# A Triple-Double Convolutional Neural Network for Panchromatic Sharpening

Tian-Jiang Zhang, Liang-Jian Deng, *Member, IEEE*, Ting-Zhu Huang, *Member, IEEE*,
Jocelyn Chanussot, *Fellow, IEEE*, and Gemine Vivone, *Senior Member, IEEE*

***Abstract*—Pansharpening refers to the fusion of a panchromatic image with a high spatial resolution and a multispectral image with a low spatial resolution, aiming to obtain a high spatial resolution multispectral image. In this paper, we propose a novel deep neural network architecture with level-domain based loss function for pansharpening by taking into account the following double-type structures, *i.e.,* double-level, double-branch, and double-direction, called as triple-double network (TDNet). By using the structure of TDNet, the spatial details of the panchromatic image can be fully exploited and utilized to progressively inject into the low spatial resolution multispectral image, thus yielding the high spatial resolution output. The specific network design is motivated by the physical formula of the traditional multi-resolution analysis (MRA) methods. Hence, an effective MRA fusion module is also integrated into the TDNet. Besides, we adopt a few ResNet blocks and some multi-scale convolution kernels to deepen and widen the network to effectively enhance the feature extraction and the robustness of the proposed TDNet. Extensive experiments on reduced- and full-resolution datasets acquired by WorldView-3, QuickBird, and GaoFen-2 sensors demonstrate the superiority of the proposed TDNet compared with some recent state-of-the-art pansharpening approaches. An ablation study has also corroborated the effectiveness of the proposed approach. The code is available at https://github.com/liangjiandeng/TDNet.**

***Index Terms*—Deep Convolutional Neural Networks, Triple-Double Network, Multi-Resolution Analysis, Multi-Scale Feature Extraction, Pansharpening, Multispectral Image Fusion, Remote Sensing.**

## I. Introduction

Remote sensing satellites are dedicated to collect image data from the Earth's surface. However, because of some constraints on the signal-to-noise ratio (SNR) for the sensor hardware, we cannot get high spatial and spectral resolutions in a unique acquisition. Thus, satellites, such as, IKONOS, GaoFen, QuickBird, and WorldView-3, usually capture images containing several spectral bands, called multispectral (MS) images, together with panchromatic (PAN) images having high spatial resolution, *i.e.*, containing many image details. Hence, the fusion of these kinds of data is often required to get very high spatio-spectral resolution products. Pansharpening is the fusion of a PAN image and an MS image to obtain the final outcome with the same spatial resolution as the PAN image and the same spectral resolution as the MS image. This research topic has been rapidly developed in recent years and has been proved to be an effective image fusion method [3]. The results of pansharpening have been widely used in ground object detection, mapping, and image data pre-processing for various high-level applications [4], [5].

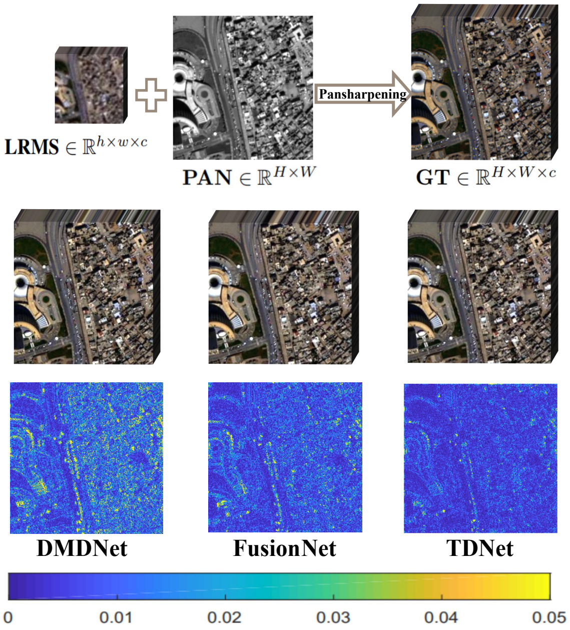

LRMS $\in \mathbb{R}^{h\times w\times c}$
PAN $\in \mathbb{R}^{H\times W}$
Pansharpening
GT $\in \mathbb{R}^{H\times W\times c}$
DMDNet
FusionNet
TDNet
0
0.01
0.02
0.03
0.04
0.05

Fig. 1. First row: pansharpening on a WorldView-3 dataset. This row includes the low spatial resolution MS (LRMS) image, the PAN image, and the desired ground truth (GT) image. Second row: the pansharpened results by three representative DL-based methods, *i.e.*, DMDNet (SAM/ERGAS/Q8=2.9355/1.8119/0.9690) [1], FusionNet (SAM/ERGAS/Q8=2.8338/1.7510/0.9714) [2], and the TDNet (SAM/ERGAS/Q8=2.7373/1.6733/0.9764). Third row: the corresponding error maps, which show that TDNet produces less errors than the other two approaches.

Over the past decades, many different approaches have been proposed for the pansharpening problem, and these techniques can be roughly divided into four categories [6]–[8], *i.e.,* component substitution (CS) methods, multi-resolution

The work is supported by National Natural Science Foundation of China (Grant No. 12171072, 61702083), and Key Projects of Applied Basic Research in Sichuan Province (Grant No. 2020YJ0216) and National Key Research and Development Program of China (Grant No. 2020YFA0714001).

T. -J. Zhang is with the Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mail: zhangtianjinguestc@163.com).

L. -J. Deng and T. -Z. Huang are with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mails: liangjian.deng@uestc.edu.cn; tingzhuhuang@126.com).

J. Chanussot is with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France. (e-mail: jocelyn.chanussot@grenoble-inp.fr).

G. Vivone is with the National Research Council - Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy (e-mail: gemine.vivone@imaa.cnr.it).

The Corresponding author: Liang-Jian Deng

analysis (MRA) methods, variational optimization (VO) approaches, and deep learning (DL) techniques, respectively. In this work, our approach is based on convolutional neural networks (CNNs), thus belonging to DL techniques. In what follows, we will introduce the representative approaches for each category.

CS-based methods are usually simple approaches belonging to traditional techniques. They project the original MS image into a transformation domain, whose purpose is to simplify the replacement of part or all the spatial information, making easier the replacement of the spatial structure components with the PAN image. It is worth mentioning that many pioneering pansharpening methods are based on the CS philosophy because approaches in this category usually have simple and efficient implementations. Some representative examples into this class are the partial replacement adaptive CS (PRACS) [9], the Gram-Schmidt (GS) spectral sharpening [10], and the band-dependent spatial-detail with local parameter estimation (BDSD) [11]. Note that the CS-based methods can generally get products with a better rendering paying it with a greater spectral distortion.

MRA methods are another class of traditional approaches whose goal is to inject the spatial details extracted from the PAN image into the MS image that is interpolated to the size of the PAN image. The MRA-based fused results are superior to those of the CS-based methods considering the spectral quality. However, these methods can easily generate artifacts, thus often introducing spatial distortion. Some methods belonging to this class are, for instance, the smoothing filter-based intensity modulation (SFIM) [12], the additive wavelet luminance proportional (AWLP) [13], the modulation transfer function generalized Laplacian Pyramid with high-pass modulation injection model (GLP-HPM) [14], and the modulation transfer function generalized laplacian pyramid with full resolution regression-based injection model (GLP-Reg) [15].

Unlike the above-mentioned traditional approaches, VO-based methods have been developed by imposing pre-specified prior terms to regularize the underlying high-resolution multispectral (HRMS) image [16]–[18]. These methods show an elegant mathematical formulation and have a good performance on spatio-spectral preservation [19]–[21] compared with some state-of-the-art CS and MRA techniques. The main drawback of VO-based methods is the heavy computational burden, including the tuning of many hyper-parameters. Therefore, CS and MRA approaches are still nowadays used for benchmarking purposes.

Recently, deep learning (DL) techniques have gained much attention due to their powerful ability to implicitly learn the priors from big data. Undoubtedly, methods based on deep learning have been widely used in the field of remote sensing images [22]–[24]. As a newly developed category to solve pansharpening, deep learning requires physical support at a higher level. The structure design is of critical importance since it is closely related to the performance gain of the model. By building a convolutional neural network (CNN) with a certain structure and functional units, (*e.g.*, deep residual network [25], multiscale and multidepth network [26]), the DL method can reproduce the nonlinear relationship between MS images, PAN images, and ideal fusion images through the training on satellite datasets. The groundbreaking attempt was made by Masi *et al.* [27], in 2016, with a three-layer CNN designed specifically for pansharpening, achieving promising results. Inspired by PNN, many researchers developed various structures relied upon CNNs. Among them, the residual module in ResNet [28] is widely used for pansharpening [1], [25], [29]. However, the learning process is difficult to be explained and the neural network often gets into the dilemma of vanishing gradient when the parameters are hard to update. In particular, some essential properties and prior information of the images, such as the uniqueness of high-frequency information, and the intrinsic relationship of the spectrum are often ignored by these types of "black box" deep models, leaving big room for further improvement. Therefore, we argue that the network framework should be designed based on some characteristics of the problem at hand underlining the unique relationships between the input images [16], [30].

In this paper, we propose a novel DL approach for pansharpening, which can exploit a multi-scale spatial details strategy, progressively injecting PAN details into the low-resolution MS image. A novel triple-double network (TDNet) structure is designed based on the MRA formulation. The main contributions of this work can be summarized as follows:

1) We propose an overall structure of the network with double-level, double-branch, and double-direction, which injects the latent multi-scale spatial details of the PAN image to the MS image in a hierarchical and bidirectional way. Under this framework, we adopted a level-domain-based loss function to pose constraints on multi-level outcomes, which ensure reasonable final fusion results.
2) Following the traditional MRA methods, an MRA block (MRAB) embedded in the TDNet structure is designed. The MRAB can better complete the extraction of structural information from the PAN image. The design of this block structure also introduces the idea of the attention mechanism, which is more flexible and robust than traditional methods.
3) Considering the pansharpening problem, which requires the injection of different objects at various scales, a multi-scale convolution kernel module is adopted to deepen and widen the proposed network to improve the capability of the nonlinear fitting. The results, shown in Fig. 1, demonstrate the superiority of the proposed method.

The remaining of this paper is organized as follows. In Sect. II, the background and related works will be briefly introduced. The proposed network is presented in Sect. III. Afterwards, the experimental results and discussion are provided in Sect. IV. Finally, conclusions are drawn in Sect. V.

## II. Notation and Related Works

### A. Notation

For convenience, the notation used throughout this paper is presented first. The low-resolution multispectral (LRMS) im-

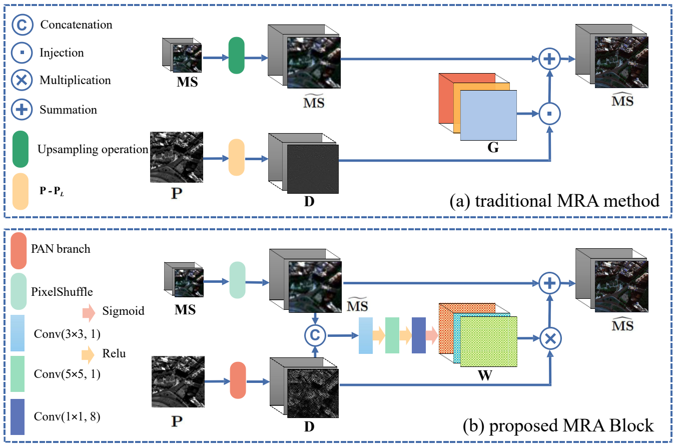


Fig. 2. (a) The diagram of traditional MRA methods. (b) MRA block (MRAB) was designed based on traditional MRA methods. Please note that the upsampling operation in (a) is a polynomial kernel with 23 coefficients [31]. 8-bands datasets are considered to define the number of convolution kernels in (b).

age and the high-resolution (HR) panchromatic (PAN) image are denoted as $\mathbf{MS} \in \mathbb{R}^{h\times w\times c}$ and $\mathbf{P} \in \mathbb{R}^{H\times W}$, respectively. The desired high-resolution multispectral (HRMS) image is defined as $\widehat{\mathbf{MS}} \in \mathbb{R}^{H\times W\times c}$. The multispectral image upsampled at PAN image scale is represented by $\widetilde{\mathbf{MS}} \in \mathbb{R}^{H\times W\times c}$, and the ground truth image is represented as $\mathbf{GT} \in \mathbb{R}^{H\times W\times c}$.

### B. Background

As introduced in Sect. I, due to limitations of hardware devices, LRMS and PAN images are only acquired. Considering the goal of pansharpening, which is to generate multispectral images with high spatial resolution, the general fusion formula can be summarized as follows,

$$\widehat{\mathbf{MS}} = \mathcal{F}_\theta(\mathbf{P},\ \mathbf{MS}), \tag{1}$$

where $\mathcal{F}_\theta(\cdot)$ is used to depict the latent relationship between the involved images. The common idea behind many pansharpening approaches (both traditional and DL-based) is to find the befitting way to characterize the relationship between the known LRMS and PAN images and the desired HRMS image.

### C. Overview of MRA Methods

Traditional MRA methods competitively perform in pansharpening. The schematic diagram of general MRA methods is shown in Fig. 2(a). It can be seen that the MRA methods have two main processes, *i.e.*, extracting spatial structure details from the PAN image, $\mathbf{P}$, and injecting information obtained from the $\mathbf{P}$ into $\widetilde{\mathbf{MS}}$ through certain strategies. The mathematical formulation of MRA methods is given by:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + \mathbf{G} \odot (\mathbf{P} - \mathbf{P_L}), \tag{2}$$

where $\mathbf{G} \in \mathbb{R}^{H\times W\times c}$ is the general form of the injection coefficient gain, $\mathbf{P_L}$ stands for the low-pass version of the PAN image $\mathbf{P}$, and $\odot$ represents the element-wise multiplication. Please, refer to [32] for more details. In (2), the spatial structure can be obtained by the difference $\mathbf{P} - \mathbf{P_L}$, where $\mathbf{P_L}$ can be obtained by different filters, see *e.g.*, [33], [34], and [35]. The related literature also presents various attempts about the detail injection process, see, *e.g.*, [36], [37]. The traditional MRA methods can preserve the spectral information, but paying it with the possible introduction of spatial distortion.

### D. CNNs for Pansharpening

Among DL methods for pansharpening, the techniques based on CNNs have been deeply explored thanks to their excellent ability in the feature extraction phase. Existing CNN-based frameworks addressing the pansharpening problem can be roughly summarized by minimizing the following loss function:

$$\min_{\Theta} \mathcal{L} = \|\mathbf{GT} - \mathcal{N}(\mathbf{P}, \mathbf{MS}; \Theta)\|, \tag{3}$$

where $\mathcal{N}(\cdot;\Theta)$ represents the functional mapping, through the unknown parameter $\Theta$, between the inputs and an ideal HRMS output, and the $\|\cdot\|$ is a function to describe the distance between the outcome of the network (HRMS) and the GT image. The basic structure for pansharpening can be expressed as follows,

$$\begin{aligned} \mathbf{C}^0 &= \{\mathbf{P},\ \mathbf{MS}\}, \\ \mathbf{C}^1 &= \sigma(\mathbf{W}^1 \otimes \mathbf{C}^0 + \mathbf{b}^1), \\ \mathbf{C}^n &= \sigma(\mathbf{W}^n \otimes \mathbf{C}^{n-1} + \mathbf{b}^n), n = 2, \cdots, L, \end{aligned} \tag{4}$$

where the initial $\{\mathbf{P},\ \mathbf{MS}\}$ generated by concatenation or other strategies is fed into the network as input. $\mathbf{C}^i$, $i = 1, 2, \cdots, L$, represents the $i$-th convolution layer with the corresponding weight $\mathbf{W}^i$ and bias $\mathbf{b}^i$, where $L$ is the total number of layers. $\sigma(\cdot)$ is usually a nonlinear activation function, *e.g.,* ReLU.

Many effective and promising CNNs are proposed for the task of pansharpening based on the above-mentioned strategy. In [27], a modified super-resolution network that maps relations through a simple three-layer convolution is proposed by Masi *et al.*. Another typical example is PanNet proposed by Yang *et al.* in [25]. It considers spectral and spatial fidelity on high-pass features and introduces the ResNet structure to deepen the given network. In [38], Yuan *et al.* propose the use of multi-scale convolution kernels to extract features on different image scales achieving satisfactory results compared with the single-scale convolution kernel. Unlike feeding together $\mathbf{P}$ and $\mathbf{MS}$ into the network, Zhang *et al.* in [39] propose a novel network architecture named BDPN, in which $\mathbf{P}$ and $\mathbf{MS}$ are processed using different branches, by exploiting a bi-directional pyramid structure.

### E. Motivation

Although various CNN-based approaches have achieved promising results, there is still room for improvements, *e.g.,* physically interpretable architectures, the use of multi-scale structures, and so forth. Recently, unlike other methods that take CNNs as black boxes, Deng *et al.* in [2] propose FusionNet inspired by traditional CS and MRA methods, which motivates us to regard the formula of traditional methods such as MRA as a guide for the design of the proposed network.

The module inspired by traditional methods can be embedded into the CNN network to have a better details extraction and injection.

Besides, existing CNN-based techniques do not fully explore and utilize the multi-scale information in the PAN and MS images loosing some possible information in the process of enhancing the LRMS image. This inspires us to focus on the information injection in hierarchical and bidirectional ways, which is the original intention of the triple-double structure.

## III. The Proposed Network

As stated before, our model is inspired by traditional MRA methods, where the spatial structure information extracted from the PAN image was added to the upsampled LRMS image. The overall flowchart of the proposed network has been shown in Fig. 3, which includes the following parts: 1) the MRA block (MRAB), whose structure is based on the MRA general formulation; 2) the multi-scale convolutional feature extraction block (MSCB) is used to further improve the quality of the fused image and to strengthen the learning potential of the network; 3) the triple-double architecture, *i.e.*, double-level, double-branch and double-direction, which can fully utilize the multi-scale information.

### A. MRAB

Let us focus on the physical MRA formula (2), in which the spatial details to be injected, *i.e.,* $\mathbf{G} \odot (\mathbf{P} - \mathbf{P_L})$, are extracted only from the PAN image with the proper injection coefficient $\mathbf{G}$. Thus, the traditional MRA approaches can equivalently be represented by the following network architecture,

$$\begin{aligned} \mathbf{D} &= \mathcal{H}(\mathbf{P}), \\ \widehat{\mathbf{MS}} &= \widetilde{\mathbf{MS}} + g(\mathbf{D}), \end{aligned} \tag{5}$$

where $\mathcal{H}(\cdot)$ is represented by the latent convolution layers, aiming to extract the details $\mathbf{D}$ from the PAN image. Besides, $g(\cdot)$ is represented by a spatial attention simulating the rule of the detail injection coefficient in (2). Furthermore, the upsampled MS image $\widetilde{\mathbf{MS}}$ can be realized by a simple PixelShuffle upsampling operation. The first formula in (5) can be viewed as the PAN spatial details, *i.e.,* $\mathbf{P} - \mathbf{P_L}$, and the second formula in (5) is equivalent to the MRA formula (2), where $g(\cdot)$ represents the nonlinear relationship among the involved images instead of a linear one as in (2). In summary, the MRA block (MRAB) consists of three parts: 1) the upsampling of the LRMS image, 2) the extraction of feature maps, and 3) the spatial attention module for detail injection. The detailed information for MRAB can be found in Fig. 2(b).

*1) Upsampling LRMS Image:* In Fig. 2(b), the first step is to upsample the original LRMS image to the same size as the GT image. In previous researches for pansharpening, the LRMS image is usually upscaled by an interpolation or a deconvolution operation. In [40], Shi *et al.* propose an efficient sub-pixel convolution operation (referred to as PixelShuffle), which learns a group of filters to upscale the low-resolution features into the high-resolution output. PixelShuffle got high performance when applied to the single image super-resolution problem [40]. Therefore, we introduce PixelShuffle into our model to upscale LRMS images to reach better performance. In particular, the feature map with $c \times r^2$ channels (where $r$ is the upscaling factor between LRMS and PAN images) is obtained through convolution, then yielding the high-resolution image by the periodic shuffling.

*2) Extracting Feature Maps:* As mentioned above, the traditional MRA methods extract details calculating the difference between the PAN image and the low-pass filtered PAN image. Thus, the final result depends on the adopted predefined filters which may mechanically discard some desired information. Thanks to the use of a convolutional layer, a set of parameters can be learned and dynamically adjusted to thoroughly explore the specific details and select expected features. Besides, to make the model adapt to the different datasets and to heradicate the misfit problem caused by the fixed filters, we extract end-to-end high-frequency information by learning the mapping $\mathcal{H}(\cdot)$ in (5). In [39], Zhang *et al.* used ResNet blocks as basic structure for feature extraction. However, to retain more information from the original images and to reduce the computational burden, we only adopted one ResNet block to form the PAN branch in TDNet.

As shown in Fig. 2(b), the extracted details from the PAN image are obtained by the PAN branch depicted in Fig. 3. The difference with the traditional MRA approaches is that the MS image is upsampled twice, *i.e.*, using a scale factor of 2 (when $r$ is equal to 4). The detailed information for the PAN branch can be found in Fig. 4.

*3) Spatial Attention Module for Detail Injection:* Recalling the original MRA formula (2) and the MRA-inspired formula (5), we can remark that the detail image $\mathbf{D}$ multiplied by $\mathbf{G}$ in (2) is equivalent to the spatial attention. Since the injection coefficient $\mathbf{G}$ is generally dependent on $\mathbf{MS}$ and $\mathbf{P}$, it motivates us to design spatial attention involving these two components. Specifically, we concatenate $\widetilde{\mathbf{MS}}$ and $\mathbf{D}$ together to carry out the convolution operation as shown in Fig. 2(b), aiming to learn a weight matrix $\mathbf{W} \in \mathbb{R}^{H \times W \times c}$ containing the sufficient features of the $\mathbf{MS}$ and the $\mathbf{P}$ images. The proposed injection strategy is to multiply the learned feature $\mathbf{D}$ obtained by the PAN branch and the weight matrix $\mathbf{W}$, then adding it to the $\widetilde{\mathbf{MS}}$ generated by PixelShuffle to yield the MRAB output.

### B. MSCB

Although the MRAB could lead to competitive outcomes with a physical interpretability, the obtained network architecture does not have deep layers, limiting the feature extraction and its nonlinear fitting abilities. Thus, we introduce a multi-scale convolutional block (denoted as MSCB) inspired by Yuan *et al.* [38] into our model to deepen the network. Fig. 5 shows the details of MSCB and its corresponding parameters.

### C. Overall Structure of Triple-Double Network

To solve the problem of a different size between LRMS and GT images, conventional methods directly upsample the LRMS image to the GT image size (usually with an upsampling by a factor 4). However, such operation can lead to spatial loss, even causing image distortion. By considering the

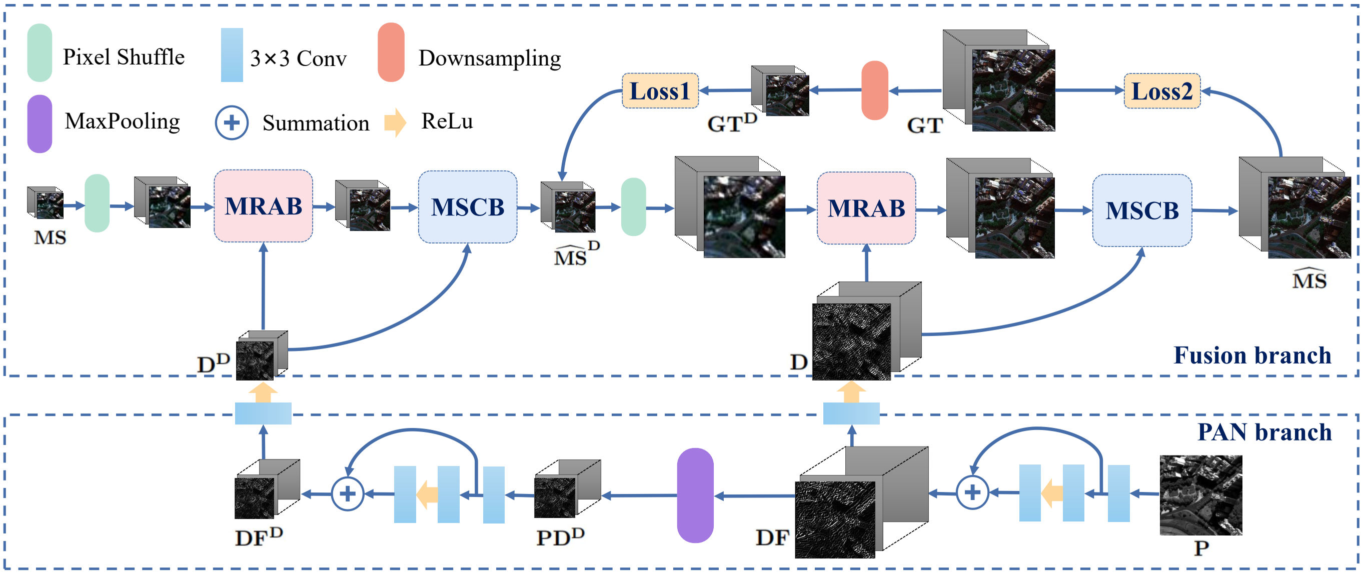


Fig. 3. Flowchart of the proposed TDNet consisting in two branches, *i.e.,* PAN branch and fusion branch. For convenience, the LRMS image and PAN image are denoted as $\mathbf{MS}$ and $\mathbf{P}$, respectively. The $\widehat{\mathbf{MS}}^{\mathbf{D}}$ is the output of the first-level fusion, and $\widehat{\mathbf{MS}}$ is the final HRMS image. $\mathbf{D}^{\mathbf{D}}$ and $\mathbf{D}$ are the output of the PAN branch. And the ground truth image and its downsampled version are denoted as $\mathbf{GT}$ and $\mathbf{GT}^{\mathbf{D}}$, respectively. The parameters and details of the network can be found in Sect. III. The number of convolution kernels used in the convolution operation is specified in Fig. 4.

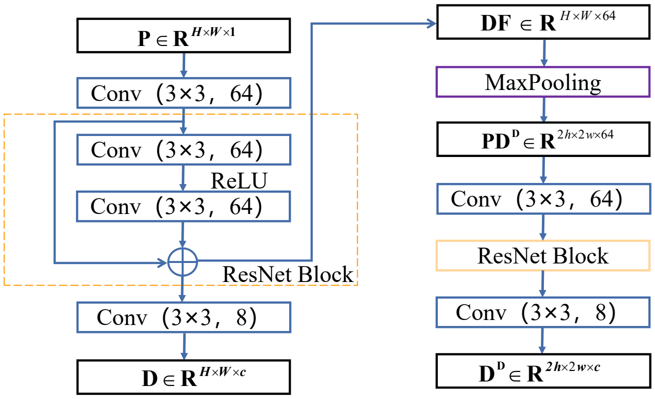

$\mathbf{P} \in \mathbf{R}^{H \times W \times 1}$
Conv (3×3, 64)
Conv (3×3, 64)
ReLU
Conv (3×3, 64)
ResNet Block
Conv (3×3, 8)
$\mathbf{D} \in \mathbf{R}^{H \times W \times c}$
$\mathbf{DF} \in \mathbf{R}^{H \times W \times 64}$
MaxPooling
$\mathbf{PD}^{\mathbf{D}} \in \mathbf{R}^{2h \times 2w \times 64}$
Conv (3×3, 64)
ResNet Block
Conv (3×3, 8)
$\mathbf{D}^{\mathbf{D}} \in \mathbf{R}^{2h \times 2w \times c}$

Fig. 4. An overview of the PAN branch, see the bottom of Fig. 3. $\mathbf{DF}$ is a feature map containing detailed information using 64 channels. $\mathbf{D}$ is a feature map with the same number of channels as $\mathbf{GT}$. The output of the MaxPooling is denoted as $\mathbf{PD}^{\mathbf{D}}$, and the feature map with detailed information with reduced size is denoted as $\mathbf{D}^{\mathbf{D}}$. Please, note that the number of convolution kernels is related to an exemplary fusion case involving a 8-bands dataset.

issues of the size difference and by fully utilizing the multi-scale information, we design the triple-double network (TDNet), *i.e.,* double-level, double-branch, and double-direction. In what follows, we will present the overall structure of the proposed TDNet. An overview of the proposed TDNet is depicted in Fig. 3.

*1) Double-branch:* From Fig. 3, it is clear that the network is divided into two branches, *i.e.,* the PAN branch and the fusion branch. The PAN branch takes the PAN image as the only input. It extracts and represents the multi-scale spatial features, which will be injected into the fusion branch for providing sufficient spatial details. The goal of the fusion branch is instead to fuse the input LRMS image and the multi-scale spatial features from the PAN branch to obtain the final HRMS image. The fusion branch contains some essential strategies mentioned before, such as, MRAB and MSCB.

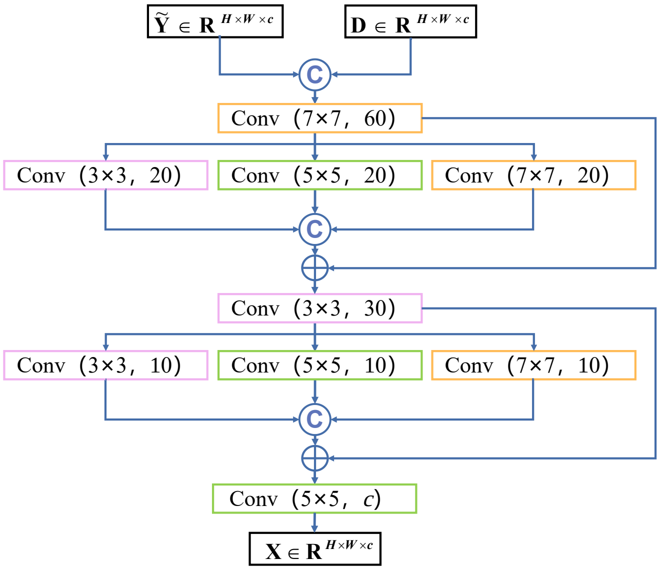

$\tilde{\mathbf{Y}} \in \mathbf{R}^{H \times W \times c}$
$\mathbf{D} \in \mathbf{R}^{H \times W \times c}$
C
Conv (7×7, 60)
Conv (3×3, 20)
Conv (5×5, 20)
Conv (7×7, 20)
C
Conv (3×3, 30)
Conv (3×3, 10)
Conv (5×5, 10)
Conv (7×7, 10)
C
Conv (5×5, $c$)
$\mathbf{X} \in \mathbf{R}^{H \times W \times c}$

Fig. 5. An overview of the MSCB. Please, note that since TDNet has a double-level structure, the diagram here refers to only the MSCB at size $H \times W$. The inputs $\tilde{\mathbf{Y}}$ and $\mathbf{D}$ are from MRAB and the feature map of the PAN branch, respectively, and the output $\mathbf{X}$ is the final pansharpened HRMS image, *i.e.,* $\widehat{\mathbf{MS}}$.

*2) Double-level:* In this work, we upsample the MS image using a two levels strategy, in which the MS image is upscaled to its double size (*i.e.,* with an upscaling factor of 2) for each level, thus exploiting the multi-scale features for pansharpening. In particular, both the PAN branch and fusion branch have the double-level structure for a better ability in resolution enhancement.

*3) Double-direction:* Due to the use of double-level, a promising strategy for fully employing multi-scale information of PAN and MS images is to design a network architecture with two directions (called double-direction). As shown in the flowchart of TDNet in Fig. 3, both the PAN branch and the fusion branch involve the double-level structure. The former downsamples the PAN image to a smaller size, and the latter one upsamples the LRMS image to a larger size. The information flows of the two branches are opposite and correspond to each other, in order to achieve the fusion of information between the branches. A similar strategy has been proven to be effective in a previous benchmark work [39].

In summary, the final architecture of the proposed TDNet has been formulated by the above-mentioned three aspects, *i.e.,* double-branch, double-level, and double-direction. Especially, double-branch takes the known LRMS image and PAN image as input to achieve the distinguishing feature representation. The double-level enables the network to exploit the multi-scale features, and the double-direction reinforces the mutual interaction between the two branches improving the performance.

## *D. Loss Fuction*

As mentioned before, our TDNet architecture contains a double-level structure, which results in two loss functions. Let $\widehat{\mathbf{MS}}^{\mathbf{D}} \in \mathbb{R}^{2h\times 2w}$ and $\widehat{\mathbf{MS}} \in \mathbb{R}^{H\times W}$ stand for the output of the first and second levels, respectively, and let $\mathbf{GT}^{\mathbf{D}} \in \mathbb{R}^{2h\times 2w}$ and $\mathbf{GT} \in \mathbb{R}^{H\times W}$ represent the GT images of the first and second levels, respectively. We define the following loss function for supervised learning for both the levels,

$$\min_{\mathbf{\Theta}} \ \mathcal{L}oss = \gamma\mathcal{L}oss_1 + (1-\gamma)\mathcal{L}oss_2, \tag{6}$$

where $\gamma \in [0, 1]$ is a constant during the training phase, and the magnitude of $\gamma$ is deeply discussed in Sect. IV. Specifically, $\mathcal{L}oss_1$ and $\mathcal{L}oss_2$ are defined as follows,

$$\mathcal{L}oss_1 = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{GT}^{\mathbf{D}}{}_i - \widehat{\mathbf{MS}}_i^{\mathbf{D}}\|_1, \tag{7}$$

$$\mathcal{L}oss_2 = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{GT}_i - \widehat{\mathbf{MS}}_i\|_1, \tag{8}$$

where $\|\cdot\|_1$ indicates the $\ell_1$ norm and $N$ is the number of training samples.

# IV. Experimental Results

This section is devoted to the demonstration of the superiority of our TDNet by comparing it with some state-of-the-art pansharpening methods on various datasets. In addition, we comprehensively discuss the structure of our TDNet to explore the inherent potential of the proposed network.

## *A. Datasets*

Three different datasets captured by three different sensors (*i.e.*, WorldView-3, GaoFen-2 and the QuickBird) are considered in this paper. The WorldView-3 works in the visible and near-infrared spectrum range, which provide eight MS-band (coastal, blue, green, yellow, red, red edge, near-infrared 1, and near-infrared 2) and a PAN channel with a spatial resolution of 1.2 m and 0.3 m, respectively. The radiometric resolution of WorldView-3 is 11 bits. Both GaoFen-2 and QuickBird provide four MS-band (red, green, blue, and near-infrared) and a PAN channel. For GaoFen-2, the spatial resolution is about 3.2 m for the MS bands and 0.8 m for the PAN channel, and the radiometric resolution is 10 bits. For QuickBird, the spatial resolution is about 2.44 m for the MS bands and 0.61 m for the PAN channel, and the radiometric resolution is 11 bits.

## *B. Implementation Details*

This section is devoted to the presentation of some implementation details related to the proposed approach.

*1) Dataset Simulation:* In this work, we mainly conduct the network training on WorldView-3 datasets, which are available to download on the website[1]. In particular, there are no GT images used as reference. Thus, the original LRMS and PAN images are simultaneously blurred and downsampled according to Wald's protocol [41]. The original LRMS image is used as reference image, *i.e.*, the GT image. We simulate the WorldView-3 dataset with 12,580 samples (also called patch pairs), each sample including PAN (with size $64\times 64$), LRMS (with size $16\times 16\times 8$), and GT (with size $64\times 64\times 8$) patches. For the 12,580 samples, we divided them into 8806/2516/1258 (70%/20%/10%) as for the training dataset, validation dataset, and testing dataset, respectively. The simulation process for the datasets is the same as in [2]. Interesting readers can refer to [2] for further details. Moreover, we also assess the performance on two 4-bands datasets, *i.e.,* QuickBird and GaoFen-2. More details about the simulation of these two datasets can be found in Sect. IV-H1.

*2) Training Platform and Parameter Setting:* The proposed network is coded with Python 3.8.0 and Pytorch 1.7.0, and is trained with NVIDIA GPU GeForce GTX 3080. We use Adam optimizer, in which the betas and weight decay are set as (0.9, 0.999) and 0, respectively, to minimize the loss function (6) by 300 epochs, and the batch size is set as 32. To achieve better performance, we set the initial learning rate as 0.01, then dynamically adjusting it to 0.001 after 220 epochs. About the hyper-parameter $\gamma$ in (6), more discussions can be found in Sect. IV-H2.

## *C. Benchmark*

Several competitive methods belonging to different pansharpening categories are employed.

- EXP: MS image interpolated by a polynomial kernel with 23 coefficients [31]
- CS methods:

[1] http://www.digitalglobe.com/samples?search=Imagery

1) GS: Gram-Schmidt sharpening approach [10].
2) PRACS: partial replacement adaptive component substitution approach [9].
3) BDSD-PC: robust band-dependent spatial-detail approach [42].

- MRA methods:
  1) SFIM: smoothing filter-based intensity modulation [12].
  2) GLP-HPM: GLP with MTF-matched filter [43] with multiplicative injection model [44].
  3) GLP-CBD: the GLP with MTF-matched filter [43] and regression-based injection model [31], [45].
  4) GLP-Reg: the GLP with MTF-matched filter [43] and full-scale regression [46][2].
- DL-based methods:
  1) PNN: pansharpening via convolutional neural networks [27][3].
  2) PanNet: convolutional neural networks for residual learning on the high-frequency domain for pansharpening [25][4].
  3) DiCNN1: detail injection based convolutional neural networks [47][5].
  4) BDPN: efficient bidirectional pyramid network for pansharpening [39][6]
  5) DMDNet: deep multi-scale detail convolutional neural networks for pansharpening [1][7].
  6) FusionNet: deep convolutional neural network inspired by traditional CS and MRA methods [2].

### D. Reduced Resolution Assessment

The reduced resolution assessment measures the similarity between the fused image and the ideal reference image (the original MS image). The similarity can be determined by the calculation of several evaluation indexes. For the reduced resolution experiments, the spectral angle mapper (SAM) [48], the dimensionless global error in synthesis (ERGAS) [49], the spatial correlation coefficient (SCC) [50], and the Q$2^n$ (Q8 for 8-band datasets and Q4 for 4-band datasets) [51] are employed as evaluation indexes. The ideal values for SAM and ERGAS are 0, whereas 1 for Q$2^n$ and SCC.

As mentioned in Sect. IV-B, we have 1258 testing samples from WorldView-3 images. We first compare the proposed TDNet with the five state-of-the-art CNN-based pansharpening approaches on the 1258 samples. From Tab. I, it is clear that the TDNet obtains the best average quantitative performance on all the metrics demonstrating the superiority of the proposed method. This can be justified because, comparing our approach with conventional CNNs for pansharpening, it utilizes multi-scale convolution kernels for a better feature extraction. Besides, comparing it with PanNet and DMDNet, which directly send the high-frequency information of PAN and MS images into the network, our TDNet adopts two branches for better exploiting the multi-scale structures of PAN and MS images. Moreover, our MRA-inspired TDNet could hold better physical meanings with respect to DiCNN. Furthermore, due to the triple-double network structure, *i.e.,* double-level, double-branch, and double-direction, the TDNet can fully utilize the latent multi-scale information of PAN and MS images, thus obtaining better results than FusionNet.

Furthermore, we generated two WorldView-3 test cases (Rio dataset and Tripoli dataset) at reduced resolution by applying Wald's protocol (please, refer to Sect. IV-B for details about Wald's protocol implementation). The GT image has a size $256 \times 256 \times 8$, as well as the LRMS and PAN images have a size $64 \times 64 \times 8$ and $256 \times 256$, respectively. Figs. 6-9 show the visual comparisons among all the 15 compared pansharpening approaches. From these figures, it is easy to note that the TDNet yields results very close to the GT image. The traditional CS and MRA techniques generate products with some obvious spatial blur, especially at near the boundaries of the buildings and/or spectral distortion. The other DL-based approaches perform better than the traditional methods. However, TDNet shows less image residuals compared with the other CNN-based methods, see Fig. 7 and Fig. 9. In Tab. II, the quantitative metrics demonstrate that TDNet still gets the best performance, assessing the superiority of our TDNet approach, which is able to reduce both spatial and spectral distortions in the fusion outcome.

### E. Full Resolution Assessment

To corroborate the results at reduced resolution, a full resolution analysis is also needed involving the original MS and PAN products. Unlike the reduced resolution test cases, we have no reference (GT) image. Thus, three widely used metrics, that do not exploit the GT image, are employed, *i.e.,* the quality with no reference (QNR) index, the spectral distortion $D_\lambda$ index, and the spatial distortion $D_s$ index [7].

Fig. 10 presents the visual comparison of all the compared pansharpening approaches on a full resolution example. In this case, the spectral quality of the image should refer to the LRMS image, and the spatial details of a high-quality fusion image should be close to the PAN image. It can be observed that result by the TDNet is the most qualified one, with sharper and clearer edges and without ghosting, blurring, *etc.* Furthermore, Tab. III reports the average performance on 50 full resolution examples. Again, TDNet obtains the best average results and the minimum standard deviations demonstrating the superiority and stability of our method on full resolution test cases.

### F. The Ablation Study

This section is devoted to ablation studies to investigate the effect of each component of the TDNet. For simplicity, we take a WorldView-3 dataset as reference. An overall performance calculated on the training/validation loss function can be found

[2] http://openremotesensing.net/kb/codes/pansharpening/

[3] Note that the given source code in Open Remote Sensing does not contain the trained models for WorldView-2 and WorldView-3, thus we re-implemented the network with default parameters in Python using Tensorflow for fair comparisons.

[4] Code link: https://xueyangfu.github.io/

[5] DiCNN1 has been implemented by ourselves with default parameters.

[6] BDPN has been implemented by ourselves with default parameters.

[7] DMDNet has been implemented by ourselves with default parameters.

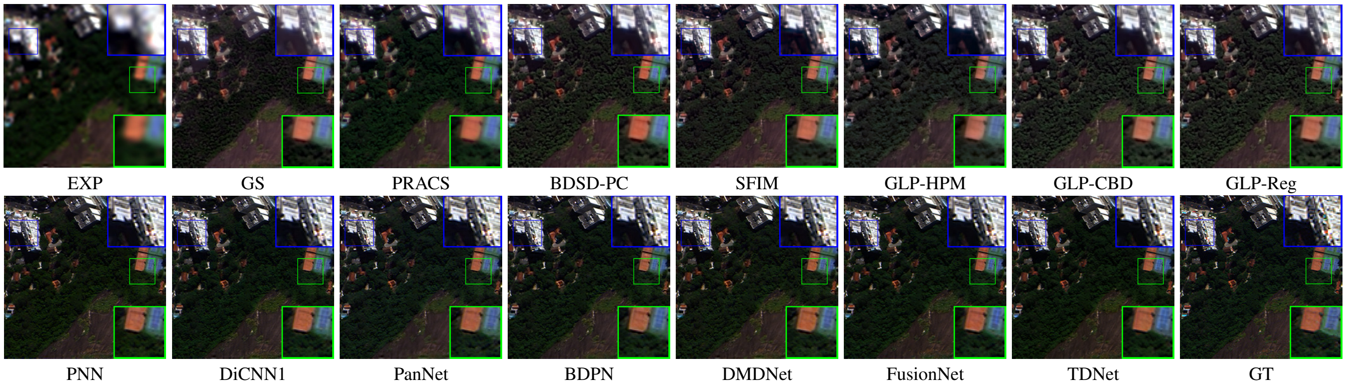


Fig. 6. Visual comparisons of all the compared approaches on the reduced resolution Rio dataset (sensor: WorldView-3).

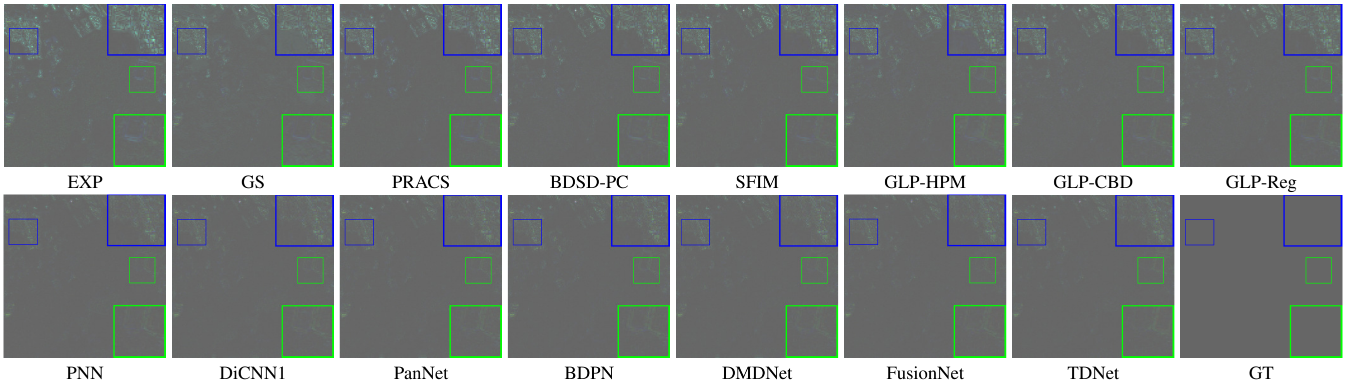


Fig. 7. The corresponding absolute error maps (AEMs) using the reference (GT) image on the reduced resolution Rio dataset (sensor: WorldView-3). For a better visualization, we doubled the intensities of the AEMs and added 0.3.

TABLE I
AVERAGE METRICS FOR ALL THE COMPARED DL-BASED APPROACHES ON 1258 REDUCED RESOLUTION SAMPLES. (BOLD: BEST; UNDERLINE: SECOND BEST)

| **Method** | *SAM* (± std) | *ERGAS* (± std) | *Q8* (± std) | *SCC* (± std) |
|---|---|---|---|---|
| **PNN [27]** | 4.4015 ± 1.3292 | 3.2283 ± 1.0042 | 0.8883 ± 0.1122 | 0.9215 ± 0.0464 |
| **DiCNN1 [47]** | 3.9805 ± 1.3181 | 2.7367 ± 1.0156 | 0.9096 ± 0.1117 | 0.9517 ± 0.0471 |
| **PanNet [25]** | 4.0921 ± 1.2733 | 2.9524 ± 0.9778 | 0.8941 ± 0.1170 | 0.9494 ± 0.0460 |
| **BDPN [39]** | 3.9952 ± 1.3869 | 2.7234 ± 1.0394 | 0.9123 ± 0.1128 | 0.9515 ± 0.0457 |
| **DMDNet [1]** | 3.9714 ± 1.2482 | 2.8572 ± 0.9663 | 0.9000 ± 0.1141 | 0.9527 ± 0.0446 |
| **FusionNet [2]** | 3.7435 ± 1.2259 | 2.5679 ± 0.9442 | 0.9135 ± 0.1122 | 0.9580 ± 0.0450 |
| **TDNet** | **3.5036 ± 1.2411** | **2.4439 ± 0.9587** | **0.9212 ± 0.1117** | **0.9621 ± 0.0440** |
| **Ideal value** | **0** | **0** | **1** | **1** |

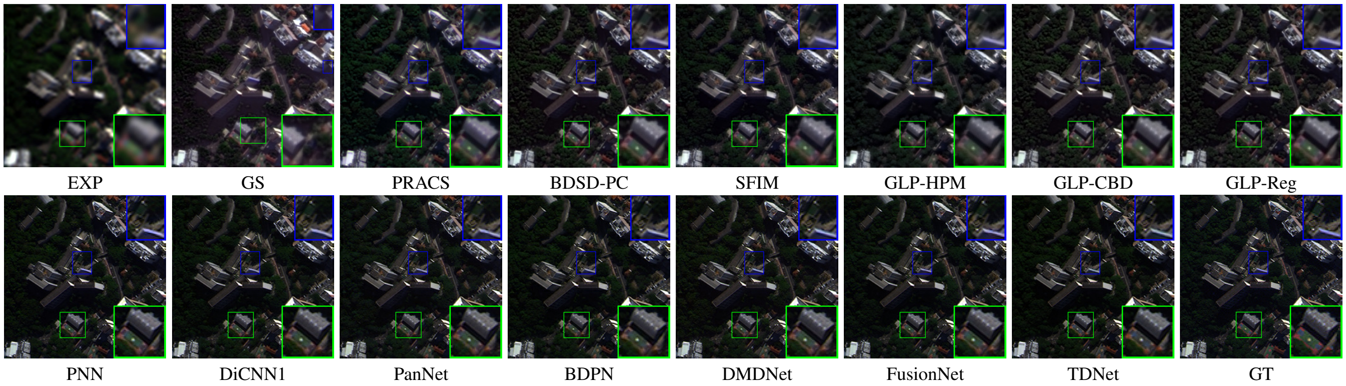


Fig. 8. Visual comparisons of all the compared approaches on the reduced resolution Tripoli dataset (sensor: WorldView-3).

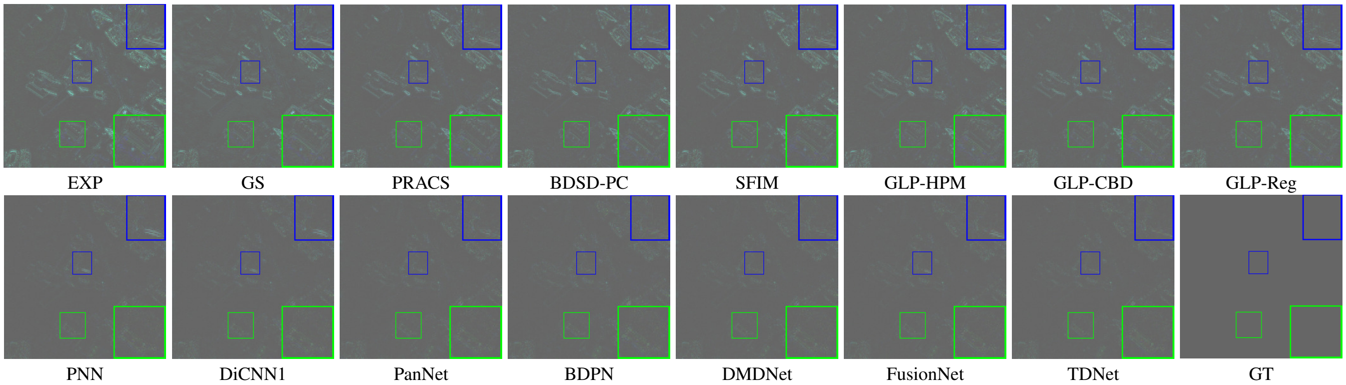


Fig. 9. The corresponding AEMs on the reduced resolution Tripoli dataset (sensor: WorldView-3). For a better visualization, we doubled the intensities of the AEMs and added 0.3.

TABLE II
QUALITY METRICS FOR ALL THE COMPARED APPROACHES ON THE REDUCED RESOLUTION RIO AND TRIPOLI DATASETS, RESPECTIVELY. (BOLD: BEST; UNDERLINE: SECOND BEST)

| Method | *(a) Rio dataset* | | | | *(b) Tripoli dataset* | | | |
|---|---|---|---|---|---|---|---|---|
| | *SAM* | *ERGAS* | *SCC* | *Q8* | *SAM* | *ERGAS* | *SCC* | *Q8* |
| **EXP** [31] | 7.0033 | 6.5368 | 0.4797 | 0.5929 | 7.8530 | 9.0903 | 0.5137 | 0.6608 |
| **GS** [10] | 8.9481 | 6.3662 | 0.6775 | 0.8666 | 8.3772 | 7.8041 | 0.7240 | 0.7291 |
| **PRACS** [9] | 7.1176 | 5.5392 | 0.6370 | 0.7380 | 7.9705 | 7.4826 | 0.7111 | 0.7951 |
| **BDSD-PC** [42] | 7.0721 | 4.9515 | 0.7164 | 0.7853 | 7.7347 | 7.0345 | 0.7304 | 0.8202 |
| **SFIM** [12] | 6.8501 | 4.9571 | 0.7558 | 0.7882 | 7.4338 | 7.0695 | 0.7514 | 0.8020 |
| **GLP-HPM** [44] | 7.2994 | 5.1185 | 0.7369 | 0.7849 | 7.9390 | 7.1489 | 0.7415 | 0.8238 |
| **GLP-CBD** [45] | 7.4053 | 5.0372 | 0.6738 | 0.7880 | 7.8051 | 6.9162 | 0.7312 | 0.8300 |
| **GLP-Reg** [46] | 7.3275 | 5.0154 | 0.6822 | 0.7889 | 7.7680 | 6.9100 | 0.7327 | 0.8298 |
| **PNN** [27] | 4.0659 | 2.7144 | 0.9487 | 0.8888 | 5.6714 | 3.5657 | 0.9435 | 0.9166 |
| **DiCNN** [47] | 3.8289 | 2.5819 | 0.9544 | 0.8895 | 5.3622 | 3.3104 | 0.9523 | 0.9348 |
| **PanNet** [25] | 3.9062 | 2.6583 | 0.9522 | 0.8814 | 4.8500 | 3.1744 | 0.9642 | 0.9190 |
| **BDPN** [39] | 4.0788 | 2.6897 | 0.9439 | 0.8969 | 5.5728 | 3.3880 | 0.9520 | 0.9379 |
| **DMDNet** [1] | 3.6917 | 2.4968 | 0.9594 | 0.8990 | 4.5649 | 2.9795 | 0.9706 | 0.9243 |
| **FusionNet** [2] | 3.5700 | 2.4346 | 0.9607 | 0.9044 | 4.3850 | 2.8533 | 0.9718 | 0.9422 |
| **TDNet** | **3.3801** | **2.3522** | **0.9648** | **0.9155** | **4.1445** | **2.7076** | **0.9761** | **0.9526** |
| **Ideal value** | **0** | **0** | **1** | **1** | **0** | **0** | **1** | **1** |

in Fig. 11 for various network structures. It is easy to show that the proposed TDNet shows the smallest loss compared with the other test cases.

*1) The Effect of MRAB:* To explore whether the MRAB contributes to the final result, we remove the MRAB from TDNet running the training with the same data and parameters. Tab. IV presents the average outcomes and the corresponding standard deviations for the TDNet with and without MRAB (w/o MRAB). It can be observed that the fusion results w/o MRAB have inferior performance on all the metrics compared with the original TDNet. This indicates that the MRAB can help the network in learning more details and features.

*2) The Effect of MSCB:* The role of MSCB in TDNet is to increase the depth and width of the network to improve the ability of feature extraction. Its innovation lies in the use of convolution kernels with different scales to learn various scales in real scenes. As shown in Fig. 12, to explore whether such MSCB can favor the fusion task, we change the original multi-scale convolution kernels (*i.e.*, 3×3, 5×5 and 7×7) to a single-scale convolution kernel (*i.e.*, $5 \times 5$). The reduced structure is called single-scale convolution block (SSCB). Tab. IV reports that the TDNet with SSCB has lower performance with respect to the original TDNet with MSCB, which verifies the benefits of using the MSCB module.

*3) The Effect of PAN Branch:* In the original network structure, we designed the PAN branch to extract the spatial features from the PAN image. These spatial features are then injected into the fusion branch. Hence, we alternatively feed the fusion branch directly using the PAN image and its downsampled version to explore whether such PAN branch can affect the final outcomes. We denoted the TDNet without PAN branch as w/o PAN branch. The quantitative results shown in Tab. IV demonstrate that the original TDNet with the PAN branch yields the best performance.

*4) The Effect of Double-level Structure:* In the work, the proposed TDNet has two levels. In each level, the MS image is upsampled to its double size. The fusion performance can benefit from the use of the double-level structure. To corroborate this point, we implemented a single-level (directly upsampling by a factor of 4) strategy using our TDNet approach. Tab. IV clearly shows that the single-level structure will significantly reduce the fusion performance demonstrating the advantages

TABLE III
AVERAGE VALUES OF QNR, $D_\lambda$ AND $D_s$ WITH THE RELATED STANDARD DEVIATIONS (STD) FOR 50 FULL RESOLUTION WORLDVIEW-3 SAMPLES. (BOLD: BEST; UNDERLINE: SECOND BEST)

| Method | *QNR* (± std) | $D_\lambda$ (± std) | $D_s$ (± std) |
|---|---|---|---|
| **EXP [31]** | 0.8078 ± 0.0673 | 0.0582 ± 0.0274 | 0.1010 ± 0.0428 |
| **GS [10]** | 0.8806 ± 0.0351 | 0.0343 ± 0.0239 | 0.0882 ± 0.0238 |
| **PRACS [9]** | 0.9204 ± 0.0172 | 0.0335 ± 0.0055 | 0.0668 ± 0.0159 |
| **BDSD-PC [42]** | 0.9063 ± 0.0231 | 0.0322 ± 0.0108 | 0.0731 ± 0.0175 |
| **SFIM [12]** | 0.8999 ± 0.0423 | 0.0371 ± 0.0207 | 0.0657 ± 0.0244 |
| **GLP-HPM [44]** | 0.8384 ± 0.0757 | 0.0332 ± 0.0162 | 0.0647 ± 0.0234 |
| **GLP-CBD [45]** | 0.8795 ± 0.0510 | 0.0418 ± 0.0210 | 0.0827 ± 0.0337 |
| **GLP-Reg [46]** | 0.8812 ± 0.0498 | 0.0408 ± 0.0205 | 0.0818 ± 0.0328 |
| **PNN [27]** | 0.9446 ± 0.0233 | 0.0255 ± 0.0138 | 0.0306 ± 0.0117 |
| **DiCNN1 [47]** | 0.9564 ± 0.0124 | 0.0231 ± 0.0113 | **0.0208** ± **0.0072** |
| **PanNet [25]** | 0.9421 ± 0.0227 | 0.0345 ± 0.0146 | 0.0242 ± 0.0107 |
| **BDPN [39]** | 0.9206 ± 0.0399 | 0.0365 ± 0.0252 | 0.0350 ± 0.0089 |
| **DMDNet [1]** | 0.9383 ± 0.0329 | 0.0309 ± 0.0162 | 0.0320 ± 0.0192 |
| **FusionNet [2]** | 0.9435 ± 0.0259 | 0.0303 ± 0.0096 | 0.0255 ± 0.0076 |
| **TDNet** | **0.9575** ± **0.0051** | **0.0209** ± **0.0079** | 0.0219 ± 0.0052 |
| **Ideal value** | **1** | **0** | **0** |

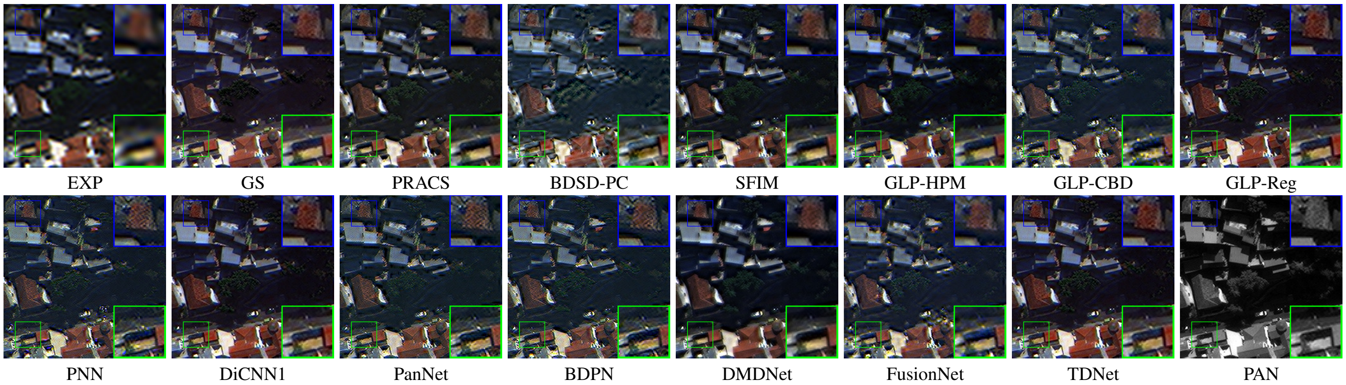


Fig. 10. Visual comparisons between the TDNet and the benchmark on the full resolution Rio dataset (sensor: WorldView-3).

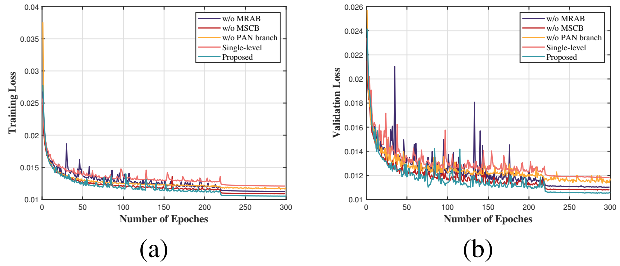


Fig. 11. Convergence curves for different network structures. (a) Training loss curves. (b) Validation loss curves. Please note that the single-level loss is calculated by (8), while the loss of the other structures is calculated by (6) (with $\gamma$ = 0.4). Please note that the learning rate for 0-220 epochs is 0.001, and for 220-300 epochs is adjusted to 0.0001.

of the double-level structure.

*5) The Effect of PixelShuffle:* In the fusion branch, we introduce PixelShuffle to upscale LRMS images, instead of the more common interpolation or deconvolution operations. To demonstrate the superiority of PixelShuffle, we replaced PixelShuffle with linear interpolation and deconvolution upsampling, denoted as TDNet(bilinear) and TDNet(Deconv), respectively. The results of the variants are reported in Tab. IV. It can be seen that TDNet performs better with the assistance of PixelShuffle.

*6) The Effectiveness of the Overall strctrure:* Compared with other DL-based methods, the results of TDNet are significantly improved. To prove the effectiveness of the TDNet structure more fairly, we reduce the number of channels in the original MSCB as showed in Fig. 12, aiming at bring the model complexity to the level of FusionNet and DMDNet. The degenerate model is denoted as TDNet(-). The results of TDNet(-) are shown in Tab. IV. Compared with the results obtained by other comparative methods shown in Tab. I. It is clear that although the performance of TDNet(-) has degraded, it still outperforms all the compared DL-based methods. This is also a proof about the superiority of the triple-double structure.

## *G. Comparison of MRAB and Traditional MRA Scheme*

An important module in our model, MRAB, is a derivative of the traditional MRA scheme. The traditional MRA scheme requires manual estimation of its injection coefficient. These handcrafted coefficient with subjective and relatively simple forms, however, always could not sufficiently and adaptively

TABLE IV

QUALITY METRICS FOR DIFFERENT NETWORK STRUCTURES ON THE REDUCED RESOLUTION 1258 DATASETS. (BOLD: BEST; UNDERLINE: SECOND BEST)

| **Method** | *SAM* (± std) | *ERGAS* (± std) | *Q8* (± std) | *SCC* (± std) |
|---|---|---|---|---|
| **w/o MRAB** | 3.7731 ± 1.2394 | 2.5608 ± 0.9443 | 0.9139 ± 0.1156 | 0.9587 ± 0.0445 |
| **SSCB** | 3.5739 ± 1.3035 | 2.4501 ± 0.9812 | 0.9199 ± 0.1241 | 0.9601 ± 0.0517 |
| **w/o PAN branch** | 3.8701 ± 1.3509 | 2.7572 ± 0.9910 | 0.9099 ± 0.1127 | 0.9577 ± 0.0443 |
| **Single-stage** | 3.9706 ± 1.2493 | 2.8432 ± 0.9677 | 0.9098 ± 0.1134 | 0.9523 ± 0.0430 |
| **TDNet(bilinear)** | 3.5197 ± 1.2567 | 2.5207 ± 0.9908 | 0.9198 ± 0.1232 | 0.9607 ± 0.0510 |
| **TDNet(Deconv)** | 3.5276 ± 1.2721 | 2.5103 ± 0.9601 | 0.9207 ± 0.1219 | 0.9610 ± 0.0456 |
| **TDNet(-)** | 3.6987 ± 1.3107 | 2.5479 ± 0.9511 | 0.9187 ± 0.1201 | 0.9607 ± 0.0457 |
| **TDNet-TMRA** | 3.9942 ± 1.8703 | 2.7812 ± 1.0123 | 0.8997 ± 0.1257 | 0.9465 ± 0.0529 |
| **TDNet** | **3.5036** ± **1.2411** | **2.4439** ± **0.9587** | **0.9212** ± **0.1117** | **0.9621** ± **0.0440** |

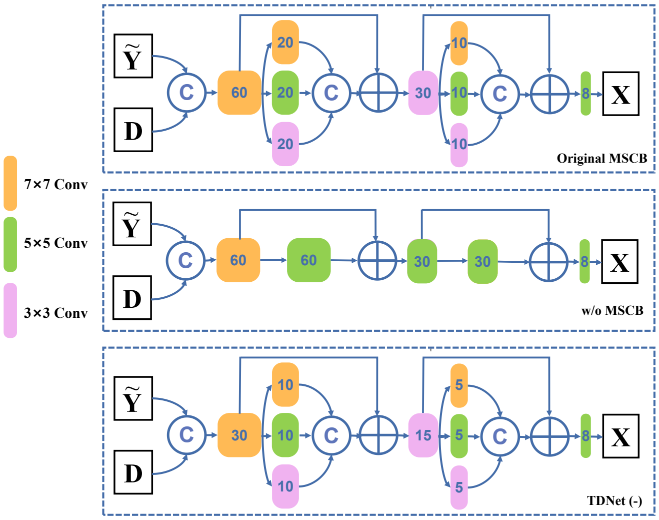


Fig. 12. TDNet with MSCB (top row) and without MSCB (bottom row, also called SSCB). Note that the numbers in color boxes mean the number of convolution kernels.

reflect complex spatial and spectral relationships among PAN image, LRMS image, and HRMS image. In addition, in the framework of deep learning, traditional schemes need to be improved to learn more discriminate representations from big data. In order to verify the rationality of our motivation, we conducted experiments to compare MRAB and traditional MRA scheme.

Specifically, one competing MRA method, GLP-HPM [44], is exploited here. Its generalized form is the same as equation (2). GLP-HPM adopts the injection model as follows:

$$\mathbf{G_k} = \frac{\widetilde{\mathbf{MS_k}}}{\mathbf{P_L}}, k = 1, ..., c. \tag{9}$$

where $\mathbf{G_k}$ represents the injection gain in the $\mathbf{k}$-th band, and the division is intended as pixel-wise. The main difference between this model and our proposed MRAB (5) is that the latter uses the injection gain learned by designed convolutional layers. By replacing the proposed MRAB as following, we can obtain the TDNet-TMRA method:

$$\begin{aligned} &\mathbf{D} = \mathcal{H}(\mathbf{P}), \\ &\widehat{\mathbf{MS_k}} = \widetilde{\mathbf{MS_k}} + \frac{\widetilde{\mathbf{MS_k}}}{\mathbf{P_L}} \odot \mathbf{D_k}, k = 1, ..., c. \end{aligned} \tag{10}$$

Table IV shows the average results produced by TDNet-TMRA and the proposed TDNet over 1258 testing samples. The numerical results on four metrics show that the MRAB is indeed effective for TDNet, while it is hard to achieve good performance by embedding a traditional MRA scheme into our network. Therefore, when combining traditional methods with DL-based methods, how to overcome the uncertainty brought by deep learning deserves more in-depth research. In this work, we provide a feasible method, *i.e.*, MRAB, uses unique network structures such as the attention mechanism to make the parameters adaptive to the data.

### *H. Discussion*

*1) Evaluation on 4-band Datasets:* The datasets used in the above experiments are all data with 8 spectral bands acquired by the same sensor (*i.e.*, WorldView-3). In this section, we will focus on assessing performance on 4-band datasets acquired by the GaoFen-2 and the QuickBird sensors. The dataset simulation for the two sensors is the same as that of the 8-band WorldView-3 datasets mentioned in Sect. IV-B. For the GaoFen-2 dataset, we downloaded the data acquired over Beijing from the website[8] and we simulated 21,607 training samples (PAN size, $64 \times 64$). Besides, 81 testing samples (PAN size, $256 \times 256$) acquired over the city of Guangzhou are used for comparison purposes. About the QuickBird test case, a large dataset acquired over the city of Indianapolis is exploited to simulate 20,685 training samples (PAN size, $64 \times 64$). Moreover, we simulated 48 testing samples with spatial size $256 \times 256$ to assess the performance for all the compared approaches. More details about the generation of these test cases can be found in [2]. In Figs. 13 and 14, we show the performance comparing all the five DL-based approaches[9]. Since it is not easy to distinguish the differences having a look at the 8-bits RGB images, we present the absolute error maps

[8] Data link: http://www.rscloudmart.com/dataProduct/sample

[9] Note that, since the traditional methods, *i.e.*, MRA and CS methods, have obtained lower performance, for the sake of brevity, we excluded them from the analysis.

TABLE V
AVERAGE ASSESSMENT OF THE COMPARED APPROACHES FOR 81 GAOFEN-2 TESTING SAMPLES AND 48 QUICKBIRD TESTING SAMPLES. (BOLD: BEST; UNDERLINE: SECOND BEST)

| | *SAM* (± std) | *ERGAS* (± std) | *Q8* (± std) | *SCC* (± std) |
|---|---|---|---|---|
| | | *Guangzhou datasets (GaoFen-2)* | | |
| **PNN** | 1.6599 ± 0.3606 | 1.5707 ± 0.3243 | 0.9274 ± 0.0202 | 0.9281 ± 0.0206 |
| **DiCNN1** | 1.4948 ± 0.3814 | 1.3203 ± 0.3543 | 0.9445 ± 0.0211 | 0.9458 ± 0.0222 |
| **PanNet** | 1.3954 ± 0.3261 | 1.2239 ± 0.2828 | 0.9468 ± 0.0222 | 0.9558 ± 0.0123 |
| **DMDNet** | 1.2968 ± 0.3156 | 1.1281 ± 0.2669 | 0.9529 ± 0.0218 | 0.9644 ± 0.0100 |
| **FusionNet** | 1.1795 ± 0.2714 | 1.0023 ± 0.2271 | 0.9627 ± 0.0167 | 0.9710 ± 0.0074 |
| **TDNet** | **1.0926 ± 0.2645** | **0.9303 ± 0.2267** | **0.9695 ± 0.0131** | **0.9750 ±0.0132** |
| | | *Indianapolis datasets (QuickBird)* | | |
| **PNN** | 5.7993 ± 0.9474 | 5.5712 ± 0.4584 | 0.8572 ± 0.1481 | 0.9023 ± 0.0489 |
| **DiCNN1** | 5.3071 ± 0.9957 | 5.2310 ± 0.5411 | 0.8821 ± 0.1431 | 0.9224 ± 0.0506 |
| **PanNet** | 5.3144 ± 1.0175 | 5.1623 ± 0.6814 | 0.8833 ± 0.1398 | 0.9296 ± 0.0585 |
| **DMDNet** | 5.1197 ± 0.9399 | 4.7377 ± 0.6486 | 0.8907 ± 0.1464 | 0.9350 ± 0.0652 |
| **FusionNet** | 4.5402 ± 0.7789 | 4.0508 ± 0.2666 | 0.9102 ± 0.1364 | 0.9547 ± 0.0457 |
| **TDNet** | **4.5047 ± 0.8022** | **3.9799 ± 0.2326** | **0.9123 ± 0.1452** | **0.9551 ± 0.0652** |
| **Ideal value** | **0** | **0** | **1** | **1** |

TABLE VI
COMPARISON OF THE NUMBER OF PARAMETERS (NOPS), THE ITERATIONS, THE TRAINING TIMES, AND THE TESTING TIMES FOR ALL THE DL-BASED APPROACHES. (THE TRAINING TIMES UNIT IS HOURS: MINUTES & THE TESTING TIMES UNIT IS SECONDS)

| | **PNN** | **DiCNN1** | **PanNet** | **BDPN** | **DMDNet** | **FusionNet** | **TDNet** |
|---|---|---|---|---|---|---|---|
| **Iterations** | $1.12 \times 10^6$ | $3 \times 10^5$ | $2.4 \times 10^5$ | $2.7 \times 10^5$ | $2.5 \times 10^5$ | $1.4 \times 10^5$ | $8.2 \times 10^4$ (300 epoches) |
| **Training times** | 25: 15 | 7: 06 | 4: 32 | 46:19 | 5: 27 | 2: 21 | 6: 30 |
| **Testing times** | 0.0778 | 0.0799 | 0.0811 | 0.0912 | 0.0852 | 0.0812 | 0.0861 |
| **NoPs** | $3.1 \times 10^5$ | $1.8 \times 10^5$ | $2.5 \times 10^5$ | $15.2 \times 10^5$ | $3.2 \times 10^5$ | $2.3 \times 10^5$ | $5.5 \times 10^5$ |

(AEMs). It is worth to be remarked that our TDNet generates more details showing less residuals. The quantitative results of Tab. V also support the conclusion that the TDNet obtains the best overall performance.

*2) The Hyper-Parameter in the Loss function:* As described in Sect. III-D, the loss function consists of two parts, in which the hyper-parameter $\gamma$ weights the two sub-loss functions. Obviously, the higher the value of $\gamma$, the more the importance to the first level. The goal is to generate the final fusion image closer to the reference image. Therefore, it is worth exploring how a change in the value of $\gamma$ can lead to better results. In our experiment, $\gamma$ is set to different values. Fig. 16 shows the changes in the ERGAS index varying $\gamma$ and increasing the epochs. When $\gamma$ = 0 or $\gamma$ = 1, the convergence is poor. Thus, we discard these two cases. Fig. 16 shows comparable results, in terms of the convergence speed and values, varying $\gamma$ (*i.e.*, assuming the values 0.2, 0.4, 0.5, 0.6, 0.8). Thus, we choose $\gamma = 0.4$ for the training of the proposed TDNet.

*3) The Computational Analysis:* Tab. VI reports the training time and the number of parameters (NoPs) for all the compared DL-based methods. The maximum number of iterations shown in the table is the optimal one for training the network. TDNet gets a relatively large amount of parameters, mainly due to the structure of the MSCB. However, the final training time of TDNet is less than that of PNN and DiCNN1 because of a less iteration number for the convergence. Besides, TDNet is able to achieve a satisfying trade-off between effectiveness and complexity. We perform the evaluation on 1258 testing samples with size 256 × 256 acquired by the WorldView-3 sensor, as described in Sect. IV-B1. Comparisons on average testing time are shown in Tab. VI. It is can be seen that the testing time of TDNet is not much longer than the compared deep learning methods, and significantly shorter than BDPN, which is also a double-level structure. Leveraging on the special structure, our network can fully fuse the complementary information from different sources in a more reasonable way, which leads to good results with the tolerable computational burden.

*4) Structure Discussion and Improvements Analysis:* In fact: the bidirectional, double-branch network structure and feature pyramid [52] has appeared in several previous significant works [39], [53], and has been proved that can implement feature extraction and image fusion hierarchically and more effectively. In particular, it is necessary to emphasize the distinction between the proposed TDNet and BDPN [39]. Aiming at making full use of the high-frequency information in PAN images, BDPN extracts the multilevel details from PAN images and directly injects them into the upsampled LRMS images. Differently, we focus more on the mapping relations among images. Specifically, the extracted high-frequency information is adopted as the input of the fusion branch in multiple stages, and the non-linear "pixel-to-pixel" mapping is learned in the designed MRAB, which ensures a reasonable fusion. Besides, we choose the multi-scale convolution module (MSCB) to be used as a component of the network, which could achieve the purpose of increasing the receptive field while avoiding deep convolution layers of the TDNet. This can also explain why the number of parameters of TDNet is

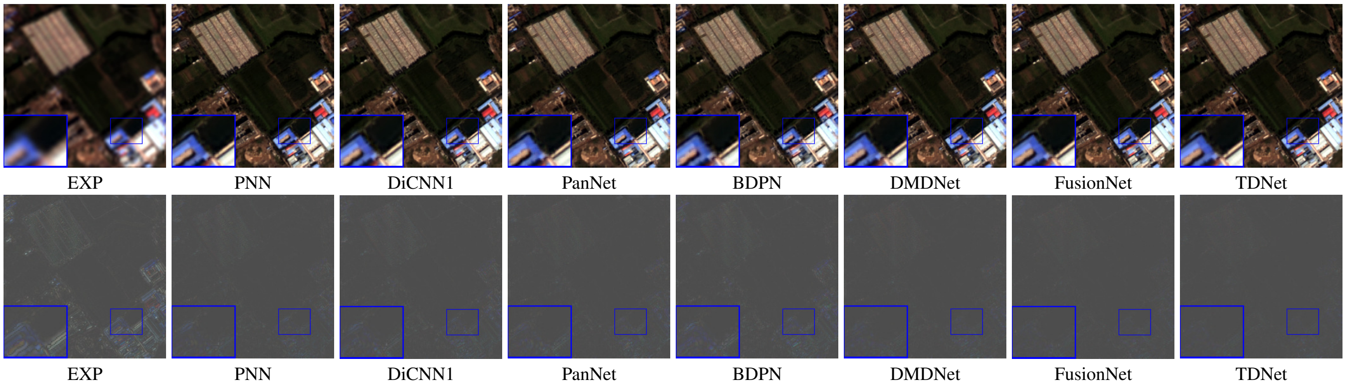


Fig. 13. Visual comparisons between the TDNet and the seven DL-based methods on the Guangzhou datasets (sensor: GaoFen-2). For orderly display, we show the GT image in Fig. 15.

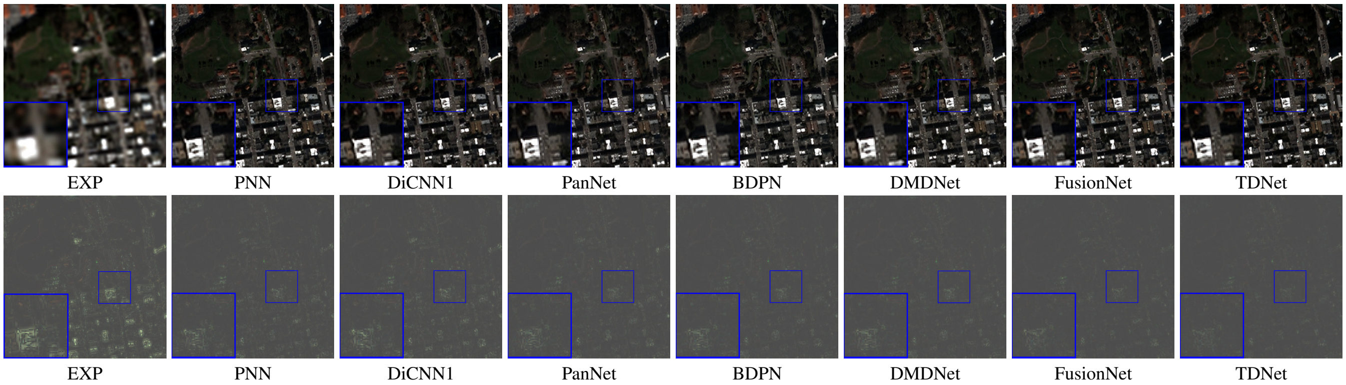
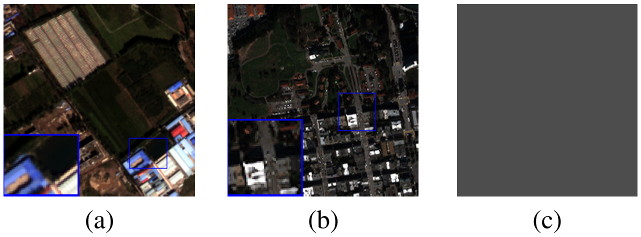


Fig. 14. Visual comparisons between the TDNet and the seven DL-based methods on the Indianapolis datasets (sensor: QuickBird). For orderly display, we show the GT image in Fig. 15.



(a) (b) (c)

Fig. 15. Reference (GT) image for 4-band datasets. (a) Guangzhou dataset (sensor: GaoFen-2) (b) Indianapolis dataset (sensor: QuickBird) (c) The AEM on GT images.

much smaller than that of the BDPN.

## V. Conclusions

In this paper, we propose a novel deep neural network architecture for pansharpening, the so-called triple-double network (TDNet), by taking into account the following three double-type structures, *i.e.*, double-level, double-branch, and double-direction. By exploiting the structure of the TDNet, the spatial details of the panchromatic image can be fully exploited and progressively injected into the LRMS image yielding a final MS image with high spatial resolution. Motivated by the traditional MRA formula, an effective MRA block was integrated into the TDNet. Furthermore, the MSCB with few ResNet blocks and some multi-scale convolution kernels was also used to deepen and widen the network, aiming to effectively enhance the feature extraction and robustness of the proposed TDNet. Extensive experiments on reduced and full resolution examples, acquired by WorldView-3, QuickBird, and Gaofen-2 sensors, demonstrate the superiority of the proposed method. In addition, several ablation studies and discussions are also conducted to corroborate the effectiveness of the proposed TDNet.

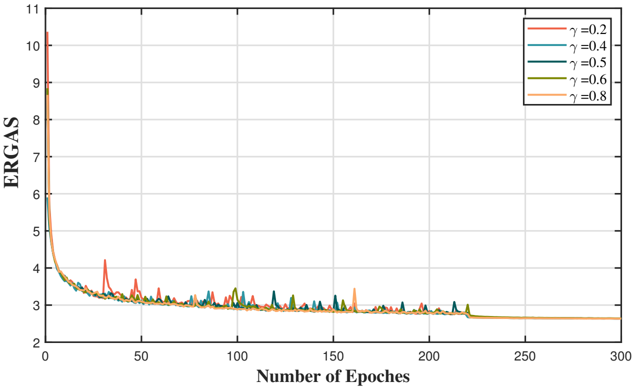


Fig. 16. ERGAS averaged on WorldView-3 test cases for $\gamma = \{0.2, 0.4, 0.5, 0.6, 0.8\}$. Since the convergences are poor when $\gamma = 0$ and $\gamma = 1$, we decided to avoid plotting them.

# References

[1] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multiscale detail networks for multiband spectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 5, pp. 2090–2104, 2020. 1, 2, 7, 8, 9, 10

[2] L. J. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6995–7010, 2021. 1, 3, 6, 7, 8, 9, 10, 11

[3] G. Vivone, P. Addesso, R. Restaino, M. Dalla Mura, and J. Chanussot, "Pansharpening based on deconvolution for multi-band filter estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 540–553, 2019. 1

[4] C. Souza, L. Firestone, L. Silva, and D. Roberts, "Mapping forest degradation in the Eastern Amazon from SPOT 4 through spectral mixture models," *Remote Sensing of Environment*, vol. 87, pp. 494–506, 2003. 1

[5] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and bayesian soft fusion," *Remote Sensing of Environment*, vol. 199, pp. 241–255, 2017. 1

[6] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 1301–1312, 2008. 1

[7] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2565–2586, 2015. 1, 7

[8] X. Meng, H. Shen, H. Li, L. Zhang, and R. Fu, "Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: Practical discussion and challenges," *Information Fusion*, vol. 46, pp. 102–113, 2019. 1

[9] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution based satellite image fusion by using partial replacement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 295–309, 2011. 2, 7, 9, 10

[10] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," 2000, uS Patent 6011875. 2, 7, 9, 10

[11] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 228–236, 2008. 2

[12] G. J. Liu, "Smoothing filter based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, pp. 3461–3472, 2000. 2, 7, 9, 10

[13] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 2376–2385, 2005. 2

[14] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *2nd GRSS/ISPRS Joint Workshop on Remote Sensing and Data Fusion over Urban Areas*. IEEE, 2003, pp. 90–94. 2

[15] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Transactions on Image Processing*, vol. 27, pp. 3418–3431, 2018. 2

[16] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Transactions on Image Processing*, vol. 22, pp. 2822–2834, 2013. 2

[17] J. Duran, A. Buades, B. Coll, and C. Sbert, "A nonlocal variational model for pansharpening image fusion," *SIAM Journal on Imaging Sciences*, vol. 7, pp. 761–796, 2014. 2

[18] 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. 2

[19] Q. Wei, N. Dobigeon, J. Y. Tourneret, J. M. Bioucas-Dias, and S. Godsill, "R-FUSE: Robust fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Signal Processing Letters*, vol. 23, pp. 1632–1636, 2016. 2

[20] L. J. Deng, G. Vivone, W. Guo, M. Dalla Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE International Conference on Image Processing (ICIP)*, 2017. 2

[21] Z. Y. Zhang, T. Z. Huang, L. J. Deng, J. Huang, X. L. Zhao, and C. C. Zheng, "A framelet-based iterative pan-sharpening approach," *Remote Sensing*, vol. 10, p. 622, 2018. 2

[22] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Information Sciences*, vol. 535, pp. 156–171, 2020. 2

[23] Y. D. Wang, L. J. Deng, T. J. Zhang, and X. Wu, "Ssconv: Explicit spectral-to-spatial convolution for pansharpening," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, p. DOI: 10.1145/3474085.3475600. 2

[24] J. F. Hu, T. Z. Huang, L. J. Deng, T. X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatio-spectral attention convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 2

[25] J. Yang, X. Fu, Y. Hu, Y. Huang, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 7, 8, 9, 10

[26] 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, pp. 978–989, 2018. 2

[27] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, p. 594, 2016. 2, 3, 7, 8, 9, 10

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[29] Z. R. Jin, L. J. Deng, T. J. Zhang, and X. X. Jin, "Bam: Bilateral activation mechanism for image fusion," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, p. DOI: 10.1145/3474085.3475571. 2

[30] M. Moller, T. Wittman, and A. L. Bertozzi, "A variational approach to hyperspectral image fusion," *The International Society for Optical Engineering (SPIE)*, vol. 7334, p. 73341E.1¨C73341E.10, 2009. 2

[31] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, pp. 2300–2312, 2002. 3, 6, 7, 9, 10

[32] G. Vivone, M. D. Mura, A. Garzelli, R. Restaino, 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, DOI: 10.1109/MGRS.2020.3019315. 3

[33] L. Alparone, B. Aiazzi, S. Baronti, and A. Garzelli, "Sharpening of very high resolution images with spectral distortion minimization," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2004. 3

[34] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 72, pp. 591–596, 2006. 3

[35] I. Amro and J. Mateos, "Multispectral image pansharpening based on the contourlet transform," *Information Optics and Photonics*, vol. 206, pp. 247–261, 2010. 3

[36] L. Wald and T. Ranchin, "Liu 'smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details'," *International Journal of Remote Sensing*, 2002. 3

[37] A. Garzelli and F. Nencini, "Interband structure modeling for pan-sharpening of very high-resolution multispectral images," *Information Fusion*, vol. 6, pp. 213–224, 2005. 3

[38] 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. 3, 4

[39] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Transactions on Geoence and Remote Sensing*, vol. 57, pp. 1–15, 2019. 3, 4, 6, 7, 8, 9, 10, 12

[40] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 1874–1883. 4

[41] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, pp. 691–699, 1997. 6

[42] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, pp. 6421–6433, 2019. 7, 9, 10

[43] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 72, pp. 591–596, 2006. 7

[44] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 930–934, 2014. 7, 9, 10, 11

[45] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRSS data fusion contest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3012–3021, 2007. 7, 9, 10

[46] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Transactions on Image Processing*, vol. 27, pp. 3418–3431, 2018. 7, 9, 10

[47] 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, pp. 1188–1204, 2019. 7, 8, 9, 10

[48] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm," *JPL Airborne Geoscience Workshop; AVIRIS Workshop: Pasadena, CA, USA*, pp. 147–149, 1992. 7

[49] L. Wald, "Data fusion: definitions and architectures: Fusion of images of different spatial resolutions," *Presses des MINES*, 2002. 7

[50] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge landsat tm and spot panchromatic data," *International Journal of Remote Sensing*, vol. 19, pp. 743–757, 1998. 7

[51] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi-/hyper-spectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, pp. 662–665, 2009. 7

[52] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 12

[53] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *European conference on computer vision*. Springer, 2016, pp. 353–369. 12