	1	0	0	1	0	0	1

TABLE IV

QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON QUICKBIRD DATA SET. OPTIMAL RESULTS ARE INDICATED IN BOLD FONT.

Method	Q4	UIQI	SAM	ERGAS	SCC	D_λ	D_S	QNR
FE-HPM [14]	.8602	.8558	1.1161	.8616	.9763	.0558	.0773	.8705
MTFGLPFS [11]	.8752	.8631	.9473	.7902	.9827	.0415	.0710	.8898
BDSD-PC [10]	.8722	.8711	.9399	.7800	.9832	**.0232**	.0592	.9195
PNN+ [16]	.8996	.9006	1.2617	1.0662	.9653	.0265	**.0351**	**.9392**
Pan-GAN [19]	.8031	.8228	1.1480	1.0168	.9612	.0531	.0617	.8887
ResTFNet [18]	.8807	.8848	.9323	.7284	.9839	.0619	.0452	.8953
SRPPNN [20]	.8512	.8485	1.2437	.9639	.9679	.0578	.1002	.8478
MPNet	**.8899**	**.8901**	**.8326**	**.6160**	**.9898**	.0347	.0684	.9002
Ideal value	1	1	0	0	1	0	0	1

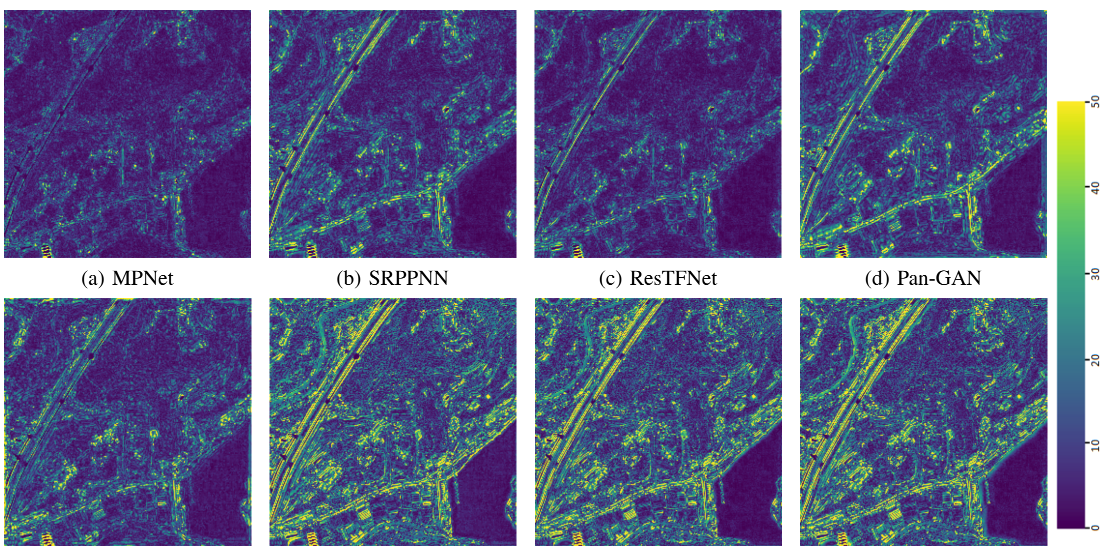
existing SOTA methods. This demonstrates that the proposed MPNet significantly beats the compared SOTA methods, and the pan-sharpened images obtained with MPNet have the least spatial and spectral distortions.



(a) Ground truth (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet

(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 9. Pan-sharpened images by different methods on the QuickBird data set.



(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN

(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 10. Residual maps between pan-sharpened images of different methods and the reference image (QuickBird images).

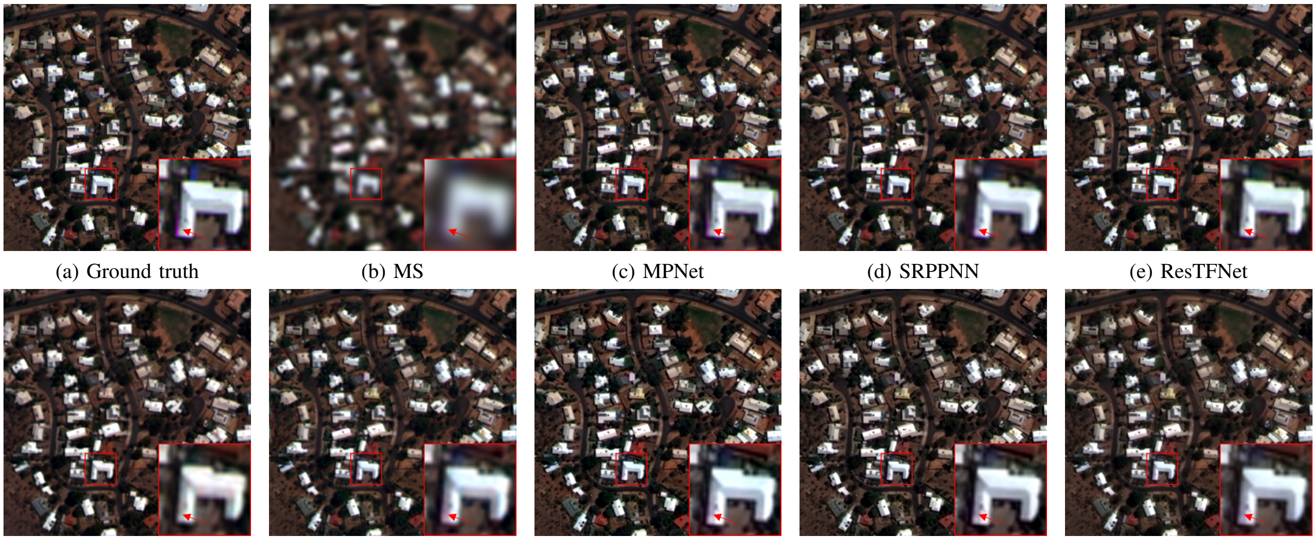
TABLE V
QUANTITATIVE EVALUATION OF DIFFERENT METHODS ON WORLDVIEW-4 DATA SET. OPTIMAL RESULTS ARE INDICATED IN BOLD FONT.

Method	Q4	UIQI	SAM	ERGAS	SCC	D_λ	D_S	QNR
FE-HPM [14]	.7790	.7852	1.9789	2.0279	.8962	.0436	.1538	.8093
MTFGLPFS [11]	.7760	.7816	2.0738	2.2135	.8958	.0402	.1528	.8132
BDSD-PC [10]	.7778	.7835	1.9693	2.0211	.8913	.0285	**.0387**	**.9331**
PNN+ [16]	.8540	.8615	1.8973	1.6915	.9267	**.0231**	.0787	.9001
Pan-GAN [19]	.7819	.8243	2.2899	2.0106	.9397	.0903	.0802	.8379
ResTFNet [18]	.8677	.8766	1.5273	1.3310	.9762	.0699	.0552	.8794
SRPPNN [20]	.8678	.8743	1.6525	1.5756	.9647	.0638	.0647	.8766
MPNet	**.8912**	**.8891**	**1.4119**	**1.2078**	**.9790**	.0364	.0600	.9067
Ideal value	1	1	0	0	1	0	0	1

D. Full-Resolution Experiments

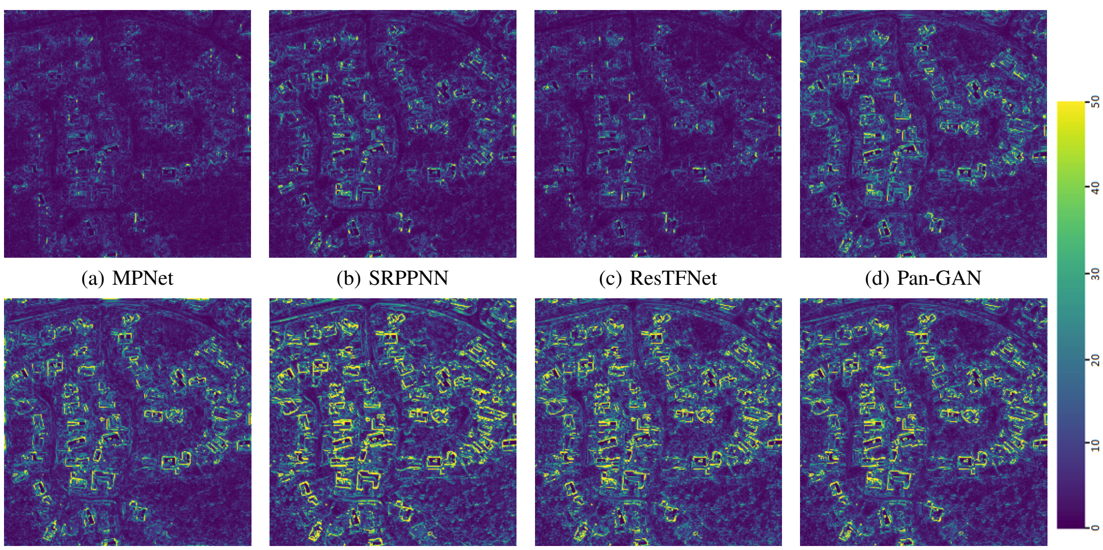
Further to the experimental results at reduced resolution level, the proposed MPNet is herein compared with the other methods at full-resolution, where the PAN and MS images of the original spatial resolutions are fused. Again, the experimental investigations are carried out via both qualitative and quantitative evaluations.

For **qualitative evaluation** on full-resolution images, the results of different methods are visualised. In particular, PAN images are shown in Fig. 13a, 15a, and 17a to inspect spatial distortion. Fig. 13b, 15b, and 17b are the corresponding MS images reflecting the spectral information. For fair and more effective comparison, a small region of all sub-images is scaled up. In addition, as shown in Figs. 18, 14, and 16, we also give a visual inspection of the detail injected into the up-sampled MS, as the quality of each pansharpening technique depends on its ability to inject high-frequency detail. From careful comparison, we can find the proposed MPNet can make full use of the spatial information embedded in the PAN image,



(a) Ground truth (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet
(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 11. Pan-sharpened images by different methods on the WorldView-4 data set.



(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN
(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 12. Residual maps between pan-sharpened images of different methods and the reference image (WorldView-4 images).

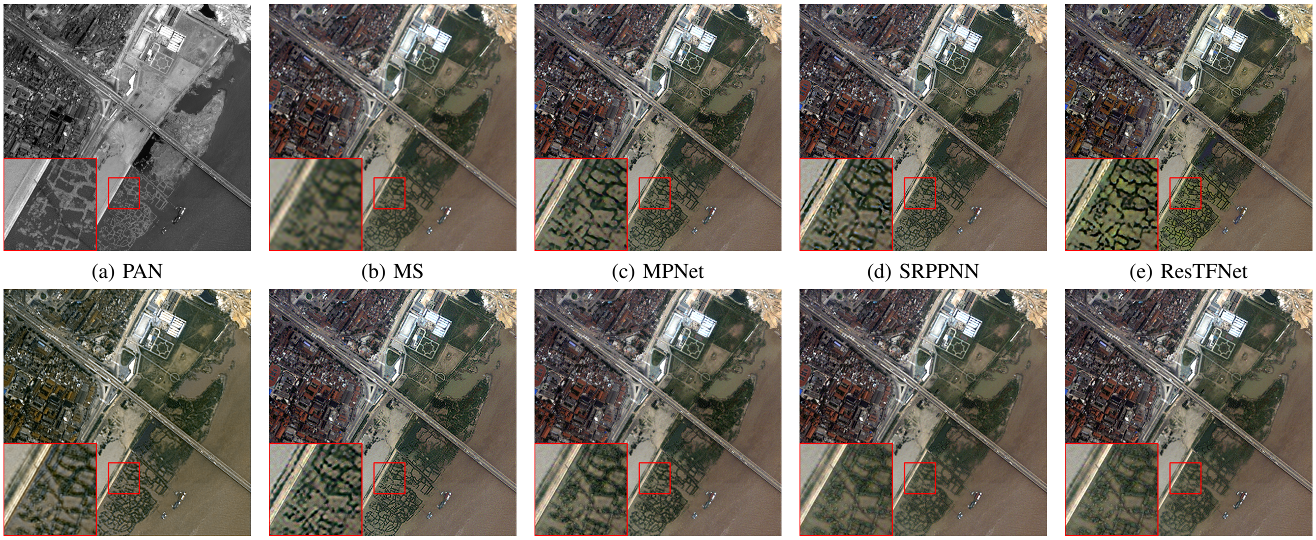
but also prevents the spectral distortion.

In terms of **quantitative evaluation**, the indices used in the previous section are not employed here since no ground truth exists. The reference-free measurement QNR [61] is used here to assess the pan-sharpened images. The QNR index is composed of two components: the spectral distortion index D_λ and the spatial distortion index D_s. Tables III-V present the comparative results, which are obtained by calculating the mean over all images on each data set. The optimal results are in highlighted bold font. It can be seen that compared with other SOTA methods, MPNet achieves competitive performance with respect to the performance indices examined. Some other methods, i.e., BDSD-PC, PNN+, outperform the proposed approach on the full-resolution data since they adapt their models on the images of the test data set. It should be noted that without experiencing the test data, our method still achieves the best full-resolution performance on the IKONOS data set, showing its effectiveness.

E. Further Evaluations

To investigate the potential of the proposed approach in more detail, a number of important further experimental studies are carried out, as discussed below. We select the results of MPNet as the benchmark in Table VI.

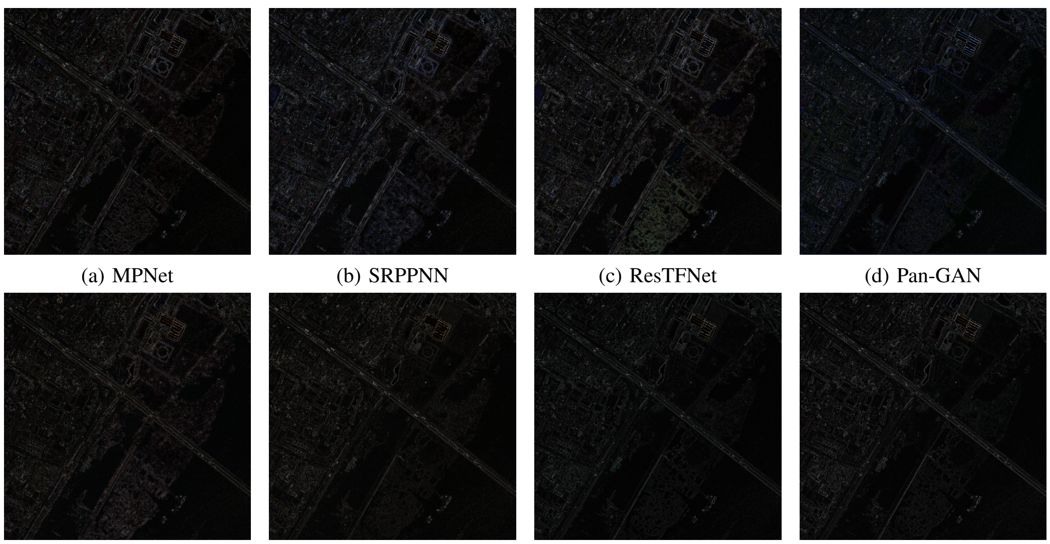
1) Effect of FEP: There are two FEPs, i.e., the PAN FEP and the MS FEP, used in MPNet. The MS FEP leverages 3D convolutions and CSSA mechanism, while the PAN FEP equips 2D convolutions and CSA mechanism. 2D convolutions are wildly used in MS pan-sharpening and the CSA mechanism have investigated in hyperspectral pan-sharpening. It is, therefore, interesting to examine the effectiveness of the heterogeneous architecture, 3D convolutions, and CSSA



(a) PAN (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet

(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 13. Pan-sharpened images by different methods at full-resolution on the IKONOS data set.



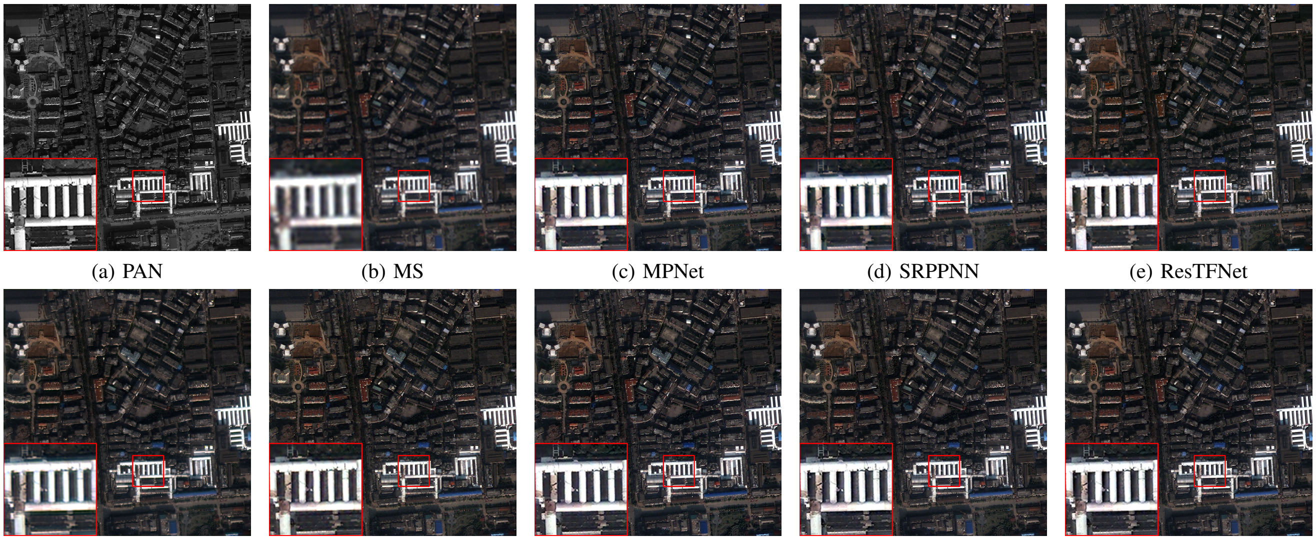
(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN

(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 14. The residuals to the LRMS image from Figure 13.

mechanism. Thus, we conducted three experiments to evaluate their effects. First, we conduct the experiment with the homogeneous architecture, where 3^2 and CSA are utilized in the MS FEP. The results show in Table VI, which incurs performance drops indicating the heterogeneous architecture is more suitable in our case. To illustrate the effect of 3D convolutions, we conduct another experiment here, where 3D $3^2 \times 1$ convolutions and CSA are incorporated. For fair comparison, an equal amount of parameters are used as with the previous experiments, with the same hyper-parameters employed in each setting. The results are reported in Table VI. As can be seen, the use of 2D 3^2 convolutions leads to performance drops, compared to the use of 3D $3^2 \times 1$ convolutions. For the sake of investigating the effect of CSSA in the proposed MPNet, we compared it with CSA. The results of this variant are listed in VI. As can be seen, adopting CSA in the FEP for MS images causes performance drops.

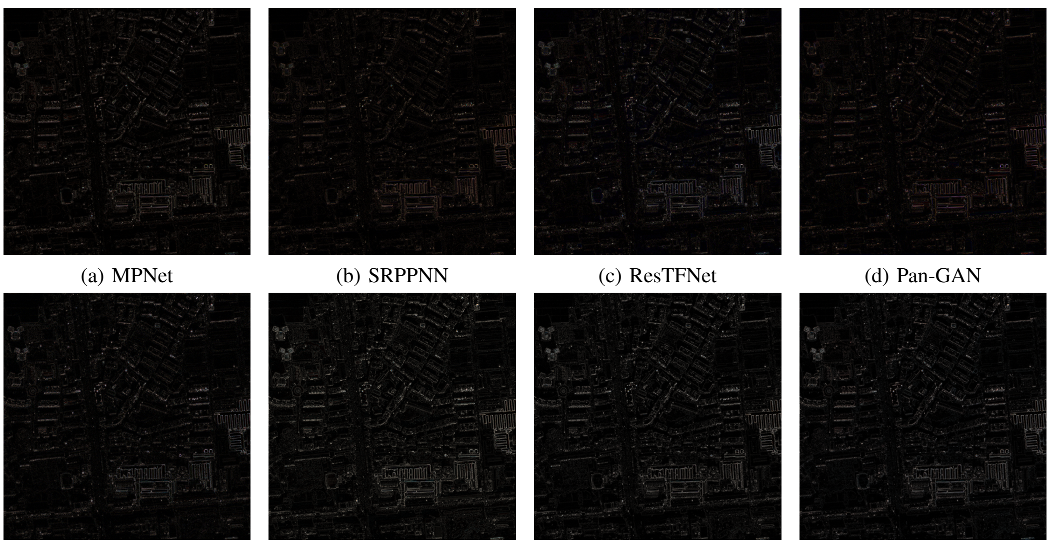
2) Effect of Hierarchical Features: To illustrate the effectiveness of hierarchical features, four additional experiments are conducted. In each experiment, a different number of levels is adopted for fusion. The results of all settings are listed in Table VI. Particularly, the level set {1, 2, 3, 4} indicates that MPNet employs four levels of features for fusion; the set {4} represents that ConvLSTM only utilizes those level-4 features; and {3,4}, {2,3,4} indicate that ConvLSTM merge the two FEPs at two levels 3,4 and at three levels 2,3,4, respectively. For each setting, the same hyper-parameters are used; for instance, when employing Adam optimizer, the learning rate, the number of epochs, and so on are each assigned the same values for different methods run. The results show that with more levels involved for fusion, the model can achieve better results.



(a) PAN (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet

(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 15. Pan-sharpened images by different methods at full-resolution on the QuickBird data set.



(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN

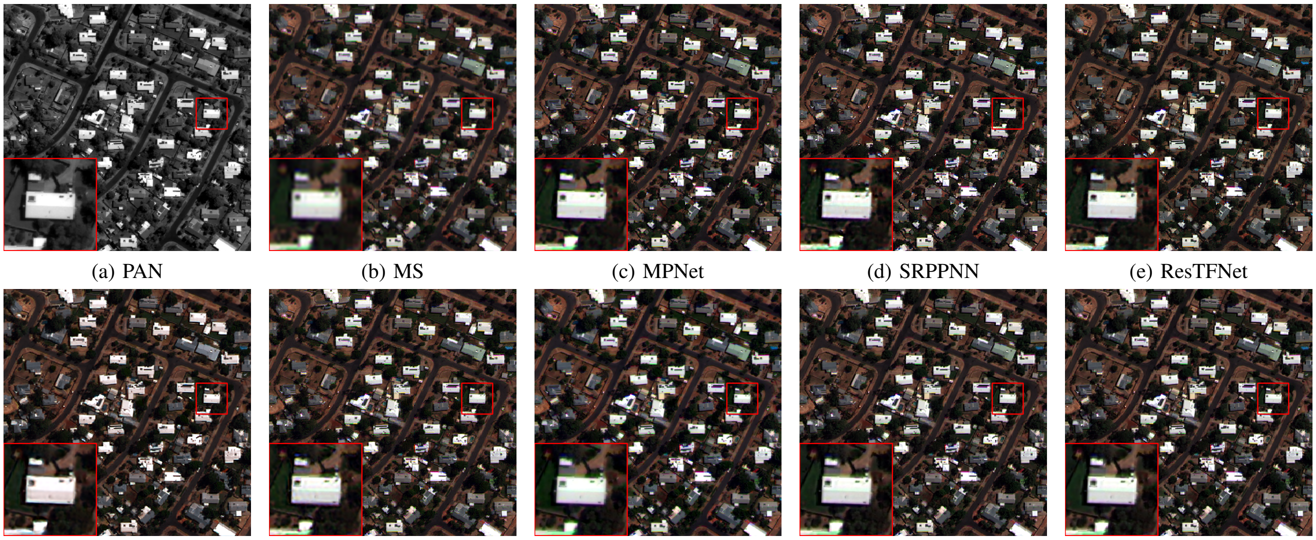
(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 16. The residuals to the LRMS image from Figure 15.

3) Effect of ConvLSTM: To evaluate the effect of the employed ConvLSTM, we compare it with several standard fusion methods, e.g., sum fusion [62], max fusion [62], prod fusion [63], and Conv fusion [62]. At the same time, ResBlock and CSSAResBlock are both investigated for fusion, where the concatenation operation is attached before these blocks. Similar to the ConvLSTM, for each of the replaced fusion operations, the fused features at the previous levels, except the last level, are fed back into the two FEPs, and the fused feature at the last level is directly injected into the reconstruction network. The results are shown in Table VI, which demonstrates the significantly superior performance of ConvLSTM.

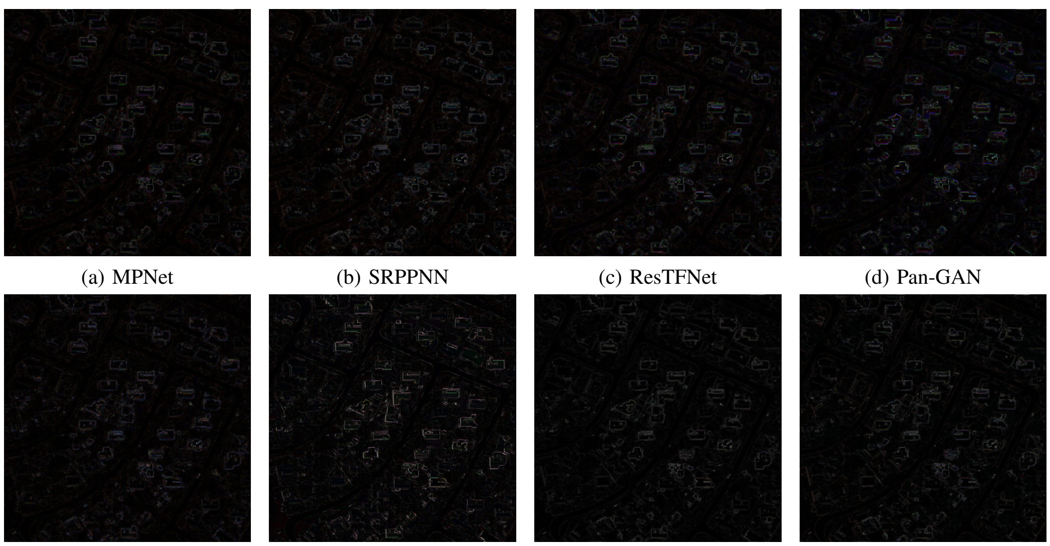
4) Effect of Attention Modules: To demonstrate the effect of attention modules in the building block of MPNet, further experiments with "ResBlock" are conducted. The results are given as the entries for the item of "ResBlock" in the "building block" section of Table VI. Compared with the MPNet, their performance declines obviously, which shows the effectiveness of the attention modules. In addition, "DenseBlock" is also utilized for comparison. The results are given as the entries for the item of "DenseBlock" in Table VI. As can be seen, this variant causes a drop in performance, which demonstrates the effect of proposed CSAResBlock and CSSAResBlock.

5) Impact of Hyper-Parameters: One of the most critical hyper-parameters in MPNet is the number of levels. It is common knowledge that the nonlinearity of CNNs can be improved by increasing the depth of a network [64]. Indeed, the concept of residual learning has been introduced to construct a very deep CNN for performance enhancement, by making full use of the high nonlinearity of deep CNN models [22]. However, a too deep CNN may lead to over-fitting with the



(a) PAN (b) MS (c) MPNet (d) SRPPNN (e) ResTFNet
(f) Pan-GAN (g) PNN+ (h) BDSD-PC (i) MTFGLPFS (j) FE-HPM

Fig. 17. Pan-sharpened images by different methods at full-resolution on the WorldView-4 data set.



(a) MPNet (b) SRPPNN (c) ResTFNet (d) Pan-GAN
(e) PNN+ (f) BDSD-PC (g) MTFGLPFS (h) FE-HPM

Fig. 18. The residuals to the LRMS image from Figure 17.

limited data and a heavy computational burden. Thus, it is also interesting to investigate the impact of the network depth (i.e., number of levels) upon the effectiveness of MPNet. For this purpose, further comparative experiments have been conducted concerning MPNets with different depths on the IKONOS data set. The results are summarised in Table VI. The proposed MPNet with 4 levels achieves superior performance when compared to the other three networks consisting of 2, 3, or 5 levels, respectively. Apart from the number of levels, the kernel size and the number of channels are important hyper-parameters. The results of models with $C = 32$, $C = 128$, and the kernel size of 5 are demonstrated, as also given in Table VI. It can be seen that the variants with $C = 32$ or $C = 128$ lead to performance drops. Although the model with the kernel size of 5 performs better, the marginal computational cost for such a small gain is unacceptable.

6) Impact of Regularization Term: In order to prevent the over-fitting problem, the regularization term is employed in this work. We present the experiments on the IKONOS data set using different λ. The test losses on the IKONOS data set are exhibited in Fig. 19. Obviously, with $\lambda = 10^{-5}$ the loss converges better than the others. In addition, the average quantitative assessments of different λ are listed in Table VI. As can be seen from Table VI, the model with $\lambda = 10^{-5}$ obtains better results than the others in terms of all objective evaluation metrics.

F. Model Complexity

We list the execution time , the training time, and the number of the trainable parameters of fusion methods, including CNN-based ones in Table VII. FE-HPM, MTFGLPFS, and BDSD-PC are performed on an Inter(R) Xeon(R) E5-

TABLE VI
EFFECT OF EACH COMPONENT IN MPNET.

		Q4	UIQI	SAM	ERGAS	SCC	QNR
MS FEP	3^2 + CSA	.8660	.8765	1.9450	1.4983	.9638	.8886
	$3^2 \times 1$ + CSA	.8902	.8942	1.6933	1.3445	.9705	.9068
	3^3 + CSA	.8972	.9004	1.6377	1.2906	.9732	.9123
Hierarchical features	{4}	.8129	.8465	2.3431	1.7583	.9536	.8435
	{3,4}	.8457	.8682	2.4762	1.6411	.9622	.8895
	{2,3,4}	.8503	.8645	2.0206	1.5690	.9607	.8805
	{1,2,3,4}	.8911	.8947	1.6938	1.3252	.9712	.9075
Fusion opeartion	Sum	.8691	.8815	1.8650	1.4522	.9660	.8707
	Max	.8811	.8881	1.7826	1.3956	.9678	.8992
	Product	.8886	.8925	1.7223	1.3584	.9696	.8916
	Conv	.8906	.8940	1.7082	1.3491	.9703	.8976
	ResBlock	.8864	.8930	1.7454	1.3530	.9701	.9014
	CSSAResBlock	.8985	.8998	1.6255	1.2934	.9729	.9129
Building block	ResBlock	.9027	.9040	1.5894	1.2621	.9745	.9139
	DenseBlock	.9059	.9069	1.5577	1.2340	.9756	.9066
Regularization term	10^{-6}	.9050	.9060	1.5692	1.2466	.9753	.9141
	10^{-4}	.8943	.8957	1.6667	1.3227	.9712	.9130
	10^{-3}	.8703	.8787	1.9048	1.4857	.9643	.9026
Number of levels	1	.8594	.8695	1.9830	1.5407	.9617	.8754
	2	.8630	.8737	1.9564	1.5188	.9627	.8967
	3	.8783	.8875	1.8389	1.4038	.9685	.8973
Kernel size	5	.9089	.9098	1.5303	1.2149	.9771	.9284
Channels	32	.8597	.8753	1.9830	1.4898	.9641	.8969
	128	.8977	.9000	1.6236	1.2868	.9730	.9102
	MPNet	.9071	.9078	1.5359	1.2237	.9763	.9223
Ideal value		1	1	0	0	1	1

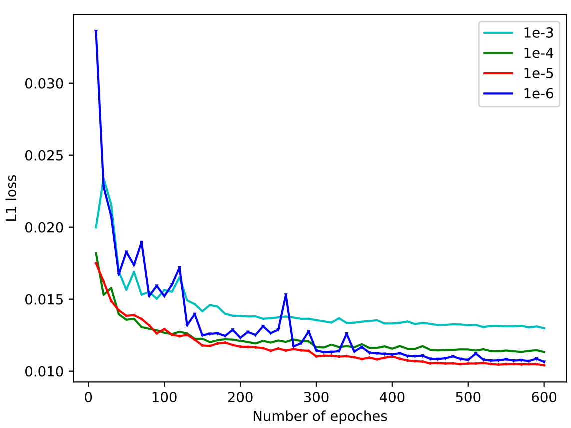


Fig. 19. Test losses of MPNet on IKONOS data set.

2620@2.10GHz via Matlab R2020b. CNN-based methods are implemented on the GPU (NVIDIA GeForce TITAN Xp) through their publicly available codes. In general, the processing speed of CNN-based methods is not slower than traditional methods because the GPU used for implementation helps improve the efficiency of CNN-based methods. MPNet requires more computational time due to 3D convolutions.

V. CONCLUSION

This paper has presented a novel MPNet for MS pan-sharpening. In stead of employing 2D CNNs for processing both PAN and MS images, a heterogeneous pair of FEPs are developed for the extraction of 2D feature maps and 3D representations from PAN and MS images, respectively. Equipped with CSA or CSSA, the FEPs can learn more informative hierarchical features. The ConvLSTM-based HFFM

TABLE VII
EXECUTION TIME, TRAINING TIME, AND NUMBER OF TRAINABLE PARAMETERS WITH OPTIMAL RESULTS INDICATED IN BOLD.

Method	Execution time	Training time	Parameters
FE-HPM [14]	0.36s(CPU)	-	-
MTFGLPFS [11]	0.12s(CPU)	-	-
BDSD-PC [10]	0.19(CPU)	-	-
PNN+ [16]	**0.01(GPU)**	22 hours	**48K**
Pan-GAN [19]	0.02(GPU)	10 hours	887K
ResTFNet [18]	0.15S(GPU)	4 hours	355K
SRPPNN [20]	0.09s(GPU)	**3 hours**	343K
MPNet	0.39s(GPU)	27 hours	952K

is developed to merge the resulting hierarchical feature extraction. Compared with SOTA methods in the literature, the proposed approach offers superior or competitive performance. For future work, it would be interesting to consider how the loss functions, e.g., perceptual loss and SSIM, employed within the current method may be optimized. Also, it would be worth investigating to improve the performance of CNN-based fusion models on real-world data via unsupervised adaptation learning.

REFERENCES

[1] E. L. Bullock, C. E. Woodcock, and P. Olofsson, "Monitoring tropical forest degradation using spectral unmixing and landsat time series analysis," *Remote Sensing of Environment*, vol. 238, p. 110968, 2020.

[2] Y. Gong, Z. Xiao, X. Tan, H. Sui, C. Xu, H. Duan, and D. Li, "Context-aware convolutional neural network for object detection in vhr remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 34–44, 2019.

[3] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for remote sensing image classification: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 6, p. e1264, 2018.

[4] H. Zhang, Y. Li, Y. Jiang, P. Wang, Q. Shen, and C. Shen, "Hyperspectral classification based on lightweight 3-d-cnn with transfer learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5813–5828, 2019.

[5] P. Zhang, Y. Bai, D. Wang, B. Bai, and Y. Li, "Few-shot classification of aerial scene images via meta-learning," *Remote Sensing*, vol. 13, no. 1, p. 108, 2021.

[6] P. Zhang, Y. Li, D. Wang, and J. Wang, "Rs-sskd: Self-supervision equipped with knowledge distillation for few-shot remote sensing scene classification," *Sensors*, vol. 21, no. 5, p. 1566, 2021.

[7] P. Zhang, Y. Bai, D. Wang, B. Bai, and Y. Li, "A meta-learning framework for few-shot classification of remote sensing scene," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4590–4594.

[8] D. Wang, Y. Li, L. Ma, Z. Bai, and J. C.-W. Chan, "Going deeper with densely connected convolutional neural networks for multispectral pansharpening," *Remote Sensing*, vol. 11, no. 22, p. 2608, 2019.

[9] D. Wang, Y. Bai, B. Bai, C. Wu, and Y. Li, "Heterogeneous two-stream network with hierarchical feature prefusion for multispectral pan-sharpening," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1845–1849.

[10] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6421–6433, 2019.

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

[12] P. Zhuang, "Pan-sharpening with a gradient domain guided image filtering prior," in *2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2019, pp. 1031–1036.

[13] P. Zhuang, Q. Liu, and X. Ding, "Pan-ggf: A probabilistic method for pan-sharpening with gradient domain guided image filtering," *Signal Processing*, vol. 156, pp. 177–190, 2019.

[14] G. Vivone, M. Simoes, M. Dalla Mura, R. Restaino, J. M. Bioucas-Dias, G. A. Licciardi, and J. Chanussot, "Pansharpening based on semiblind deconvolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 4, no. 53, pp. 1997–2010, 2015.

[15] G. Vivone, L. Alparone, A. Garzelli, and S. Lolli, "Fast reproducible pansharpening based on instrument and acquisition modeling: Awlp revisited," *Remote Sensing*, vol. 11, no. 19, p. 2315, 2019.

[16] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[17] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1188–1204, 2019.

[18] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, 2020.

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

[20] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

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

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

[23] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5449–5457.

[24] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.

[25] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5549–5563, 2019.

[26] S. Ji, C. Zhang, A. Xu, Y. Shi, and Y. Duan, "3d convolutional neural networks for crop classification with multi-temporal remote sensing images," *Remote Sensing*, vol. 10, no. 1, p. 75, 2018.

[27] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[28] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 11, no. 5, pp. 1656–1669, 2018.

[29] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "Psgan: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[30] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[31] Z. Bai, Y. Li, M. Woźniak, M. Zhou, and D. Li, "Decomvqanet: Decomposing visual question answering deep network via tensor decomposition and regression," *Pattern Recognition*, p. 107538, 2020.

[32] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[33] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convlstm for video salient object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 715–731.

[34] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *European conference on computer vision*. Springer, 2016, pp. 816–833.

[35] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," *arXiv preprint arXiv:1709.06316*, 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[38] Y. Zheng, J. Li, Y. Li, J. Guo, X. Wu, and J. Chanussot, "Hyperspectral pansharpening using deep prior and dual attention residual network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8059–8076, 2020.

[39] Q. Li, Q. Wang, and X. Li, "Exploring the relationship between 2d/3d convolution for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[40] T. Uemori, A. Ito, Y. Moriuchi, A. Gatto, and J. Murayama, "Skin-based identification from multispectral image data using cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 349–12 358.

[41] Y. Chen, C. Li, P. Ghamisi, C. Shi, and Y. Gu, "Deep fusion of hyperspectral and lidar data for thematic classification," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2016, pp. 3591–3594.

[42] L. Mou and X. X. Zhu, "Learning to pay attention on spectral domain: A spectral attention module-based convolutional network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 110–122, 2020.

[43] B. Aiazzi, S. Baronti, M. Selva, and L. Alparone, "Bi-cubic interpolation for shift-free pan-sharpening," *ISPRS journal of photogrammetry and remote sensing*, vol. 86, pp. 65–76, 2013.

[44] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sensing*, vol. 9, no. 1, p. 67, 2017.

[45] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional lstm based spectral-spatial feature learning for hyperspectral image classification," *Remote Sensing*, vol. 9, no. 12, p. 1330, 2017.

[46] F. Zhou, R. Hang, Q. Liu, and X. Yuan, "Hyperspectral image classification using spectral-spatial lstms," *Neurocomputing*, vol. 328, pp. 39–47, 2019.

[47] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity." in *AAAI*, 2020, pp. 12 797–12 804.

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

[49] Y. Xing, M. Wang, S. Yang, and L. Jiao, "Pan-sharpening via deep metric learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 165–183, 2018.

[50] X. Meng, Y. Xiong, F. Shao, H. Shen, W. Sun, G. Yang, Q. Yuan, R. Fu, and H. Zhang, "A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 18–52, 2020.

[51] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 5, pp. 1037–1041, 2015.

[52] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *American Society for Photogrammetry and Remote Sensing*, vol. 63, no. 6, pp. 691–699, 1997.

[53] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, "A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods," *IEEE Geoscience and Remote Sensing Magazine*, 2020.

[54] N. Ketkar, "Introduction to pytorch," in *Deep learning with python*. Springer, 2017, pp. 195–208.

[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[56] Y. Zeng, W. Huang, M. Liu, H. Zhang, and B. Zou, "Fusion of satellite images in urban area: Assessing the quality of resulting images," in *2010 18th International Conference on Geoinformatics*. IEEE, 2010, pp. 1–4.

[57] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[58] P. E. Dennison, K. Q. Halligan, and D. A. Roberts, "A comparison of error metrics and constraints for multiple endmember spectral mixture analysis and spectral angle mapper," *Remote Sensing of Environment*, vol. 93, no. 3, pp. 359–367, 2004.

[59] E. Ayhan and G. Atay, “Spectral and spatial quality analysis in pan sharpening process,” *Journal of the Indian Society of Remote Sensing*, vol. 40, no. 3, pp. 379–388, 2012.

[60] J. Zhou, D. Civco, and J. Silander, “A wavelet transform method to merge landsat tm and spot panchromatic data,” *International journal of remote sensing*, vol. 19, no. 4, pp. 743–757, 1998.

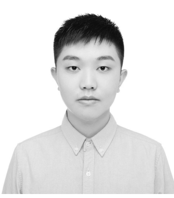
[61] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, “Multispectral and panchromatic data fusion assessment without reference,” *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 193–200, 2008.

[62] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutional two-stream network fusion for video action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1933–1941.

[63] Y. Xing, S. Yang, Z. Feng, and L. Jiao, “Dual-collaborative fusion model for multispectral and panchromatic image fusion,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[64] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *European conference on computer vision*. Springer, 2014, pp. 184–199.

Dong Wang Dong Wang received his Master’s degree in college of information engineering, the Northwest A&F University in 2018. He is currently pursuing the Ph.D. degree with the School of Computer Science, Northwestern Polytechnical University, Xi’an, China. His current research interests include pan-sharpening and few-shot learning.



Yunpeng Bai received his Master’s degree in information systems from Melbourne School of Engineering, the University of Melbourne in 2020. Since 2018, his research interests are in computer vision, with a focus on applied artificial intelligence and deep learning. As a graduate student, he completed a research project on urban planning and object detection in collaboration with the University of Cambridge. Since then he has co-authored multiple papers on remote sensing satellite image analysis and target detection.

Chanyue Wu received her Master’s degree in college of information engineering, the Northwest A&F University in 2017. Since 2015, her research interests are in point cloud denoising, image processing, and deep learning.

Ying Li received the Ph.D. degree in circuits and systems from Xidian University, Xi’an, China, in 2002.

Since 2003, she has been with the School of Computer Science, Northwestern Polytechnical University, Xi’an, where she is currently a Full Professor. Her current research interests include computation intelligence, image processing, and pattern recognition. She has published extensively in the above areas.

Changjing Shang received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1995.

She was with Heriot-Watt University, Edinburgh, U.K.; Loughborough University, Loughborough, U.K.; and Glasgow University, Glasgow, U.K. She is currently a University Research Fellow with the Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth, U.K. She has published extensively, and supervised more than ten Ph.Ds/PDRAs in the areas of pattern recognition, data mining and analysis, space robotics, and image modeling and classification.

Qiang Shen received the Ph.D. degree in computing and electrical engineering from Heriot-Watt University, Edinburgh, U.K., in 1990, and the D.Sc. degree in computational intelligence from Aberystwyth University, Aberystwyth, U.K., in 2013.

He is appointed as chair of computer science and the Pro Vice-Chancellor for Faculty of Business and Physical Sciences, Aberystwyth University. He has authored two research monographs and over 400 peer-reviewed papers.

Dr. Shen was a recipient of the Outstanding Transactions Paper Award from the IEEE.