THREE-BRANCH MULTILEVEL ATTENTIVE FUSION NETWORK FOR HYPERSPECTRAL PANSHARPENING

Peiyan Guan, Edmund Y. Lam

Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam, Hong Kong SAR, China

IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-2792-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/IGARSS46834.2022.9883218

ABSTRACT

In this paper, we propose a three-branch multilevel attentive fusion network (TMA-Net) for hyperspectral pansharpening, which aims to merge low-resolution hyperspectral images (LR-HSIs) and high-resolution panchromatic images (HR-PANs) to obtain HSIs with high resolution. We construct three branches to extract rich features of the two images and the correlation between them, which enables us to capture abundant useful information for pansharpening. We merge the multilevel features extracted by each branch in multiple steps to fully fuse the useful information. An attentive fusion module (AFM) is designed to guide the fusion procedure. It explores the relation between different features and employs attention mechanism to refine them adaptively. The experimental results illustrate the superiority of the TMA-Net.

***Index Terms*—** Hyperspectral pansharpening, three-branch, multilevel fusion, attention mechanism.

1. INTRODUCTION

Due to the hardware limitation, hyperspectral imaging technique capturing rich spectral information usually acquires images with low spatial resolution, which restricts its' application. Hyperspectral pansharpening is a promising technique to generate high-resolution hyperspectral images (HR-HSIs) by fusing low-resolution hyperspectral images (LR-HSIs) and high-resolution panchromatic images (HR-PANs). Over the years, various traditional methods belonging to different classes have been proposed [1, 2]. However these methods generally suffer from great spectral and spatial distortion.

Recently, deep learning algorithms have shown prominent performance in image processing [3–6]. Inspired such success, many methods applying deep networks in pansharpening have been proposed [7, 8]. Some of them concatenate the two images as the input to the network to restore HR-HSIs [9, 10], which is called early fusion. Some other methods propose to first extract abstract features of the two images separately and then merge them at a high level [11], which is called late fusion. However, the two kinds of methods do not investigate the distinct characteristics of the two images and the correlation between them concurrently, which result in undesired distortion. Another drawback of these methods is that they only integrate the images in one step, which causes insufficient fusion of essential information on various levels. In addition, distinguishing the useful contents from the redundant ones of different features is also crucial for accurate fusion.

To solve these problems, we propose a three-branch multilevel attentive fusion network (TMA-Net) for hyperspectral pansharpening. The TMA-Net employs three branches which take the HR-PAN, LR-HSI and the concatenation of them as input, respectively, to investigate both intra-image features and inter-image correlation. In order to utilize more useful contents of different inputs, we fuse the features from the three branches in multiple levels instead of only the highest level. Moreover, we design an attentive fusion module (AFM) to merge the features from different branches more accurately. The AFM learns the guidance which identifies the common and complementary information between the features. It then uses attention mechanisms to enhance the informative components and weakens the less useful ones under the guidance while fusing the features. Experimental results verify the effectiveness of our method.

2. METHODOLOGY

2.1. Problem statement

Given a LR-HSI $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ with C spectral bands and $H \times W$ pixels and an HR-PAN $\mathbf{P} \in \mathbb{R}^{rH\times rW\times 1}$, where r denotes the scale ratio of spatial domain between the two images, our work aims to fuse them to obtain an HR-HSI $\mathbf{Y} \in \mathbb{R}^{rH\times rW\times C}$ with both high spatial and spectral resolution. We propose the TMA-Net to achieve this goal.

2.2. Network architecture

The overall architecture of the TMA-Net is presented in Fig. 1. It employs a three-branch structure to extract the features of different inputs. The features of the three branches

The work is supported in part by the Research Grants Council of Hong Kong (GRF 17200019, 17201620, 17200321) and the University of Hong Kong (104005864).

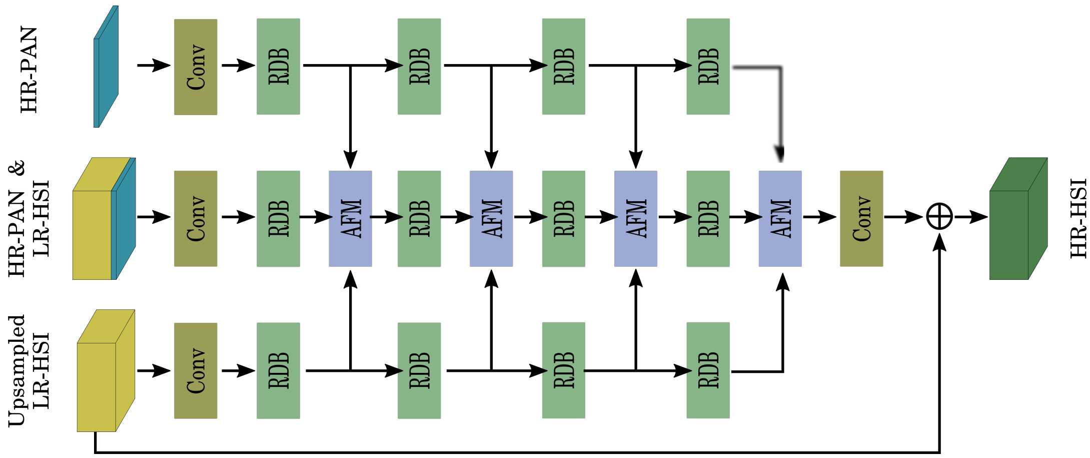

Fig. 1. Architecture of the TMA-Net.

are fused in multiple levels and the fusion feature of the highest level is used to reconstruct the HR-HSI.

Specifically, the top branch takes the HR-PAN as the input and explore the intrinsic characteristics of this image. It employs a convolution layer and four stacked residual dense blocks (RDBs) to extract features of the HR-PAN from low- to high-level. The RDB adopts the same design as those in [12]. The convolution layers in the RDBs have 64 filters, each of which has a size of 3×3 and a stride of 1. The bottom branch has the same structure as the top branch but uses the upsampled LR-HSI as the input, which has the same size of the HR-HSI. This branch aims to capture the hierarchical features of the LR-HSI. Compared to the other two branches, the middle branch contains four additional attentive fusion modules (AFMs). The input of this branch is the concatenation of the upsampled LR-HSI and HR-PAN. Apart from exploring the correlation between the two images, the middle branch merges features from the three branches in multiple steps with the AFMs. Employing three branches which take different images as the inputs enables our method to explicitly leverage both the intra-image characteristics and the inter-image correlation. Besides, the different features are combined in multiple levels, which enables us to fuse as much useful information as possible for the reconstruction.

After the feature fusion procedure, a convolution layer is used to map the fused features to the residual of the HR-HSI. This layer thus has the same number of filters as the band number C. The size and stride of these filters are also 3×3 and 1. By adding the residual and the upsampled LR-HSI, we are capable of obtaining the desired HR-HSI. Such global residual learning scheme avoids the gradient vanishing issue in deep network training.

2.3. Attentive fusion module

The AFM learns the relationship between the features from different branches and employs attention mechanisms to

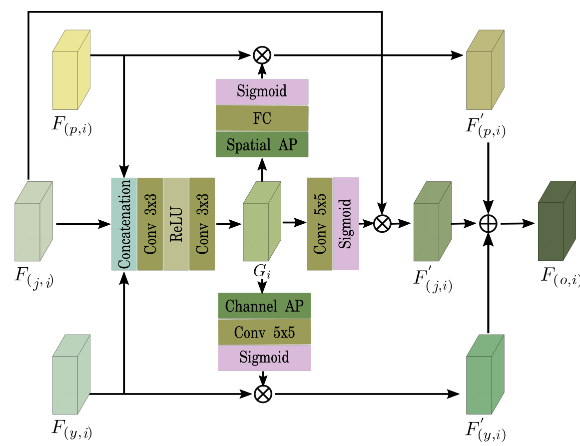

Fig. 2. Architecture of the AFM.

adaptively fuse the features. The structure of the AFM is shown in Fig. 2. Let $F_{(p,i)}$, $F_{(j,i)}$ and $F_{(y,i)}$ represent the features extracted by the ith RDB in the top, middle and bottom branches, respectively, which have the same size of $H \times W \times 64$. The three features contain information of different inputs. Thus, we use the concatenation of them to acquire the guidance with two convolution layers as

$$G_i = W_2(\delta(W_1[F_{(p,i)}, F_{(j,l)}, F_{(y,i)}])), \quad (1)$$

where $[\cdot,\cdot,\cdot]$ denotes the concatenation of the features and δ is the ReLU function. The guidance G_i helps us identify the complementary and common information between the three features.

As shown in Fig. 2, we then use different functions to generate a spectral attention map $A_{(p,i)} = f_{\theta_p}(G_i) \in \mathbb{R}^{1\times 1\times 64}$, a spectral-spatial attention map $A_{(j,i)} = f_{\theta_j}(G_i) \in \mathbb{R}^{H\times W\times 64}$ and a spatial attention map $A_{(y,i)} = f_{\theta_y}(G_i) \in \mathbb{R}^{H\times W\times 1}$ based on the guidance, where $f_{\theta_p}(\cdot)$ consists of a spatial-wise average pooling (Spatial AP) layer, a fully connected (FC) layer and a sigmoid function, $f_{\theta_j}(\cdot)$ comprises of a 5×5 convolution layer and a sigmoid function, f_{θ_y} contains a channel-wise averaging pooling (Channel AP) layer, a 5×5 convolution layer and a sigmoid function.

The attention maps enhance the useful parts and suppress the uninformative ones of the extracted features, which is beneficial to improving the fusion accuracy. We use the attention maps as the adaptive weights for the features and obtain the fused feature $F_{(o,i)}$ with the additive operation as

$$F_{(o,i)} = A_{(p,i)} \circ F_{(p,i)} + A_{(j,i)} \circ F_{(j,i)} + A_{(y,i)} \circ F_{(y,i)}, \quad (2)$$

where $\circ$ is the element-wise multiplication. Note that $A_{(p,i)}$ and $A_{(y,i)}$ are first broadcasted to the size of $H \times W \times 64$. The addition is simple and efficient, which reduces the computation burden greatly.

2.4. Loss function

After designing the network architecture, we define the loss function for the network training. Compared to the commonly-used ℓ_2 norm, the ℓ_1 norm is able to generate sharper and clearer images. Thus, we adopt the ℓ_1 norm as the loss function for the TMA-Net, which is defined as

$$\ell_1 = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{F}(\mathbf{P}_i, \mathbf{X}_i; \Theta) - \mathbf{Y}_i|, \quad (3)$$

where N is the number of training samples and $\mathcal{F}(\cdot;\Theta)$ denotes the TMA-Net with parameters Θ.

3. EXPERIMENTAL RESULTS

3.1. Dataset and implementation details

The effectiveness of our method is validated on the Washington DC Mall dataset, which consists of 191 bands except the water absorption bands. The top left area with a size of 1216×304 is used for our experiment. We subdivide the cropped parts into 16 non-overlapping patches. 11 patches of them are randomly selected for the network training while the rest 5 comprise the testing dataset. The original image is used as the ground truth HR-HSI. The LR-HSI is generated by Gaussian bluring the HR-HSI first and then downsampling it by the scale ratio $r = 8$. In addition, we use the average of all visual bands of the HR-HSI as the HR-PAN. We set the patch size of HR-HSI to 64, the batch size to 16, the learning rate to 2×10^{-4} and the training epoch to 300. Our method is implemented with Pytorch and trained on an Nvidia RTX 2060s graphics card.

3.2. Evaluation metrics

Four metrics are employed to assess the performance of different methods quantitatively, including spectral angle mapper (SAM) [13], root-mean-square error (RMSE), erreur relative globale adimensionnelle de synthèse (ERGAS) [14] and structure similarity index (SSIM). SAM and SSIM measure the spectral and spatial distortion, respectively. RMSE and ERGAS assess the global quality of the reconstructed image. The ideal values for SSIM, SAM, ERGAS and RMSE are 1, 0, 0 and 0, respectively.

Table 1. Quantitative results of each method on the Washington DC Mall dataset. The best value is in bold.

Methods	SSIM	SAM	RMSE	ERGAS
PCA [1]	0.8774	12.06	0.0243	7.259
Indusion [2]	0.9277	7.578	0.0171	5.162
MSDCNN [9]	0.9537	5.666	0.0128	4.100
DARN [10]	0.9597	5.479	0.0119	3.672
TFNet [11]	0.9583	5.592	0.0124	3.856
HyperPNN [7]	0.9583	5.827	0.0126	3.997
TMA-Net	**0.9738**	**4.046**	**0.0092**	**2.745**

3.3. Results and analysis

We compare the performance of the TMA-Net with two traditional methods, including PCA [1] and Indusion [2], and four deep learning methods, including MSDCNN [9], DARN [10], TFNet [11] and HyperPNN [7]. We implement the traditional methods with publicly available Matlab codes and re-implement the deep learning methods with Pytorch. These methods are evaluated on the same dataset.

The quantitative results are presented in Table 1. As can be seen, the deep learning methods outperform the traditional ones. Despite both belonging to early fusion methods, MSDCNN shows worse results than DARN, which shows that the network depth and structure also play an important role in the fusion performance. In contrast, the TMA-Net surpasses the other deep learning methods significantly, which demonstrates the superiority of employing three-branches to investigate the both the intra- and inter-image characteristics and fusing the features in multiple levels.

We select a patch "WDC_16" from the testing dataset to present the visual comparison of different methods as shown in Fig. 3. One region of interest (ROS) is also zoomed in for better demonstration. As can be seen, the fused images of PCA, TFNet and HyperPNN show clear color distortion while Indusion produces images with serious ringing artifacts. By contrast, the TMA-Net generates HR-HSIs with slightest color aberration and structure distortion, which is consistent with the quantitative results. Moreover, as shown in the magnified regions marked by yellow boxes, our method recovers sharp edges and contours faithfully to the reference HR-HSI while other methods present blurry results, which verifies the effectiveness of the TMA-Net on recovering spatial structures and texture details.

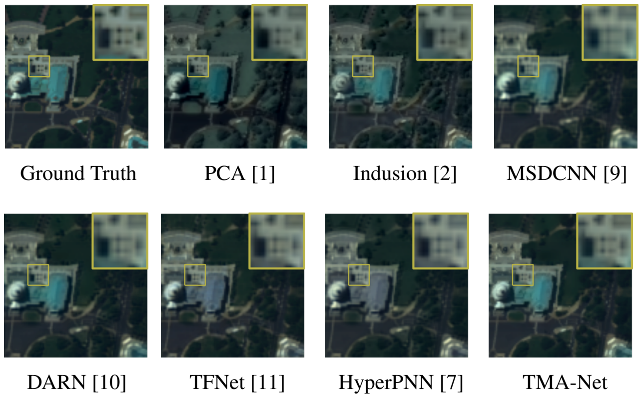

Fig. 3. Qualitative results of each method over the Washington DC Mall dataset.

4. CONCLUSION

In this work, we propose the TMA-Net equipped with AFMs for hyperspectral pansharpening. By using three branches, the intra- and inter-image information is fully investigated. Merging the multilevel features of the three branches promotes the fusion of more useful contents. In each fusion step, an AFM is employed to fuse the features more accurately. With the guidance of the AFM, the desired components are differentiated from the less useful ones. Our method outperforms the state-of-the-art methods greatly.

5. REFERENCES

[1] Vijay P Shah, Nicolas H Younan, and Roger L King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[2] Muhammad Murtaza Khan, Jocelyn Chanussot, Laurent Condat, and Annick Montanvert, "Indusion: fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geoscience and Remote Sensing Letters*, vol. 5, no. 1, pp. 98–102, January 2008.

[3] Nan Meng, Hayden Kwok-Hay So, Xing Sun, and Edmund Lam, "High-dimensional dense residual convolutional neural network for light field reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 873–886, March 2021.

[4] Kai Zhang, Wangmeng Zuo, and Lei Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4608–4622, September 2018.

[5] Yanmin Zhu, Chok Hang Yeung, and Edmund Y Lam, "Microplastic pollution monitoring with holographic classification and deep learning," *Journal of Physics: Photonics*, vol. 3, no. 2, pp. 024013, April 2021.

[6] Zhenbo Ren, Hayden K-H So, and Edmund Y Lam, "Fringe pattern improvement and super-resolution using deep learning in digital holography," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 11, pp. 6179–6186, April 2019.

[7] Lin He, Jiawei Zhu, Jun Li, Antonio Plaza, Jocelyn Chanussot, and Bo Li, "HyperPNN: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 8, pp. 3092–3100, August 2019.

[8] Peiyan Guan and Edmund Y Lam, "Multistage dual-attention guided fusion network for hyperspectral pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, October 2021.

[9] Qiangqiang Yuan, Yancong Wei, Xiangchao Meng, Huanfeng Shen, and Liangpei 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, March 2018.

[10] Yuxuan Zheng, Jiaojiao Li, Yunsong Li, Jie Guo, Xianyun Wu, and Jocelyn 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, November 2020.

[11] Xiangyu Liu, Qingjie Liu, and Yunhong Wang, "Remote sensing image fusion based on two-stream fusion network," *Information Fusion*, vol. 55, pp. 1–15, March 2020.

[12] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu, "Residual dense network for image super-resolution," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 2472–2481.

[13] Roberta H Yuhas, Alexander FH Goetz, and Joe W Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, June 1992, vol. 1, pp. 147–149.

[14] Lucien Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Third Conference "Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images"*, 2000, pp. 99–103.