MULTI-TEMPORAL SAR IMAGE DESPECKLING BASED A CONVOLUTIONAL NEURAL NETWORK

Chenxia Zhou[1], Jie Li[2, *], Huanfeng Shen[1, 3], Qiangqiang Yuan[2, 3]

[1] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China @whu.edu.cn
[2] School of Geodesy and Geomatics, Wuhan University, Wuhan, China @whu.edu.cn
[3] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China @whu.edu.cn

Youth Forum

KEY WORDS: Multi-temporal SAR Despeckling, Speckle, Convolutional Neural Network, Spatio-temporal Information

ABSTRACT:

Speckle noise is an intrinsic property of Synthetic Aperture Radar (SAR) imagery, which affects the quality of image. Single-temporal despeckling methods usually pay attention to the utilization of spatial information, but sometimes due to lack of sufficient information, the despeckling image is too smooth or losses some information about edge details. However, multi-temporal SAR images can provide extra information for despeckling resulting in better performance. Therefore, in this paper, we proposed a novel multi-temporal SAR despeckling method based a convolutional neural network (MSAR-CNN) embedded temporal and spatial attention (TSA) module to deeply mine the spatial and temporal correlation of multitemporal SAR images. The whole network, which is end-to-end trained with simulate realistic SAR data, consists of several residual blocks. In addition, the simulated and real-data experiments demonstrate that the proposed MSAR-CNN outperforms most of the mainstream methods in both the quantitative evaluation indexes and visual effects.

1. INTRODUCTION

Synthetic aperture radar (SAR) can be capable of all-time and all-weather observation, which provides conditions for obtaining long time-series images of the same area. Thus, the application of multi-temporal SAR images is emerging with the launch of more SAR satellites, such as forest and disaster monitoring (Rauste et al., 2005; and Bovolo et al, 2007), land-cover classification (Dobson et al., 1995), and glaciers and snow analysis (Fallourd et al., 2011; and Nagler et al., 2000). However, speckle is generated by the coherent processing of radar signals, which affects the SAR images of scene interpretation and automatic analysis. Therefore, before SAR images are applied, the speckle suppression operation should be done.

In the past few decades, most SAR despeckling methods focus on utilizing the redundancy of neighbouring or nonlocal spatial information on a single temporal image. Although these methods keep a balance between speckle reduction and spatial resolution degradation, sometimes the lack of sufficient similar spatial information leads to poor robustness. However, multi-temporal images provide additional time dimensional information to supplement spatial information. At first, the multi-temporal despeckling methods only process images in time dimension like unbiased temporal average filter (UTA) (Lee et al., 1991). Up to now, part of the spatial or temporal dimension denoising methods are extended to the spatio-temporal joint dimension. The following three are typical. Firstly, three-dimensional adaptive neighbourhood filter (3D-ANF) is a classical spatio-temporal filter, which determines the spatio-temporal adaptive neighbourhoods by statistic information in the local 3D patch of the center pixel (Ciuc et al., 2001). Two-step multi-temporal nonlocal mean method (2S-PPB) consists of a temporal averaging step (the first step), and a spatial denoising step (the second step) (Su et al., 2014). Lastly, multi-temporal SAR block-matching in 3D (MSAR-BM3D) expands spatial grouping into spatio-temporal grouping, as well as four-dimensional collaborative filtering (Chierchia et al., 2017b). Recently, some other novel methods have been proposed, such as ratio-based SAR despeckling method (RABASAR) (Zhao et al., 2019) and a scattering covariance matrix of image patch for multi-temporal SAR image despeckling (SCM-MSAR) (Ma et al. 2019). These methods combine spatio-temporal information to present a more effective despeckling result than single-temporal methods in spatial resolution preservation. But if the significant change of multi-temporal SAR images exists, these methods may introduce the error information to the despeckling results.

Recently, deep convolutional neural network (CNN) has performed well in SAR despeckling domain (Chierchia et al., 2017a; Wang et al., 2017; and Zhang et al., 2018). Compared to traditional methods, the deep learning based SAR despeckling methods can fit the non-linear relationship more accurately between the speckle image and noise-free image because of the deep structure. But these methods are limited to single-temporal SAR despeckling, and the redundant information among multi-temporal SAR images is not exploited. In this paper, we aim to combine spatio-temporal information with deep CNN to get higher spatial resolution multi-temporal SAR images. Therefore, we proposed a combining spatio-temporal residual network for multi-temporal SAR image despeckling. The model consists of several residual blocks (He et al., 2016) embedding a fusion module known as temporal and spatial attention (TSA). TSA (Wang et al., 2019) is an important module, which consists of temporal attention and spatial attention, and helps aggregate information across the features of each time image. Firstly, the temporal attention aims at computing the element-wise correlation between the target time image and each time image in feature level. Then each temporal feature is weighed by the normalized correlation coefficient at each location by element-wise product. A convolutional layer is used for fusing the convolved weighted features from all times. On the basis of the temporal fusion, spatial attention is applied to adaptively rescale

*Corresponding author

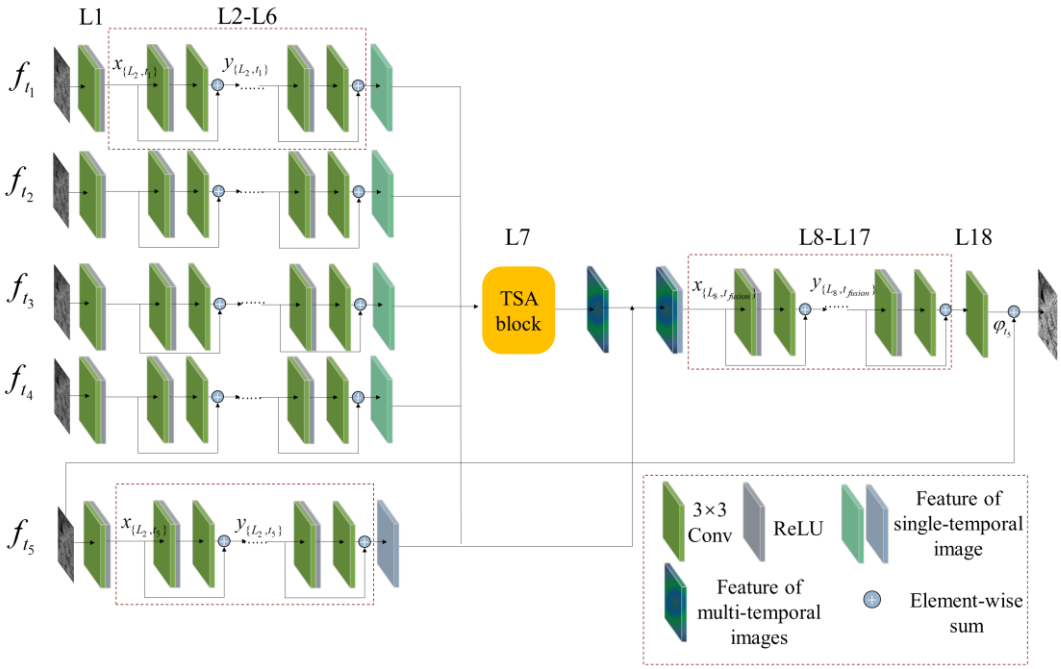


Fig. 1 the overall Frameworks of MSAR-CNN

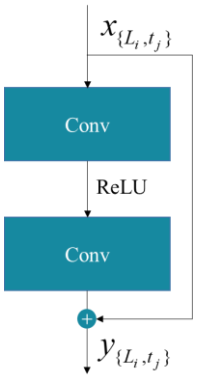


Fig. 2 Residual Block.

the feature at each location in each channel to deeply mine the cross-channel and spatial feature.

The remainder of this paper is organized as follows. Section 2 describes our proposed MSAR-CNN model. Experimental results and some relevant discussions are demonstrated in Section 3. The conclusions are finally summarized in Section 4.

2. PROPOSED METHOD

2.1 Framework of MSAR-CNN

The overall framework of the proposed MSAR-CNN is shown in Fig.1. Given five different multi-temporal SAR images t_{1-5} as inputs, we denote the last image t_5 as the target image and the others as assistant images. Firstly, the feature of the target image and assistant images are respectively extracted by five parallel structures which consist of a convolutional layer along with rectified linear unit (ReLU) activation function and five residual blocks. And then TSA is used to fusion the spatio-temporal information. Ten series residual blocks with a convolution in the last act as reconstruction layers on the concatenated features of fusion and target. Lastly, the despeckling target SAR image is generated by the sum of the output residual and the input target image. The detailed configuration of the proposed MSAR-CNN is provided in Table 1.

	Layers	Filter size	Filters
L1	Conv+ReLU	1×3×3×5	32
L2-L6	Residual blocks	32×3×3×5	32
L7	Temporal and spatial attention block	32×3×3×5	32
L8-L17	Residual blocks	64×3×3×1	64
L18	Conv	64×3×3×1	1

Table 1. Detailed configurations of the proposed MSAR-CNN

2.2 Residual Learning

The residual learning is an effective strategy to improve the performance of the network and speed up the training when the network is deeper. The key point is the shortcut connection which makes new features easier to extract on the stack layers. Here, we introduce two different residual learning strategies respectively in image level and feature level.

Therefore, to overcome the difficulty of the common deep network in approximating identical mappings by stacked flat structures, we consider restoring the residual speckle noise image

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-5-2020-101-2020 | © Authors 2020. CC BY 4.0 License.

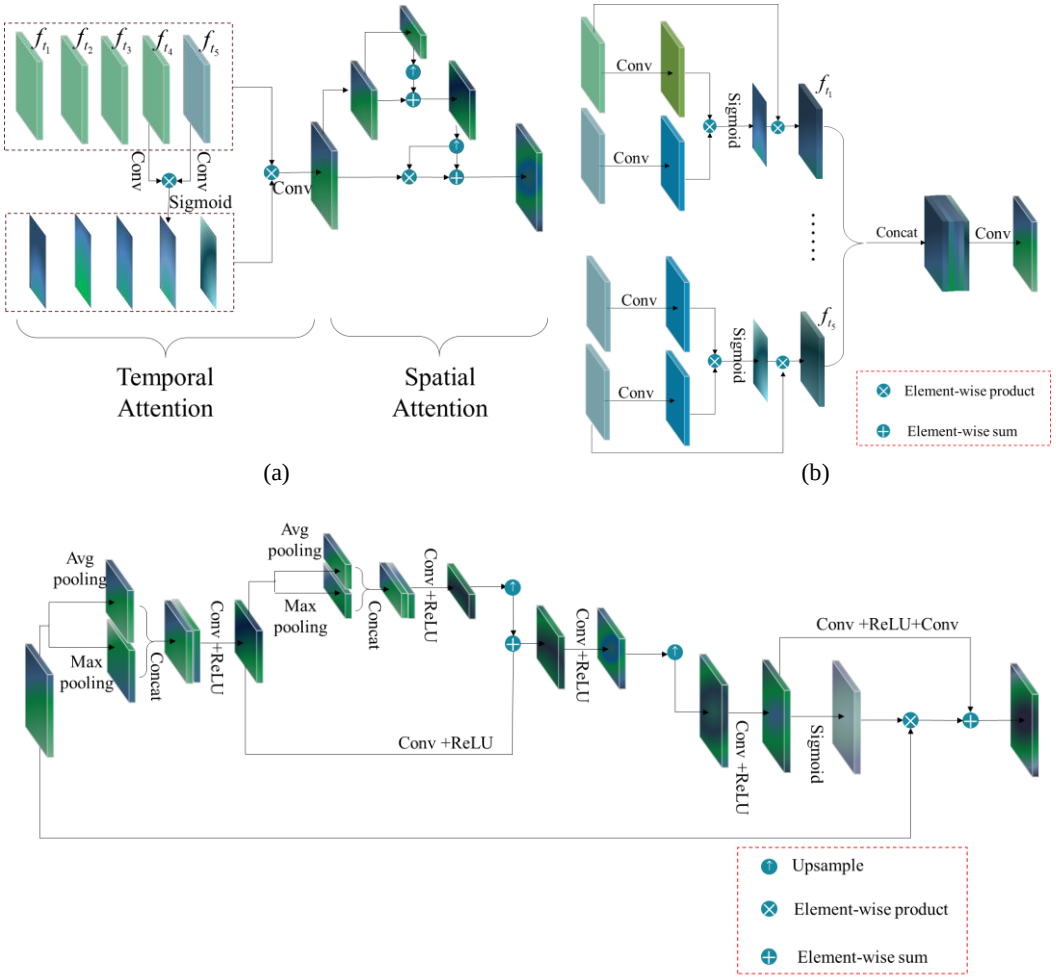


Fig. 3 the Frameworks of Temporal and Spatial Attention (TSA). (a) the overall structure for TSA (b) Temporal Attention. (c) Spatial Attention

by skip connection in image level. As we all know speckle noise is described by the multiplicative noise model

$$f = uv \tag{1}$$

where f and u are, respectively, the contaminated image and clean image, the speckle noise v is assumed to be statistically independent with $E(v)=1$ and stationary variance δ_v^2. However, the multiplicative noise can be translated into the flowing additional equation

$$f = u + (v-1)u \tag{2}$$

Therefore, the residual speckle noise φ is defined as

$$\varphi = u - f \tag{3}$$

where φ is also donated as $(1-v)u$, which is the additional single-dependent noise with zero-mean and nonstationary variance related to u.

For the proposed model, given multi-temporal data training pairs $\{\hat{u}_{t_5}, f_t, u_{t_5}\}_N$, f_t is the input five temporal speckle images, u_{t_5} represents the clean image of the target time phase as the label, $\hat{u}_{t_5}$ is the corresponding despeckling image. and N donates the number of data pairs. The output residual speckle noise is defined as

$$\varphi_{t_5} = u_{t_5} - f_{t_5} \tag{4}$$

The mean-square error is set as loss function, formulated as

$$loss(\Theta) = \frac{1}{2N}\sum_{i=1}^{N}\left\| Net(f_t^{\,i}, \Theta) - \varphi_{t_5}^{\,i} \right\|_2^2 \tag{5}$$

where Θ is the parameters of the network.

Furthermore, to be better utilizing and mining the character of different temporal SAR images and avoiding the vanishing gradient problem, the basic structures of the proposed MSAR-CNN are the residual blocks as shown in Fig. 1 respectively stacked in the layers 2-6 and layers 8-17. Fig. 2 shows the building block of residual learning, which is defined as

$$y_{\{L_i,t_j\}} = \Re(x_{\{L_i,t_j\}}, \{w,b\}) + x_{\{L_i,t_j\}} \tag{6}$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-5-2020-101-2020 | © Authors 2020. CC BY 4.0 License.

where $x_{\{L_i,t_j\}}$ and $y_{\{L_i,t_j\}}$ represent the input and output features of the residual block, when $L_i = L_2,\cdots,L_6$, $t_j = t_1,\cdots,t_5$ and $L_i = L_8,\cdots L_{17}$, $t_j = t_{fuison}$, $\Re$ is the function consisted of two convolutions embedded one ReLU, $\{w,b\}$ are the parameters of the two convolutions.

2.3 Spatio-temporal information fusion

As mentioned in section 1, spatio-temporal redundant information can effectively improve the SAR image despeckling performance owning to the high correlation and similarity in different temporal images. The spatial relation and temporal relation are critical in fusion since they directly determine the performance of multi-temporal despeckling algorithms. Because SAR is sensitive to geometric structures, the difference between two short period images is increased and the non-linear relation of multi-temporal SAR images is more complex. However, the existing traditional methods have a limitation on fitting the more complex non-linear relationship resulting in the spatial resolution reduction or detail loss. Therefore, in the proposed MSAR-CNN model, the temporal and spatial attention (TSA) module (Wang et al., 2019) are used for more accurately fusing spatio-temporal information shown in Fig. 3(a) which is embedded in the middle of the whole network to calculate the temporal relation and spatial relation in feature level rather than image level. As can be seen in Fig. 3(a), temporal relation is firstly calculated by temporal attention, and then spatial relation is found by spatial attention based on the fused features in the temporal dimension.

The temporal attention block aims to compute the similarity between each temporal image and target temporal image in the feature level in Fig. 3(b). Thus, for each temporal image $t \in [1:5]$, the element-wise correlation is calculated by the element-wise product between the features of each image and target image. Then, the element-wise correlation is normalized by sigmoid function to compute the similarity distance h, which is formulated as

$$h(F_t, F_5) = Sigmoid(\theta_t(F_t)^T \cdot \theta_5(F_5)) \quad (7)$$

where F_5 is the feature of the target SAR image, F_t represents the features of each temporal SAR image, θ_t and θ_5 respectively equal the parameter of embedded convolutions.

Secondly, the similarity distance h is multiplied in a pixel-wise manner to the original feature F_t. Lastly, a convolutional operation is used to fusion all attention-modulated features $\tilde{F}_t$.

$$\tilde{F}_t = h(F_t, F_5) \cdot F_t \quad (8)$$

$$F_{\text{fusion}} = Conv([\tilde{F}_1, \tilde{F}_2, \cdots, \tilde{F}_5) \quad (9)$$

where F_{fusion} denotes the fused features.

In temporal attention, the fused features are got by a pixel-wise manner in temporal scale regardless of the spatial scale. The goal of spatial attention is to correct weights at each location in each channel features. Thus, a three-level pyramid structure is employed to extend the attention receptive field by pooling (mean pooling and maximum pooling) and convolution shown in Fig. 3(b). Then, the spatial-attention-modulated feature is upsampled to element-wise add to the original features for fusing spatial features (Wang et al., 2018).

3. EXPERIMENTAL RESULTS

In this section, we present the simulated and real experimental results of the proposed MSAR-CNN model to verify the effectiveness. We compare the performance of our model with four different multi-temporal despeckling methods: UTA (Lee et al., 1991), nonlocal temporal filter (NLTF) (Chierchia et al., 2017b), MSAR-BM3D (Chierchia et al., 2017b), and RABASAR (Zhao et al., 2019). For all the compared method, the parameters are set as suggested in the referenced paper. Five different temporal images are used except MSAR-BM3D which uses four different temporal images because the time series must be equal to a power of two.

3.1 Training data and parameters setting

Most deep learning based SAR despeckling method used the optical image as the training set, but the differences do exist between the two data even though the optical image is transformed to SAR amplitude image. Therefore, turning to the MSAR-CNN, the images in training dataset as label are calculated by the arithmetic mean of the long-time series SAR images. Here, we select 50 images (size of 8000 × 8000) stride by 10 from 100 Sentinel-1 amplitude images of the city Wuhan in china to produce five temporal noise-free SAR images. Then the five temporal images are concatenated and cropped to 400 images size of 400 × 400 × 5 for the label of the training set. The training label set is divided into four parts of 100 images each, respectively multiplying different strength gamma-distributed speckle noise with the equivalent number of looks (ENL) of 1, 2, 4, and 8.

Then, these training data are then cropped in each patch size as 40×40, with the stride equal to 10, with the ADAM gradient-based optimization method (Kingma et al., 2014), mini-batches of 64 patches. Training proceeds for 100 epochs with initial learning rate 0.001, and after 20 epochs, the learning rate is reduced through being multiplied by a descending factor of gamma = 0.1. We implement the different models in the PyTorch framework and train the models with an NVIDIA Quadro P4000 GPU.

Looks	1		2	
Index	PSNR	EPD-ROA	PSNR	EPD-ROA
UTA	52.340	1.039	54.240	1.012
NLTF	49.424	1.123	52.310	1.034
MSAR-BM3D	37.200	1.334	46.430	1.168
RABASAR	**56.896**	**0.974**	**57.207**	0.977
MSAR-CNN	55.076	1.046	54.903	**1.005**

Table 2 Average quantitative assessment result of test dataset (16 simulated realistic SAR images) with single and two looks

3.2 Simulated experiments

The simulated test SAR data were produced in the same way as the training data. We randomly selected a testing set of 16 (size of 500 × 500 × 5) Sentinel-1 images of the city Wuhan differing from the training data. The pick signal to noise ratio (PSNR, as higher as possible) and the edge-preservation degree based on the ratio of average (EPD-ROA, as closer to 1 as possible) are used

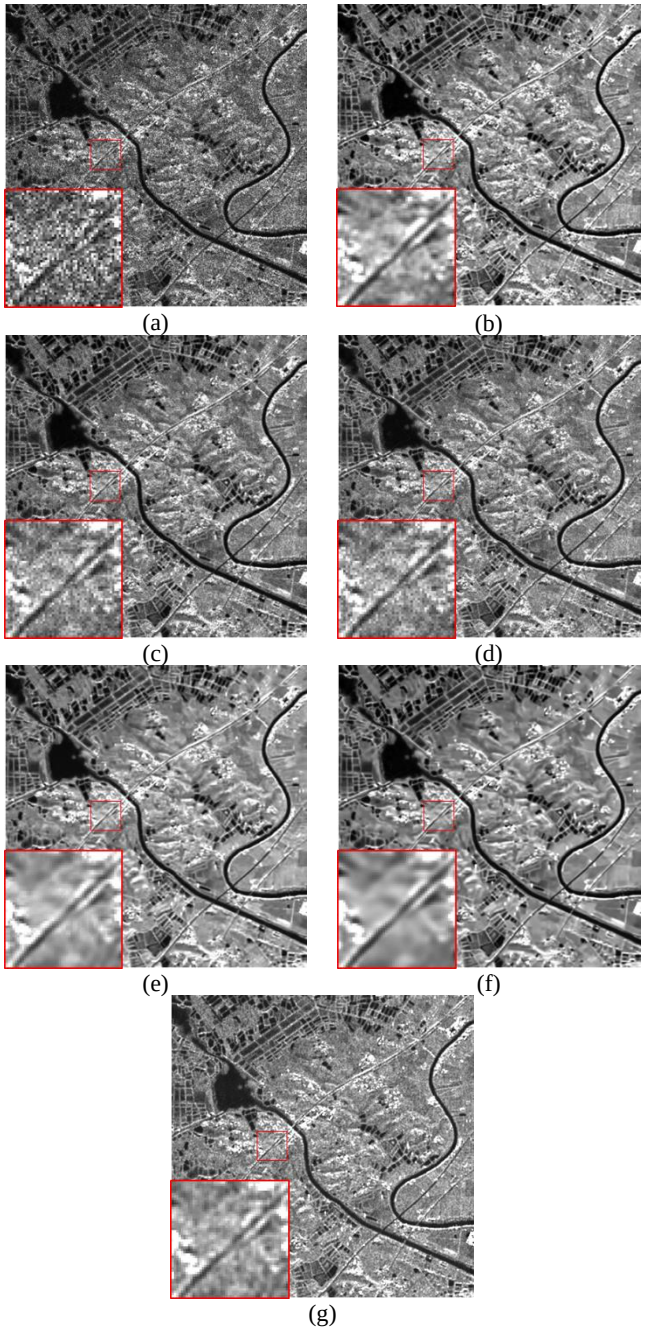
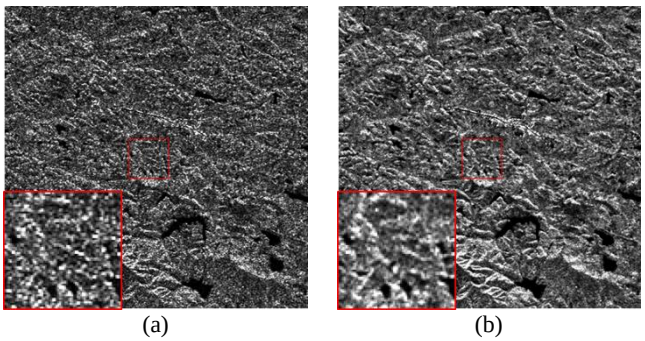


Fig. 4 Despeckled images for the simulate realistic image. (a) Noisy image. (b) Truth image. (c) UTA. (d) NLTF. (e) MSAR-BM3D. (f) RABASAR. (g) the proposed MSAR-CNN



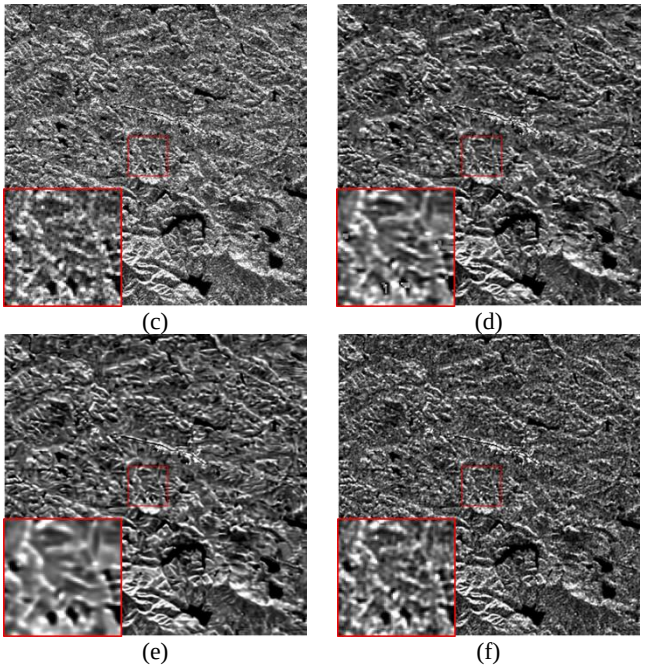


Fig. 5 Despeckled images for the Sentinel SLC image. (a) Noisy image. (b) UTA. (c) NLTF. (d) MSAR-BM3D. (e) RABASAR. (f) the proposed MSAR-CNN.

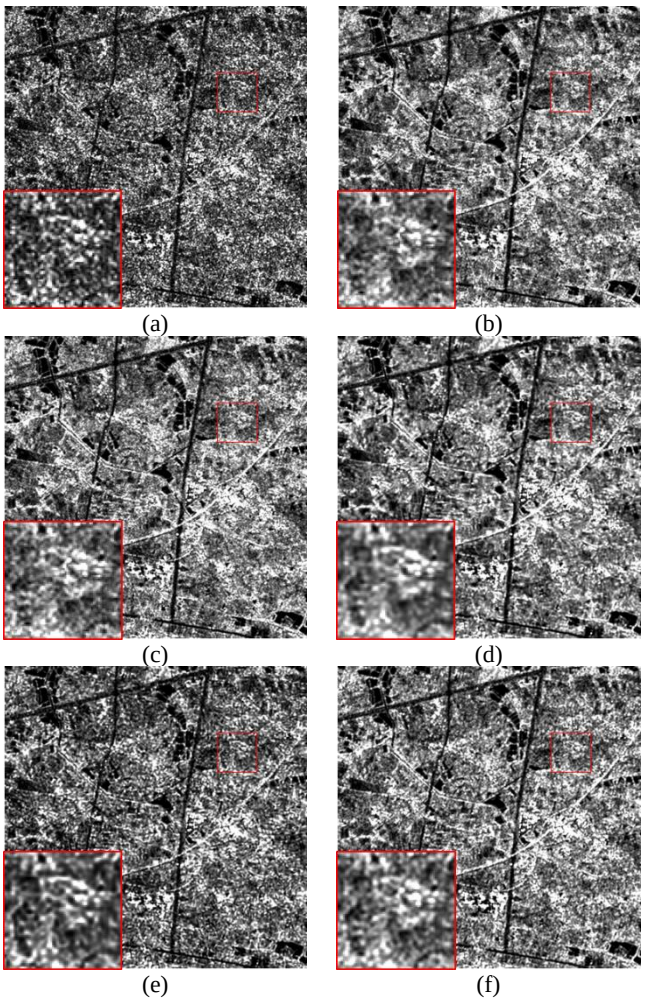


Fig. 6 Despeckled images for the Sentinel-1 GRD-HR image. (a) Noisy image. (b) UTA. (c) NLTF. (d) MSAR-BM3D. (e) RABASAR. (f) the proposed MSAR-CNN

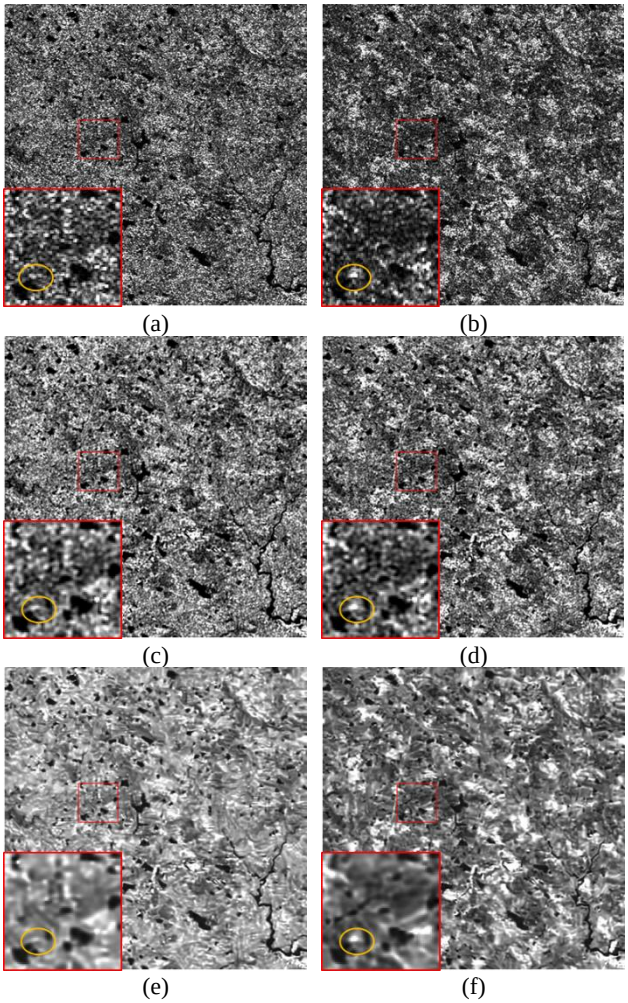


Fig. 7 The despeckling results of two different temporal Sentinel-1 SLC images. (a) Noisy image of September 17, 2019. (b) Noisy image of October 11, 2019. (c) and (d) the proposed MSAR-CNN despeckling results of (a) and (b). (e) and (f) RABASAR despeckling results of (a) and (b).

to verify the performance of the proposed MSAR-CNN. PSNR and EPD-ROA are formulated as

$$PSNR(x,y) = 10\log_{10}\frac{\max(y)^2}{MSE(x,y)} \tag{10}$$

$$EPD-ROA = \frac{\sum_{i}^{m} |I_{D1}(i)/I_{D2}(i)|}{\sum_{i}^{m} |I_{O1}(i)/I_{O2}(i)|} \tag{11}$$

where x and y are, respectively, the despeckled image and the reference image; i is the index set of the SAR image, $I_{D1}(i)$ and $I_{D2}(i)$ respectively represent the adjacent pixel values in the horizontal and vertical directions of the despeckled image, and $I_{O1}(i)$ and $I_{O2}(i)$ represent the adjacent pixel values in the horizontal and vertical directions of the reference clean image.

Table 2 lists the average quantitative evaluation results for the test dataset with single and two looks, with the best performance marked in bold and the second-best underlined. Furthermore, for comprehensive evaluation, the despeckling results of a simulated image with 2 looks are shown in Fig. 4. The visual result of MSAR-BM3D is not worse than other traditional methods, while the quantitative assessment is the worst. It may be related to the artifacts shown in part despeckling results of MSAR-BM3D, and it is also verified in the real experiments. For the UTA and NLTF results, residual speckle is the main problem. Although The best quantitative assessment is got by RABASAR, the despeckling image is over-smoothing. However, the details are important for the subsequent application and analysis of SAR image. The better edge preservation result can be got by the proposed MSAR-CNN, which is consistent with the quantitative assessment. Therefore, on the whole, the proposed MSAR-CNN result provides the most similar performance with the truth, even though some residual noise may exist.

3.3 Real experiments

For the real experiments, to present the comprehensive comparison, different noise strength SAR images are selected. There are, respectively, Sentinel-1 single look complex (SLC, ENL = 1) and ground range detected high resolution (GRD-HR, ENL = 4.4) images of the city of Wuhan, cropped to 500 × 500, differing from the training dataset.

Here, the SLC images in a mountain area of August 24, September 5 and 29, October 11 in 2019 are used as auxiliary images to despeckle the SLC image of September 17. Fig. 5 presents the results of different multitemporal despeckling methods. The obvious residual speckle is apparent with UTA and NLTF. The performances of MSAR-BM3D and RABASAR are better. Over-smoothing is still existing in the results of MSAR-BM3D and RABASAR, while the RABASAR shows lesser details than others. Compared to other traditional methods, the proposed MSAR-CNN method provides a satisfying denoising result since it leads to a good balance between noise reduction and spatial resolution degradation especially preservation of point.

For GRD-HR data of Sentinel-1, we select the flat area of the images, where the auxiliary images are dated on October 21, November 2 and 26, and December 8, 2017, and the target image is dated November 14, 2017. In Fig. 6, RABASAR gives the best result of the traditional methods especially lying in the balance despeckling performance. UTA and NLTF are lacking in the preservation of point-like targets. And MSAR-BM3D loss some detail like edge and texture. Relatively speaking, the proposed MSAR-CNN retains more details than RABASAR and MSAR-BM3D, and both the retention of original information and the removal of noise perform well. To sum up, the adaptive despeckling ability of the proposed MSAR-CNN method is the best.

3.4 Temporal information preservation

For verifying the ability of temporal information preservation, we select the five temporal Sentinel-1 SLC images of August 24, September 5, 17, and 29, October 11 in 2019 as input. And then the despeckling results of September 5 and October 11 are outputted, shown in Fig. 7.

In Figs. 7(a) and 7(b), the two temporal images are very different because of the changing of geotexture and radiation along with time. Generally, both RABASR and the proposed MSAR-CNN method can effectively handle the changes due to the higher similarity between the despeckling results and the original noise images. However, from the zoomed images in red rectangular Fig. 7, especially the small building shown in yellow circle, it can be

observed that the results of the proposed MSAR-CNN method are more similar with the original target temporal information than RABASAR. Therefore, the proposed MSAR-CNN effectively fuses the temporal and spatial information.

4. CONCLUSIONS

In this paper, we proposed a new multi-temporal despeckling method based a convolutional neural network. Since the whole network was trained with arithmetic mean SAR image, it generated a reasonable despeckling results. In addition, owning to the utilization of residual learning strategy and TSA module, the visual results of the proposed showed the better balanced performance on detail preservation and speckle reduction compared to other traditional methods. The future work will be devoted to introduce a recursive network architecture and update the training dataset to reduce residual speckle noise and further improve the quantitative assessment.

ACKNOWLEDGEMENTS

Acknowledgements of support in part by the National Natural Science Foundation of China under Grant 61671334 and 41701400.

REFERENCES

Bovolo, F., Bruzzone, L., 2007. A split-based approach to unsupervised change detection in large-size multitemporal images: Application to tsunami-damage assessment. *IEEE Trans. Geosci. Remote Sens.* 45, 1658-1670.

Chierchia, G., Cozzolino, D., Poggi, G., Verdoliva, L., Ieee, 2017a. SAR image despeckling through convolutional neural network. *2017 Igarss*, 5438-5441.

Chierchia, G., El Gheche, M., Scarpa, G., Verdoliva, L., 2017b. Multitemporal SAR Image Despeckling Based on Block-Matching and Collaborative Filtering. *IEEE Trans. Geosci. Remote Sens.* 55, 5467-5480.

Ciuc, M., Bolon, P., Trouve, E., Buzuloiu, V., Rudant, J.P., 2001. Adaptive-neighborhood multitemporal synthetic speckle removal in aperture radar images. *Appl. Opt.* 40, 5954-5966.

Dobson, M.C., Ulaby, F.T., Pierce, L.E., 1995. Land-cover classification and estimation of terrain attributes using synthetic aperture radar. *Remote Sens. Environ.* 51, 199-214.

Fallourd, R., Harant, O., Trouve, E., Nicolas, J.M., *et. al*, 2011. Monitoring Temperate Glacier Displacement by Multi-Temporal TerraSAR-X Images and Continuous GPS Measurements. *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.* 4, 372-386.

He, K., Zhang, X., Ren, S., Sun, J., Ieee, 2016. Deep Residual Learning for Image Recognition, 2016 *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, pp. 770-778.

Kingma, D. P. and Ba, J., 2014. Adam: A method for stochastic optimization. [Online]. Available: https://arxiv.org/abs/1412.6980

Lee, J.S., Grunes, M.R., Mango, S.A., 1991. Speckle reduction in Multipolarization, Multifrequency SAR imagey. *IEEE Trans. Geosci. Remote Sens*. 29, 535-544.

Ma, X. and Wu, P., 2019. Multitemporal SAR Image Despeckling Based on a Scattering Covariance Matrix of Image Patch. *Sensors*, 19(14), p.3057.

Nagler, T., Rott, H., 2000. Retrieval of wet snow by means of multitemporal SAR data. *IEEE Trans. Geosci. Remote Sens.* 38, 754-765.

Rauste, Y., 2005. Multi-temporal JERS SAR data in boreal forest biomass mapping. *Remote Sens. Environ.* 97, 263-275.

Su, X., Deledalle, C.-A., Tupin, F., Sun, H., 2014. Two-Step Multitemporal Nonlocal Means for Synthetic Aperture Radar Images. *IEEE Trans. Geosci. Remote Sens.* 52, 6181-6196.

Wang, P.Y., Zhang, H., Patel, V.M., 2017. SAR Image Despeckling Using a Convolutional Neural Network. *IEEE Signal Process Lett.*, 1763-1767.

Wang, X., Chan, K. C. K., Yu, K., Dong, C., Loy, C. C., 2019. Edvr: video restoration with enhanced deformable convolutional networks [Online]. Available: https://arxiv.org/pdf/1905.02716v1.

Wang, X., Yu, K., Dong, C., Loy, C.C., Ieee, 2018. Recovering Realistic Texture in Image Super-resolution by Deep Spatial Feature Transform, 2018 *In Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, pp. 606-615.

Zhang, Q., Yuan, Q.Q., Li, J., Yang, Z., Ma, X.S., 2018. Learning a Dilated Residual Network for SAR Image Despeckling. *Remote Sens.* 10, 18.

Zhao, W., Deledalle, C.-A., Denis, L., Maitre, H., Nicolas, J.-M., Tupin, F., 2019. Ratio-Based Multitemporal SAR Images Denoising: RABASAR. *IEEE Trans. Geosci. Remote Sens.* 57, 3552-3565.