deSpeckNet: Generalizing Deep Learning-Based SAR Image Despeckling

Adugna G. Mullissa, *Member, IEEE*, Diego Marcos, Devis Tuia, *Senior Member, IEEE*, Martin Herold, and Johannes Reiche

***Abstract*— Deep learning (DL) has proven to be a suitable approach for despeckling synthetic aperture radar (SAR) images. So far, most DL models are trained to reduce speckle that follows a particular distribution, either using simulated noise or a specific set of real SAR images, limiting the applicability of these methods for real SAR images with unknown noise statistics. In this article, we present a DL method, deSpeckNet,[1] that estimates the speckle noise distribution and the despeckled image simultaneously. Since it does not depend on a specific noise model, deSpeckNet generalizes well across SAR acquisitions in a variety of landcover conditions. We evaluated the performance of deSpeckNet on single polarized Sentinel-1 images acquired in Indonesia, The Democratic Republic of Congo, and The Netherlands, a single polarized ALOS-2/PALSAR-2 image acquired in Japan and an Iceye X2 image acquired in Germany. In all cases, deSpeckNet was able to effectively reduce speckle and restore the images in high quality with respect to the state of the art.**

***Index Terms*— Convolutional neural network (CNN), deep learning (DL), speckle, synthetic aperture radar (SAR).**

I. Introduction

THE recent availability of global Earth observation synthetic aperture radar (SAR) data, for instance from the Sentinel-1 SAR satellites, has been a game-changer for large scale, all-weather, day/night monitoring of land surfaces. However, the applicability of these data sets has been limited by the presence of speckle. Speckle is inherent in all SAR images as they are acquired by a coherent active microwave imaging system. Speckle occurs when the backscattered signal from independent targets is coherently superimposed within each resolution cell. Depending on the size of the resolution cell, the superimposition of these signals results in interference whose effect is observed as speckle. Speckle is true scattering information collected from a target, but for image processing purposes it is often considered as noise. To improve the radiometric quality of SAR images before analysis, speckle has to be reduced. In fact, speckle reduction has been an active research area since the advent of airborne and spaceborne imaging SAR sensors in the 1970s [1].

Earlier speckle filtering methods focused on spatially adaptive filters. Most of these were based on pixel intensity statistics determined on a local neighborhood window. This class of filters operated in a sliding window fashion, where the pixel to be filtered is the center pixel in the moving window. The boxcar filter [2] is the simplest spatial filter that estimates the mean value of all the pixels in the moving window. The boxcar filter was effective in reducing speckle in homogeneous regions at the cost of resolution. The Lee filter [3] reduced the impact of the loss in resolution by estimating the minimum mean square error (MSE) in a neighboring window. The Frost filter [4] applied exponential weighting and damping factor to control the amount of filtering in a low pass filtering setting. These methods improved the preservation of features in speckle filtering, however, they introduced artifacts along feature boundaries. To overcome this problem, Lee *et al.* [5] proposed to select similar pixels by using a series of edge aligned nonrectangular windows, Vasile *et al.* [6] used intensity-driven neighborhood region growing based on the image intensity and Lee *et al.* [7] proposed selecting similar neighboring pixels based on scattering characteristics. In addition, Deledalle *et al.* [8] used nonlocal means with weighted maximum likelihood estimation to reduce speckle, and the 3-D block matching approach (BM3D) [9], which groups image patches into 3-D arrays based on their similarity and performs estimations into a 2-D image array from the grouped blocks.

A second family of approaches exploits the wavelet transform of the single look image in log form. Notable works involve the wavelet Bayesian denoising that is introduced in [10], based on Markov random fields (MRFs). Mahdianpari *et al.* [11] introduced a SAR image despeckling method that is based on adaptive Gauss-MRF. In [12], a homomorphic wavelet maximum *a posteriori* (MAP) filter was introduced improving the performance of the original Gamma-MAP speckle filter.

A third family of approaches has recently started to attract attention: thanks to the advent of powerful computation capability, significant advances have been made using deep

Manuscript received February 27, 2020; revised May 29, 2020, August 31, 2020, and November 19, 2020; accepted November 30, 2020. This work was supported in part by the Global Forest Watch-World Resources Institute (GFW-WRI) Radar for Detecting Deforestation (RADD) Project and in part by the U.S. Government SilvaCarbon Program. *(Corresponding author: Adugna G. Mullissa.)*

Adugna G. Mullissa, Diego Marcos, Martin Herold, and Johannes Reiche are with the Laboratory of Geo-information Science and Remote Sensing, Wageningen University, 6700 AA Wageningen, The Netherlands (e-mail: adugna.mullissa@wur.nl; diego.marcos.gonzalez@wur.nl; martin.herold@wur.nl; johannes.reiche@wur.nl).

Devis Tuia is with the ECEO Laboratory, Swiss Federal Institute of Technology (EPFL), CH-1951 Sion, Switzerland (e-mail: devis.tuia@epfl.ch).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2020.3042694.

Digital Object Identifier 10.1109/TGRS.2020.3042694

[1]https://github.com/adugnag/deSpeckNet

learning (DL) methods to perform image denoising tasks. The most notable difference between these methods and those described earlier is that DL based methods learn a suitable denoising function based on pairs of noisy and clean images, instead of using a predefined function. Chierchia *et al.* [13] implemented SAR-convolutional neural network (CNN) by adopting the concept of residual learning and deep CNN proposed in [14] for additive white Gaussian noise reduction. In SAR-CNN, the input SAR images are transformed to the homomorphic form and used for training the SAR-CNN network. The network uses a temporally averaged image as a clean reference label, i.e., as a proxy of the speckle-free reference image. Once the network is trained, the prediction image is transformed back to the original image domain by using an exponential function. Zhang *et al.* [15] used a dilated residual network (DRN) using skip connections to train a deep neural network for SAR image despeckling. The marked difference between SAR-CNN and SAR-DRN was the usage of real SAR images in SAR-CNN, whereas SAR-DRN: 1) was trained on simulated images; 2) exploited residual connections; and 3) processed images in their native form. Recent works focused on combining loss functions with different purposes: for example, Vitale *et al.* [16] uses simultaneously an MSE loss that reconstructs the noise-free image and a Kulback–Lieber loss to reconstruct the distribution of the speckle. Furthermore, Pan *et al.* [17] deal with the unknown noise statistics in SAR images by embedding a CNN model for additive white Gaussian noise reduction with a Multichannel Logarithm with Gaussian denoising (MuLoG) algorithm for multiplicative noise, first introduced in [18].

In a supervised learning paradigm, the quality of the prediction depends on the quality of the reference labels used for training. For generating the reference labels, two approaches have been followed in the literature. The first is the model-based simulation from known speckle statistics [19]. This approach relies on artificially simulating the speckle noise and adding it to an optical image (referred to as a clean image) by following an additive white Gaussian noise model [14] or a multiplicative noise model [19]. This approach has three major drawbacks.

1) It requires to assume an *a priori* speckle noise model based on a Gamma probability distribution and a multiplicative noise model. This poses a problem for adapting the network to data that does not follow the same noise model. A good example for this is the difference in the noise statistics between SAR images with different resolution, band, and landcover types, such as the intensity images of high-resolution single look Iceye X2 *X*-band SAR images and medium resolution preprocessed and multilooked Sentinel-1 C-Band ground range detected (GRD) images. Therefore, a model trained on the single look SAR data would not necessarily adapt to a preprocessed SAR image, as illustrated in Fig. 1. Hence, it is imperative to design a model that is robust to changes in speckle-noise statistics.
2) The artificial simulation and addition of noise to an optical image does not represent the true appearance of real SAR images. This is exemplified in the representation of deterministic scatterers in the SAR images. This phenomenon is not captured by an artificial simulation of noise in an optical image. Hence, a model trained on these simulated images cannot recognize these features when applied to a true SAR image.
3) The white noise property of the simulated images does not represent the scatterer-dependent spatial distribution of the noise in semiheterogeneous and heterogeneous media. Hence, a network trained on homogeneous noise statistics tends to over smooth features in heterogeneous scenes, resulting in suboptimal results.

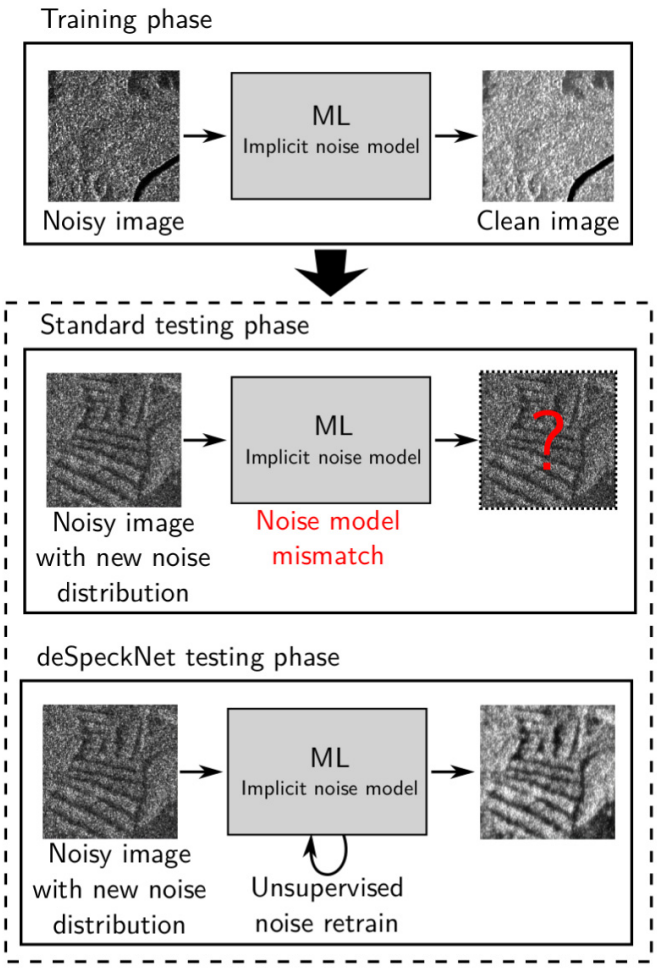
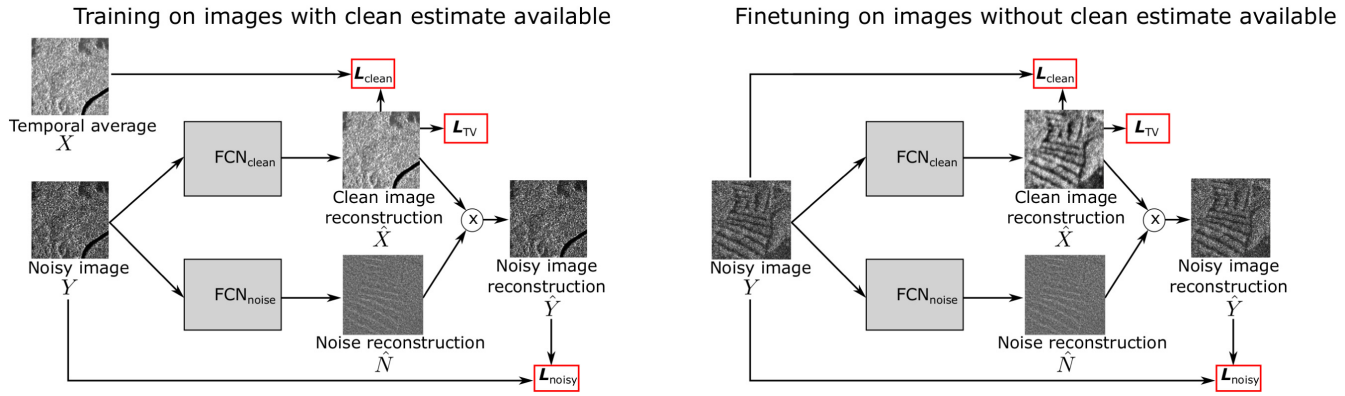


Fig. 1. When a machine learning method is applied to image denoising, it implicitly learns a noise model from the training examples. If there is a mismatch between this distribution and that of the test images the performance can drop substantially. In this work, we propose a method to readjust the noise model to the test images without any additional clean ground-truth images.

These problems led to the second approach: to use real SAR images as labels. However, to obtain a noise-free training label for real SAR images is impossible, because speckle is inherent in all SAR images. One solution proposed in the literature is selecting an area where there is a multitemporal image stack with as little temporal change as possible and taking a temporal average to estimate a noise-free image to be used as a proxy for the reference label. This approach has been demonstrated to provide good results [13]. However, when the scene under investigation is nonstationary in time, taking



(a) Phase 1 (b) Phase 2

Fig. 2. Two phases of deSpeckNet. (a) Model is first trained with an image for which a clean estimate exists (e.g., a temporal average). Three losses are used: an MSE based loss L_{Clean}, L_{TV} and L_{Noisy} reconstruction losses, for the clean and the noisy (original) images, respectively. (b) When moving to an image for which no clean reference is available, an MSE based loss L_{Clean} and a TV loss L_{TV} on the clean image reconstruction and an MSE based loss L_{Noisy} for the noisy image reconstruction.

the temporal average will result in erroneous estimation of the reference label, limiting the applicability of this method. Hence, it is imperative to design a network that can generalize well in areas where a temporal average label is not available.

In this article, we propose a DL pipeline named deSpeckNet that can despeckle SAR images with unknown noise statistics. Initially, deSpeckNet is trained using a temporally averaged SAR image as a reference label. In this first step, the model simultaneously estimates the noise-free image and the noise component. In the second step, this model is subsequently fine-tuned to fit any type of SAR image acquired over any type of cover conditions without using any clean reference labels, i.e., in an unsupervised way. To show the versatility of the approach, the proposed method is evaluated on SAR images acquired in Indonesia, the Democratic Republic of Congo (DRC), The Netherlands, Japan, and Germany and across several SAR sensor configurations involving different noise models.

This article is organized as follows. Section II describes the proposed methodology. The data sets used are described in Section III. Section IV describes the experimental settings, whereas Section V provides results and discussions. Conclusions are presented in Section VI.

II. deSpeckNet

In a distributed medium, an SAR image with a fully developed speckle is assumed to follow a multiplicative speckle model [20]:

$$Y = XN \tag{1}$$

where Y is the observed SAR intensity image, X is the underlying radar reflectivity of the scene, which can be viewed as hypothetically noise free intensity image (since SAR images cannot exist without speckle) and N is the speckle image. The random speckle noise follows a Gamma probability density function:

$$p(N|L) = \frac{1}{\Gamma(L)} L^L N^{L-1} e^{-NL}, \quad N \geq 0, \ L \geq 1 \tag{2}$$

where Γ is a Gamma function and L is interpreted as the number of looks of the SAR image. For single look SAR image ($L = 1$), (2) simplifies to an exponential distribution. If the SAR image is in the amplitude domain, N is characterized by a Nakagami probability density function

$$p(N|L) = \frac{1}{\Gamma(L)} 2L^L N^{2L-1} e^{N^2-L}, \quad N \geq 0, \ L \geq 1. \tag{3}$$

In the single look amplitude case, (3) simplifies to the Rayleigh probability density function.

The proposed DL-based despeckling method (deSpeckNet) consists of two phases: the first follows a supervised learning paradigm by using a temporally averaged SAR image as a reference label. The second phase consists of unsupervised fine-tuning that learns to adapt to a new noise distribution. To this end, we design the architecture of deSpeckNet to follow the multiplicative noise model defined in (1).

A. Architecture

We use a Siamese architecture to estimate the noise-free image ($\hat{X}$) and the estimated noise ($\hat{N}$) separately (Fig. 2). We adopted this architecture to provide the possibility to tune the network to the noise statistics of other SAR images from a different region or from a different sensor. The two identical branches estimate the clean image $\hat{X}$ (FCN$_{\text{clean}}$) and the noise $\hat{N}$ (FCN$_{\text{noise}}$). With those two components, we reconstruct the input noisy image ($\hat{Y}$) using (1).

Both (FCN$_{\text{noise}}$) and (FCN$_{\text{clean}}$) consist of four main building blocks namely convolution [21], batch normalization [22], nonlinear activation [23] and loss function. The architecture for deSpeckNet does not use any pooling layers to avoid using upsampling layers to reconstruct the images to their original sizes [19], as these lay additional computational burden. Instead, we opted to maintain the sizes of feature maps in the intermediate layers and increase the depth of the network.

To train the network, we apply three types of loss functions, the MSE-based L_{Clean}, L_{Noisy} losses, and a total variation (TV)

loss L_{TV}, combined as

$$\mathrm{Loss} = L_{\mathrm{Clean}} + L_{\mathrm{Noisy}} + L_{\mathrm{TV}}. \tag{4}$$

In the first, supervised training phase [Fig. 2(a)], we apply an MSE-based L_{Clean} loss between the clean label X and reconstructed clean image $\hat{X}$

$$L_{\mathrm{Clean}}(\mathbf{X}, \hat{\mathbf{X}}) = \mu \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2 \tag{5}$$

where n is the number of pixels in a training patch and μ is the weight assigned to the loss. Once $\hat{N}$ and $\hat{X}$ are reconstructed in the network, we apply an elementwise multiplication following the multiplicative noise model used for SAR images (1). The reconstructed noisy image $\hat{Y}$ is finally compared to the input noisy SAR image Y using another MSE loss, L_{Noisy} (Fig. 2)

$$L_{\mathrm{Noisy}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \xi \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{6}$$

Here, ξ is the weight assigned to the loss. The usage of this second loss function is important for providing a learning signal in the second phase, when no temporal average is available. This approach makes deSpeckNet different from the other DL based approaches for denoising SAR images.

In the second unsupervised fine-tuning phase [Fig. 2(b)], since a temporally averaged image is assumed to be unavailable, we use the input noisy image as a reference label and down-weight the L_{Clean} loss by a small value μ_2

$$L_{\mathrm{Clean},2}(\mathbf{Y}, \hat{\mathbf{X}}) = \mu_2 \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{X}_i)^2. \tag{7}$$

This has the effect of maintaining the solution close to the original image so that spatial structures are preserved while the other losses smooth (L_{TV}) and denoise (L_{Noisy}).

For smoothing, we used a TV [24] loss L_{TV} in order to encode a smoothness prior on the clean image

$$L_{\mathrm{TV}}(\hat{\mathbf{X}}) = \lambda \sum_{i} |\nabla \hat{X}_i|. \tag{8}$$

The L_{TV} loss minimizes the absolute differences between neighboring pixel-values, enforcing smoothness while preserving edges.

III. Data Sets

To evaluate the performance of deSpeckNet we used a Sentinel-1 GRD image time series with 23 images acquired over Pegunungan Barisan, Sumatra, Indonesia, to synthesize the reference labels and train the initial model. Since deSpeckNet is designed to adjust to different noise levels in different SAR images, the decision to train the model based on Sentinel-1 GRD is based on convenience for synthesizing a reference label image with a fewer number of images. To demonstrate the performance of deSpeckNet in despeckling SAR images without a temporally averaged label, and obtained using different sensors and across multiple regions, we fine-tune the model on the following (Fig. 3).

1) Images from different geographical regions and landcover types, we used three study areas, each are composed of a single Sentinel-1 GRD image acquired over the Kindu area in the DRC, the city of Utrecht and the region of Flevoland in The Netherlands, respectively.
2) Images from different sensors and landcover types. We used an ALOS-2/PALSAR-2 image acquired over Fujiyama, Japan, and an Iceye X2 image acquired near the city of Kiel in Germany.

A. Sentinel-1 Data

The Sentinel-1 GRD images used in the experiments are acquired in the interferometric wide (IW) swath mode with a technique known as terrain observation with progressive scan (TOPS). They were acquired in C-band for both single and dual polarization. The GRD images were multilooked to five looks in the range direction and projected to ground range using an Earth ellipsoid model by the data provider. The Sentinel-1 data sets used in this article are acquired from four regions (Table I).

1) *Indonesia:* The training image for deSpeckNet was acquired over the Pegunungan Barisan area in Sumatra, Indonesia. It consisted of an image with 1682×2300 pixels. The multitemporal images used to synthesize the training labels were acquired from July 5, 2018, to April 19, 2019. To assess the performance of deSpeckNet in tuning the network for a different region, we used a monotemporal images acquired on July 5, 2017. The area is mostly covered by oil palm plantations and forests. There are hardly any urban region within the area.
2) *DRC:* To assess the performance of deSpeckNet in tuning the network for a different region, we used a Sentinel-1 image acquired over Kindu in the DRC. The image was acquired on August 26, 2017, and it consists of an image with 1001×1001 pixels. This test area is mostly covered by primary forest with some bare soil. To synthesize the clean reference label for assessing the performance of deSpeckNet, we used 29 multitemporal images acquired over the same area from August 26, 2017, to August 9, 2018.
3) *Netherlands-Utrecht:* To assess the performance of deSpeckNet in tuning the network for a different region and landcover type, we used a Sentinel-1 image acquired over the city of Utrecht in The Netherlands. The image scene is dominated by urban areas. The images were acquired on October 11, 2018, consisting of 1360×2087 pixels. These images were selected to demonstrate the generalization capability of deSpeckNet in urban regions. We also used 22 multitemporal images acquired from October 11, 2018, to July 2, 2019, to evaluate the performance of deSpeckNet quantitatively.
4) *Netherlands-Flevoland:* To demonstrate the performance of deSpeckNet in a temporally unstable region, we used a Sentinel-1 image acquired over the Dutch region of Flevoland. The image was acquired on October 11, 2018, and is 1821×1204 pixels wide. Since it is an agricul-

TABLE I
ACQUISITION PARAMETERS FOR THE SENTINEL-1 TRAINING AND TEST IMAGES

Parameter	Indonesia	DRC	Utrecht	Flevoland
Polarization	VH	VH	VV	VH
Product	GRD	GRD	GRD	GRD
Acquisition mode	IW	IW	IW	IW
Resolution	$10\text{m} \times 10\text{m}$	$10\text{m} \times 10\text{m}$	$10\text{m} \times 10\text{m}$	$10\text{m} \times 10\text{m}$
Incidence angle	40.2^0	41.2^0	40.8^0	40.8^0
Orbit	Descending	Descending	Descending	Descending
Dates	July 05, 2018	August 26, 2017	October 11, 2018	October11, 2018
	April 19, 2019	August 9, 2018	July 2, 2019	-
Multi-temporal images	23	29	22	1

TABLE II
ACQUISITION PARAMETERS FOR THE ALOS-2-PALSAR-2 AND ICEYE X-2 TEST IMAGES

Parameter	ALOS-2	Iceye
Polarization	VH	VV
Sensor	PALSAR-2	X2
Product	L1.5	SLC
Acquisition mode	Strip map	Strip map
Resolution	$2.5\text{m} \times 2.5\text{m}$	$0.71\text{m} \times 1.48\text{m}$
Incidence angle	23.6^0	20.4^0
Orbit	Descending	Descending
Dates	June 06, 2014	April 29, 2019
Multi-temporal images	1	1

tural area, the temporally stable backscatter assumption could not be fulfilled to synthesize the reference labels and this case is assessed only qualitatively.

B. ALOS-2/PALSAR-2 Image

The ALOS2-PALSAR2 image is acquired in stripmap mode and is also multilooked two times in the azimuth direction and is projected to ground range using an Earth ellipsoid model by the data provider. The ALOS-2/PALSAR-2 sensor acquires data in L-band for both single, dual and quad polarization data (Table II). The monotemporal test image is acquired over the Fujiyama area in Japan on June 06, 2014. It consists of an image with 1060×1601 pixels. The area is a natural environment covered by forests and some bare areas. We selected this sensor and image to demonstrate the performance of deSpeckNet in adapting to a new geographic region and new sensor. It was not possible to freely acquire multitemporal images to synthesize the temporally averaged labels for quality evaluation, as it is a commercial sensor. Therefore, and as in the Flevoland case, we assess the results only qualitatively.

C. Iceye X2 Image

The Iceye X2 sensor acquires data in X-band for single polarization data in Strip map mode (Table II). The image is acquired near the city of Kiel in Northern Germany and is 1287×958 pixels. The area is a mixed scene of agricultural and urban regions. We selected this region to demonstrate the performance of deSpeckNet when applied to a high resolution image in a different region and landcover type. In this scene, it was also not possible to freely acquire multitemporal images to synthesize the temporally averaged labels for quality evaluation, as it is a commercial sensor.

IV. EXPERIMENTAL SETUP

A. Label

We use the reference label preparation method suggested in [13] that uses a large stack of multitemporal images to create the label clean image. Since this reference label is prepared from real SAR images, it represents the properties of real SAR features. To achieve the best results, the patches selected for training have to be stable in time, i.e., the scene must not have large temporal variation. This can be ensured by using the standard deviation of intensity for each pixel in the image as

$$Z = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \tilde{x})^2}{N-1}}. \tag{9}$$

Here, x is the pixel intensity, $\tilde{x}$ is the mean of the temporal pixel intensity value, N is the number of images in the temporal stack. If the standard deviation of pixel intensity values over the period is above a threshold ($\nu = 0.1$), that pixel in the temporally averaged image is masked. In places where we have a low standard deviation below a predefined threshold we can apply a temporal averaging on the image as

$$\mathrm{x} = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i, & \text{if } Z \leq 0.1 \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

In this way, we can prepare a temporal average of patches to be used as reference labels, whose number is detailed in Table I. In cases where the number of available multitemporal images is limited, a spatial multilooking is recommended to suppress residual speckle noise in the synthesized reference label image [25]. In our experiments, we have used 23 to 28 multitemporal Sentinel-1 GRD images that were originally multilooked five times, so we did not perform additional multilooking to synthesize the reference label images.

B. Training

To investigate the performance of deSpeckNet, we trained the network in a temporally stationary scene (S1-Indonesia) and to investigate the tuning capability of the designed architecture we applied it to an image acquired in different geographic regions and landcover types and sensors. In both

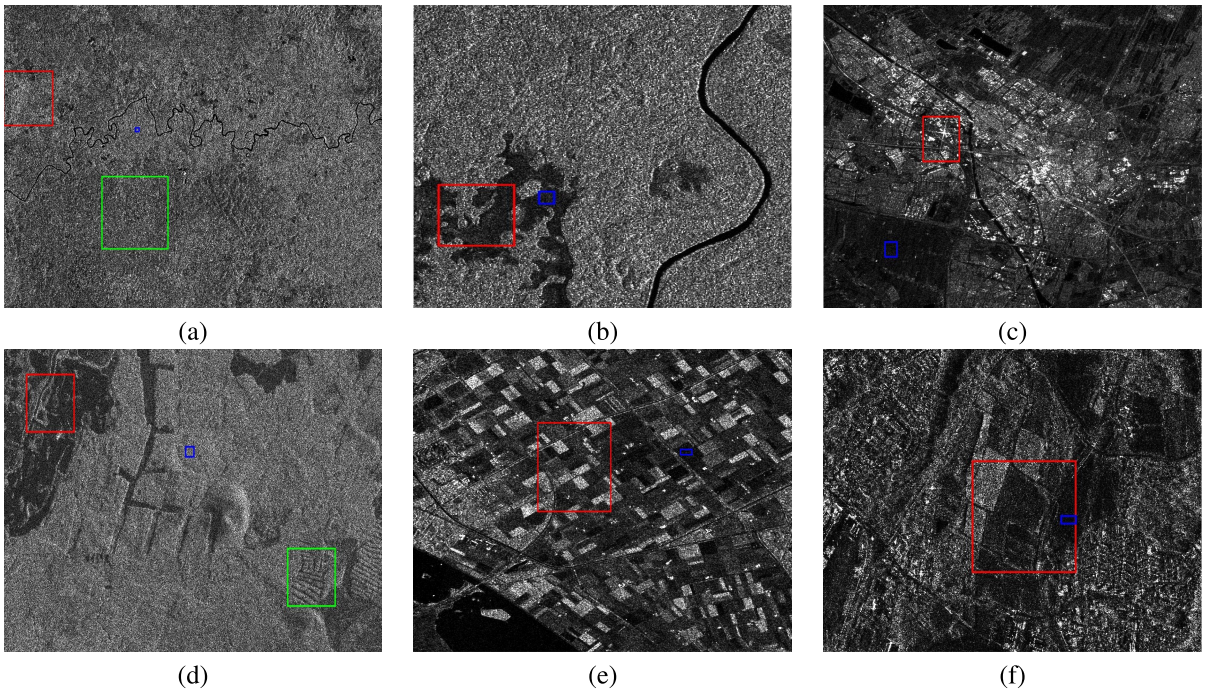


Fig. 3. Input images used to test deSpeckNet. The area in the red and green boxes was used for qualitative comparison of the methods, whereas the area in blue was used to estimate quality metrics such as ENL and coefficient of variation (Cx). (a) Indonesia. (b) DRC. (c) The Netherlands-Utrecht. (d) Japan. (e) The Netherlands-Flevoland. (f) Germany. All the test images used multilooked images except the Iceye image in Germany which was a single look image.

TABLE III
ARCHITECTURE OF THE FCN BLOCKS OF DESPECKNET

Layer	Module type	Dimension
Conv1	Conv ReLU	$3 \times 3 \times 1 \times 64$
Conv2 ($\times$15)	Conv BN ReLU	$3 \times 3 \times 64 \times 64$
Prediction	Conv	$3 \times 3 \times 64 \times 1$

$\text{FCN}_{\text{clean}}$ and $\text{FCN}_{\text{noise}}$, we used 17 blocks consisting of convolution, batch normalization and nonlinear activation, determined empirically. The details of each block are shown in Table III.

To train deSpeckNet in the initial phase, we prepared the input noisy images and their corresponding reference labels into 40×40 patches. We created an overall 117 888 patches for training. Since, deSpeckNet is designed to fine-tune images with different noise statistics than what it was trained on, there is no need for a training set that represents well the test set, and thus, the diversity of the training set becomes less of a limiting factor. We used a batch size of 128 so at every epoch the network used 921 iterations. We trained the network for a total of 30 epochs using a learning rate between 10^{-3} and 10^{-4} by decreasing the learning rate by 0.002 every 10 epochs. In the initial training phase, we set the weight (λ) of the L_{TV} to zero and the L_{Clean} was given a μ of 1 and ξ of L_{Noisy} was set to 0.01. To fine-tune the network for new regions or a new set of data we used the same learning rate as the initial training phase for one epoch. In the fine-tuning phase, for the Sentinel-1 GRD images we set the λ of the L_{TV} to 0 and the L_{Clean} was given a μ of 10^{-2} and ξ of L_{Noisy} was set to 1. For single look data sets such as the Iceye X2 images we set the λ of the L_{TV} to 10^{-4} and the L_{Clean} was given a μ of 10^{-2} and ξ of L_{Noisy} was set to 1. In both cases, we used an Adam optimizer [26] whose decay rate was fixed at 0.9. We used early stopping in fine-tuning the model to reconstruct the new images without reference labels. To properly evaluate the performance of the network we trained the network ten times from random seeds using the improved Xavier weight initialization [27].

To train deSpeckNet, we used the MatConvNet framework [28] in a MATLAB 2018a environment run on a Linux operating system with Intel Xeon(R) E-2176M CPU and Quadro P2000 GPU.

C. Quality Metrics

In test areas where we have multitemporal images and where the assumption of temporal stationarity was fulfilled, we used the temporally averaged images as validation ground-truth images to derive quality metrics. The quality metrics used to evaluate the performance of deSpeckNet in the presence of full reference label image are the peak signal to noise ratio (PSNR), structural similarity index (SSIM), despeckling gain (DG), edge preservation index (EPI) and the strong scatterer preservation index (C_{NN}). PSNR estimates the quality of the reconstructed noise-free image resemblance to the reference data, in this case, the temporally averaged Sentinel-1

image as

$$\text{PSNR} = 20\ \log_{10}\left(\frac{\hat{X}_{\max}}{\sqrt{\text{MSE}}}\right). \tag{11}$$

Here, $\hat{X}_{\max}$ is the maximum power given as 255. We do this by converting the 32-bit data to 8-bit data. The MSE is computed between the label (X) and the reconstructed images ($\hat{X}$) given as $\text{MSE} = \text{E}[(\hat{X} - X)^2]$.

SSIM estimates the structural similarity between the label (X) and the reconstructed image ($\hat{X}$) as

$$\text{SSIM}(\hat{X}, X) = \frac{(2\mu_{\hat{X}}\mu_X + c_1)(2\sigma_{\hat{X}X} + c_2)}{\left(\mu_{\hat{X}}^2 + \mu_X^2 + c_1\right)\left(\sigma_{\hat{X}}^2 + \sigma_X^2 + c_2\right)} \tag{12}$$

where μ_X is the mean of image X and σ_X is its standard deviation.

The DG estimates the speckle rejection ability of a particular despeckling method [29]. Therefore, a large DG value indicates a higher speckle removal ability. DG is estimated as follows:

$$\text{DG} = 10\log_{10}\left(\frac{\text{MSE}(\hat{X}, Y)}{\text{MSE}(\hat{X}, X)}\right). \tag{13}$$

To evaluate edge preservation, we use the EPI [30], [31]. EPI is derived by first defining the gradient preservation (GP) index, which is the ratio between the gradient values in the filtered intensity image ($\hat{X}$) and the gradient of the reference image (X)

$$\text{GP} = \frac{\sum \nabla_{\hat{X}}}{\sum \nabla_X}. \tag{14}$$

Here, ∇ is the Sobel gradient operator. EPI is calculated by projecting the GP values in the interval [0, 1] using a triangular equation as

$$\text{EPI} = \begin{cases} 1 - |1 - \text{GP}|, & \text{if GP} < 2 \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

The strong scatterer preservation index (C_{NN}) estimates the strong scatterer preservation ability of a particular filter [29]. Similarly, a higher C_{NN} indicates a higher preservation of a strong scatterer. C_{NN} is given as

$$C_{\text{NN}} = 10\log_{10}\frac{X_{\text{CF}}}{X_{\text{NN}}} \tag{16}$$

where X_{CF} is the intensity observed at the strong scatterer site and X_{NN} is the average intensity of the surrounding eight pixels.

In test regions where a temporal average image is not available, as a metric for quality of performance, we used visual inspection as a qualitative measure and the equivalent number of looks (ENL) and the coefficient of variation (C_x) derived in a homogeneous region as a quantitative measure for comparison. The ENL is derived by taking the ratio between the mean square (μ^2) and variance (σ^2) of a homogeneous region in the image as

$$\text{ENL} = \frac{\mu_{\hat{X}}^2}{\sigma_{\hat{X}}^2} \tag{17}$$

whereas C_x is a measure for the preservation of texture in the filtered image and is derived in a homogeneous region by taking the ratio of the standard deviation with the mean intensity, given as

$$C_x = \frac{\sigma_{\hat{X}}}{\mu_{\hat{X}}}. \tag{18}$$

As can be seen from (17) in uniform regions the filter that achieves the smallest variance in pixel intensities, i.e., best despeckling performance will result in the highest ENL and lowest C_x values.

V. Results and Discussion

To assess the performance of deSpeckNet, we compared it with the improved Lee-sigma filter [32], SAR-BM3D [9], an unsupervised method based on block-matching, and SAR-CNN [13], a supervised CNN-based method, both qualitatively and quantitatively. We selected SAR-CNN because it is designed to be trained on real SAR images as opposed to the methods trained on simulated noise and optical images.

A. Evaluation on the Same Region Used for Training

In the Indonesia image, a forested landscape for which a temporal average is available as clean reference, improved Lee sigma filter failed to remove speckle and to preserve the subtle features in the image. SAR-BM3D, being an unsupervised method that does not make use of the reference image and benefits from highly structured elements in the image to perform nonlocal matching, performed suboptimally in preserving edges and removing noise from homogeneous regions due to the lack of strong structures. deSpeckNet and SAR-CNN performed similarly since they are both trained with supervision on this image. Both are better than SAR-BM3D in preserving features and removing noise from homogeneous regions (Fig. 4 and Table IV).

B. Tests on Other Sentinel Scenes

1) Qualitative Results: On the remaining images no reference clean image is used to train SAR-CNN nor deSpeckNet. In the case of SAR-CNN, we apply the model that was trained on the Indonesia image [two image subset are presented in (Fig. 4)]. The Lee sigma filter and SAR-BM3D, being unsupervised, work on the same setting as in the Indonesia image. On the DRC image (Fig. 5), which is similar in nature to the one over Indonesia, both the improved Lee sigma, SAR-BM3D still perform suboptimally in preserving features and removing noise from homogeneous regions or preserving subtle features. SAR-CNN filtered the DRC image using the noise model learned from the Indonesia image. Although both are obtained with Sentinel-1, the differences in the noise distribution are large enough to result in blurred edges and unevenly filtered noise from homogeneous regions. deSpeckNet, however, was able to be tuned to the DRC image without using any new clean reference labels [Fig. 5(e)], removing much of the noise from homogeneous regions while preserving the edges between regions.

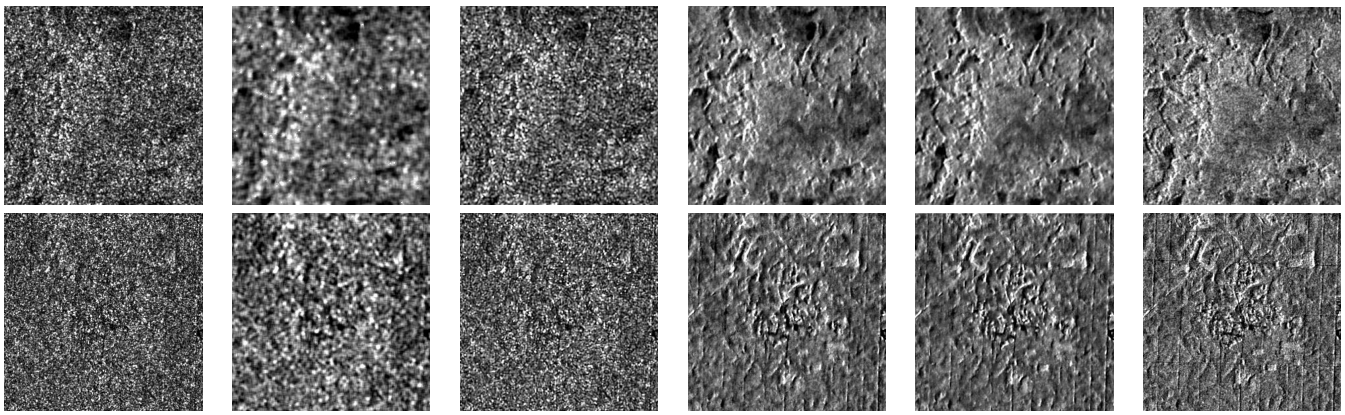


Fig. 4. Despeckling result for different baseline methods and deSpeckNet in the Indonesia Sentinel-1 image. We show a 300 × 300 and 400 × 400 image patch for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. (f) Temporal average image used to estimate PSNR/SSIM/DG. Some residual noise can be observed in the temporal average image because we averaged 23 images. However, due to the MSE loss function, its effect on the training performance is negligible.

TABLE IV

QUANTITATIVE QUALITY METRICS FOR ALL THE TEST IMAGES. SAR-CNN (*TEST) REFERS TO SAR-CNN FINE TUNED ON A TEMPORALLY AVERAGED IMAGE FOR THE TARGET SCENE AND IT IS USED AS AN UPPER BOUND. ALL RESULTS REPORT AN AVERAGE OVER TEN RUNS AND THE CORRESPONDING STANDARD DEVIATION. THE LEE SIGMA AND SAR-BM3D ARE NOT INITIALIZED RANDOMLY. HENCE, THE UNCERTAINTIES ARE NOT SHOWN

Test area	Metric	Lee Sigma	SAR-BM3D	SAR-CNN	deSpeckNet	SAR-CNN (*test)
Indonesia	PSNR	39.10	36.14	45.67 ± 0.07	**45.70** ± 0.05	45.67 ± 0.07
	SSIM	0.93	0.86	0.97 ± 0.00	**0.97** ± 0.00	0.97 ± 0.00
	DG	4.76	1.77	11.49 ± 0.07	**11.51** ± 0.05	11.49 ± 0.07
	EPI	0.20	0.86	0.90 ± 0.02	**0.92** ± 0.01	0.90 ± 0.02
	ENL	30.51	11.67	105.35 ± 4.53	**114.51** ± 7.12	105.35 ± 4.53
	Cx	0.18	0.29	**0.09** ± 0.002	0.09 ± 0.003	0.09 ± 0.002
DRC	PSNR	37.34	35.52	38.45 ± 0.00	**39.33** ± 0.01	39.91 ± 0.01
	SSIM	0.89	0.85	0.90 ± 0.00	**0.92** ± 0.00	0.92 ± 0.00
	DG	3.3	1.47	4.45 ± 0.06	**5.3** ± 0.008	5.82 ± 0.04
	EPI	0.31	0.84	0.74 ± 0.01	**0.92** ± 0.002	0.91 ± 0.01
	ENL	41.53	11.04	27.72 ± 5.26	**118.21** ± 17.99	150.29 ± 38.93
	Cx	0.18	0.30	0.18 ± 0.026	**0.09** ± 0.008	0.07 ± 0.01
NL-Utrecht	PSNR	26.89	28.02	19.88 ± 1.75	**28.14** ± 0.32	29.97 ± 0.50
	SSIM	0.77	0.81	0.66 ± 0.00	**0.84** ± 0.00	0.85 ± 0.00
	DG	-0.3	0.82	-7.25 ±1.87	**0.94** ± 0.32	1.72 ± 0.52
	EPI	0.53	**0.97**	0.88 ± 0.01	0.87 ± 0.03	0.97 ± 0.02
	ENL	100.66	80.47	57.36 ± 20.69	**270.20** ± 98.90	454.54 ± 139.52
	Cx	**0.009**	0.15	0.13 ± 0.017	0.062 ± 0.0087	0.04 ± 0.007
	C_{NN}	**2.84**	2.57	2.83 ± 0.41	2.08 ± 0.14	2.84 ± 0.08
Japan	ENL	60.83	42.45	8.66 ± 0.21	**114.95** ± 25.39	-
	Cx	0.11	0.10	0.31 ± 0.008	**0.09** ± 0.004	-
NL-Flevoland	ENL	41.40	12.12	5.56 ± 2.59	**132.37** ± 40.05	-
	Cx	0.15	0.28	0.44 ± 0.07	**0.08** ± 0.008	-
Germany	ENL	29.36	5.69	3.43 ± 0.13	**65.86** ± 1.18	-
	Cx	0.18	0.41	0.53 ± 0.00	**0.12** ± 0.00	-

In the Netherlands-Utrecht image (Fig. 6), the improved Lee sigma filter overfiltered and distorted all features in the image. Whereas, SAR-BM3D performed better than in the previous images as the image contrast was higher than the Indonesia and DRC case. However, it suffered from overfiltering, resulting in the smoothing out of subtle features. SAR-CNN resulted in suboptimal results as it failed to adequately remove the noise from homogeneous regions and improve the overall signal to noise ratio of the image. deSpeckNet succeeded in preserving features and adequately removing noise from homogeneous regions [Fig. 6(d)] when compared to SAR-CNN. The advantage of deSpeckNet was also demonstrated when applied to the image in the Netherlands-Flevoland. Here, the improved Lee sigma filter overfiltered the scene resulting in blurred features and SAR-BM3D was not able to remove the speckle maintaining the noisy appearance, whereas SAR-CNN indiscriminately filtered the image distorting many of the image features. On the contrary, deSpeckNet removed the speckle

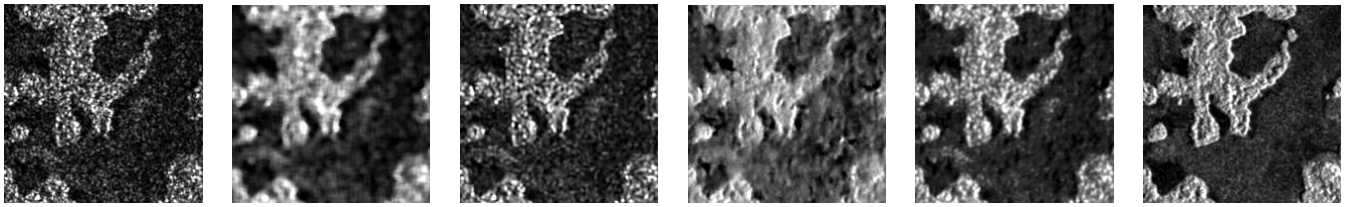


Fig. 5. Despeckled images in the DRC Sentinel-1 image. We show a 200 × 200 image patch for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. (f) Temporal average image used to estimate PSNR/SSIM/DG. This area is selected to demonstrate the generalization capability of the model in a similar landcover types from what it was trained on but different geographic region.

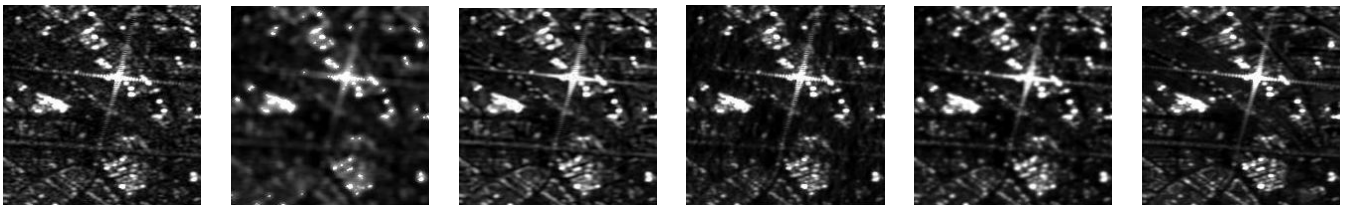


Fig. 6. Despeckled images in the Netherlands-Utrecht Sentinel-1 image. We show a 200 × 200 tile for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. (f) Temporal average image used to estimate PSNR/SSIM/DG. This area is selected to demonstrate the scaleability of the model in a similar sensor type from what it was trained on but different geographical regions and landcover types. The temporal average image displayed is used only for deriving the quality metrics. [In Fig. 6(f), for visualization purposes we did not mask pixels that did not fulfill the temporal standard deviation criteria.]

Fig. 7. Despeckled images in the Netherlands-Flevoland Sentinel-1 image. We show a 350 × 350 image patch for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. This area is selected to demonstrate the generalization capability of the model in a similar sensor type from what it was trained on but different landcover type.

from homogeneous regions while maintaining subtle features in the image [Fig. 7(e)].

2) Quantitative Results: The capability of deSpeckNet is further exemplified by the improvement of the quantitative metrics of PSNR, SSIM, DG, EPI and ENL when compared with the improved Lee sigma, SAR-BM3D, and SAR-CNN. The ENL is estimated in a minimum window size of 25 × 55 pixel window for the Flevoland test area and a maximum of 80 × 115 pixel window for the Japan test area Fig. 3. In the Indonesia and DRC test areas deSpeckNet achieved the highest PSNR, SSIM, DG, EPI, and ENL values than SAR-BM3D and SAR-CNN (Table IV). This trend was slightly changed when comparing the quantitative results from the Netherlands-Utrecht image. Here, SAR-BM3D achieved a slightly higher EPI and C_{NN} when compared with deSpeckNet. This is due to the fact that SAR-BM3D is better adapted to highly structured images with distinct strong scatterers, such as this urban scene. However, deSpeckNet results in a higher PSNR, SSIM and DG value than SAR-BM3D, which suggests that the latter might be incurring in inconsistencies in filtering, to which the localized EPI and C_{NN} metric is less sensitive. SAR-CNN achieved overall low values in all compared metrics due to its inability to adapt to new noise distributions. A similar trend was also observed in the Netherlands-Flevoland image (Table IV), where deSpeckNet achieved significantly higher ENL values and the smallest C_x than all the other baseline methods. The improved Lee sigma filter achieved the lowest PSNR, SSIM, DG, EPI, and ENL values in all test images showing a sharp contrast between the traditional localized speckle filters and machine learning-based filters. To establish the upper bound for tuning, we did a supervised tuning of SAR-CNN in the DRC and Netherlands-Utrecht image by using the temporally averaged image. The SAR-CNN supervised fine-tuning achieved a mean PSNR of 39.91, SSIM of 0.92 and DG of 5.82 in the DRC and a mean PSNR of 29.97, SSIM of 0.85 and DG of 1.72 in

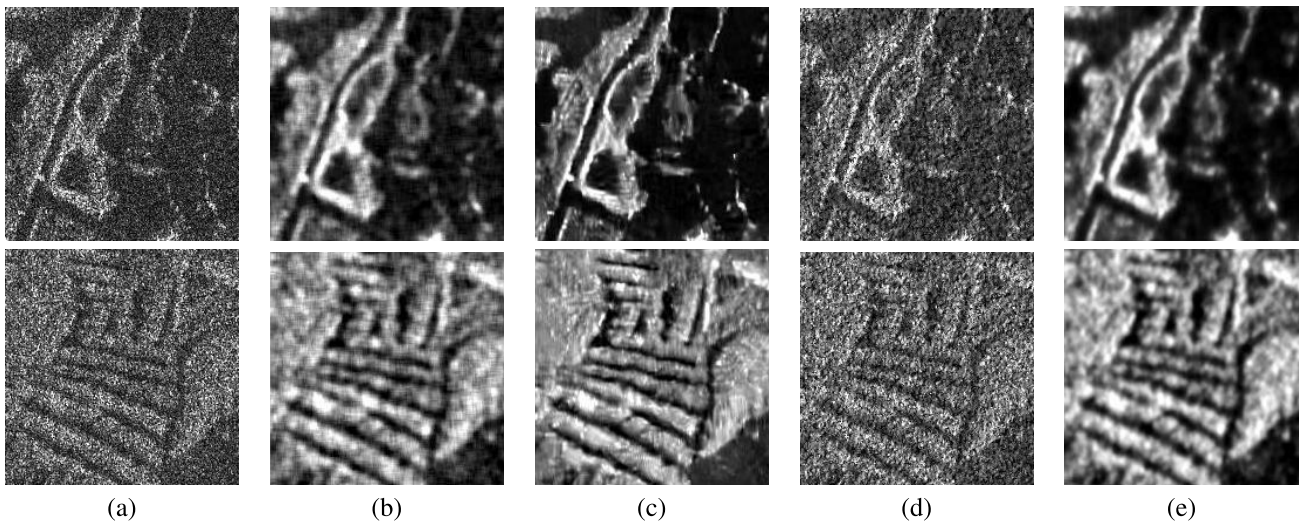


Fig. 8. Despeckled images in the Japan ALOS2-PALSAR2 image. We show two 200 × 200 image patches for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. This area is selected to demonstrate the generalization capability of the model in a similar landcover types from what it was trained on but different sensor.

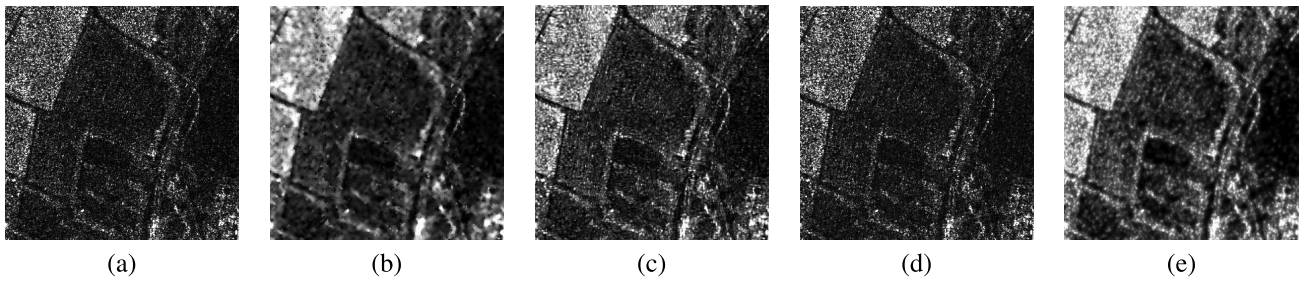


Fig. 9. Despeckled images in the Kiel Iceye x-2 image. We show a 350 × 350 image patch for (a) input noisy image, (b) Lee-sigma, (c) SAR-BM3D, (d) SAR-CNN, and (e) deSpeckNet. This area and data type is selected to demonstrate the generalization capability of the model in a different sensor types, geographic region and landcover type.

the Netherlands-Utrecht image (Table IV), which was slightly higher than the values achieved by deSpeckNet. Note that these results show that deSpeckNet, ever without using supervised tuning on the test image, is able to reach an equivalent performance of a network tuned in a clean test image, which, the more often, is not available.

C. Tests on Other Sensors

To further demonstrate the capability of deSpeckNet in generalizing to new sensors, we used the ALOS-2 PALSAR-2 image acquired in Japan. Visually the performance of SAR-BM3D was better than the Indonesia and DRC images, due to sharper contrast and structure in the scene. However, there was a severe loss of resolution and texture in the image that resulted from overfiltering and some spurious details were hallucinated by the filter as a result of destruction of the texture in the image. The improved Lee sigma filter blurred all features achieving suboptimal results and SAR-CNN had a suboptimal performance due to the difference in the noise distribution. In contrast, deSpeckNet was able to improve the SAR signal to noise ratio while preserving subtle features in the image (Fig. 8). Same conclusions were reached when considering to the high resolution Iceye X2 image in Germany. In this case also, deSpeckNet performed better at removing noise from homogeneous regions (Table IV) while preserving subtle features in the image (Fig. 9).

D. Noise Estimation

One of the advantages of deSpeckNet is the estimation of the speckle noise distribution. This plays an important role in tuning the model to a different set of images. This can be confirmed by investigating the probability density function of the estimated noise along with the parameters that define the noise probability density function. We do so by fitting a distribution (2) to the estimated noise image. As can be observed from (Fig. 10), the probability density function of the noise follows a Gamma distribution for all test areas except the Iceye Germany test area. Here, as opposed to the other test case we used a single look SAR image hence the speckle noise distribution was not a Gamma distribution as the other test cases but an exponential pdf and deSpeckNet was able to estimate the noise pdf accurately Fig. 10(f). In addition, we also compared the reference image ratio (Y/X with the ratio image ($Y/\hat{X}$) estimated by deSpeckNet and the other methods (Fig. 11). From Fig. 11, we can clearly see that deSpeckNet inherently have a limitation in yielding uncorrelated speckle in the image, which is manifested by residual image structures in the ratio images. These artifacts in the ratio

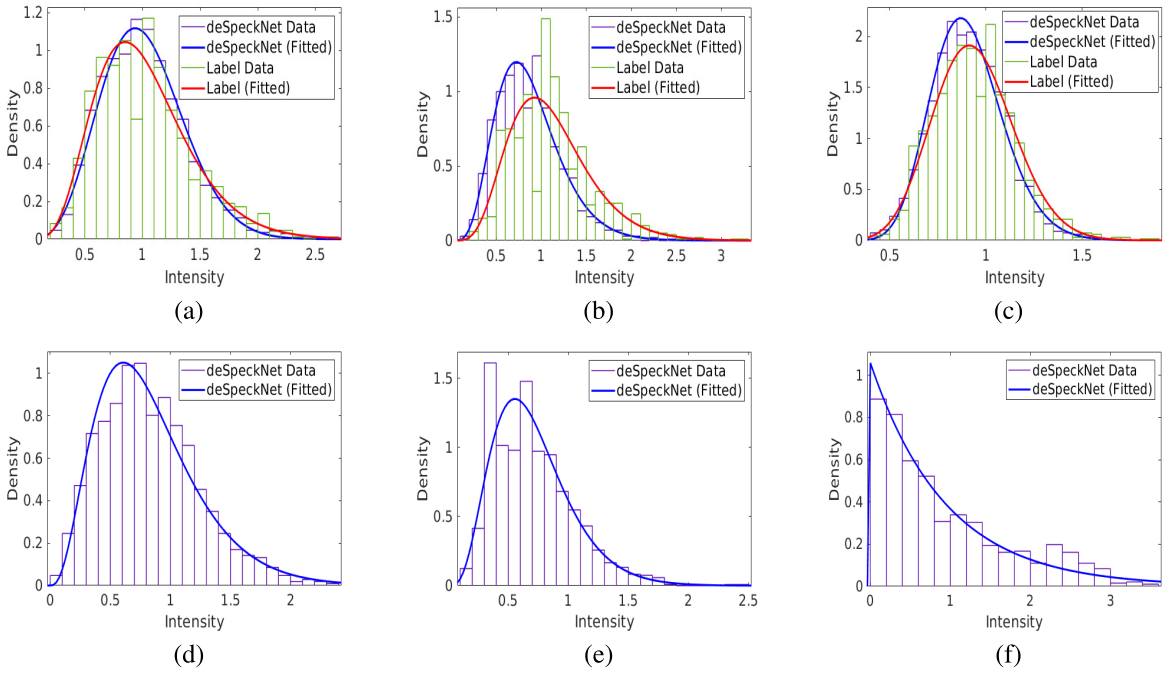


Fig. 10. Noise distribution estimated by deSpeckNet and the speckle noise derived from the reference data for each of the test areas. (a) and (b) DRC. (c) The Netherlands-Utrecht. (d) Japan. (e) The Netherlands-Flevoland. (f) Germany. All the test areas were from multilooked images except the Iceye image in Germany which was a single look image.

TABLE V

COMPARISON OF MEAN (MOR) AND VARIANCE (VOR) FOR THE RATIO IMAGE ($Y/\hat{X}$) SYNTHESIZED BY THE DIFFERENT METHODS AND THE REFERENCE RATIO IMAGE (Y/X) THAT IS SYNTHESIZED FROM THE TEMPORALLY AVERAGED REFERENCE IMAGE

Test area	Metric	Lee Sigma	SAR-BM3D	SAR-CNN	deSpeckNet	Reference
Indonesia	MoR	0.98	0.95	1.03	1.05	1.05
	VoR	0.13	0.03	0.15	0.16	0.18
DRC	MoR	0.98	0.95	1.01	0.93	1.02
	VoR	0.15	0.05	1.14	0.10	0.17
NL-Utrecht	MoR	0.98	0.97	1.41	0.96	0.95
	VoR	0.04	0.01	0.07	0.02	0.04
Japan	MoR	0.99	0.89	0.93	0.96	-
	VoR	0.23	0.16	0.07	0.23	-
NL-Flevoland	MoR	0.97	0.96	0.91	0.83	-
	VoR	0.11	0.02	6.37	0.18	-
Germany	Mor	0.98	0.78	1.11	1.04	-
	VoR	0.48	0.15	0.003	0.54	-

images are the result of using input images as a reference label in phase II to preserve features in the filtered output. To get a deeper insight into the performance of the methods we used the mean of ratio (MoR) and variance of ratio (VoR) metric to compare the performance of the baseline methods. In the Indonesia, DRC and Netherlands-Utrecht images, deSpeckNet provides the closest estimate of MoR and VoR estimates to that of the reference ratio image derived from the temporally averaged reference image X (Table V).

E. Computational Considerations

The overall computational cost of deSpeckNet when training the initial model for 30 epochs was 31.8 h. This was twice the computational burden of SAR-CNN, due to the duplicity of the CNN blocks in the Siamese architecture. In contrast, SAR-BM3D took 16 min to denoise the input image in Indonesia. However, when tuning the model to new images the model was able to be tuned within one epoch for the DRC and Japan image and two epochs for the Netherlands image. This amounted to 7.7×10^{-4} seconds per pixel to tune the model, as the computational burden depends on the dimension of the input image. When testing the model on an image it had a computational burden of 2.69×10^{-5} s per pixel. All training was performed on GPU whereas, due to memory limitations, all testing was performed using CPU.

Even though the computational cost of deSpeckNet was relatively heavy in the initial training phase, it was able to be tuned to new images with a relatively small amount of time. This makes deSpeckNet particularly desirable for operational applications that require fast response such as near-real time change monitoring applications.

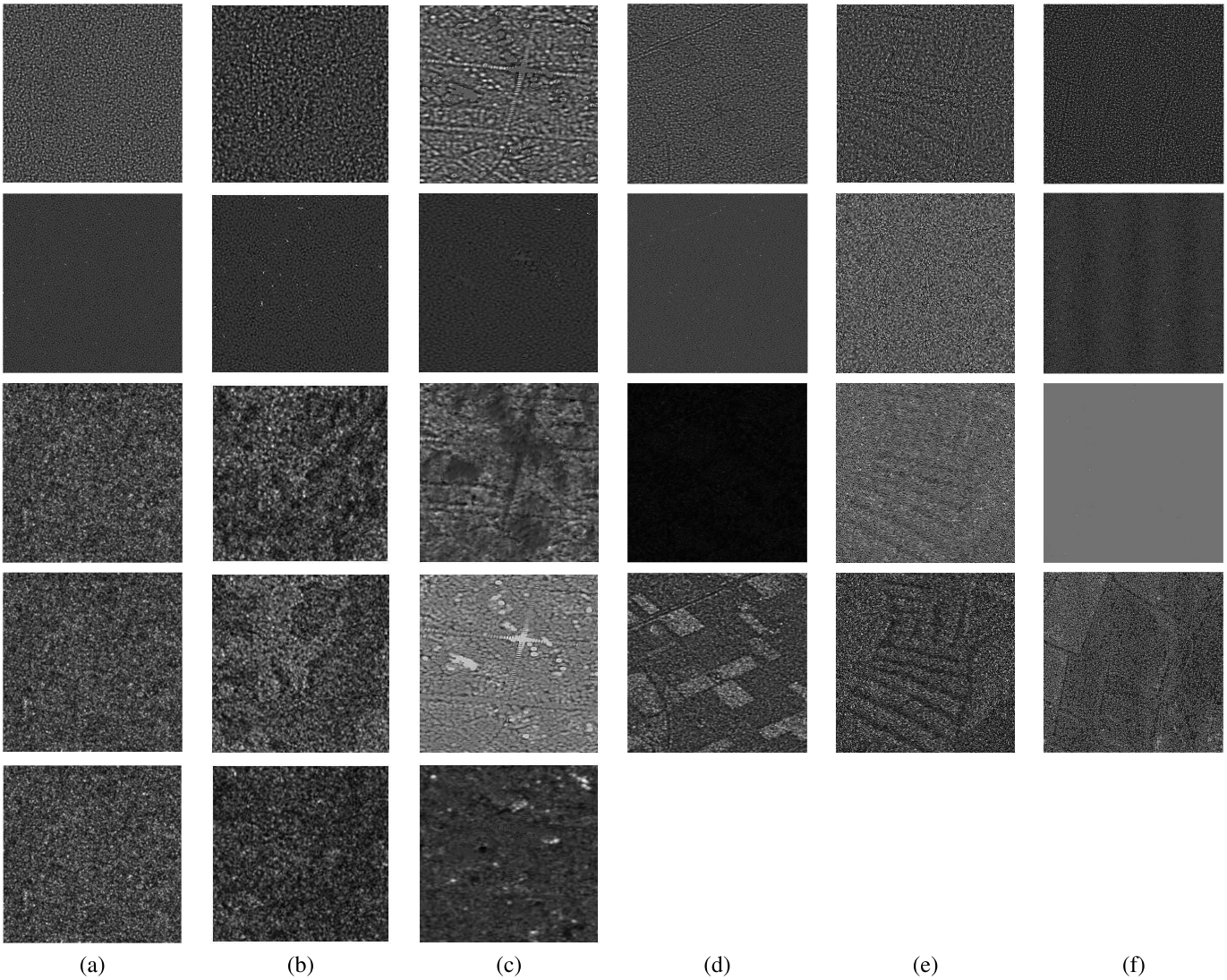


Fig. 11. Comparison of ratio images ($Y/\hat{X}$) for (Top row) Improved Lee Sigma, (Bottom row) SAR-BM3D, SAR-CNN, deSpeckNet, and reference ratio (Y/X). (a) Indonesia, (b) Congo, (c) The Netherlands-Utrecht, (d) The Netherlands-Flevoland, (e) Japan, and (f) Germany. Since no clean image is available for (d)–(f), the last row is empty for those images. To ease visual comparison, the ratio images were rescaled to the range [0.5 1.5].

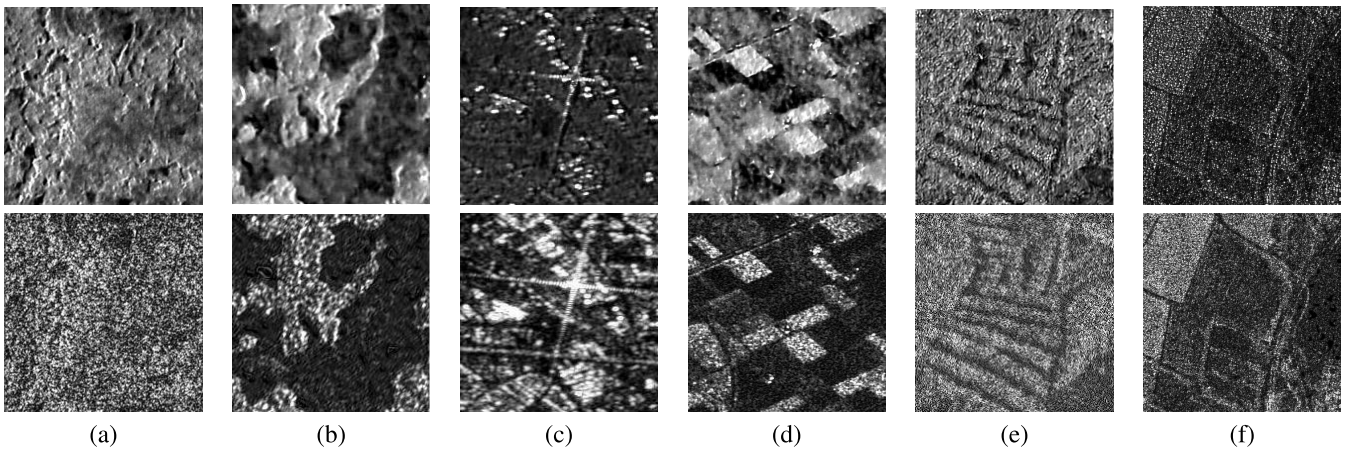
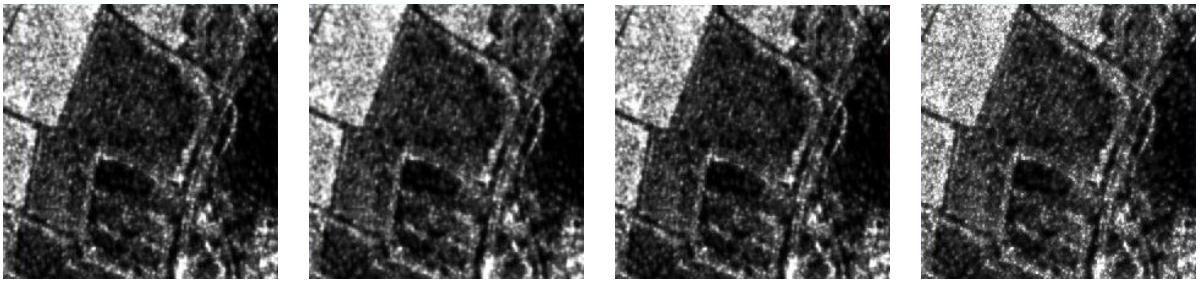


Fig. 12. Comparison of processing output from (Top) phase one only and (Bottom) phase two only. (a) Indonesia. (b) Congo. (c) The Netherlands-Utrecht. (d) The Netherlands-Flevoland. (e) Japan. (f) Germany.

F. Ablation Study

In this section, we study the importance of the building blocks of the proposed architecture. We first show the necessity of the two phases, the training on the stack of multitemporal images (phase 1) and the unsupervised fine-tuning on the test image (phase 2), with respect to the full



(a) (b) (c) (d)

Fig. 13. Effect of TV loss in tuning a single look Iceye image. We show a 350×350 image patch for (a) tuning with $\lambda = 10^{-3}$, ENL $= 182.38$, (b) tuning with $\lambda = 10^{-5}$, ENL $= 165.7$, (c) tuning with $\lambda = 10^{-6}$, ENL $= 164.11$, and (d) tuning with $\lambda = 0$, ENL $= 98.66$. Tuning is done in one epoch.

TABLE VI

RESULTS OF DESPECKNET, WHEN USING (TOP) PHASE 1 ONLY, (MIDDLE) PHASE 2 ONLY, AND (BOTTOM) BOTH PHASES 1 AND 2 (PROPOSED METHOD)

	Image	PSNR	SSIM	DG	EPI	ENL	Cx	C_{NN}
Phase 1 only	Indonesia	45.76	0.97	11.51	0.93	101.83	0.09	-
	DRC	38.45	0.90	4.42	0.65	26.03	0.19	-
	NL-Utrecht	18.80	0.63	-8.39	0.76	62.90	0.12	2.50
	NL-Flevoland	-	-	-	-	14.98	0.44	-
	Japan	-	-	-	-	10.71	0.3	-
	Germany	-	-	-	-	3.92	0.5	-
Phase 2 only	Indonesia	35.45	0.83	1.0	0.75	10.36	0.31	-
	DRC	35.32	0.84	1.27	0.66	56.58	0.13	-
	NL-Utrecht	24.39	0.69	-3.26	0.52	1.78	0.74	1.41
	NL-Flevoland	-	-	-	-	12.83	0.27	-
	Japan	-	-	-	-	6.38		
	Germany	-	-	-	-	1.47		
Phase 1 and 2	Indonesia	45.76	0.97	11.51	0.92	101.83	0.09	-
	DRC	39.32	0.92	5.3	0.91	121.10	0.09	-
	NL-Utrecht	27.60	0.83	0.94	0.84	242.61	0.062	2.08
	NL-Flevoland	-	-	-	-	127.49	0.08	-
	Japan	-	-	-	-	158.21	0.09	-
	Germany	-	-	-	-	65.22	0.12	-

TABLE VII

PSNR, SSIM, EPI AND ENL COMPUTED USING DIFFERENT WEIGHTS FOR THE RESPECTIVE LOSS FUNCTIONS FOR TRAINING THE NETWORK USING THE IMAGE IN INDONESIA. HERE μ IS THE WEIGHT OF L_{Clean}, λ IS THE WEIGHT OF L_{TV} AND ξ IS THE WEIGHT OF L_{Noisy}

μ	λ	ξ	PSNR	SSIM	DG	EPI	ENL	Cx
1	10^{-2}	10^{-2}	38.62	0.94	4.59	**0.95**	62.65	0.12
1	0	10^{-2}	**45.76**	**0.97**	11.51	0.93	101.83	0.09
1	0	10^{-1}	45.66	0.97	11.48	0.91	111.61	0.09
1	0	1	44.23	0.96	9.99	0.80	**165.76**	**0.07**
1	10^{-3}	10^{-3}	45.41	0.97	11.21	0.91	112.49	0.09
1	0	10^{-3}	45.70	0.97	11.52	0.90	97.71	0.10
1	10^{-5}	10^{-3}	45.64	0.97	11.45	0.94	112.83	0.09
1	10^{-7}	10^{-3}	45.66	0.97	11.47	0.93	103.91	0.09

model using both (Table VI and Fig. 12). From these results, its clear that the combination of the two phases is crucial to achieving higher performance in despeckling. Next, we performed a series of ablation studies focused on the loss functions. As shown in (Tables VII and VIII) the usage of L_{TV} in the initial training phase was not important to remove some artifacts and blurring effects found in the reconstructed image when the L_{Clean} loss is applied (Table VII). However, when tuning the model on noisier, single look images, its presence was important to further smoothen noisy homogenous regions in the image to be tuned by forcing the $\text{FCN}_{\text{noise}}$ part of the network in estimating the speckle component in the image. This is exemplified by the increase in the ENL when applying the TV loss on an Iceye single look image (Fig. 13). To evaluate the necessity of applying a loss function in the $\text{FCN}_{\text{clean}}$ side of the network, we removed both the L_{Clean} and the L_{TV} to train the network but it failed to achieve the output demonstrated when using, only the L_{Clean} loss.

In general, deSpeckNet achieved success in generalizing to different areas. It achieved higher success when applied to an

TABLE VIII

PSNR, SSIM, DG, EPI, ENL AND CX COMPUTED USING DIFFERENT WEIGHTS FOR THE RESPECTIVE LOSS FUNCTIONS FOR TUNING THE NETWORK USING THE TEST IMAGE IN THE DRC. HERE μ IS THE WEIGHT OF L_{Clean}, λ IS THE WEIGHT OF L_{TV} AND ξ IS THE WEIGHT OF L_{Noisy}

μ	λ	ξ	PSNR	SSIM	DG	EPI	ENL	Cx
10^{-2}	10^{-2}	1	26.55	0.07	-33.93	0.23	8.43	-0.01
10^{-2}	0	1	**39.32**	**0.92**	**5.3**	**0.92**	121.1	**0.09**
1.5×10^{-2}	10^{-5}	1	39.07	0.91	5.05	0.61	133.89	0.08
1.5×10^{-2}	10^{-4}	1	38.65	0.91	4.63	0.63	70.85	0.11
1.5×10^{-2}	10^{-6}	1	39.02	0.91	4.99	0.61	147.66	0.08
1.5×10^{-3}	10^{-6}	1	38.04	0.89	4.01	0.80	**227.46**	0.06
1.5×10^{-1}	10^{-4}	1	37.24	0.88	3.20	0.89	33.92	0.17
1.5	10^{-3}	1	36.90	0.87	2.86	0.78	42.84	0.15

image in both rural and urban scenes. This is attributed to the multiplicative model assumption enforced in deSpeckNet and the use of the input noisy image as a reference with a small weight.

VI. CONCLUSION

We have presented a method, deSpeckNet, that is able to learn a speckle noise model suitable for effective despeckling without the need of any reference clean image nor any assumptions on the noise distribution other than the multiplicative noise model. Our experiments on a wide variety of SAR images, obtained with different sensors and over different regions, confirm the robustness of deSpeckNet.

The proposed deSpeckNet proved to be effective in reducing speckle noise while preserving the image quality with minimal unsupervised fine-tuning. It was also able to adapt to all the tested SAR images regardless of resolution, acquisition parameters or geographical region, providing better despeckling results than state-of-the-art methods and equaling performance obtained by CNN models optimized with temporally averaged images, generally unavailable, at test time. For future work, we plan on improving the loss functions that encode the assumption on the clean image to improve the performance of deSpeckNet in mixed urban and rural scenes where both strong deterministic and distributed targets exist.

REFERENCES

[1] L. J. Porcello, N. G. Massey, R. B. Innes, and J. M. Marks, "Speckle reduction in synthetic-aperture radars," *J. Opt. Soc. Amer.*, vol. 66, no. 11, pp. 1305–1311, Nov. 1976.

[2] J.-S. Lee, "Digital image enhancement and noise filtering by use of local statistics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–2, no. 2, pp. 165–168, Mar. 1980.

[3] J.-S. Lee, K. W. Hoppel, S. A. Mango, and A. R. Miller, "Intensity and phase statistics of multilook polarimetric and interferometric SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 5, pp. 1017–1028, 1994.

[4] V. S. Frost, J. A. Stiles, K. S. Shanmugan, and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–4, no. 2, pp. 157–166, Mar. 1982.

[5] J.-S. Lee, M. R. Grunes, and G. de Grandi, "Polarimetric SAR speckle filtering and its implication for classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 5, pp. 2363–2373, Dec. 1999.

[6] G. Vasile, E. Trouve, J.-S. Lee, and V. Buzuloiu, "Intensity-driven adaptive-neighborhood technique for polarimetric and interferometric SAR parameters estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 6, pp. 1609–1621, Jun. 2006.

[7] M. R. Grunes, D. L. Schuler, E. Pottier, and L. Ferro-Famil, "Scattering-model-based speckle filtering of polarimetric SAR data," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 1, pp. 176–187, Jan. 2006.

[8] C.-A. Deledalle, L. Denis, F. Tupin, A. Reigber, and M. Jager, "NL-SAR: A unified nonlocal framework for resolution-preserving (Pol)(In)SAR denoising," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2021–2038, Apr. 2015.

[9] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[10] D. Espinoza Molina, D. Gleich, and M. Datcu, "Evaluation of Bayesian despeckling and texture extraction methods based on Gauss–Markov and auto-binomial gibbs random fields: Application to TerraSAR-X data," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 2001–2025, May 2012.

[11] M. Mahdianpari, B. Salehi, and F. Mohammadimanesh, "The effect of PolSAR image de-speckling on wetland classification: Introducing a new adaptive method," *Can. J. Remote Sens.*, vol. 43, no. 5, pp. 485–503, Sep. 2017.

[12] A. Lopes, E. Nezry, R. Touzi, and H. Laur, "Maximum a posteriori speckle filtering and first order texture models in SAR images," in *Proc. 10th Annu. Int. Symp. Geosci. Remote Sens.*, 1990, pp. 2409–2412.

[13] G. Chierchia, D. Cozzolino, G. Poggi, and L. Verdoliva, "SAR image despeckling through convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5438–5441.

[14] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[15] Q. Zhang, Q. Yuan, J. Li, Z. Yang, and X. Ma, "Learning a dilated residual network for SAR image despeckling," *Remote Sens.*, vol. 10, no. 2, p. 196, Jan. 2018.

[16] S. Vitale, G. Ferraioli, and V. Pascazio, "A new ratio image based CNN algorithm for SAR despeckling," 2019, *arXiv:1906.04111*. [Online]. Available: http://arxiv.org/abs/1906.04111

[17] T. Pan, D. Peng, W. Yang, and H.-C. Li, "A filter for SAR image despeckling using pre-trained convolutional neural network model," *Remote Sens.*, vol. 11, no. 20, p. 2379, Oct. 2019.

[18] C.-A. Deledalle, L. Denis, S. Tabti, and F. Tupin, "MuLoG, or how to apply Gaussian denoisers to multi-channel SAR speckle reduction?" *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4389–4403, Sep. 2017.

[19] F. Lattari, B. Gonzalez Leon, F. Asaro, A. Rucci, C. Prati, and M. Matteucci, "Deep learning for SAR image despeckling," *Remote Sens.*, vol. 11, no. 13, p. 1532, Jun. 2019.

[20] F. Ulaby and D. Long, *Microwave Radar and Radiometric Remote Sensing*. Norwood, MA, USA: Artech House, 2015.

[21] A. G. Mullissa, C. Persello, and A. Stein, "PolSARNet: A deep fully convolutional network for polarimetric SAR image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 12, pp. 5300–5309, Dec. 2019.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[23] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8609–8613.

[24] G. Aubert and J.-F. Aujol, "A variational approach to removing multiplicative noise," *SIAM J. Appl. Math.*, vol. 68, no. 4, pp. 925–946, Jan. 2008.

[25] W. Zhao, C.-A. Deledalle, L. Denis, H. Maitre, J.-M. Nicolas, and F. Tupin, "Ratio-based multitemporal SAR images denoising: RABASAR," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3552–3565, Jun. 2019.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[27] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[28] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[29] G. Di Martino, M. Poderico, G. Poggi, D. Riccio, and L. Verdoliva, "Benchmarking framework for SAR despeckling," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1596–1615, Mar. 2014.
[30] A. G. Mullissa, V. Tolpekin, and A. Stein, "Scattering property based contextual PolSAR speckle filter," *Int. J. Appl. Earth Observ. Geoinfor.*, vol. 63, pp. 78–89, Dec. 2017.
[31] S. Foucher and C. López-Martínez, "Analysis, evaluation, and comparison of polarimetric SAR speckle filtering techniques," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1751–1764, Apr. 2014.
[32] J.-S. Lee, J.-S. Lee, J.-H. Wen, T. L. Ainsworth, K.-S. Chen, and A. J. Chen, "Improved sigma filter for speckle filtering of SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 1, pp. 202–213, Jan. 2009.

Adugna G. Mullissa (Member, IEEE) received the Ph.D. degree in radar remote sensing from the University of Twente, Enschede, The Netherlands, in December 2017.

From September 2016 to January 2017, he was a Visiting Scientist at the Lyle School of Civil Engineering, Purdue University, West Lafayette, IN, USA. From January 2018 to January 2019, he was a Researcher at the University of Twente, investigating deep-learning methodologies for crop classification using polarimetric SAR data. He is currently a Researcher on near real-time deforestation monitoring using SAR images and machine learning at the Laboratory of Geoinformation Science and Remote Sensing, Wageningen University, Wageningen, The Netherlands. His research interests include microwave remote sensing, pattern recognition, and machine learning.

Diego Marcos received the M.Sc. degree in computational sciences and engineering from EPFL, Lausanne, Switzerland, in 2014. He developed his Ph.D. between the universities of Zurich and Wageningen on the interface between remote sensing and computer vision.

He is a Post-Doctoral Researcher with Wageningen University, Wageningen, The Netherlands. His main research interests are interpretable machine learning and its application to the environmental sciences.

Devis Tuia (Senior Member, IEEE) received the Ph.D. degree from the University of Lausanne, Lausanne, Switzerland, in 2009.

He held a post-doctoral position at the University of Valéncia, Valencia, Spain, the University of Colorado, Boulder, CO, USA, and EPFL Lausanne, Lausanne. From 2014 to 2017, he was an Assistant Professor with the Department of Geography, University of Zurich, Zurich, Switzerland. He was then a Professor of geoinformation science at Wageningen University, Wageningen, The Netherlands. Since 2020, he has been an Associate Professor at EPFL. He is interested in algorithms for information extraction and data fusion of remote sensing images using machine learning. More info on https://sites.google.com/site/devistuia/

Martin Herold was born in 1975. He received the Ph.D. degree from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2004, and the Habilitation degree on a topic on operational global land cover observation and assessments from Friedrich-Schiller-Universität Jena (FSU), Jena, Germany, in 2009.

He is the Chair for geoinformation science and remote sensing at Wageningen University, Wageningen, The Netherlands. He is an expert in the development and implementation of land change and biomass monitoring systems using novel earth observation technologies and approaches and in application contexts of the UNFCCC and the Sustainable Development Goals. He has published more than 200 scientific papers, has been a lead author for refining the IPCC GPG for GHG inventories (2018/2019), and is recognized as a Web of Science/Publons Highly Cited Researcher in 2019 and 2020. He enjoys supervising students and evolving scientists and supporting capacity development initiatives for moving innovative satellite and ground-based approaches into sustainable and climate-smart land use practice.

Johannes Reiche received the Ph.D. degree from Wageningen University, Wageningen, The Netherlands, in 2015.

He is an Assistant Professor in radar remote sensing at the Laboratory of Geo-information Science and Remote Sensing, Wageningen University. His research interest is on utilizing radar remote sensing to unravel human activities and dynamics of forest ecosystems with a strong focus on multisensor methods, machine learning, and near real-time change monitoring.

Dr. Reiche is an Advisory Board Member of the World Resource Institute Global Forest Watch Program and a Scientific Member of the JAXA Kyoto and Carbon Initiative and Global Forest Observation Initiative.