Journal of
Applied Remote Sensing

RemoteSensing.SPIEDigitalLibrary.org

Learning synthetic aperture radar image despeckling without clean data

Gang Zhang
Zhi Li
Xuewei Li
Yiqiao Xu

SPIE.

Gang Zhang, Zhi Li, Xuewei Li, Yiqiao Xu, “Learning synthetic aperture radar image despeckling without clean data,” *J. Appl. Remote Sens.* **14**(2), 026518 (2020), doi: 10.1117/1.JRS.14.026518

Learning synthetic aperture radar image despeckling without clean data

Gang Zhang,[a,*] Zhi Li,[a] Xuewei Li,[b] and Yiqiao Xu[a]
[a]Space Engineering University, Huairou District, Beijing, China
[b]Beijing University of Posts and Telecommunications, Department of Information and Communication Engineering, Haidian District, Beijing, China

Abstract. Speckle noise can reduce the image quality of synthetic aperture radar (SAR) and make interpretation more difficult. Existing SAR image despeckling convolutional neural networks require quantities of noisy–clean image pairs. However, obtaining clean SAR images is very difficult. Because continuous convolution and pooling operations result in losing many informational details while extracting the deep features of the SAR image, the quality of recovered clean images becomes worse. Therefore, we propose a despeckling network called multiscale dilated residual U-Net (MDRU-Net). The MDRU-Net can be trained directly using noisy–noisy image pairs without clean data. To protect more SAR image details, we design five multiscale dilated convolution modules that extract and fuse multiscale features. Considering that the deep and shallow features are very distinct in fusion, we design different dilation residual skip connections, which make features at the same level have the same convolution operations. Afterward, we present an effective $L_{\_\text{hybrid}}$ loss function that can effectively improve the network stability and suppress artifacts in the predicted clean SAR image. Compared with the state-of-the-art despeckling algorithms, the proposed MDRU-Net achieves a significant improvement in several key metrics. [DOI: 10.1117/1.JRS.14.026518]

Keywords: synthetic aperture radar; image despeckling; dilated convolution; residual skip connection.

Paper 190981 received Dec. 10, 2019; accepted for publication May 19, 2020; published online Jun. 3, 2020.

1 Introduction

Synthetic aperture radar (SAR)[1] is an active Earth observation system deployed on aircraft, satellites, or other flight platforms. Compared with the optical and infrared systems, SAR can provide all-time, all-weather, high-resolution, and wide-swath observation. It also has a certain ability to penetrate the Earth's surface and discover underground targets. Therefore, the SAR has advantages in disaster monitoring,[2] environmental monitoring,[3] ocean surveillance,[4] resource exploration,[5] surveying, and military applications.

However, due to the imaging mechanism of SAR, a large amount of speckle noise exists in the observed SAR images.[6] Speckle noise[7] is a kind of random multiplicative noise, which is formed by the mutual interference of radar echo phase. The speckle noise in SAR images appears as granular noise or black-and-white noise. The speckle noise in single-look SAR images follows a Gaussian distribution with zero mean,[8] while the speckle noise in multilook SAR images follows a gamma distribution with unit mean and variance $1/\sqrt{L}$,[9] and in the $1/\sqrt{L}$, the L is the number of looks. The existence of speckle noise reduces the resolution of SAR image and masks the detailed structure of targets. Because the image details are masked, the accuracy of SAR image classification,[10] segmentation,[11] and change detection[12] is reduced. In addition, speckle noise will also bring great difficulty to the phase unwrapping in SAR interferometry, which will affect the accuracy of interferometry.[13] Therefore, removing speckle noise is a significant research in SAR image field.

[*] Address all correspondence to Gang Zhang, E-mail: gangzhang1989@126.com

To remove the speckle noise from SAR images, many approaches have been proposed. The spatial filters[14–17] are first applied to remove the speckle noise in SAR images, but the edges of the filtered image are smooth. To solve this problem, the spatial filters are improved in two aspects. One is to use different filters for different scenes.[18,19] The other is to design adaptive sliding window filters.[20,21] The transform domain filters (TDFs) mainly include wavelet domain filters[22,23] and post-wavelet domain filters.[24–26] Although the despeckling performance of the TDFs is significantly higher than spatial filters, the complexity of TDF is very high. The filters based on the Markov random field model[27] can remove the speckle noise in the spatial and transform domains, but they require a lot of prior knowledge of SAR images and speckle noise. Owing to simple ideas and superior performance of nonlocal mean (NLM) filters,[28–31] NLM filters have been widely used to reduce speckle noise. But the filtered images will contain artificial textures because of the block effect.

With the development of convolutional neural networks (CNNs), some researchers[32–35] have tried to use CNN to complete image despeckling tasks. However, these CNN methods still have some problems. First, CNN-based despeckling methods require a large number of the noisy–clean image pairs, where clean images are used as labels. But the clean SAR images are difficult to obtain. To construct the noisy–clean image pairs, they[32–35] usually add simulated speckle noise to the optical images. The predicted clean SAR images contain optical interference. In fact, they do not use the real SAR images, and all training data are generated by optical images. This approach cannot be applied to actual work. Second, to preserve more image details, they[33,34] cropped a large SAR image into many small patches. But the cropping operation will destroy the structure, texture, and other information of the image. Therefore, to address these problems, inspired by noisy-to-noisy paradigm,[36] we propose a network called multiscale dilated residual U-Net (MDRU-Net). The MDRU-Net is an improved version of U-Net.[37] The MDRU-Net can be trained directly by using noisy–noisy image pairs. Unlike the previously mentioned despeckling CNN methods,[32–35] which use small patches (i.e., 40×40) as input, the input size of the MDRU-Net is 256×256.

Our main contributions are listed as follows:

- To solve the problem of lack of clean SAR images, we put forward MDRU-Net. The network does not require clean SAR images during training, and its input is noisy–noisy SAR image pairs.
- We design a multiscale dilated convolution (MDC) module that uses multiple dilated convolutions to extract and fuse multiscale features for protecting more SAR image details.
- To reduce the difference between the shallow and deep features in fusion, we plan five different dilation residual skip (DRS) connections to narrow the distinctness.
- We propose an effective loss function called $L_{\_\text{hybrid}}$ loss function to suppress artifacts and improve the stability of the network.
- We do extensive experiments on the UC Merced land-use (UCML), SEN-1, and SEN-2 datasets. The experimental results show that the proposed MDRU-Net can obtain good SAR image despeckling effect without clean data.

The remainder of this paper is structured as follows. The related work is briefly reviewed in Sec. 2. In Sec. 3, the proposed methods are illustrated in detail. Experiment settings and results are presented in Sec. 4. Conclusions are given in Sec. 5.

2 Related Work

2.1 *Convolutional Neural Network for Synthetic Aperture Radar Image Despeckling*

With the gradual maturity of CNN, intelligent applications of SAR are made possible. However, speckle noise is a major obstacle affecting the intelligent interpretation of SAR images. How to use CNN to effectively and quickly remove speckle noise becomes the primary task of intelligent interpretation. Chierchia et al.[32] first proposed a despeckling CNN for SAR images (SAR-CNN). The SAR-CNN was inspired by the denoising CNNs,[38] which worked very well in reducing additive white Gaussian noise. However, the SAR-CNN adopted a coupled logarithm and exponential transforms in the process of removing speckle noise. So it is not an end-to-end learning network. To solve this problem, Wang et al.[33] designed an image despeckling CNN (ID-CNN), which consisted of eight convolutions and a division residual layer. Zhang et al.[34] presented an SAR image despeckling network with dilated residual structure (SAR-DRN). They adjusted the dilation rate of the dilated convolution to increase the network receptive field and capture more image details. Francesco et al.[35] utilized U-Net to remove the speckle noise of SAR images and they demonstrated the performance of the skip connection.

2.2 *Dilated Convolution*

In the image semantic segmentation task, to aggregate multiscale context information without losing image resolution, Yu and Koltun[39] developed a convolutional network module called dilated convolution. The dilated convolution can increase network receptive field without increasing the weight. Figure 1 illustrates the operation of the dilated convolution on the feature map. The size of the feature map is 9×9 and the red dots represent the original weight of the kernel. The yellow blank blocks represent the expanded weight with the value of 0.

Liu et al.[40] planned a multibranch residual module with dilated convolutions to extract multiscale features so that the classification and identification of spacecraft electronic load signals can be solved. Yang et al.[41] designed an end-to-end dilated inception network (DINet) to predict visual saliency maps. The dilated inception module of the DINet used dilated convolutions with different dilation rates in parallel, which not only can significantly reduce the computational load but also can enrich the diversity of the receptive field in the features. Zhang et al.[42] presented a multiscale single-image super-resolution network with dilated convolutions. This network effectively increased the receptive field of the network by adjusting the dilation rate. In the SAR image despeckling task, the SAR-DRN[34] only utilized seven dilated convolutions and its despeckling performance exceeds the SAR-CNN[32] with 17 traditional convolutions.

2.3 *Skip Connection*

In CNN, continuous convolution and pooling operations are used to extract deep features. As a result, much detail information of the image is lost. To solve this problem, many methods[32–35] obtain small patches through the cropping operation, and these small patches are used as training data. However, the cropping operation can destroy the structure, texture information of the image. The raised skip connection[43] can enforce the networks to reconsider low-level features, which are going to fade away when the low-level features feed forward. Qi et al.[44] proposed a convolutional encoder–decoder network with skip connections to improve the predictive performance of the saliency maps. They used a skip connection between the encoder and the decoder to transfer the hierarchical features. Tong et al.[45] designed a dense skip connection in a very deep network. The dense skip connection not only alleviated the problem of gradient disappearance but also accelerated the efficiency of super-resolution image reconstruction. Ronneberger et al.[37] presented a U-Net to obtain very accurate segmentation results. Francesco et al.[35] conducted a skip connection ablation experiment on the SAR images. They found that the greater the degree

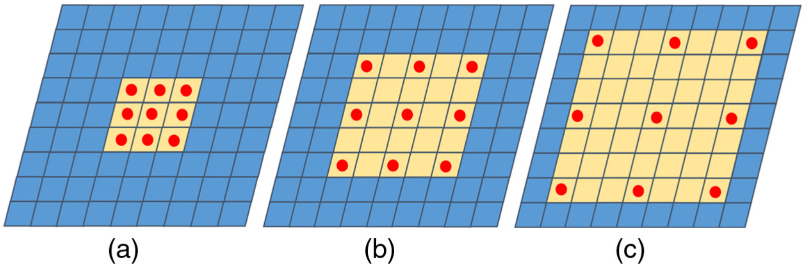

Fig. 1 The operation of the dilated convolution on the feature map: (a) the dilation rate is 1 and the kernel size is 3×3; (b) the dilation rate is 2 and the kernel size is enlarged to 5×5; and (c) the dilation rate is 3 and the kernel size is broadened to 7×7.

of compression of the input image, the more important the skip connections are, and the more obvious the despeckling effect of the SAR image is.

2.4 *Loss Function*

The selection of the loss function affects the convergence speed and the optimization degree. Zhao et al.[46] demonstrated the effect of the loss function in detail. In the SAR image despeckling task, most research studies[32,34,35] use the mean squared error (MSE) as the loss function. The ID-CNN[33] used a mixture loss function, which includes the MSE and the total variation (TV) loss function.[47]

The MSE loss function is a differentiable convex function that enables the network to achieve the global optimality. But it has the following drawbacks. First, MSE can penalize the noise outliers too much, which will easily cause the CNN exploding gradient problem. Second, if MSE is used as the loss function, the predicted clean image will have artifacts. Assume that the noisy–clean image pairs are $\{x_i, y_i\}_N$, $i = 0,1,2, \cdots, N-1$, where N expresses the total number of training image pairs. The x_i and y_i are the noisy image and the clean image, respectively. The size of x_i and y_i is $W \times H$. Here x_i can be written as

$$x_i = y_i \times n_i, \tag{1}$$

where n_i is the speckle noise. The predicted image of the despeckling CNN can be expressed as

$$\hat{x}_i = \mathcal{F}(x_i; \Phi), \tag{2}$$

where $\mathcal{F}(x_i; \Phi)$ is the despeckling CNN and the Φ is the weight of the despeckling CNN. Therefore, the MSE loss function of the noisy–clean method can be written as follows:

$$L^c_{\text{MSE}} = \frac{1}{N}\sum_{i=0}^{N-1}\left[\frac{1}{W \times H}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}(\hat{x}_{i_{w,h}} - y_{i_{w,h}})^2\right], \tag{3}$$

where $\hat{x}_{i_{w,h}}$ and $y_{i_{w,h}}$ represent the pixel value in the (w, h) position, respectively. The $\hat{x}_i$ and y_i are i'th predicted image and clean image, respectively.

The mean absolute error (MAE) loss function is a nonconvex function and its optimization process is a suboptimization. Compared with the MSE, it is less penalizing the noise outliers. The MAE loss function of the noisy–clean method can be given as

$$L^c_{\text{MAE}} = \frac{1}{N}\sum_{i=0}^{N-1}\left(\frac{1}{W \times H}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}|\hat{x}_{i_{w,h}} - y_{i_{w,h}}|\right). \tag{4}$$

It can be seen from Eq. (4) that the derivative of MAE at 0 is not unique, which can affect the stability of the network.

The TV loss function[47] is a regular term loss function that reduces the difference between adjacent pixels to ensure the smoothness of the image. It can be formulated as follows:

$$L_{\text{TV}} = \sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\sqrt{p^2(\hat{x}_i) + q^2(\hat{x}_i)}, \tag{5}$$

where p and q are computed as

$$\begin{aligned} p(\hat{x}_i) &= \hat{x}_{i_{w+1,h}} - \hat{x}_{i_{w,h}}, \\ q(\hat{x}_i) &= \hat{x}_{i_{w,h+1}} - \hat{x}_{i_{w,h}}, \end{aligned} \tag{6}$$

where $\hat{x}_{i_{w+1,h}}$ and $\hat{x}_{i_{w,h+1}}$ are the neighboring pixel values of the $\hat{x}_{i_{w,h}}$.

3 Proposed Method

3.1 *Noisy-to-Noisy Training*

Lehtinen et al.[36] had demonstrated that denoising networks can be learned by mapping a noisy image to another noisy image. The performance of a denoised network trained with noisy–noisy image pairs is similar to that of noisy–clean image pairs. This study is significant for the speckle noise suppression in SAR images.

The previous despeckling CNN methods use MSE and MAE. Equations (3) and (4) represent MSE and MAE, respectively. The noisy-to-noisy training requires the following loss functions:

$$L^n_{\mathrm{MSE}} = \frac{1}{N}\sum_{i=0}^{N-1}\left\{\frac{1}{W\times H}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}[\mathcal{F}(y_i\times n_{i1};\Phi)_{w,h}-(y_i\times n_{i2})_{w,h}]^2\right\}, \tag{7}$$

$$L^n_{\mathrm{MAE}} = \frac{1}{N}\sum_{i=0}^{N-1}\left\{\frac{1}{W\times H}\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}||[\mathcal{F}(y_i\times n_{i1};\Phi)_{w,h}-(y_i\times n_{i2})_{w,h}]||\right\}, \tag{8}$$

where n_{i1} and n_{i2} are two independent noise samples. Whether it is noisy–clean training or noisy–noisy training, their optimization process is to minimize the loss function.

When the proposed MDRU-Net is trained, the input of the network is a pair of noisy–noisy SAR images. The first noisy image is a real SAR image, and the second noisy image is a corrupted SAR image. The corrupted SAR image is simulated by adding 4-look speckle noise to the real SAR image. When MDRU-Net is tested, only one real SAR image is needed (see Sec. 4.1, for detailed usage of our model on different datasets).

3.2 *Multiscale Dilated Residual U-Net Architecture*

Figure 2 displays the architecture of the MDRU-Net in detail. The MDRU-Net consists of an encoder (left side), a decoder (right side), and multiple DRS connections (middle). The encoder acts as a feature extractor, which extracts deep semantic features of the SAR image through continuous convolutions, MDC modules, and max-pooling operations. The decoder is responsible for fusing the deep and shallow semantic features and using the fused features to gradually restore a clean SAR image. The DRS connection is used between the encoder and the decoder. The DRS connection copies the low-level features of the encoder to the decoder. The number of network features is listed in Table 1, where w, h, and c represent the width, height, and channel of the feature map, respectively.

The encoder of the MDRU-Net consists of two convolutions, five MDC modules, and five max-pooling layers. The two convolutions, Conv1_1 and Conv1_2, are $3\times 3\times 64$ convolutions.

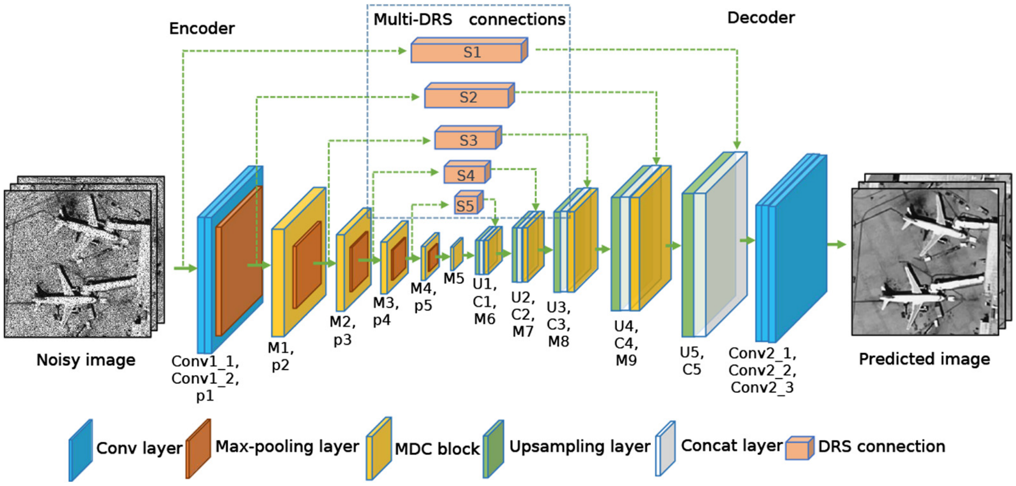

Fig. 2 The detailed architecture of the MDRU-Net.

Table 1 The features of input and output for each layer.

Layers	Input ($w \times h \times c$)	Output ($w \times h \times c$)
Conv1_1	$256 \times 256 \times 1$	$256 \times 256 \times 64$
Conv1_2	$256 \times 256 \times 64$	$256 \times 256 \times 64$
M1	$128 \times 128 \times 64$	$128 \times 128 \times 160$
M2	$64 \times 64 \times 160$	$64 \times 64 \times 256$
M3	$32 \times 32 \times 256$	$32 \times 32 \times 192$
M4	$16 \times 16 \times 192$	$16 \times 16 \times 256$
M5	$8 \times 8 \times 256$	$8 \times 8 \times 256$
M6	$16 \times 16 \times 480$	$16 \times 16 \times 256$
M7	$16 \times 16 \times 320$	$32 \times 32 \times 192$
M8	$32 \times 32 \times 256$	$64 \times 64 \times 256$
M9	$64 \times 64 \times 320$	$128 \times 128 \times 160$
Conv2_1	$256 \times 256 \times 224$	$256 \times 256 \times 64$
Conv2_2	$256 \times 256 \times 64$	$256 \times 256 \times 32$
Conv2_3	$256 \times 256 \times 32$	$256 \times 256 \times 1$

The five max-pooling layers, p1–p5, are to sample the SAR image and obtain hierarchical features step by step. The output of p5 is the deep semantic features of the SAR image. The kernel size of each max-pooling layer is 2×2 and the stride is 2. The five MDC modules, M1–M5, are the proposed MDC module for extracting and fusing multiscale semantic features. The output of the encoder is the 8×8 deep semantic feature maps.

The decoder of the MDRU-Net is composed of five upsampling layers, five concat layers, four MDC modules, and three convolutions. The upsampling layers, U1–U5, are bilinear interpolation. The upsampling layer is used to extend the feature map. The scaling factor of each upsampling layer is 2. The five concat layers, C1–C5, are used to fuse the shallow and deep semantic features in the channel dimensions. The shallow features come from the encoder and are passed to the decoder by the DRS connections. The deep features come from the output of the last layer (M5) in the encoder. The four MDC modules, M6–M9, are used to blend deep and shallow semantic features of SAR images. The three convolutions are Conv2_1, Conv2_2, and Conv2_3. The Conv2_1 and Conv2_2 are traditional convolutions and the size of their kernel is 3×3. The last layer, the Conv2_3, is the output of the MDRU-Net. Its output is a predicted clean SAR image with the size of 256×256.

In MDRU-Net, five DRS connections are used. The DRS connection is used to reduce the difference between the different level features in the network and to copy the shallow features of the encoder to the decoder.

Note that the M1–M9 are the proposed MDC module and will be demonstrated in Sec. 3.3. The detail description of the DRS connection can be seen in Sec. 3.4. Except for Conv2_3, the traditional convolutions and dilated convolutions in MDRU-Net are followed by a rectified linear unit layer.

3.3 *Multiscale Dilated Convolution Modules*

Many methods have been used to make full use of the image information or features to improve the performance, such as increasing network depth,[48] increasing network width,[49] or applying the new loss function.[46] However, they did not take into consideration that objects in the image were similar in different regions. Furthermore, in most image denoising CNNs, they use traditional convolution to extract the semantic features of an image. Once the network structure is determined, the receptive field of the network is fixed. Therefore, we design the MDC module to address these problems. The MDC module consists of multiple dilated convolutions with different dilation rates and a sum-fusion layer. The multiple dilated convolutions are used to extract multiscale features and the sum-fusion layer is used to fuse multiscale features. The fusion method splices on the channel dimension. It is worthy to note that the MDC module improves the receptive field of the network without increasing the network parameters. At the same time, different dilation rates can be set where the large dilation rate allows the network to capture global features and the small dilation rate is used to capture local features.

In the MDRU-Net, nine MDC modules are used. We design five different structures. The five structures of the MDC modules are displayed in Fig. 3. The configuration of the MDC modules is listed in Table 2, where *m*, *r*, and channels mean the number of the dilated convolutions, dilation rates, and channels of the dilated convolution, respectively. In the MDC modules, as the SAR image features become smaller and smaller, the smaller dilation rate is used to focus on the local features.

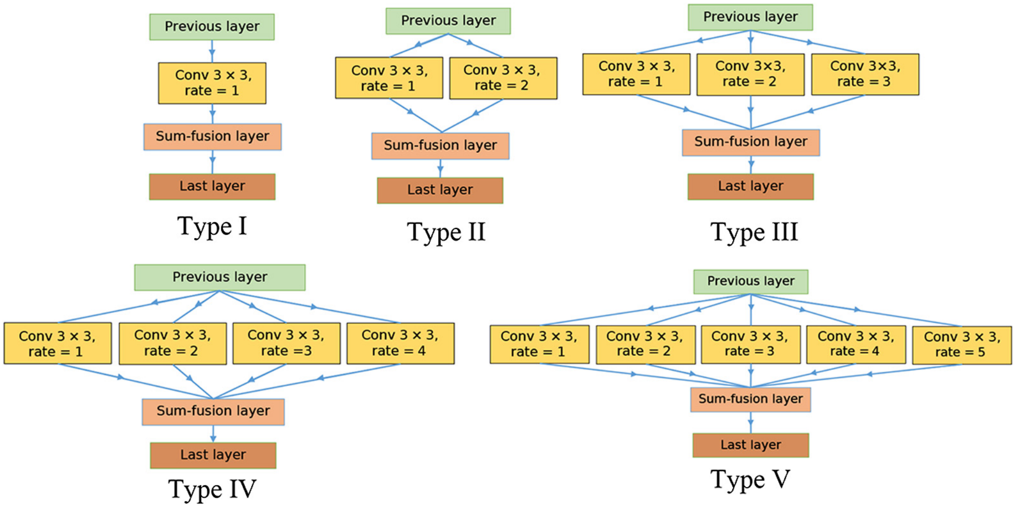

Fig. 3 The detailed structure of the MDC modules in the MDRU-Net.

Table 2 The detailed configuration of the MDC modules in the MDRU-Net.

	Name	Structure	*m*	*r*	Channels
Encoder	M1	Type V	5	1, 2, 3, 4, 5	32, 32, 32, 32, 32
	M2	Type VI	4	1, 2, 3, 4	64, 64, 64, 64
	M3	Type III	3	1, 2, 3	64, 64, 64
	M4	Type II	2	1, 2	128, 128
	M5	Type I	1	1	256
Decoder	M6	Type II	2	1, 2	128, 128
	M7	Type III	3	1, 2, 3	64, 64, 64
	M8	Type VI	4	1, 2, 3, 4	64, 64, 64, 64
	M9	Type V	5	1, 2, 3, 4, 5	32, 32, 32, 32, 32

3.4 *Dilation Residual Skip Connections*

The skip connection not only pass the detailed information of the image[44] but also speed up the training.[45] In many literatures with skip connection,[37,43–45,48] they directly combine shallow and deep features and do not consider the difference between the two features. To reduce the difference, we design a new skip connection structure called DRS connection. The DRS connection consists of dilated convolution and residual block, which are called dilated residual (DR) block. The DR block is composed of a 3×3 traditional convolution and a 3×3 dilated convolution. The dilation rate of the dilated convolution is α. The value of α is an positive integer and can be set to any value. Note that the value of α is limited by the input feature size of the dilated convolution. For example, if the input feature size of dilated convolution is 11×11 and the original kernel size of dilated convolution is 3×3, the value of the dilation rate ranges from 1 to 5. The dilation rate can be written as

$$\alpha = \mathrm{int}\left(\frac{K_d - 1}{K_o - 1}\right), \tag{9}$$

where int represents the round operation, K_d is the dilated kernel size, and K_o is the original kernel size. If the value of α exceeds 5, the dilated convolution loses its effectiveness.

When α is 1, the dilated convolution is the same as traditional convolution, so the dilated convolution cannot increase the receptive field of the network. As α increases, the receptive field of the network will gradually increase. As the receptive field increases, the network can cover more image information. In this way, the networks can pay more attention to the global features of the image, and the despeckling performance of the network can increase.

There are five DRS connections in the MDRU-Net, which are called S1–S5. Each DRS connection contains one or more DR blocks. The detailed configuration of the five DRS connections is shown in Table 3, where Conn.1 is the input of the DRS connection and Conn.2 represents the output of the DRS connection in the MDRU-Net. The blocks represents the number of DR blocks in the DRS connection. In all DR blocks, the number of convolutional channels is 32. Figure 4 displays the framework of the S1 connection in detail, where α is 1, 2, 3, 4, and 5 in the five DR blocks. The S1 is used between the encoder and the decoder in the MDRU-Net. By performing a continuous convolution operation on the input noise image, the S1 connection can effectively reduce the difference between the shallow features and the deep features.

3.5 $L_{\_\mathrm{hybrid}}$ *Loss Function*

We have discussed the *MSE*, MAE, and TV loss function and knew their strengths and weaknesses. The two noisy SAR images are x_{i1} and x_{i2}, the clean SAR image is y_i, and the predicted clean SAR image is $\mathcal{F}(x_{i1};\Phi)$. The proposed $L_{\_\mathrm{hybrid}}$ loss function can be given as

$$\mathrm{L}_{\mathrm{hybrid}} = \begin{cases} \frac{1}{2}C_{\mathrm{MSE}} + C_{\mathrm{TV}}, & C_{\mathrm{MAE}} \le \eta \\ \eta C_{\mathrm{MAE}} + C_{\mathrm{TV}}, & \text{otherwise} \end{cases}, \tag{10}$$

where C_{MSE}, C_{MAE}, and C_{TV} can be written as

Table 3 The detailed configuration of the DRS connections in the MDRU-Net.

Skip layer	Conn.1	Conn.2	Blocks	α
S1	Noisy image	C5	5	1, 2, 3, 4, 5
S2	p1	C4	4	1, 2, 3, 4
S3	p2	C3	3	1, 2, 3
S4	p3	C2	2	1, 2
S5	p4	C1	1	1

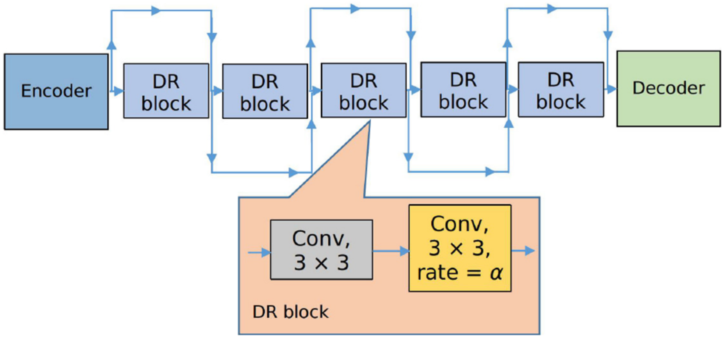

Fig. 4 The detailed framework of the S1 connection. The gray 3×3 is a traditional convolution. The yellow 3×3 is the dilated convolution with $rate = \alpha$.

$$C_{\mathrm{MSE}}(x_{i1}, x_{i2}) = \frac{1}{W \times H} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} [\mathcal{F}(x_{i1}; \Phi)_{w,h} - x_{i2_{w,h}}]^2, \tag{11}$$

$$C_{\mathrm{MAE}}(x_{i1}, x_{i2}) = \frac{1}{W \times H} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} |\mathcal{F}(x_{i1}; \Phi)_{w,h} - x_{i2_{w,h}}|, \tag{12}$$

$$C_{\mathrm{TV}}(x_{i1}) = \frac{1}{W \times H} L_{\mathrm{TV}}[\mathcal{F}(x_{i1}; \Phi)], \tag{13}$$

where L_{TV} is given in Eq. (5). The η is a variable. The value of η can be set as

$$\eta = \frac{1}{n} \sum_{k=0}^{n-1} \left[\frac{1}{W \times H} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} |\mathcal{F}(x_{k1}; \Phi)_{w,h} - x_{k2_{w,h}}| \right], \tag{14}$$

where n represents the batch size and x_{k1} and x_{k2} are the k'th image pair in each batch size. The $L_{\_\mathrm{hybrid}}$ loss function can improve the stability and generalization ability of the network.

4 Experimental Evaluation

In this paper, the experiments have been performed on a personal computer with Ubuntu 16.04. The hardware is an Intel Xeon(R) CPU E5-2620v3, an NVIDIA Quadro M6000 24GB GPU, and 48 GB of RAM. The software tool is PyCharm, the version of Python is Python 3.6, and the deep learning framework is TensorFlow 1.10.

4.1 *Datasets*

Three public datasets, UCML,[50] SEN-1,[51] and SEN-2[51] datasets, are used to demonstrate the performance of the proposed methods.

The UCML dataset[50] is composed of the optical remote sensing images. UCML is released by the UC Merced computer vision laboratory. The dataset is obtained from the large-scale U.S. Geological Survey national map urban area imagery series, and the dataset contains 21 scene data for research purposes. Each scene has 100 images and the size of each image is 256 × 256 × 3.

The SEN1-2 dataset[51] consists of SEN-1 and SEN-2 datasets and it is the optical-SAR image pairs generated from the Sentinel-2 and Sentinel-1 satellites. It has 282,384 images. These images are land scenes in spring, summer, autumn, and winter. We divide the SEN1-2 dataset into a real SAR image subdataset (SEN-1) and an optical image subdataset (SEN-2). The two subdatasets have 141,192 images, respectively. The image size of the SEN-1 is 256 × 256 × 1 and the image size of the SEN-2 is $256 \times 256 \times 3$. In our experiments, to ensure the fairness of the experiment, 2100 images are randomly extracted from the SEN-1, named mini SEN-1 (mSEN-1). Meanwhile, 2100 images are randomly extracted from the SEN-2, named mini SEN-2 (mSEN-2).

Next, according to the noisy–noisy training method, the training image pairs are constructed.

4.1.1 *Training data of the simulated synthetic aperture radar images*

We use two datasets, UCML and mSEN-2, as the simulated SAR images to demonstrate the despeckling performance of the proposed methods. First, we process the images of the UCML dataset into grayscale images. Then, we randomly divide 2100 images of the UCML dataset into 1400 images as the training set, 200 images as the validation set, and 500 images as the testing set. Finally, we add two kinds of simulated speckle noise to each image of the training set and obtain the training image pairs $\{x_{i1}, x_{i2}\}_N$. The x_{i1} and x_{i2} are all noisy images. The x_{i1} is the input image and the x_{i2} is the ground-truth image (the noisy image). The mSEN-2 has the same processing method as UCML dataset. An example of training image processing samples for UCML and mSEN-2 datasets is shown in Fig. 5.

4.1.2 *Training data of the real synthetic aperture radar images*

We used the mSEN-1 dataset as the real SAR images to verify the despeckling performance of the proposed methods. First, we randomly selected 1400 images from the 2100 images in the mSEN-1 dataset as the training set, 200 images as the validation set, and 500 images as the testing set. Then, we corrupted the training set and generated the noisy–noisy image pairs for training. The real SAR training image pairs are $\{x_{i1}, x_{i2}\}_N$. The x_{i1} represents the real SAR image and is used as the input image of the networks. The x_{i2} implies the corrupted image of the mSEN-1 dataset and is used as the ground-truth image (the noisy image) of the networks. In our experiments, the corrupted method is to add simulated speckle noise to real SAR images. As shown in Fig. 6, an example of the processed mSEN-1 dataset is listed. The reference image of the mSEN-1 is the grayscale image of the mSEN-2.

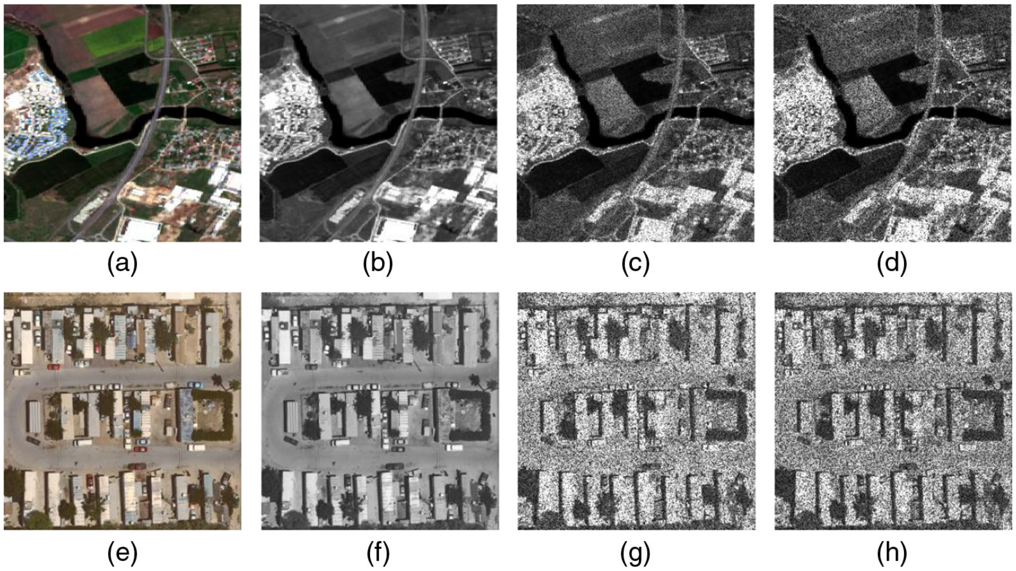

Fig. 5 An example of the samples after processing on the mSEN-2 and UCML datasets. (a)–(d) The optical, grayscale, simulated input (x_{i1}), and simulated ground-truth (x_{i2}) images on the mSEN-2 dataset, respectively. (e)–(h) The optical, grayscale, simulated input (x_{i1}), and simulated ground-truth (x_{i2}) images on the UCML dataset, respectively.

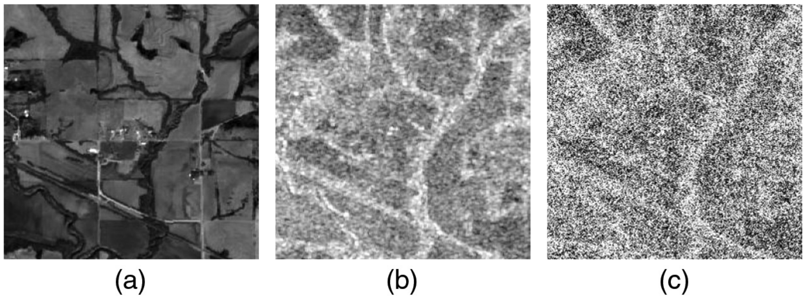

Fig. 6 An example of the training data after processing on the mSEN-1 dataset: (a) the reference image, (b) the real SAR image, and (c) the corrupted image, respectively.

4.2 *Quality Assessment Criteria*

To evaluate the despeckled SAR images, we choose the signal-to-noise ratio (SNR), the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM),[52] the despeckling gain (DG),[53] the equivalent number of looks (ENL)[34] and the edge preservation index (EPI) as assessment criteria.

The SNR is the ratio of signal strength to noise intensity. Assume that x_i and y_i are the noisy image and the clean (reference) image, respectively. The output of the despeckling network is $\mathcal{F}(x_i;\Phi)$. Let $\mathcal{F}(x_i;\Phi)=\hat{x}_i$ mean that $\hat{x}_i$ is the despeckled image. SNR is defined as

$$\mathrm{SNR}=\frac{1}{M}\sum_{i=0}^{M-1}\left[10\log_{10}\frac{\sum_{w=0}^{W-1}\sum_{h=0}^{H-1}\hat{x}_{i_{w,h}}^{2}}{W\times H\times C_{\mathrm{MSE}}(\hat{x}_i,y_i)}\right], \tag{15}$$

where $C_{MSE}(\cdot)$ is given in Eq. (11), and the M is the number of testing set.

The PSNR is the most widely used objective measure of image quality. It represents the ratio between the maximum signal power and the noise power. The PSNR measures the similarity between the despeckled image and the reference image. The PSNR is written as

$$\mathrm{PSNR}=\frac{1}{M}\sum_{i=0}^{M-1}\left[10\log_{10}\frac{\mathrm{MAX}_{y_i}^{2}}{C_{\mathrm{MSE}}(\hat{x}_i,y_i)}\right], \tag{16}$$

where MAX_{y_i} is the largest pixel value in the image y_i. For example, the maximum pixel value of the grayscale image is 255.

The SSIM[52] measures the similarity of the image structure between the despeckled image and the reference image. It is not affected by changes in contrast and brightness. The value range of the SSIM is [0, 1]. The SSIM is expressed as

$$\mathrm{SSIM}=\frac{1}{M}\sum_{i=0}^{M-1}\mathrm{SSIM}_i, \tag{17}$$

where SSIM_i can be written as

$$\mathrm{SSIM}_i=\frac{(2\mu_{\hat{x}_i}\mu_{y_i}+C_1)(2\sigma_{\hat{x}_i y_i}+C_2)}{(\mu_{\hat{x}_i}^2+\mu_{y_i}^2+C_1)(\sigma_{\hat{x}_i}^2+\sigma_{y_i}^2+C_2)}, \tag{18}$$

where $\mu_{\hat{x}_i}$, $\sigma_{\hat{x}_i}$, μ_{y_i}, and σ_{y_i} represent the mean and standard deviation of the images $\hat{x}_i$ and y_i, respectively. The $\sigma_{\hat{x}_i y_i}$ is the covariance of the images $\hat{x}_i$ and y_i. The C_1 and C_2 are constants, and the role of C_1 and C_2 is to avoid SSIM calculation errors when the mean and standard deviation of the image are both 0.

The DG[53] is a new paradigm for the objective assessment of SAR despeckling methods and its calculation requires a noisy image, a despeckled image, and a reference image. The DG can be given as

$$\mathrm{DG} = \frac{1}{M}\sum_{i=0}^{M-1}\left[10\log_{10}\frac{C_{\mathrm{MSE}}(x_i, y_i)}{C_{\mathrm{MSE}}(\hat{x}_i, y_i)}\right]. \tag{19}$$

From the calculation formulas of the above four assessment criteria, they all need a reference image. However, the mSEN-1 dataset is a real SAR image dataset. It lacks clean image as the reference image when calculating the indices. Therefore, in order to objectively evaluate the despeckling performance of the real SAR images, the clean grayscale images of mSEN-2 are used as the reference images for mSEN-1.

The ENL[34] is a common indicator, and it is used to evaluate the speckle noise intensity of SAR images. The ENL can be defined as

$$\mathrm{ENL} = \frac{1}{M}\sum_{i=0}^{M-1}\frac{\mu_{\hat{x}_i}^2}{\sigma_{\hat{x}_i}^2}. \tag{20}$$

The EPI is used to evaluate the edge preservation ability of the despeckled image in the horizontal or vertical directions. The value range of EPI is [0, 1]. The higher the *EPI* value, the stronger the edge preservation ability of despeckling network is. The EPI can be written as

$$\mathrm{EPI} = \frac{\sum_{i=0}^{M-1}|(\mathrm{DN}_{V1}-\mathrm{DN}_{V2})|_{\hat{x}_i}}{\sum_{i=0}^{M-1}|(\mathrm{DN}_{V1}-\mathrm{DN}_{V2})|_{x_i}} + \frac{\sum_{i=0}^{M-1}|(\mathrm{DN}_{H1}-\mathrm{DN}_{H2})|_{\hat{x}_i}}{\sum_{i=0}^{M-1}|(\mathrm{DN}_{H1}-\mathrm{DN}_{H2})|_{x_i}}, \tag{21}$$

where $|\cdot|$ represents the absolution operation. The DN_{V1} and DN_{V2} are the pixel values of adjacent pixels on the vertical direction, respectively. The DN_{H1} and DN_{H2} are the pixel values of adjacent pixels on the horizontal direction, respectively.

4.3 *Implementation Details*

We use the prepared noisy–noisy image pairs to train network and use the Adam algorithm[54] as an update algorithm for network parameters. The Adam algorithm is a stochastic optimization method proposed by Diederik and Jimmy,[54] which is integrated in many deep learning platforms such as TensorFlow, Caffe, and PyTorch. In Adam algorithm, there are three main parameters, which are $\beta 1$, $\beta 2$, and ϵ. In our experiments, we used default values of three parameters provided by Adam algorithm.[54] The default values are $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\epsilon = 10^{-8}$.

The learning rate is not fixed and has a smooth reduction in our experiments. Assume that the maximum number of training iterations is I and the current number of iterations is t. The current learning rate (cur_lr) is represented as

$$\mathrm{cur\_lr} = \begin{cases} \mathrm{lr}\left\{0.5 + \cos\left[\frac{t-I(1-\xi)}{2\xi I}\pi\right]\right\}^2, & t \geq I\xi \\ \mathrm{lr}, & t < I\xi \end{cases}, \tag{22}$$

where lr is the initial learning rate. The ξ is a constant with a range [0, 1], and ξ controls the starting position where the learning rate start to decrease.

In our experiments, I is 50,000, lr is set to 0.0001, ξ is set to 0.3, and the batch size (n) is 4. The explanation for choosing these values of the above parameters is as follows. When we train the despeckling network, I is set to 100,000, and a model is saved for every 5000 iterations. By testing each model, we find that the network has converged at 50,000 iterations, and the test results are optimal. Therefore, the maximum number of iterations is set to 50,000.

In the selection of the lr, we test 0.01, 0.001, and 0.0001. When lr is 0.01, the loss value of network appears NaN (Not a Number). When lr is 0.001, the loss value oscillates shapely. When lr is set to 0.0001, the loss value can quickly converge.

With the deepening of the network training, the optimization of the network requires a smaller learning rate. The ξ is a parameter that controls the start position where the learning rate starts to decrease. By observing the change of the loss value, the loss value begins to oscillate up and down slightly at 15,000 iterations, and the loss value does not decrease.

Therefore, we set ξ to 0.3. After 15,000 iterations, the learning rate starts to decrease and the loss value starts to decrease.

The n is set based on the GPU memory size. When n is set to 5, there is insufficient memory during network training. When n is set to 4, the network can train normally. It is worth noting that the larger the n, the better the despeckling performance of the network is.

4.4 *Experimental Results and Analysis*

4.4.1 *Despeckling performance of the U-Net*

To prove that the U-Net[37] can use the noisy–noisy training method to remove speckle noise in the simulated and real SAR images, we first calculate the values of the SNR, PSNR, SSIM, and ENL of the reference image and the input image. These values are the result of the model "No." As shown in Eqs. (19) and (21), the calculation of the DG and EPI indicators requires a despeckled image. Therefore, the values of *DG* and EPI are not given in model "No." Then, in order to adapt to the three datasets, we make three major modifications to the original U-Net.[37] The first one is to modify the input image size of the original U-Net from 572 × 572 to 256 × 256. The second one is to change all convolutions from unpadded to padding. The third one is to remove the cropping operation. It is worth noting that the U-Net mentioned later represents the modified U-Net. Finally, we use the constructed noisy–noisy image pairs $\{x_{i1}, x_{i2}\}_N$ to train the U-Net and obtain the despeckling models. The experimental results of the U-Net on the three testing sets are shown in Table 4, where ↑ means that the larger the value, the stronger the despeckling ability of the network is. The No means directly calculate the values of input image and reference image without using any despeckling method. From the experimental results, it can be seen that the U-Net using noisy–noisy training method can effectively remove the speckle noise and improve the quality to some extent in the simulated and real SAR images.

4.4.2 *Ablation experiment of the multiscale dilated convolution modules*

We have demonstrated that the U-Net can indeed remove the speckle noise without clean SAR data in simulated and real SAR images. However, the despeckling performance of the U-Net is limited. To improve the despeckling performance of the U-Net and verify the proposed MDC module, we replace the convolutions in U-Net with the MDC modules. It is worthy to note that the first and last convolutions in U-Net are reserved. The MDC modules are illustrated in detail in Sec. 3.3. The experimental results on the three testing sets are shown in Table 5. The bold black body is the better experimental results and the MDCs represent that MDC modules are used in U-Net. It can be seen from the experimental results that the MDC modules used in the U-Net can greatly improve the despeckling performance. Compared to the U-Net, the PSNR of the three testing sets increased by 5.602, 1.441, and 5.002 dB, respectively. The other assessment criteria have increased too.

Table 4 The experimental results of the U-Net on the UCML, mSEN-1, and mSEN-2 datasets.

Dataset	Model	SNR ↑	PSNR ↑	SSIM ↑	DG ↑	ENL ↑	EPI ↑
UCML	No	17.883	20.321	0.434	—	4.068	—
	U-Net	19.976	25.140	0.665	6.412	14.405	0.683
mSEN-1	No	–0.010	9.663	0.034	—	3.897	—
	U-Net	7.263	15.830	0.237	6.167	10.135	0.589
mSEN-2	No	15.428	20.529	0.465	—	4.063	—
	U-Net	18.708	26.374	0.760	6.736	14.135	0.652

Table 5 The experimental results of the MDC modules on the UCML, mSEN-1, and mSEN-2 datasets.

	Model	SNR ↑	PSNR ↑	SSIM ↑	DG ↑	ENL ↑	EPI ↑
UCML	No	17.883	20.321	0.434	—	4.068	—
	U-Net	19.976	25.140	0.665	6.412	14.405	0.683
	MDCs	**25.972**	**30.742**	**0.864**	**10.420**	**16.230**	**0.775**
mSEN-1	No	−0.010	9.663	0.034	—	3.897	—
	U-Net	7.263	15.830	0.237	6.167	10.135	0.589
	MDCs	**8.704**	**17.271**	**0.328**	**7.608**	**12.882**	**0.666**
mSEN-2	No	15.428	20.529	0.465	—	4.063	—
	U-Net	18.708	26.374	0.760	6.736	14.135	**0.652**
	MDCs	**22.807**	**31.376**	**0.892**	**10.837**	**15.882**	0.634

Bold values represent better experimental results.

4.4.3 *Ablation experiment of the dilation residual skip connections*

To demonstrate the despeckling performance of the DRS connection, we replace the skip connections in U-Net with the DRS connections. The proposed DRS connections have been introduced in Sec. 3.4. The experimental results on the three testing sets are shown in Table 6. The DRSs represent that DRS connections are used in U-Net. By comparing the experimental results, it can be seen that the DRS connections significantly improve the despeckling ability of the U-Net in simulated and real SAR images. The main reason is that the proposed DRS connection allows each level of semantic features to experience the same number of convolution operations. The DRS connection can improve the fusion efficiency when fusing features and increase the despeckling performance of the despeckling networks. Therefore, the DRS connection can effectively decrease the difference between deep and shallow semantic features and help the despeckling network models to improve the ability of removing speckle noise. Compared with the experimental results of the U-Net, the UCML dataset increased 5.679 dB, the mSEN-1 dataset increased 1.335 dB, and the mSEN-2 dataset increased 4.957 dB on the PSNR.

Table 6 The experimental results of the DRS connections on the UCML, mSEN-1, and mSEN-2 datasets.

Dataset	Model	SNR ↑	PSNR ↑	SSIM ↑	DG ↑	ENL ↑	EPI ↑
UCML	No	17.883	20.321	0.434	—	4.068	—
	U-Net	19.976	25.140	0.665	6.412	14.405	0.683
	DRSs	**25.626**	**30.819**	**0.866**	**10.886**	**16.391**	**0.790**
mSEN-1	No	−0.010	9.663	0.034	—	3.897	—
	U-Net	7.263	15.830	0.237	6.167	10.135	0.589
	DRSs	**8.598**	**17.165**	**0.321**	**7.502**	**13.114**	**0.683**
mSEN-2	No	15.428	20.529	0.465	—	4.063	—
	U-Net	18.708	26.374	0.760	6.736	14.135	0.652
	DRSs	**22.768**	**31.331**	**0.892**	**10.792**	**15.889**	**0.658**

Table 7 The experimental results of the $L_{\_\text{hybrid}}$ loss function on the UCML, mSEN-1, and mSEN-2 datasets.

Dataset	Model	SNR ↑	PSNR ↑	SSIM ↑	DG ↑	ENL ↑	EPI ↑
UCML	No	17.883	20.321	0.434	—	4.068	—
	MSE	19.976	25.140	0.665	6.412	14.405	0.683
	MAE	21.129	26.327	0.714	6.006	14.376	0.678
	$L_{\_\text{hybrid}}$	**25.816**	**31.022**	**0.871**	**10.702**	**15.299**	**0.690**
mSEN-1	No	−0.010	9.663	0.034	—	3.897	—
	MSE	7.263	15.830	0.237	6.167	10.135	0.589
	MAE	7.872	16.139	0.220	6.476	10.452	0.531
	$L_{\_\text{hybrid}}$	**8.912**	**17.479**	**0.337**	**7.816**	**11.872**	**0.664**
mSEN-2	No	15.428	20.529	0.465	—	4.063	—
	MSE	18.708	26.374	0.760	6.736	14.135	0.652
	MAE	19.907	27.474	0.771	6.937	14.657	**0.686**
	$L_{\_\text{hybrid}}$	**22.748**	**31.315**	**0.892**	**10.776**	**16.872**	0.633

4.4.4 *Despeckling performance of the $L_{\_\text{hybrid}}$ loss function*

To explain the improvement brought by the $L_{\_\text{hybrid}}$ loss function, we first use MSE and MAE loss functions to train the U-Net, respectively. Then, we replace the MAE and MSE loss with the $L_{\_\text{hybrid}}$ loss function. The detailed analysis of the $L_{\_\text{hybrid}}$ loss function can be found in Sec. 3.5. The experimental results obtained on the three testing sets are shown in Table 7. We find that the MAE loss function has better despeckling performance than the MSE, while $L_{\_\text{hybrid}}$ loss function has higher despeckling performance than the MAE.

4.4.5 *Despeckling performance of the multiscale dilated residual U-Net*

In this section, we verify the despeckling performance of the proposed MDRU-Net for simulated and real SAR images. The training data of the MDRU-Net are noisy–noisy image pairs.

Table 8 The experimental results of the MDRU-Net on the UCML, mSEN-1, and mSEN-2 datasets.

	Model	SNR ↑	PSNR ↑	SSIM ↑	DG ↑	ENL ↑	EPI ↑
UCML	No	17.883	20.321	0.434	—	4.068	—
	U-Net	19.976	25.140	0.665	6.412	14.405	0.683
	MDRU-Net	**26.329**	**31.616**	**0.883**	**11.295**	**16.152**	**0.774**
mSEN-1	No	−0.010	9.663	0.034	—	3.897	—
	U-Net	7.263	15.830	0.237	6.167	10.135	0.589
	MDRU-Net	**9.109**	**19.376**	**0.467**	**9.013**	**13.822**	**0.767**
mSEN-2	No	15.428	20.529	0.465	—	4.063	—
	U-Net	18.708	26.374	0.760	6.736	14.135	0.652
	MDRU-Net	**23.331**	**31.900**	**0.902**	**11.361**	**17.822**	**0.735**

The image size is 256×256. The detailed architecture of the MDRU-Net has been introduced in Sec. 3.2. The MDRU-Net uses the $L_{\_\text{hybrid}}$ loss function. The experimental results of MDRU-Net on the three testing sets are shown in Table 8. By comparing the experimental results, the PSNR values improved 6.476, 3.546, and 5.526 dB in the three testing sets, respectively.

4.5 *Compared with the State-of-the-Art Despeckling Methods*

To compare the despeckling performance with the MDRU-Net, we select the refined Lee filter[10] (RLF), the improved sigma filter[17] (ISF), the probabilistic patch-based (PPB) filter,[30] the three-dimensional block matching (BM3D) filter for SAR image despeckling (SAR-BM3D),[28] the SAR-CNN,[32] and the SAR-DRN.[34] Note that the RLF, ISF, PPB, and SAR-BM3D are the traditional despeckling algorithms and are widely used to filter SAR images. The SAR-CNN and SAR-DRN are the state-of-the-art despeckling CNNs for SAR images and their training data are the noisy–clean image pairs.

Table 9 The comparative experimental results of airplane, highway, and buildings.

Method	Scene	PSNR ↑	SSIM ↑	ENL ↑	EPI ↑
RLF	Airplane	24.23	0.723	18.387	0.614
ISF		25.46	0.750	19.112	0.621
PPB		24.98	0.743	—	—
SAR-BM3D		27.17	0.800	—	—
SAR-CNN		27.89	0.801	—	—
SAR-DRN		28.01	0.819	—	—
MDRU-Net (N2C)		31.72	0.846	**19.533**	**0.749**
MDRU-Net (Ours)		**31.82**	**0.848**	19.528	0.747
RLF	Buildings	29.18	0.795	17.499	0.647
ISF		29.78	0.811	17.993	0.667
PPB		29.50	0.871	—	—
SAR-BM3D		31.36	0.902	—	—
SAR-CNN		31.63	0.901	—	—
SAR-DRN		31.78	0.901	—	—
MDRU-Net (N2C)		32.38	0.903	19.591	0.732
MDRU-Net (Ours)		**32.39**	**0.904**	**19.596**	**0.751**
RLF	Highway	24.56	0.714	18.152	0.608
ISF		24.86	0.733	18.321	0.631
PPB		24.90	0.764	—	—
SAR-BM3D		26.41	0.834	—	—
SAR-CNN		26.48	0.834	—	—
SAR-DRN		26.53	0.836	—	—
MDRU-Net (N2C)		32.79	0.849	**22.186**	0.850
MDRU-Net (Ours)		**32.87**	**0.852**	22.185	**0.851**

To ensure the fairness of the experiment, according to the method of selecting training data by SAR-DRN,[34] we randomly select 400 images from the UCML dataset[50] as training data, and then perform data augmentation on the selected images. The final training data are 1600 images by rotating, flipping, and mirroring. To construct the noisy–noisy training image pairs, we add simulated speckle noise to the clean images. The difference from SAR-DRN[34] is that the training image pairs of the MDRU-Net are noisy–noisy image pairs and the image size is 256×256. After training, the SAR image despeckling model is obtained. During the testing phase, we use the airplane, highway, and buildings as our testing set. The testing set is the same as that used in SAR-DRN.[34] We only compared the case where the speckle noise level is 8. In the selected three scenes, we add simulated speckle noise to them. The experimental results of airplane, highway, and buildings are shown in Table 9. The meaning of MDRU-Net (N2C) is that the input of the training network is noisy–clean image pairs, and the MDRU-Net (Ours) is the method proposed in this paper to use noisy–noisy image pairs during training.

From the experimental results, it can be found that even if the noisy–noisy image pairs are used to train the MDRU-Net, the despeckling performance is significantly higher than other algorithms. On the PSNR, the MDRU-Net obtained 31.82, 32.34, and 32.87 dB in airplane, buildings, and highway scenes, respectively. Compared with the SAR-DRN,[34] the MDRU-Net (Ours) increased approximately 3.81, 0.56, and 6.34 dB on the three scenes, respectively.

In addition, by comparing the experimental results of MDRU-Net (N2C) and MDRU-Net (Ours), the despeckling performance of MDRU-Net (N2C) and MDRU-Net (Ours) is very close. Therefore, the MDRU-Net (Ours) is recommended to remove the speckle noise in real SAR images.

5 Conclusion

In this paper, the despeckling network MDRU-Net for SAR images is proposed. The MDRU-Net can use the noisy–noisy image pairs to train in the absence of clean SAR images.

The MDRU-Net consists of an encoder, a decoder, and multiple DRS connections. The encoder acts as a feature extractor, which extracts deep semantic features of the SAR images. The decoder is responsible for restoring a clean SAR image. To protect more details of SAR images for extracting deep semantic SAR features, the MDC module is designed. MDC module is used in the encoder and decoder. The MDC module has five types and contains multiple dilated convolutions. However, there is a great difference between shallow and deep semantic features in fusion. To reduce the difference between the two semantic features, the DRS connection is raised. The DRS connection not only reduces the difference but also protects more important details. To make up for the drawbacks of MAE and MSE, we propose the $L_{\text{\_hybrid}}$ loss function. The $L_{\text{\_hybrid}}$ loss function not only improves the stability of the despeckling network but also suppresses the artifacts in predicted clean SAR images. We do extensive experiments on the simulated and real SAR images. The experimental results illustrate that the proposed method achieves state-of-the-art despeckling performance in several key metrics.

References

1. A. Moreira et al., “A tutorial on synthetic aperture radar,” *IEEE Geosci. Remote Sens. Mag.* **1**, 6–43 (2013).
2. U. Risa, S. Toshikazu, and T. Katsumi, “An efficient orthorectification of a satellite SAR image used for monitoring occurrence of disaster,” in *Int. Workshop Adv. Image Technol.*, pp. 1–4 (2018).
3. I. Katherine et al., “Assessing single-polarization and dual-polarization TerraSAR-X data for surface water monitoring,” *Remote Sens.* **10**, 949 (2018).
4. Z. Lin et al., “Squeeze and excitation rank faster R-CNN for ship detection in SAR images,” *IEEE Geosci. Remote Sens. Lett.* **16**(5), 751–755 (2019).
5. B. Bauer-Marschallinger et al., “Toward global soil moisture monitoring with Sentinel-1: harnessing assets and overcoming obstacles,” *IEEE Trans. Geosci. Remote Sens.* **57**(1), 520–539 (2019).

6. J. S. Lee et al., "Speckle filtering of synthetic aperture radar images: a review," *Remote Sens. Rev.* **8**(4), 313–340 (1994).
7. F. Argenti et al., "A tutorial on speckle reduction in synthetic aperture radar images," *IEEE Geosci. Remote Sens. Mag.* **1**(3), 6–35 (2013).
8. C. Khatri, "Classical statistical analysis based on a certain multivariate complex Gaussian distribution," *Ann. Math. Stat.* **36**, 98–114 (1965).
9. F. T. Ulaby et al., "Textural information in SAR images," *IEEE Trans. Geosci. Remote Sens.* **GE-24**(2), 235–245 (1986).
10. J. S. Lee, M. R. Grunes, and G. D. Grandi, "Polarimetric SAR speckle filtering and its impact on classification," in *IEEE Int. Geosci. and Remote Sens. Symp. Proc.*, Vol. 2, pp. 1038–1040 (1997).
11. V. Kharchenko, N. Kuzmenko, and I. Ostroumov, "An investigation of synthetic aperture radar speckle filtering and image segmentation considering wavelet decomposition," in *Eur. Microwave Conf. Central Europe*, pp. 398–401 (2019).
12. J. Fritz and V. Chandrasekar, "The impact of adaptive speckle filtering on multi-channel SAR change detection," in *IEEE Int. Geosci. and Remote Sens. Symp.*, Vol. 4, pp. 561–564 (2008).
13. C. Deledalle, L. Denis, and F. Tupin, "MuLoG: a generic variance-stabilization approach for speckle reduction in SAR interferometry and SAR polarimetry," in *IEEE Int. Geosci. and Remote Sens. Symp.*, pp. 5816–5819 (2018).
14. J. S. Lee, "Speckle analysis and smoothing of synthetic aperture radar images," *Comput. Graphics Image Process.* **17**, 24–32 (1981).
15. D. T. Kuan et al., "Adaptive noise smoothing filter for images with signal-dependent noise," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-7**(2), 165–177 (1985).
16. V. S. Frost et al., "A model for radar images and its application to adaptive digital filtering of multiplicative noise," *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-4**(2), 157–166 (1982).
17. J. S. Lee et al., "Improved sigma filter for speckle filtering of SAR imagery," *IEEE Trans. Geosci. Remote Sens.* **47**, 202–213 (2009).
18. A. Lopes, R. Touzi, and E. Nezry, "Adaptive speckle filters and scene heterogeneity," *IEEE Trans. Geosci. Remote Sens.* **28**(6), 992–1000 (1990).
19. C. Wang and R. Wang, "Multi-model SAR image despeckling," *Electron. Lett.* **38**(23), 1425–1426 (2002).
20. H. Zhao et al., "SAR image despeckling based on adaptive neighborhood window and rotationally invariant block matching," in *IEEE Int. Conf. Signal Process., Commun. and Comput.*, pp. 515–520 (2014).
21. F. Lang, J. Yang, and D. Li, "Adaptive-window polarimetric SAR image speckle filtering based on a homogeneity measurement," *IEEE Trans. Geosci. Remote Sens.* **53**(10), 5435–5446 (2015).
22. P. A. A. Penna and N. D. A. Mascarenhas, "SAR speckle nonlocal filtering with statistical modeling of HAAR wavelet coefficients and stochastic distances," *IEEE Trans. Geosci. Remote Sens.* **57**(9), 7194–7208 (2019).
23. R. Farhadiani, S. Homayouni, and A. Safari, "Hybrid SAR speckle reduction using complex wavelet shrinkage and non-local PCA-based filtering," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**(5), 1489–1496 (2019).
24. F. Zakeri and M. J. V. Zoej, "Adaptive method of speckle reduction based on curvelet transform and thresholding neural network in synthetic aperture radar images," *J. Appl. Remote Sens.* **9**(1), 095043 (2015).
25. S. Jafari and S. Ghofrani, "Using two coefficients modeling of nonsubsampled Shearlet transform for despeckling," *J. Appl. Remote Sens.* **10**(1), 015002 (2016).
26. H. Wu et al., "Denoising method based on intrascale correlation in nonsubsampled contourlet transform for synthetic aperture radar images," *J. Appl. Remote Sens.* **13**(4), 046503 (2019).
27. M. Soccorsi, D. Gleich, and M. Datcu, "Huber–Markov model for complex SAR image restoration," *IEEE Geosci. Remote Sens. Lett.* **7**(1), 63–67 (2010).
28. S. Parrilli et al., "A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage," *IEEE Trans. Geosci. Remote Sens.* **50**(2), 606–616 (2012).

29. S. Vitale et al., "Guided patchwise nonlocal SAR despeckling," *IEEE Trans. Geosci. Remote Sens.* **57**(9), 6484–6498 (2019).
30. C. Deledalle, L. Denis, and F. Tupin, "Iterative weighted maximum likelihood denoising with probabilistic patch-based weights," *IEEE Trans. Image Process.* **18**(12), 2661–2672 (2009).
31. C. Deledalle et al., "Exploiting patch similarity for SAR image processing: the nonlocal paradigm," *IEEE Signal Process. Mag.* **31**(4), 69–78 (2014).
32. G. Chierchia et al., "SAR image despeckling through convolutional neural networks," in *IEEE Int. Geosci. and Remote Sens. Symp.*, pp. 5438–5441 (2017).
33. P. Wang, H. Zhang, and V. M. Patel, "SAR image despeckling using a convolutional neural network," *IEEE Signal Process. Lett.* **24**(12), 1763–1767 (2017).
34. Q. Zhang et al., "Learning a dilated residual network for SAR image despeckling," *Remote Sens.* **10**, 196 (2018).
35. L. Francesco et al., "Deep learning for SAR image despeckling," *Remote Sens.* **11**, 1532 (2019).
36. J. Lehtinen et al., "Noise2noise: learning image restoration without clean data," CoRR abs/1803.04189 (2018).
37. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," CoRR abs/1505.04597 (2015).
38. K. Zhang et al., "Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.* **26**(7), 3142–3155 (2017).
39. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," CoRR abs/1511.07122 (2015).
40. Y. Liu et al., "MRD-NETS: multi-scale residual networks with dilated convolutions for classification and clustering analysis of spacecraft electrical signal," *IEEE Access* **7**, 171584–171597 (2019).
41. S. Yang et al., "A dilated inception network for visual saliency prediction," *IEEE Trans. Multimedia* 1–1 (2019).
42. Z. Zhang, X. Wang, and C. Jung, "DCSR: dilated convolutions for single image super-resolution," *IEEE Trans. Image Process.* **28**(4), 1625–1635 (2019).
43. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," CoRR abs/1411.4038 (2014).
44. F. Qi et al., "A convolutional encoder–decoder network with skip connections for saliency prediction," *IEEE Access* **7**, 60428–60438 (2019).
45. T. Tong et al., "Image super-resolution using dense skip connections," in *IEEE Int. Conf. Comput. Vision*, pp. 4809–4817 (2017).
46. H. Zhao et al., "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imaging* **3**(1), 47–57 (2017).
47. M. Javanmardi et al., "Unsupervised total variation loss for semi-supervised deep learning of semantic segmentation," CoRR abs/1605.01368 (2016).
48. X. J. Mao, C. Shen, and Y. B. Yang, "Image denoising using very deep fully convolutional encoder–decoder networks with symmetric skip connections," CoRR abs/1603.09056 (2016).
49. C. Chong and X. Zengbo, "Aerial-image denoising based on convolutional neural network with multi-scale residual learning approach," *Information* **9**, 169 (2018).
50. Y. Yi and N. Shawn, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, pp. 270–279 (2010).
51. S. Michael, H. Lloyd, and Z. Xiao, "The Sen1-2 dataset for deep learning in SAR-optical data fusion," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **IV-1**, 141–146 (2018).
52. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Thirty-Seventh Asilomar Conf. Signals, Syst. Comput.*, Vol. 2, pp. 1398–1402 (2003).
53. G. D. Martino et al., "Benchmarking framework for SAR despeckling," *IEEE Trans. Geosci. Remote Sens.* **52**(3), 1596–1615 (2014).
54. P. K. Diederik and L. B. Jimmy, "Adam: a method for stochastic optimization," CoRR abs/1412.6980 (2014).

Gang Zhang is a PhD student at the Space Engineering University. He received his master's degree from the Xidian University and his bachelor's degree from Xi'an Polytechnic University. His research interests include synthetic aperture radar image processing, pattern recognition, and deep learning.

Zhi Li is a professor at the Space Engineering University. He received his PhD from the Institute of Geology of the China Earthquake Administration in 2003. He received his BE and master's degrees from the National University of Defense Technology in 1994 and 1997, respectively. He has authored or co-authored more than 60 papers and 12 books. His research interests include space system applications and artificial intelligence.

Xuewei Li is a PhD student at Beijing University of Posts and Telecommunications. She received her bachelor's and master's degrees from Xi'an University of Technology. Her current research interests include image aesthetic assessment, image processing, and machine learning.

Yiqiao Xu is a PhD student at the Space Engineering University. He received his master's degree from the Electronic Engineering Institute and his bachelor's degree from the Anhui University of Science and Technology. His research interests include signal processing, image pattern recognition, and deep learning.