HYPERSPECTRAL IMAGE SUPER-RESOLUTION VIA MULTI-DOMAIN FEATURE LEARNING

*Qiang Li, Qi Wang *, Xuelong Li*

School of Computer Science and School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China

ABSTRACT

Hyperspectral image super-resolution (SR) methods are continually being refreshed due to deep neural networks. Despite this, the existing works barely explore more spatial information using mixed 2D/3D convolution. Moreover, they do not make full use of multi-domain features to realize information complementation. To tackle these challenges, we propose a hyperspectral image SR approach via multi-domain feature learning. To be specific, a multi-domain feature learning strategy using 2D/3D unit is presented to explore spatial and spectral information by alternate manner. To recover the more details, the edge body generation mechanism (EBGM) is introduced to learn the high frequency information, which generates the edge prior. Besides, the multi-domain feature fusion (MDFF) is designed to fully integrated hierarchical knowledge from different 2D/3D units, leading to further achieve information complementation. Experiments demonstrate that our approach attains the better performance over the state-of-the-art methods.

Index Terms— Hyperspectral image, super-resolution, multi-domain feature learning, edge body generation

1. INTRODUCTION

Hyperspectral image has rich spectral information, which can reflect the subtle spectral properties of the target in detail. Combined with related image processing tasks, specific spectral features are selected according to target attributes, which is helpful for accurate analysis. Due to physical limitations on spectral sensor, hyperspectral image presents lower resolution in spatial domain than in spectral domain. Hence, more efforts have been proposed to address super-resolution (SR), which refers to restoring a LR hyperspectral image to a high-resolution (HR) hyperspectral image.

Hyperspectral image SR methods using convolutional neural network (CNN) have shown great success in the recently. According to the type of convolution used by the networks, it can be roughly divided into three categories, namely, 2D convolutional network, 3D convolutional network, and mixed 2D/3D convolutional network. As for the models adopting 2D convolution [1, 2], they views each band of the hyperspectral image as an image to design the network. This helps to explore spatial features, but does not take advantage of rich spectral information. Considering this issue, various networks using regular 3D convolution have been proposed [3]. Compared with 2D convolutional networks, these algorithms significantly improve the performance. However, the regular 3D convolution yields a large number of parameters compared with 2D convolution, which is not conducive to designing deeper networks for limitation hardware condition. To address this problem, the researchers exploit separable 3D convolution [4] to build the model [5], which greatly reduces unaffordable memory usage.

Since the contents of two adjacent spectra are usually similar [6], the constructed model should more emphasis on the analysis of spatial domain in feature learning. Thus, Li *et al.* first propose a parallel structure SR network via mixed 2D/3D convolution (MCNet) [7]. Although the method produces the superior performance over the state-of-the-art approaches, the parallel structures leads to module redundancy. With respect to the mixed convolution, there has been little research into this in existing works. Importantly, previous studies do not make full use of multi-domain features to realize information complementation. To address these challenges, we propose a hyperspectral image SR approach via multi-domain feature learning, which is shown in Fig. 1. In summary, our main contributions of this paper are as follows:

- A novel multi-domain feature learning strategy using 2D and 3D unit is proposed by alternate manner, which can effectively explore multi-domain knowledge by sharing spatial information.
- The edge body generation mechanism is introduced to explicitly learn edge feature representation, which provides edge prior for high-quality reconstruction, so as to recover the more details.
- The multi-domain feature fusion is designed to adaptively preserve the accumulated features for different deep 2D/3D unit. It is conducive to the integration of spectral and spatial information.

* Qi Wang is the corresponding author. This work was supported by the National Key R&D Program of China under Grant 2018YFB1107403, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.

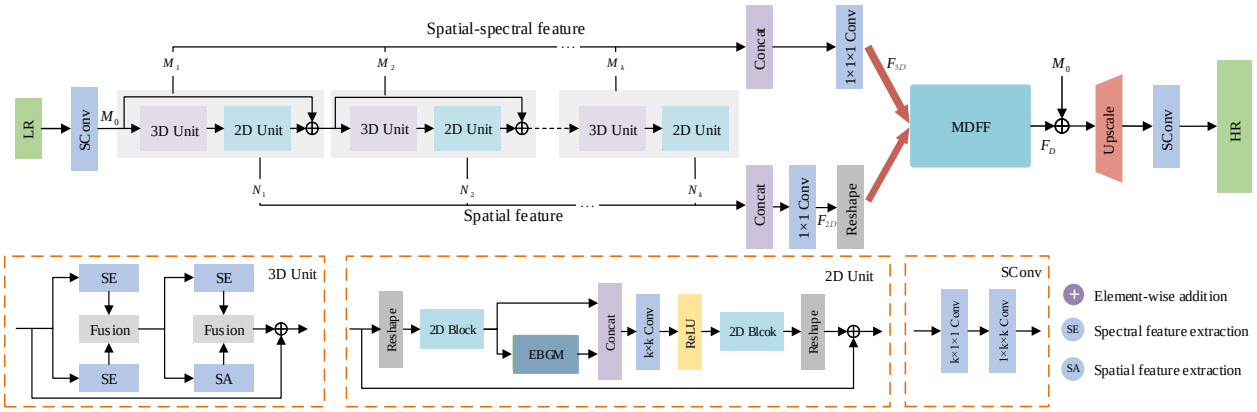

Fig. 1. The overall architecture of the proposed network for hyperspectral image SR.

2. METHODOLOGY

In this section, we describe the proposed method in details, including network architecture, multi-domain feature learning, and edge body generation mechanism.

2.1. Network Architecture

Now we present the proposed architecture. Suppose $I_{LR} \in R^{L\times W\times H}$ is the LR hyperspectral image, where L is the total bands. W and H denote the width and height of hyperspectral image, respectively. Since 3D convolution can analyze information other than spatial dimension, we adopt separable 3D convolution [4] (it is defined as *SConv*) to extract the initial features M_0. Then, these features are fed into the main module that contains 2D and 3D unit. To make full use of all the hierarchical features generated by different 2D and 3D units, the multi-domain feature fusion (MDFF) is designed to adaptively preserve the accumulated features, i.e.,

$$F_{3D} = f_D(Concat[M_1, ..., M_k]), \quad (1)$$

$$F_{2D} = f_D(Concat[N_1, ..., N_k]), \quad (2)$$

where $Concat$ is utilized to concatenate the different deep information, and $f_D(\cdot)$ is the function to reduce the feature channel. As the size of feature maps obtained by the two is different, we reshape one of them. With respect with specific operation, it will be described in Section 2.2. Finally, the fused features are produced by

$$F_D = f_D(Concat[w_1 * F_{3D}, w_2 * Reshape(F_{2D})]). \quad (3)$$

Through this way, it can effectively integrate spectral and spatial information, leading to achieve deep complementation.

After extracting features in the LR space, we exploit a transposed convolution layer to upscale them in the HR space according to r, which is followed by a *SConv*. The super-resolved hyperspectral image $I_{SR} \in R^{L\times rW\times rH}$ is achieved by

$$I_{SR} = f_{SD}(f_{up}(F_D + M_0, r), \quad (4)$$

where $f_{up}(\cdot)$ and $f_{SD}(\cdot)$ are the functions for upscaling and reconstruction, respectively.

2.2. Multi-domain feature learning

The benefit of using spectral information of hyperspectral image is to improve the spatial performance. Under the condition that the spectral information can be extracted, how to combine the 2D/3D convolution to increase spatial exploration still needs more research efforts. Therefore, we develop a novel structure that appears alternately through 2D and 3D unit, which can effectively explore multi-domain knowledge by sharing spatial information.

With respect to 3D unit, we employ the kernels $k \times 1 \times 1$ and $1 \times k \times k$ to explore the spatial and spectral information, i.e., *SE* and *SA*. After obtaining independent features, how to fuse them is one of the key skills to improve the performance. Currently, the existing works nearly employ addition to fuse them. In our work, three fusion strategies are adopted to analyze the performance impact of the model, whose results are shown in Section 3.3.

To apply 2D convolution after 3D unit, it is necessary to reshape the feature maps. Concretely, suppose that the size of feature maps is $N \times C \times L \times W \times H$, where B is the batch size, and C denotes the number of filters. To transform it, we handle the each band individually, i.e, the channel L and N are integrated together in this process. Different from the structure in MCNet [7], 2D unit add more 2D blocks, which consists of two 2D convolutions with the kernel $k \times k$, ReLU function, and residual connection. Meanwhile, the edge body generation mechanism (EBGM) is embedded to explicitly learn edge feature representation, which provides edge prior for high-quality reconstruction.

Compared with the network in which 2D units are replaced by 3D units, the alternate structure not only effectively

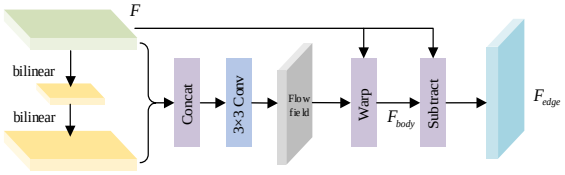

Fig. 2. The architecture of edge body generation.

add the learning ability in spatial domain, but also can reduce the number of parameters. Importantly, this alternate approach can also integrate multi-domain knowledge by sharing spatial information, thus improving feature learning.

2.3. Edge body generation mechanism

For hyperspectral image SR task, the introduction of high-frequency information is helpful to restore the texture details. A natural way is to utilize off-the-shelf edge detectors, such as Canny, Sobel, etc., to retain image edge information. However, these methods use the binarization measurement, which can easily lead to the loss of image features and the existence of false edge. Inspired by optical flow in semantic segmentation [8], the edge body generation mechanism (EBGM) is adopted to generate the edge prior.

Without loss of generality, the feature map F from 2D block can be divided into the body part F_{body} and the edge part F_{edge}, which describe the smooth structure and sharper details, respectively. Suppose they satisfy the addition, i.e,

$$F = F_{body} + F_{edge}. \tag{5}$$

Usually, the LR feature map contains more body part, so we first exploit the bilinear method to downsample the original feature map. Subsequently, it is upsampled to the same size as the original feature map F by bilinear method, and the two are concatenated together. These features are input in the convolution layer to predict the flow field. Finally, the original feature map F is deformed by flow field to obtain the body part, which can be formulated as

$$F_{body}(x) = \sum_{x \in N(x_l)} w_x F(x), \tag{6}$$

where w_x is the weight calculated from the flow field. $F(x)$ is the corresponding pixel feature. $N(x_l)$ represents the pixel involved in the calculation. Finally, the edge prior F_{edge} is acquired by subtraction with original feature map F.

3. EXPERIMENT

3.1. Dataset

To verify the performance the proposed method, two public datasets are adopted, including CAVE[1] and Chikusei[2]. With regard to CAVE dataset that contains 32 images, we select 80% of samples as training set and the rest as the test set. As for Chikusei dataset, it is hyperspectral remote sensing data. The image in the top left $128 \times 2000 \times 2335$ is selected to train, and the rest of image is used to test. Given a training set, we randomly crop 24 and 108 patches for CAVE and Chikusei dataset, respectively. These patches are augmented by rotation, flip, etc. Then, the LR hyperspectral images with the size of $L \times 32 \times 32$ are obtained by downsamping these patches using bicubic interpolation.

[1]http://www1.cs.columbia.edu/CAVE/databases/multispectral/
[2]http://naotoyokoya.com/Download.html

Table 1. Study of spatial and spectral fusion in 3D unit.

Fusin Method	PSNR	SSIM	SAM	#Params.
Add	39.140	0.9320	3.242	1826k
Max	39.273	0.9321	3.200	1826k
Concat + Conv	39.117	0.9317	3.210	1893k

3.2. Implementation Details and Evaluation Metrics

During training the network, we fix the size k and number of filters as 3 and 64, except for upsampling and convolution layer in EBGM. As for EBGM, the size and number of filter is set to 3 and 2. The model is optimized by Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ to minimize L1 loss function. The initial learning rate is set to 10^{-4} for all layers, which decreases by a half at every 35 epochs.

To evaluate SR performance, three metrics are utilized, including Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), and Spectral Angle Mapper (SAM).

3.3. Study of Spectal and Spatial Fusion

In our work, three fusion ways are adopted to analyze the influence of the performance. Given two inputs X and Y, they represent the extracted spatial and spectral features, respectively. Suppose they are all $N \times C \times L \times W \times L$ in size. Conv fusion is to stack X and Y together along second channel, and then convolve the stacked data with a group of filters with $1 \times 1 \times 1$ to obtain fused result. Table 1 show the performance of different fusion strategies for $\times 4$ SR on CAVE dataset. As seen from the table, max fusion achieves good results, especially in PSNR. Therefore, max fusion is utilized to evaluate the performance for the rest of the paper.

3.4. Ablation Study

To demonstrate the effectiveness of main parts, including EBGM and MDFF, the ablation study is performed for $\times 4$ SR on CAVE dataset by setting different combinations, which is depicted in Table 2. We can find that when any one part is added into the network, three evaluation metrics are improved to a certain extent, compared with those before the attachment. When both EBGM and MDFF exist, all values

Table 2. Ablation study about different combinations

Part	Different combinations of part			
EBGM	×	×	√	√
MDFF	×	√	×	√
PSNR	39.052	39.125	39.115	39.273
SSIM	0.9317	0.9317	0.9318	0.9321
SAM	3.242	3.216	3.204	3.200

Table 3. Quantitative evaluation on CAVE dataset.

Method	PSNR	SSIM	SAM	#Params
SSPSR[1]	38.366	0.9227	3.484	12875k
3D-FCNN [3]	37.626	0.9195	3.360	39k
MCNet [7]	39.026	0.9319	3.292	2174k
Ours	**39.273**	**0.9321**	**3.200**	1826k

attains better results than the above combinations. The above analysis reveals the effectiveness of each part.

3.5. Comparisons with the State-of-the-art Methods

We evaluate the performance of the proposed approach using three existing methods for $\times 4$ SR on two datasets. As shown in Tables 3-4, the proposed method achieves the comparable or even better results among these competitors. Compared with these methods, the designed model could explore more multi-domain knowledge by alternate 2D and 3D unit. Moreover, the EBGM can generate edge prior, which enables the model to focus on high-frequency information to recover details. We also show the visual results on CAVE dataset in Fig. 3. To clearly exhibit the difference with ground-truth, the absolute error map is given by selecting 10-th band. As seen in figure, our method yields the shallow textures in some areas. Besides, one pixel position is selected to analyze the spectrum difference. The curve generated by the proposed approach are closer to the ground-truth in most cases.

4. CONCLUSION

In this paper, we develop a hyperspectral image SR approach using multi-domain feature learning. To fully utilize complementary information between domains, a novel multi-domain feature learning strategy 2D and 3D unit is proposed by alternate manner. Moreover, the edge body generation mechanism is introduced to learn edge prior, which helps to restore texture details. The experiments demonstrate that our method obtains the better performance over the existing approaches.

5. REFERENCES

[1] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Trans. Comput. Imaging*, vol. 6, pp. 1082–1096, 2020.

Table 4. Quantitative evaluation on Chikusei dataset.

Method	PSNR	SSIM	SAM	#Params.
SSPSR[1]	39.775	0.9450	7.890	13546k
3D-FCNN [3]	39.060	0.9364	10.681	39k
MCNet [7]	**40.516**	0.9507	11.087	2174k
Ours	40.483	**0.9512**	**7.369**	1826k

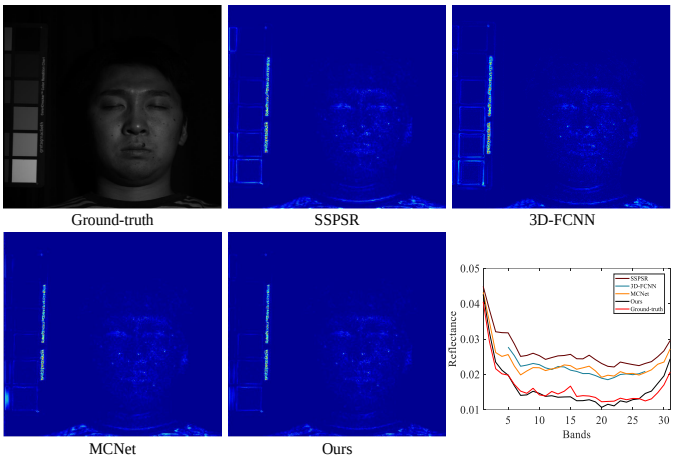

Fig. 3. Visual comparisons with different algorithms.

[2] J. Hu, X. Jia, Y. Li, G. He, and M. Zhao, "Hyperspectral image super-resolution via intrafusion network," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 10, pp. 7459–7471, 2020.

[3] S. Mei, X. Yuan, J. Ji, Y. Zhang, S. Wan, and Q. Du, "Hyperspectral image spatial super-resolution via 3D full convolutional neural network," *Remote Sens.*, vol. 9, pp. 1139, 2017.

[4] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 318–335.

[5] J. Li, R. Cui, B. Li, Y. Li, S. Mei, and Q. Du, "Dual 1D-2D spatial-spectral cnn for hyperspectral image super-resolution," in *Proc. Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3113–3116.

[6] Q. Wang, Q. Li, and X. Li, "A fast neighborhood grouping method for hyperspectral band selection," *IEEE Trans. Geosci. Remote Sensing*, 2020.

[7] Q. Li, Q. Wang, and X. Li, "Mixed 2D/3D convolutional network for hyperspectral image super-resolution," *Remote Sens.*, vol. 12, no. 10, pp. 1660, 2020.

[8] X. Li, A. You, Z. Zhu, H. Zhao, M. Yang, K. Yang, and Y. Tong, "Semantic flow for fast and accurate scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 435–452.