SATELLITE IMAGE FUSION AND DENOISING USING CONVOLUTIONAL NEURAL NETWORKS

Shih-Shuo Tung*[1] , Li-Fen Huang[2], Nai-Yu Chen[2], and Wen-Laing Hwang[1]
[1]Instutute of Information Science, Academia Sinica
[2]Taiwan Space Agency

KEY WORDS: Image Fusion, Convolutional Neural Network, Image Denoising

ABSTRACT: Due to the optical errors in the imaging system, the received remote sensing image is blur and noisy. Although some image techniques can restore the image, the different algorithms generate different type of restorations. In this paper, there are two restorations, one is noisy but with rich texture information and another is smooth but noise-free. The goal of this paper is to fuse these two restored images to a sharp and noise-free image. However, the traditional non-learning methods are time consuming and no training dataset for this task. In order to accelerate the processing, the neural network based method is applied. In this paper, we proposed a dataset for training and a convolutional neural network to fuse images where each image was decomposed into three channels, high-frequency, structure, and texture components. The decomposed high-frequency channel contains most noise information. In order to have high quality dataset, denoising is very important. Here, the LISTA (Learned Iterative Shrinkage and Thresholding Algorithm) method was applied to do denoising. According to the three decomposed channels, three rules were designed to fusion. The high quality dataset can be constructed by the previous processing. Therefore, the convolutional neural network learning can be achieved through the dataset. In the experiments, the proposed method achieved the significant performance when comparing to state-of-the-art methods. For a 12000x12000 PAN satellite image fusion, the processing time is about 2 minutes for GeForce RTX 2080 graphics card. If higher level devices, the cost time will decrease.

1. INTRODUCTION

Due to the optical errors in the imaging system, the received remote sensing image is blur and noisy. In order to restore the image, some image processing techniques are applied. However, different restoration algorithms have different advantages and disadvantages. There is no one method can restore the blur and noisy image in perfect quality. One method maybe can restore the image with sharp edges or textures but the restoration is also accompanied by the noise. On the other hand, another restored image is noise-free but the edges and textures are not sharp. It is a trade-off between the sharpness and noise.

In order to solve the problem, fusion of these two types of restoration is a solution to have a sharp and noise-free image. In recent years, there are two categories of fusion with denoising, one is optimization based (Li, 2021) method and another is learning based (Ulyanov, 2018; Uezato, 2020) method. The non-learning method uses only the mathematical formulations to formulate the fusion problem and solve the problem with optimization methods. These methods can do fusion without dataset but the time cost of computation is usually very high. However, the learning based methods have the advantages in time computing and high performance when the models were trained from high quality datasets. However, there are no datasets obtained for our applications. Usually, the fusion applications are multi-focus fusion, infrared visible image fusion, multi-modal image fusion, and multi-exposure image fusion.

In this paper, our goal is to do fusion with denoising in a fast time. Therefore, we must applied the learning based method to fusion. However, there are two difficulties, the first is the high quality dataset collected, and the second is the suitable simple and effective neural network. As we mentioned before, there are no datasets can be obtained for satellite images fusion with denoising. Here, we applied the three-layer decomposition method (Li, 2021) with unrolling method (Gregor, 2010), also called LISTA (Learned Iterative Shrinkage and Thresholding Algorithm), to generate the dataset. In (Zhang, 2020), the IFCNN network is simple and effective, it can also be applied to our model although the original IFCNN is applied to other fusions. With the dataset and network, the fusion can be achieved with a high quality result and low computation time.

There are two contributions of this paper, one is to generate the dataset and another is to do fusion with little computation cost through the CNN (convolutional neural network). The remainder of this paper is organized as follows. In Section 2, we present a review of related works. Section 3 outlines our approaches including the dataset generation and network construction. Section 4 shows the performance of the proposed method. Conclusions are presented in Section5.

2. RELATED WORKS

Three-layer decomposition based method (Li, 2021) decomposed each input image into high and low frequency images respectively. The noise information is usually preserved in the high frequency image. Then, each low frequency image is decomposed into structure and texture images through the interval gradient filter. In the structure image, it contains only the strength information of the pixel. In the texture image, it contains only the edge and texture information. The low frequency image can be obtained by

$$\arg\min_{X_m^l} \|X_m - X_m^l\|_F^2 + \beta(\|g_a * X_m^l\|_F^2 + \|g_b * X_m^l\|_F^2), \tag{1}$$

where X_m is the input image, X_m^l is the low frequency image, g_a and g_b are $[1 - 1]$ and $[1 - 1]^{\mathrm{T}}$ respectively, and m=1, 2. The high frequency image X_m^h, structure image $X_m^{l,s}$, and texture image $X_m^{l,t}$ can be obtained by

$$X_m^h = X_m - X_m^l, \tag{2}$$

$$X_m^{l,s} = IGF(X_m^l), \tag{3}$$

$$X_m^{l,t} = X_m^l - X_m^{l,s}, \tag{4}$$

where IGF is interval gradient filter. The fusion procedure is to fuse these three decomposed images from input images.

In the high frequency image, the aim is to reduce the noise and preserve the high frequency information. The sparse representation is first applied to denosing,

$$\min_{\alpha_m^r} \|\alpha_m^r\|_0 \quad s.t. \ \|v_m^r - D\alpha_m^r\|_2^2 < \epsilon_m^r, \tag{5}$$

where α_m^r is the sparse coefficient, v_m^r is the vector at rth block of the mth image, D is the dictionary, and ϵ_m^r is the error. Second, the fused sparse coefficient α_F^r and high frequency image block v_F^r can be obtained by

$$\alpha_F^r = \alpha_{\hat{m}}^r, \ \hat{m} = \arg\max_m \{\|\alpha_m^r\|_1 | m = 1,2\}, \tag{6}$$

$$v_F^r = D\alpha_F^r. \tag{7}$$

The fusion of structure images $F^{l,s}$ can be obtained by

$$F^{l,s} = F_{\bar{m}}^{l,s}, \ \bar{m} = \arg\max_m \left\{\left\|X_m^{l,s}\right\|_1 | m = 1,2\right\}. \tag{8}$$

The fusion of texture image $F^{l,t}$ can be obtained by

$$F^{l,t} = F_{\tilde{m}}^{l,t}, \ \tilde{m} = \arg\max_m \left\{NSF(X_m^{l,t}) | m = 1,2\right\}, \tag{9}$$

$$NSF = \sqrt{RF^2 + CF^2}, \tag{10}$$

$$RF = \sqrt{\frac{1}{\hat{M} \times \hat{N}} \sum_{a=1}^{\hat{M}} \sum_{b=1}^{\hat{N}} [I(a,b) - I(a,b-1)]^2}, \tag{11}$$

$$CF = \sqrt{\frac{1}{\hat{M} \times \hat{N}} \sum_{a=1}^{\hat{M}} \sum_{b=1}^{\hat{N}} [I(a,b) - I(a-1,b)]^2}. \tag{12}$$

The final fused image is determined by summing the three fused images, by summing Eqs. (7), (8), and (9). The overview of the method is shown in Figure 1. In the sparse representation Eq. (5), the conventional optimization method is OMP (orthogonal matching pursuit) (Pati, 1995). The method is simple but iterative. If getting high quality reconstruction, it needs more iterations. If the iterations can be unrolled into several layers in a network (Gregor, 2010), the computation time becomes lower. Our approach is to replace the sparse representation by a neural network to accelerate the procedure to generate the training dataset.

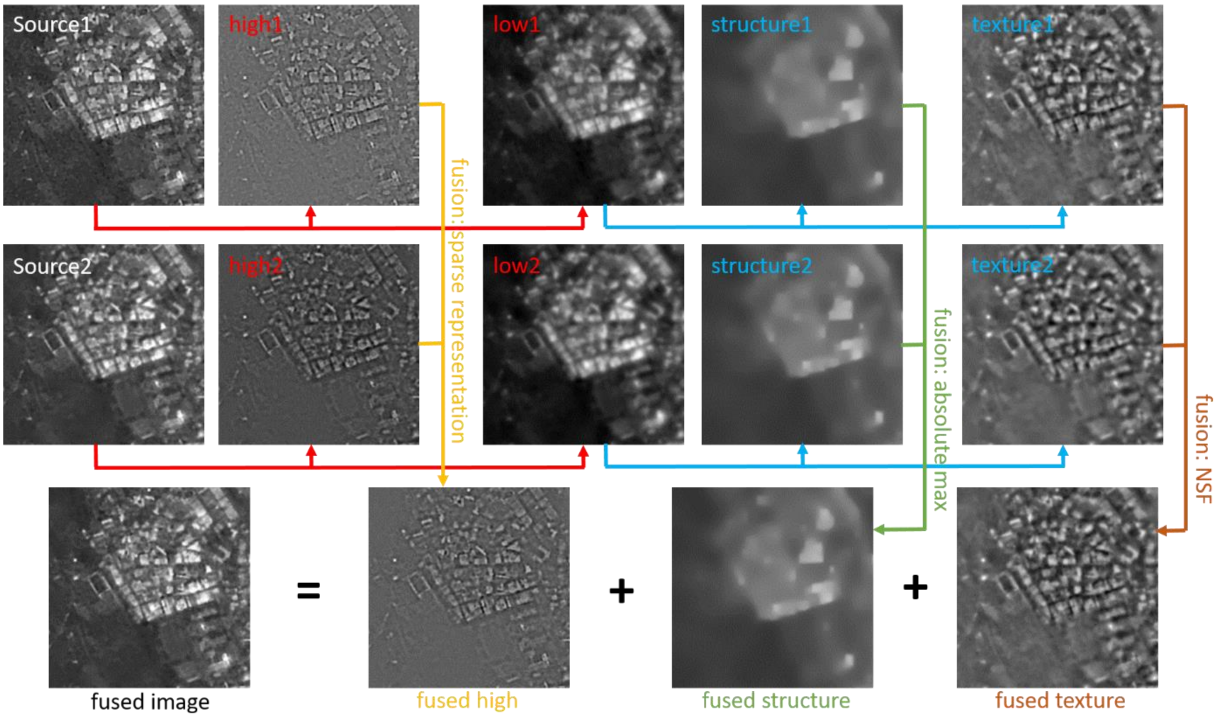


Figure 1. The overview of the three-layer decomposition based method

3. PROPOSED METHODS

In this paper, we proposed the dataset generation method and a CNN fusion model. In the first part, the dataset was generated from the improved three-layer decomposition based method by unrolling. In the second part, a CNN was proposed and each input image is a three-channel image including high frequency, structure, and texture channels.

3.1 Dataset generation

As mentioned in the related works, the OMP algorithm is an iterative method and time consuming. Here, we adopted the unrolling method in sparse representation to acceleration. If the processing time can be reduced, we can apply the method to fusion to construct the dataset. First, we relax the sparse representation problem with L1 norm as

$$\min_{\alpha} \|\alpha\|_1 \quad s.t. \ \|y - D\alpha\|_2^2 < \epsilon, \tag{13}$$

and it can also be reformulated as

$$\hat{\alpha} = \text{argmin}_{\alpha} \frac{1}{2} \|D\alpha - y\|_2^2 + \lambda\|\alpha\|_1. \tag{14}$$

The equation can be solved by gradient descent and soft-thresholding via iterations,

$$\hat{\alpha}_{t+1} = h_{\lambda/c}\left(\hat{\alpha}_t - \frac{1}{c} D^T(D\hat{\alpha}_t - y)\right), \quad \hat{\alpha}_0 = 0, \tag{15}$$

where $h_{\lambda/c}$ is the soft-thresholding function, λ/c is the threshold, and c is a constant,

$$[h_\theta(u)]_i = sign(u_i)\max(|u_i| - \theta,\ 0). \tag{16}$$

Eq. (15) can also be rewritten as

$$\hat{\alpha}_{t+1} = h_{\lambda/c}\left(\left(I - \frac{1}{c} D^T D\right)\hat{\alpha}_t + \frac{1}{c} D^T y\right), \quad \hat{\alpha}_0 = 0. \tag{17}$$

Let $S = \left(I - \frac{1}{c} D^T D\right)$ and $W_e = \frac{1}{c} D^T$ be the fixed variables, the sparse coefficients can be obtained by iterative process. The method is called ISTA (Iterative Shrinkage and Thresholding Algorithm). Compared to the OMP method, it is simple.

From Eq. (17), S and W_e will be used once and pass the soft-thresholding for each iteration. So, if through N iterations, the procedure will repeat N times. If we let the numbers of iteration as the layers in a network, the sparse coefficients can be obtained without many iterations. It can be obtained only through the fixed layers network. Therefore, the cost of time computation will decrease and the result will be more stable. The concept is to transfer the iterations to learnable layers, it is called unrolling or unfolding. The learning method is called LISTA (Learned ISTA) (Gregor, 2010). The differences between ISTA and LISTA are shown in Figure 2. The iterations in ISTA is transferred to layers in LISTA.

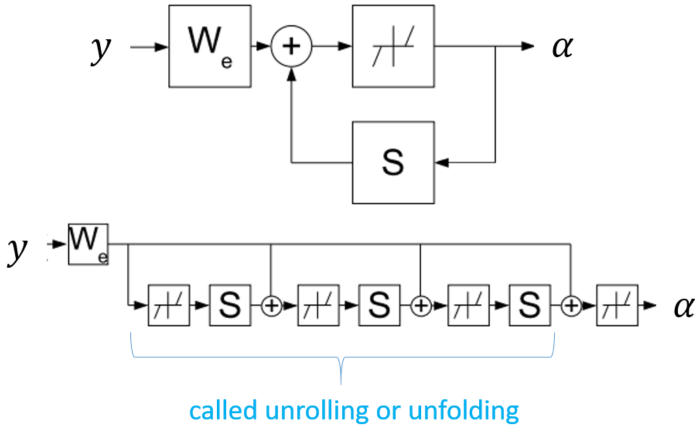


Figure 2. ISTA and LISTA configurations

In LISTA, S and W_e are all learnable variables, and the loss function is defined as

$$Loss = \frac{1}{2}\|D\alpha - y\|^2 + \lambda\|\alpha\|_1. \tag{18}$$

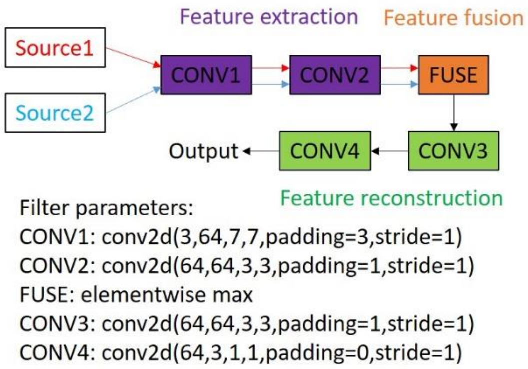
If the number of iteration is fixed, the layers of the network is also fixed. According to the observation y, dictionary D, and parameter λ, the network can be optimized by back-propagation.

Combine Eqs. (18), and (6)-(9), the fused image can be generated. The dataset can be generated and collected through this improved three-layer decomposition based method. Due to the acceleration, the construction of the dataset is feasible. According to our evaluation, the 1000×1000 high frequency image fusion from the original method is about 258 seconds and the LISTA method is about 2.73 seconds. The improvement is considerable.

3.2 CNN fusion model

When the dataset was collected from the LISTA fusion approach, the next step is to learn the CNN fusion model. The IFCNN (Zhang, 2020) model was composed with four convolution layers and one fusion layer. It is not only simple but also effective and suitable to our application. The model can be divided into three parts, feature extraction, feature fusion, and feature reconstruction. The feature extraction was composed with two convolution layers. The parameters in the first convolution layer were fixed and those were from the first layer of ResNet101 trained from ImageNet. This procedure can ensure the model extract the fine features and save the training time. The second convolution layer is learned from random initial. In the fusion layer, we choose the maximum element of the input features. This operation is flexible to different size or number of input features. In the feature reconstruction layer, two convolution layers were used to reconstruct the fusion image from the fused features. The architecture and setting of parameters are shown in Figure 3.

Due to the fixed first convolution layer, the inputs of the model are the three-channel images. In PAN satellite images, we decompose each input image into three-layer decompositions, high frequency, structure, and texture channels, from Eqs. (1)-(4). These three channels represent the three important features of each input image. The concept of the fusion model is to replace the fusion procedure from Eqs. (5)-(9) by learning approach. When doing fusion, the inference time is fewer than the LISTA fusion approach. For a 12000×12000 PAN image fusion under the RTX 2080 graphic card, the processing time of the CNN model is under 2 minutes and the LISTA method is under 7.5 minutes.



Filter parameters:
CONV1: conv2d(3,64,7,7,padding=3,stride=1)
CONV2: conv2d(64,64,3,3,padding=1,stride=1)
FUSE: elementwise max
CONV3: conv2d(64,64,3,3,padding=1,stride=1)
CONV4: conv2d(64,3,1,1,padding=0,stride=1)

Figure 3. The architecture and parameters of the CNN model

4. EXPERIMENTS

In the experiments, we first analyse the role of parameter λ in Eq. (18). Then, we compare the fusion results from the original three-layer decomposition fusion, LISTA fusion, and CNN fusion approaches. The cost of computation time and features of each approaches are also in comparisons.

4.1 Parameter analysis

The parameter λ in Eq. (18) is determined from the given input. If λ is larger, the coefficient is sparser, and the reconstructed image will be smoother. On the other hand, if λ is smaller, the coefficient is more dense, there are more details in the reconstructed image but accompanied by the noise. This is a trade-off by choosing the value of λ. Figure 4 is the simulation with different values of λ for the same input image.

If we know the noise level of the input image, we can set the value of λ systematically. If the noise level is high, the setting of λ is large. Otherwise, if the noise level is low, the setting of λ is small. By experiments, we observe the total variation ratio of the two inputs and then deduce the decision boundary of λ. From the ration, we define four noise levels within different number of iteration, light, medium, heavy, and super heavy, respectively. The number of iteration means the number of layers in LISTA network. These settings are detailed in Table 1.

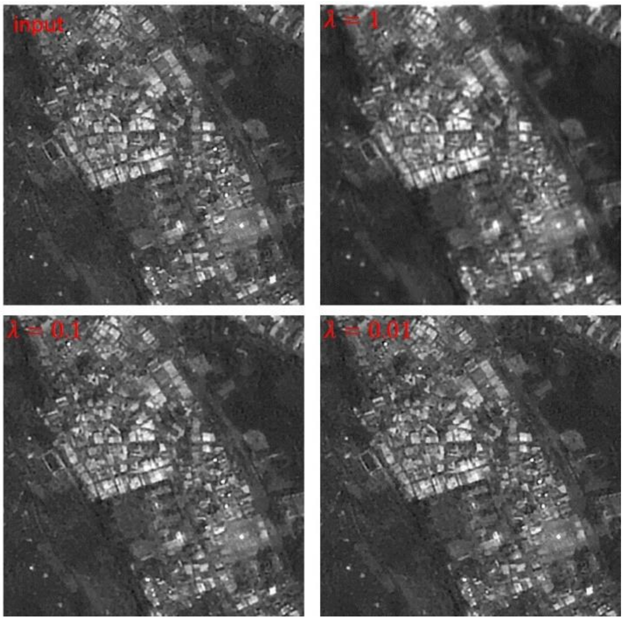


Figure 4. Results with different values of parameter λ

Table 1. Settings of parameter λ

number of iteration = 5, c=50		
ratio	noise level	λ value
1≤ratio<1.3	light	$\lambda = 0.07$
1.3≤ratio<1.6	medium	$\lambda = 0.1$
1.6≤ratio<2	heavy	$\lambda = 1$
ratio≥2	super heavy	$\lambda = 10$
number of iteration = 10, c=50		
ratio	noise level	λ value
1≤ratio<1.3	light	$\lambda = 0.05$
1.3≤ratio<1.6	medium	$\lambda = 0.1$
1.6≤ratio<2	heavy	$\lambda = 1$
ratio≥2	super heavy	$\lambda = 10$
number of iteration = 15, c=50		
ratio	noise level	λ value
1≤ratio<1.3	light	$\lambda = 0.01$
1.3≤ratio<1.6	medium	$\lambda = 0.05$
1.6≤ratio<2	heavy	$\lambda = 1$
ratio≥2	super heavy	$\lambda = 10$

4.2 Computation time

We compare the time cost of these three fusions, original three-layer decomposition fusion (Li, 2021), proposed LISTA fusion, and CNN fusion. The LISTA fusion and CNN fusion are implemented in Python and can be accelerated in GPU. First, we compare the processing time of each part in the three-layer based methods. Table 2 shows the computation time at each step under the NVIDIA GeForce GTX 1060 graphic card. The table shows that the high frequency image fusion in (Li, 2021) is time consuming compared to other parts in the algorithm. The time cost is about 26 times to the sum of other parts with the image size of 1000×1000, 500×500, and 300×300. However, the time cost of high frequency image fusion in LISTA approach is lower.

Table 3 shows the comparisons of LISTA fusion and CNN fusion under GTX 1060 and RTX 2080 graphic cards. These two methods are in different types of fusion. The model of CNN fusion is learned from the results of LISTA fusion. The advantage of CNN fusion can be observed obviously. For a 12000×12000 PAN satellite image, the inference time of CNN fusion is less than two minutes with RTX 2080 graphic card. If higher level graphic card, the inference time will become lower. Table 4 shows the computation time for 12000×12000 high frequency image fusion in LISTA fusion method with different number of iteration under RTX 2080 graphic card. The time cost becomes higher when the number of iteration becomes larger.

Table 2. Computation time under the NVIDIA GeForce GTX 1060 graphic card

image size	1000*1000	500*500	300*300
noise level estimation	0.0190s	0.0068s	0.0028s
high-low decomposition	0.1132s	0.0418s	0.0129s
structure-texture	4.0227s	1.0273s	0.2182s
fuse high	258.8690s	62.3386s	19.4272s
fuse structure	0.0065s	0.0028s	0.0013s
fuse texture	5.4545s	1.3913s	0.4916s
high vs. sum of others	About 26 times	About 25 times	About 26 times
LISTA: fuse high	2.73s	0.84773s	0.28225s

Table 3. Comparisons of LISTA fusion and CNN fusion

PAN 12000*12000	GTX 1060	RTX 2080
LISTA fusion		
3 layer decompositions	642.87s (CPU)	77.32s
fuse high	378.68s	257.25s (iteration=10)
fuse structure	1.86s	1.70s
fuse texture	42.422s	42.80s
others	21.96s	5.65s
total	1069.8s	384.15s
CNN		
total	691.62s	105.67s

Table 4. Comparisons of different number of iteration

RTX 2080 in fuse high with PAN 12000*12000		
number of iteration=5	number of iteration =10	number of iteration =15
224.49s	257.25s	308.48s

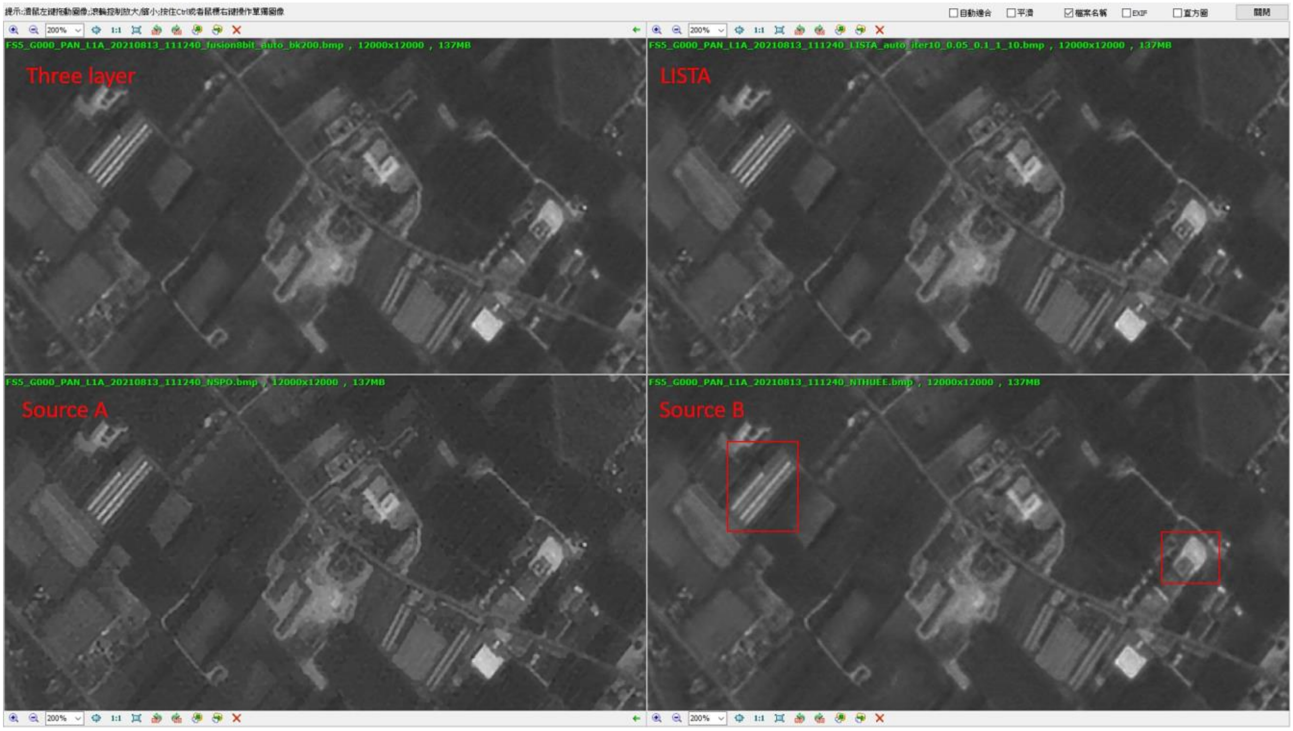
4.3 Fusion results

In the fusion simulations, we do fusion of two input source images. Source A has more edge and texture information but with noise and Source B is smooth and nearly noiseless. The unrolling layer of LISTA fusion is set as 15, and the patch size is 8×8. The training data for CNN fusion is the result from the LISTA fusion. The fusion results from three-layer decomposition fusion and LISTA fusion are shown in Figure 5, and the results from LISTA fusion and CNN fusion are shown in Figure 6. All the three fusion results preserve the edges and textures with little noise or nearly noise-free. There are little differences can be distinguished.

Table 5 makes a comparisons of the features from each fusion approach including the time cost, sparse denoising, unrolling iteration, learning based method, supervised learning, and GPU acceleration. Due to the source code of the three-layer decomposition fusion (Li, 2021) is Matlab, the time cost is higher. The LISTA fusion and CNN fusion are flexible with Python and can be accelerated with GPU.

5. CONCLUSIONS

In order to solve the trade-off between the sharpness and noise, the fusion of images is a solution to the problem. In this paper, we proposed an improvement of the three-layer decomposition fusion by using the unrolling approach which is called LISTA fusion. The goal of LISTA fusion is to accelerate the three-layer decomposition fusion and to generate the dataset for learning. The CNN fusion is also proposed to do fusion in a faster way by training the results from LISTA fusion. Both the proposed fusion methods have high quality results. The time cost of the CNN fusion is also acceptable in real applications.



提示:滑鼠左鍵拖動圖像;滾輪控制放大/縮小;按住Ctrl或者鼠標右鍵操作單獨圖像

□自動適合 □平滑 ☑檔案名稱 □EXIF □直方圖 關閉

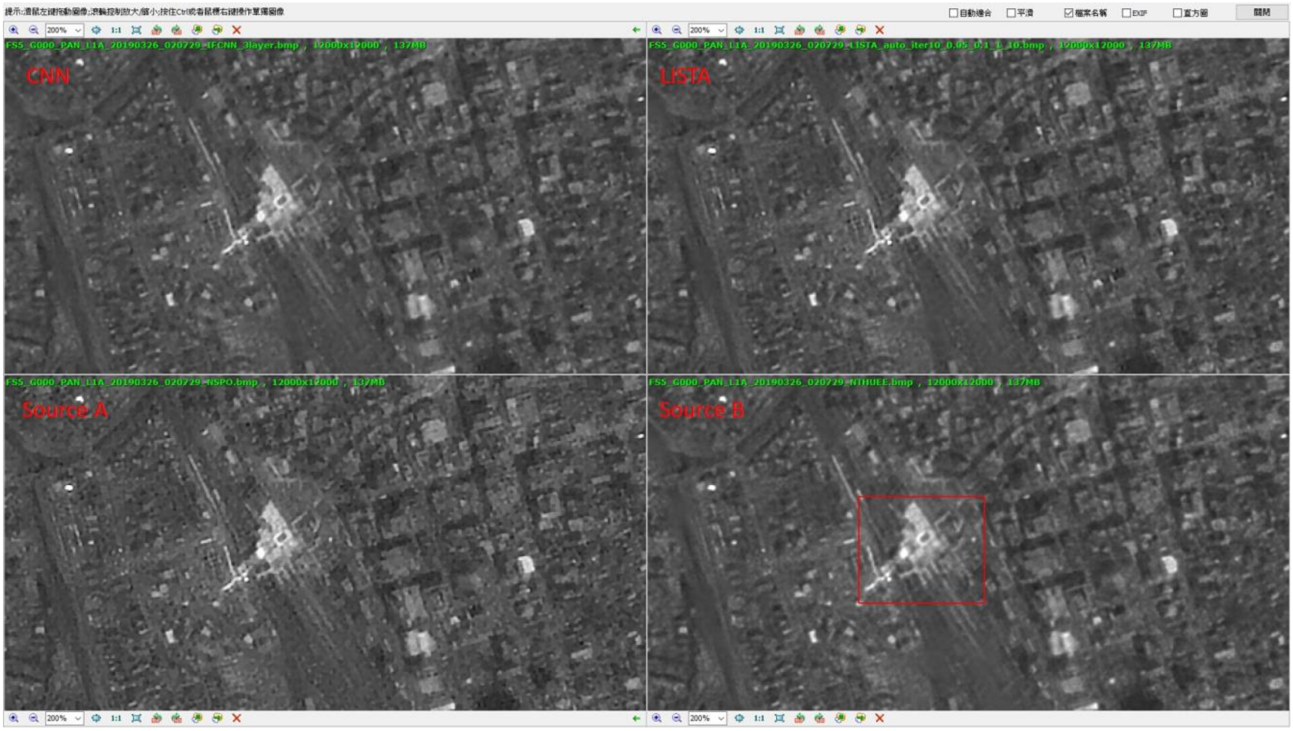
FS5_G000_PAN_L1A_20210813_111240_fusion8bit_auto_bk200.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20210813_111240_LISTA_auto_iter10_0.05_0.1_1_10.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20210813_111240_NSPO.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20210813_111240_NTHUEE.bmp , 12000x12000 , 137MB

Figure 5. Results from three-layer decomposition fusion and LISTA fusion



提示:滑鼠左鍵拖動圖像;滾輪控制放大/縮小;按住Ctrl或者鼠標右鍵操作單獨圖像

□自動適合 □平滑 ☑檔案名稱 □EXIF □直方圖 關閉

FS5_G000_PAN_L1A_20190326_020729_IFCNN_3layer.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20190326_020729_LISTA_auto_iter10_0.05_0.1_1_10.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20190326_020729_NSPO.bmp , 12000x12000 , 137MB

FS5_G000_PAN_L1A_20190326_020729_NTHUEE.bmp , 12000x12000 , 137MB

Figure 6. Results from CNN fusion and LISTA fusion

Table 5. Comparisons of each fusion method

method	three-layer	LISTA fusion	CNN fusion
time cost	About 5~8 hr	Under 7.5 min	Under 2 min
sparse denoising	Yes	Yes	No
LISTA based	No	Yes	No
learning based	No	Yes	Yes
supervised	No	No	Yes
Python	No	Yes	Yes

6. REFERENCES

Gregor, K., and LeCun, Y., 2010. Learning Fast Approximations of Sparse Coding. ICML.

Li, X., Zhou, F., and Tan, H.,2021. Joint image fusion and denoising via three-layer decomposition and sparse representation. *Knowledge-Based System,* 224, pp. 107087.

Pati, Y., Rezaiifar, R., and Krishnaprasad, P., 1995. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, 1, pp. 1–3.

Uezato, T., Hong, D., Yokoya, N., and He, W., 2020. Guided deep decoder: unsupervised image pair fusion. ECCV.

Ulyanov, D., Vedaldi, A., and Lempitsky, V., 2018. Deep Image Prior. CVPR.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., and Zhang, L., 2020. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion,* 54, pp. 99-118.