Multi-temporal speckle reduction with self-supervised deep neural networks

Inès Meraoumia, Emanuele Dalsasso, Loïc Denis, Rémy Abergel, Florence Tupin

Multi-temporal speckle reduction with self-supervised deep neural networks

Inès Meraoumia, Emanuele Dalsasso, Loïc Denis, *Senior Member, IEEE,* Rémy Abergel, and Florence Tupin *Senior Member, IEEE,*

***Abstract*—Speckle filtering is generally a prerequisite to the analysis of synthetic aperture radar (SAR) images. Tremendous progress has been achieved in the domain of single-image despeckling. Latest techniques rely on deep neural networks to restore the various structures and textures peculiar to SAR images. The availability of time series of SAR images offers the possibility of improving speckle filtering by combining different speckle realizations over the same area.**

The supervised training of deep neural networks requires ground-truth speckle-free images. Such images can only be obtained indirectly through some form of averaging, by spatial or temporal integration, and are imperfect. Given the potential of very high quality restoration reachable by multi-temporal speckle filtering, the limitations of ground-truth images need to be circumvented. We extend a recent self-supervised training strategy for single-look complex SAR images, called MERLIN, to the case of multi-temporal filtering. This requires modeling the sources of statistical dependencies in the spatial and temporal dimensions as well as between the real and imaginary components of the complex amplitudes.

Quantitative analysis on datasets with simulated speckle indicates a clear improvement of speckle reduction when additional SAR images are included. Our method is then applied to stacks of TerraSAR-X images and shown to outperform competing multi-temporal speckle filtering approaches.

The code of the trained models and supplementary results are made freely available at https://gitlab.telecom-paris.fr/ring/multi-temporal-merlin/.

***Index Terms*—SAR, image despeckling, deep learning, self-supervised training.**

I. Introduction

Earth Observation requires diverse information that can be captured with complementary remote sensing systems. Synthetic Aperture Radar (SAR) is an active sensor widely used in applications ranging from ocean and forest monitoring, land use and human activity monitoring, to the estimation of digital elevation models [1].

However, interpreting SAR images is particularly challenging because of the presence of strong fluctuations in the backscattered intensities: due to the coherent sum of the contributions of all scatterers located within the same resolution cell, constructive or destructive interferences occur, leading to the so-called *speckle* phenomenon. SAR image analysis is greatly simplified when speckle fluctuations are reduced in a pre-processing step.

Speckle reduction has been tackled by various approaches, from methods based on the selection of pixels with similar intensities [2], to techniques based on wavelet decompositions [3] or non-local filtering [4], and, more recently, significant progress was achieved with deep neural networks [5]–[7].

Training a neural network for speckle reduction requires the definition of a loss function which reflects the performance of the network on the training set. In a *supervised training* setting, this loss function characterizes the proximity of the network prediction to a ground truth image: the speckle-free image corresponding to the ideal output. Building a training set with matching pairs of speckle-free and corrupted SAR images is difficult. Starting from a corrupted SAR image, creating the corresponding speckle-free ground truth has no ideal solution (in fact, this is our ultimate goal). Speckle-free images can be approximated either by another modality (e.g., optical remote sensing developed in [8], natural images) or by computing the temporal mean of a long time series of SAR images [9]. Once a speckle-free image is selected, a corrupted version can be produced by drawing samples from a theoretical distribution of speckle. To prevent any domain shift between the training and testing phases, speckle simulation has to accurately capture the actual speckle fluctuations observed in SAR images, in particular its spatial correlations. An alternative is to define a *self-supervised training* loss, i.e., a loss function relating the network estimation to other observations. SAR2SAR [10] extends to speckle reduction the Noise2Noise principle [11]: the denoised image should be close, on average, to other independent noisy observations of the same scene. Because of changes occurring between image acquisitions, special care must be taken to compensate these changes. NL-SAR-DL [12] also exploits multiple images of the same area to perform a self-supervised training with a weighted loss to dismiss changed areas. Single-image self-supervised training is also possible. Speckle2Void [13] follows the blind spot methodology introduced in [14] that excludes the central pixel from the network estimation in order to drive the training step (by minimizing the statistical distance between the speckled central pixel and the despeckled network prediction based solely on the surrounding area). Rather than spatially splitting the input image into blind spots and surrounding areas, MERLIN [15] splits the Single Look Complex (SLC) input image into the real and imaginary parts to define the self-supervised loss.

Time series offer more information to reduce speckle fluctu-

I. Meraoumia, E. Dalsasso and F. Tupin are with LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France, e-mail: forename.name@telecom-paris.fr.

L. Denis is with Univ Lyon, UJM-Saint-Etienne, CNRS, Institut d Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France, e-mail: loic.denis@univ-st-etienne.fr.

R. Abergel is with the laboratoire MAP5, UMR CNRS 8145, Université Paris Cité, France, e-mail: remy.abergel@parisdescartes.fr.

ations than a single SAR image. Multi-temporal averaging can be largely improved by compensating for changes, as proposed in Quegan filter [16], up to a limit depending on the length of the time series and the quality of single-image restorations used for change suppression. Successful single-image despeckling techniques have been extended to multi-temporal data: SAR-BM3D [17], based on collaborative filtering of blocks of similar patches, also considers patches located at other dates in the multi-temporal extension [18]; the two-step multi-temporal non-local means [19] perform weighted averages along the temporal or spatial dimensions based on patch similarities [20]. RABASAR [21] proposes to compute first a "super-image" by temporally multi-looking the image stack (this super-image has almost no residual speckle fluctuations) and then process ratio images in which only speckle and changes with respect to the super-image are remaining. The content of these ratio images is largely simplified and thus easier to restore. The final despeckled images are obtained after multiplication by the super-image. To despeckle the ratio image, a deep neural network such as SAR2SAR can be used [22]. A drawback of ratio-based processing is that the lowest-contrasted structures present either in the speckled image or in the super-image might be improperly restored, leading to the suppression of these details or the apparition of a "ghost" structure leaking from the super-image.

Our contributions: We show how a deep neural network can be trained end-to-end to produce a despeckled image from a time series of co-registered SAR images. This is made possible by the use of a self-supervised loss function [15], bypassing the impossibility to access to high-quality ground truth images. Compared to simpler strategies based only on a single date enriched by a higher signal-to-noise multi-temporal average (a super-image) [21], we feed the network with all available dates. This leaves all freedom to the network to perform optimal temporal combinations, leading to improved restorations even when only a few additional images are included.

Our method is grounded on a generative model of speckle that accounts for fully-developed speckle areas, the presence of dominant scatterers due to man-made structures, interferometric coherence (both temporal and geometrical decorrelation phenomena), and the spatial correlations induced by the SAR transfer function.

The theoretical framework of the method is developed in section II. A numerical study is then performed on data with simulated speckle to characterize the performance of the method. The approach is then tested on stacks of TerraSAR-X Stripmap images.

II. Proposed approach: multi-temporal MERLIN

To derive a self-supervised training strategy in the context of multi-temporal filtering, we start by building a generative model of speckle in paragraph II-A. We then discuss in paragraph II-B conditions under which a component of the reference date, statistically independent from the rest of the data, can be set aside in order to drive the training of deep neural networks. In section II-C we describe our unsupervised training strategy and the network architecture choices.

Scalar and vector notations:		
j	$\mathbb{C}$	imaginary unit
$\boldsymbol{z}$	$\mathbb{C}^{TN}$	representation of a stack of T N-pixels images
$\boldsymbol{z}(\cdot,k)$	$\mathbb{C}^{T}$	vector of values at pixel k
$\boldsymbol{z}(t,\cdot)$	$\mathbb{C}^{N}$	t-th image of the stack
$\boldsymbol{z}_t$	$\mathbb{C}^{N}$	t-th image of the stack (compact notation)
$\boldsymbol{z}_{\text{ref}}$	$\mathbb{C}^{N}$	image at date t_{ref}, the date to restore
Scene parameters:		
$\boldsymbol{d}$	$\mathbb{C}^{TN}$	dominant scatterers
$\boldsymbol{r}$	$\mathbb{R}_{+*}^{TN}$	reflectivities of speckled areas
Speckle field:		
$\boldsymbol{\epsilon}$	$\mathbb{C}^{TN}$	uncorrelated speckle
$\boldsymbol{\Gamma}_k$	$\mathbb{C}^{T\times T}$	speckle coherence matrix at pixel k
$\mathbf{L}_k$	$\mathbb{C}^{T\times T}$	correlating operator such that $\mathbf{L}_k\mathbf{L}_k^\dagger=\boldsymbol{\Gamma}_k$
$\mathbf{L}$	$\mathbb{C}^{TN\times TN}$	correlating operator for the full stack
Complex amplitudes on the radar antenna:		
$\boldsymbol{s}$	$\mathbb{C}^{TN}$	complex amplitude of the speckled component
$\boldsymbol{z}$	$\mathbb{C}^{TN}$	resultant complex amplitude: $\boldsymbol{z}=\boldsymbol{s}+\boldsymbol{d}$
$\tilde{\boldsymbol{z}}$	$\mathbb{C}^{TN}$	complex amplitude including SAR system effects
Acquisition specific parameters:		
$\boldsymbol{\varphi}_t$	$\mathbb{C}^{N}$	atmospheric, topographic, and displacement phase effects at each pixel of the t-th image
$\boldsymbol{\psi}_t$	$\mathbb{C}^{N}$	phase ramp corresponding to the spectrum shift due to angular discrepancies
$\mathbf{Q}$	$\mathbb{C}^{N\times N}$	SAR response (spectral apodization and 0-padding)
$\mathbf{H}_t$	$\mathbb{C}^{N\times N}$	SAR response (spectral apodization, 0-padding+shift)
Pre-processing step to enforce statistic independence:		
$\dot{\boldsymbol{z}}$	$\mathbb{C}^{TN}$	complex amplitudes with recentered power spectrum
$\gamma_{ij}(k)$	$\mathbb{C}$	complex correlation coefficient (i.e., coherence) between $\dot{\boldsymbol{z}}(t_i,k)$ and $\dot{\boldsymbol{z}}(t_j,k)$
$\mathbf{W}_k$	$\mathbb{C}^{2\times 2}$	whitening matrix at pixel k
$\mathbf{W}$	$\mathbb{C}^{2N\times 2N}$	whitening operator for a pair of images
$\mathring{\boldsymbol{z}}$	$\mathbb{C}^{TN}$	complex amplitudes after whitening
Self-supervised training:		
$\mathring{\boldsymbol{a}}_{\text{ref}}$	$\mathbb{C}^{N}$	real part of pre-processed image at date t_{ref}
$\mathring{\boldsymbol{b}}_{\text{ref}}$	$\mathbb{C}^{N}$	imaginary part of pre-processed image at date t_{ref}
$\mathcal{L}_{\text{MERLIN}}$		self-supervised loss function
$\tilde{\boldsymbol{r}}_{\text{ref}}$	$\mathbb{R}_{+*}^{N}$	low-pass filtered reflectivities at date t_{ref}
$\dot{\boldsymbol{d}}_{\text{ref}}$	$\mathbb{C}^{N}$	low-pass filtered dominant scatterers at date t_{ref}

Table I
MAIN NOTATIONS AND CORRESPONDING DIMENSIONS.

A. Generative speckle model of multi-temporal SLC stacks

The ability to partition the data into two mutually independent sets is central to our self-supervised training strategy. It is thus necessary to model the different sources of speckle correlations arising in multi-pass SAR imaging. If the images are acquired in interferometric conditions, then the speckle remains partially coherent from one pass to the next. Otherwise, the speckle is fully decorrelated and multi-temporal filtering can be very effective.

We consider here a more general speckle model than in [15], to account for the mix present in SAR images between (i) areas that follow Goodman's fully developed model (coherent summation of many similar elementary phasors), composed of rough surfaces and scattering volumes, and (ii) regions where the complex amplitude is mainly defined by the magnitude and phase of dominant scatterers. To include both phenomena, we model a stack $\boldsymbol{z}\in\mathbb{C}^{TN}$ of T SLC SAR images, each with N pixels, as the superimposition of two components: a *speckle component* $\boldsymbol{s}\in\mathbb{C}^{TN}$, driven by a reflectivity map $\boldsymbol{r}\in\mathbb{R}_{+*}^{TN}$,

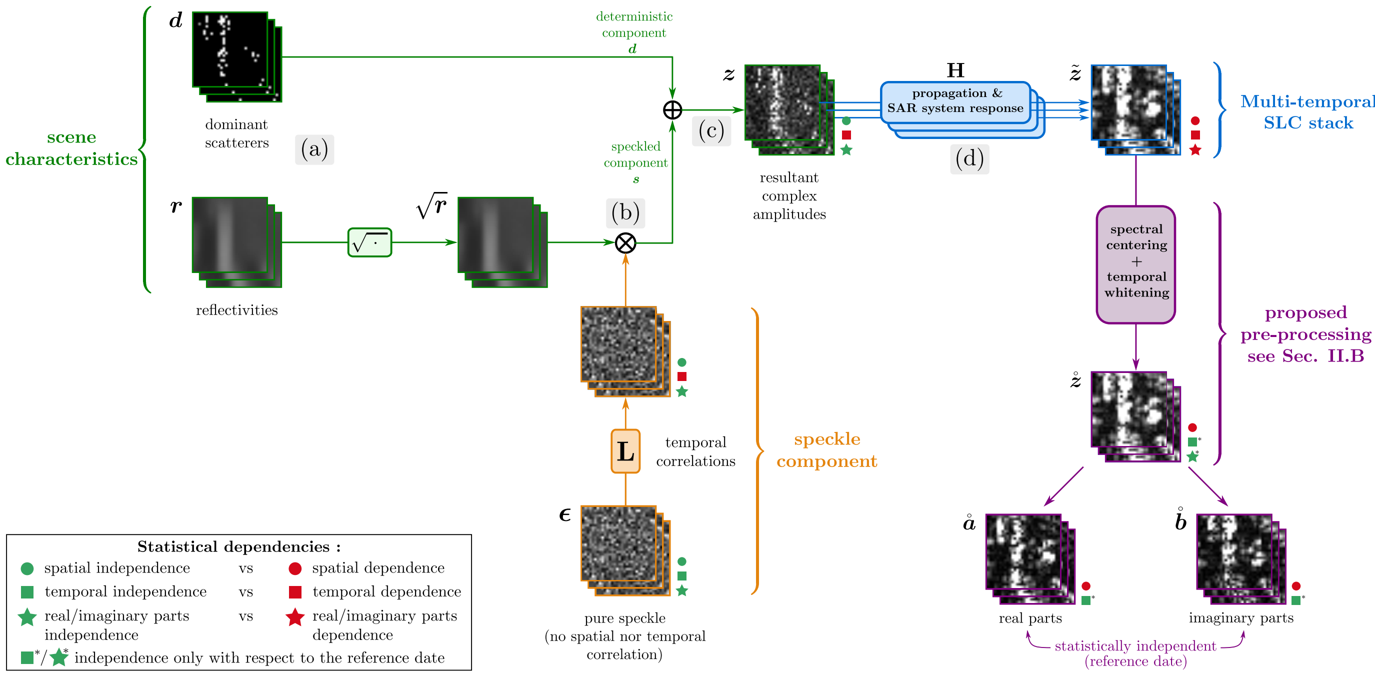

Figure 1. Generative model of speckle in multi-temporal SLC stacks of SAR images.

and the *dominant scatterers component* $\boldsymbol{d} \in \mathbb{C}^{TN}$, see Figure 1(a).

In the following, the multi-temporal stacks will be represented in the form of a column vector (e.g., $\boldsymbol{z} \in \mathbb{C}^{TN}$), by concatenation of the T images, and both the image at date t (noted $\boldsymbol{z}(t,\cdot) \in \mathbb{C}^{N}$, or $\boldsymbol{z}_t$ in compact form) and the vector of complex amplitudes at pixel k for all dates (noted $\boldsymbol{z}(\cdot,k) \in \mathbb{C}^{T}$) will be considered. A permutation matrix $\boldsymbol{\Pi}$ can be applied to transform the vector $\boldsymbol{z}$ from an ordering according to a scan of all pixels for each date, one date after another, to an ordering according to a scan of all dates for a given pixel, before moving to the next pixel:

$$\boldsymbol{\Pi}\boldsymbol{z} = \boldsymbol{\Pi}\begin{pmatrix}\boldsymbol{z}(t_1,\cdot)\\ \vdots \\ \boldsymbol{z}(t_T,\cdot)\end{pmatrix} = \begin{pmatrix}\boldsymbol{z}(\cdot,k_1)\\ \vdots \\ \boldsymbol{z}(\cdot,k_N)\end{pmatrix}. \tag{1}$$

Table I summarizes the main notations used in the paper.

According to Goodman's model [23], the *speckle component* $\boldsymbol{s}(\cdot,k) \in \mathbb{C}^{T}$ at pixel k follows a complex circular Gaussian distribution $\mathcal{N}_c(\boldsymbol{\Sigma}_k)$ defined by

$$\mathrm{p}(\boldsymbol{s}(\cdot,k)|\boldsymbol{\Sigma}_k) = \frac{1}{\pi^T \det(\boldsymbol{\Sigma}_k)} \exp\left[-\boldsymbol{s}(\cdot,k)^{\dagger}\,\boldsymbol{\Sigma}_k^{-1}\,\boldsymbol{s}(\cdot,k)\right], \tag{2}$$

where $\cdot^{\dagger}$ denotes the conjugate transpose; $\boldsymbol{\Sigma}_k$ is the speckle covariance matrix at pixel k. Note that $\boldsymbol{\Sigma}_k = \mathrm{diag}(\sqrt{\boldsymbol{r}(\cdot,k)})\boldsymbol{\Gamma}_k\mathrm{diag}(\sqrt{\boldsymbol{r}(\cdot,k)})$ with $\boldsymbol{\Gamma}_k$ the coherence matrix (the entries verify $|\boldsymbol{\Gamma}_k(t_i,t_j)| \leq 1$ for all t_i and t_j and $\boldsymbol{\Gamma}_k(t,t) = 1$ for all t), $\boldsymbol{r} \in \mathbb{R}_{+*}^{TN}$ is the vector of reflectivities, and the square root is applied entry-wise. The coherence matrices characterize how the temporal evolution of the scene decorrelates the speckle. Starting from a pure speckle $\boldsymbol{\epsilon}_k \in \mathbb{C}^{T}$, with no correlation along the spatial and the temporal axis ($\boldsymbol{\epsilon}_k \sim \mathcal{N}_c(\mathbf{I})$), a multiplication by the matrix $\mathbf{L}_k$, where $\mathbf{L}_k\mathbf{L}_k^{\dagger} = \boldsymbol{\Gamma}_k$ (e.g., $\mathbf{L}_k$ is a Cholesky factor of coherence matrix $\boldsymbol{\Gamma}_k$), gives a random vector that follows the distribution $\mathcal{N}_c(\boldsymbol{\Gamma}_k)$. Thus, the speckled component can be generated from $\boldsymbol{\epsilon}_k$ (Figure 1(b)):

$$\boldsymbol{s}(\cdot,k) = \mathrm{diag}(\sqrt{\boldsymbol{r}(\cdot,k)})\mathbf{L}_k\boldsymbol{\epsilon}_k\,. \tag{3}$$

The vector $\boldsymbol{s} \in \mathbb{C}^{TN}$ that concatenates all T images one after another can be obtained by

$$\boldsymbol{s} = \begin{pmatrix}\boldsymbol{s}(t_1,\cdot)\\ \vdots \\ \boldsymbol{s}(t_T,\cdot)\end{pmatrix} = \mathrm{diag}(\sqrt{\boldsymbol{r}})\,\underbrace{\boldsymbol{\Pi}^{-1}\begin{pmatrix}\mathbf{L}_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathbf{L}_N\end{pmatrix}\boldsymbol{\Pi}}_{\mathbf{L}}\,\boldsymbol{\epsilon}\,. \tag{4}$$

The covariance matrix of the *speckled component* $\boldsymbol{s}$ is block diagonal after a proper permutation

$$\mathrm{Cov}[\boldsymbol{s}] = \mathrm{diag}(\sqrt{\boldsymbol{r}})\boldsymbol{\Pi}^{-1}\begin{pmatrix}\boldsymbol{\Gamma}_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \boldsymbol{\Gamma}_N\end{pmatrix}\boldsymbol{\Pi}\mathrm{diag}(\sqrt{\boldsymbol{r}})\,, \tag{5}$$

which shows that correlations are only along the temporal axis of the spatio-temporal stack.

The *dominant scatterers component* $\boldsymbol{d} \in \mathbb{C}^{TN}$ contains non zero values only at pixels with dominant scatterers. Such scatterers may appear or disappear at some point in the time series.

The SLC amplitudes of the scene $\boldsymbol{z}$ then correspond to the superimposition of the two components: $\boldsymbol{z} = \boldsymbol{s} + \boldsymbol{d}$, see Figure 1(c). We model the effects of the atmospheric phase, the topographic (and possibly displacement) phase of the speckle component [24], and the spectral response of the SAR system as follows (Figure 1(d))

$$\tilde{\boldsymbol{z}} = \begin{pmatrix} \tilde{\boldsymbol{z}}(t_1,\cdot) \\ \vdots \\ \tilde{\boldsymbol{z}}(t_T,\cdot) \end{pmatrix} = \begin{pmatrix} \mathbf{H}_1 \mathrm{diag}(\exp(j\boldsymbol{\varphi}_1)) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{H}_T \mathrm{diag}(\exp(j\boldsymbol{\varphi}_T)) \end{pmatrix} \boldsymbol{z}\,, \tag{6}$$

where $\tilde{\boldsymbol{z}}$ is the complex amplitude that includes these effects, $\mathbf{H}_t \in \mathbb{C}^{N\times N}$ is the SAR response for the t-th acquisition, and $\boldsymbol{\varphi}_t = \boldsymbol{\varphi}_{\mathrm{atmo}_t} + \boldsymbol{\varphi}_{\mathrm{topo}_t} + \boldsymbol{\varphi}_{\mathrm{disp}_t} \in \mathbb{C}^N$ combining the different sources of phase modification. The spectral response of the SAR system is generally identical for all passes, up to a 2D shift due to angular discrepancies (incidence and possibly squint angle differences between acquisitions). Linear operators $\mathbf{H}_t$, $1 \leq t \leq T$, can thus be written $\mathbf{H}_t = \mathrm{diag}(\exp(-j\boldsymbol{\psi}_t))\mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\psi}_t))$, where $\mathbf{Q} \in \mathbb{R}^{N\times N}$ is the real-valued operator (in spatial domain), corresponding to a spectral response (in Fourier domain) that is symmetrical and centered on the 0 frequency (i.e., 0 Doppler), and the phase vector $\boldsymbol{\psi}_t$ is the 2D ramp corresponding to this 2D shift in Fourier domain (accounting for the angular discrepancies at pass t). The complex amplitudes of the t-th pass can be rewritten

$$\tilde{\boldsymbol{z}}_t = \mathrm{diag}(\exp(-j\boldsymbol{\psi}_t))\mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_t + j\boldsymbol{\psi}_t))\boldsymbol{z}_t\,. \tag{7}$$

The linear operator $\mathbf{Q}$ accounts for the spectral apodization introduced to reduce the sidelobes of strong scatterers and a possible over-sampling (0-padding in Fourier domain), both inducing a low-pass filtering effect on SAR images that does not depend on t.

Since the multi-temporal stack $\tilde{\boldsymbol{z}}$ is generated from $\boldsymbol{\epsilon}$ through a series of linear operations, $\tilde{\boldsymbol{z}}$ is also Gaussian distributed with a mean equal to $\tilde{\boldsymbol{d}}$, where for each date t the subvector $\tilde{\boldsymbol{d}}(t,\cdot) \in \mathbb{C}^N$ is equal to $\mathbf{H}_t\mathrm{diag}(\exp(j\boldsymbol{\varphi}_t))\boldsymbol{d}_t$, the low-pass filtered dominant scatterers component, and a covariance given at the bottom of page 5. Complex values in $\tilde{\boldsymbol{z}}$ are both spatially and temporally correlated.

B. Achieving statistical independence of the real/imaginary component at date t_{ref}

The principle of the self-supervised training proposed in [15], called MERLIN, consists of splitting the real and imaginary components of a single-date SLC image and exploiting their statistical independence. Two differents tasks can be considered when extending speckle reduction to multi-temporal stacks: (i) the multiple-input single-output (MISO) framework where multiple dates are provided in input but only a single image at a reference date t_{ref} is restored; (ii) the multiple-input multiple-output (MIMO) framework that restores at once all the dates provided in the input multi-temporal stack. In the following, we follow the MISO approach depicted in Figure 2 for two reasons:

- the requirement of *statistical independence* with respect to the inputs of the network is easier to achieve when a single output is considered;
- in order for a MIMO network to output very different images in case of large changes, several *independent paths* must emerge within the network architecture, which requires a *huge network capacity* (i.e., many parameters) [25] and a *careful initialization* to avoid getting stuck in poor quality local minima during training (as observed in our preliminary experiments).

In our MISO multi-temporal approach, we provide the network with the multi-temporal SLC stack of T images where the real part (or imaginary part) of the reference date t_{ref} is excluded. This excluded component is then used to supervise the training under the assumption that it is statistically independent from the inputs (where the reflectivities $\boldsymbol{r}$ and dominant scatterers $\boldsymbol{d}$ are considered deterministic and only the speckle $\boldsymbol{\epsilon}$ is random). Two preprocessing steps are required to ensure this independence.

First, the shift of the SAR system response in the spectral domain at date t_{ref} induces correlations between real and imaginary components at this date[1]. A simple pre-processing step can be applied to recenter the spectrum of the image at the reference date around the 0 frequency by multiplication by the 2D phase ramp $\exp(j\boldsymbol{\psi}_{t_{\mathrm{ref}}})$. In order to preserve interferometric coherence, we apply the same spectral shift to all dates (so that the relative shift between Fourier spectra remains unchanged). We denote the centered complex amplitudes by $\dot{\boldsymbol{z}}$, defined by

$$\forall t,\, \dot{\boldsymbol{z}}(t,\cdot) = \mathrm{diag}(\exp(j\boldsymbol{\psi}_{t_{\mathrm{ref}}}))\tilde{\boldsymbol{z}}(t,\cdot) \tag{8}$$

where the phase ramp $\boldsymbol{\psi}_{t_{\mathrm{ref}}}$ required to recenter the spectrum can be estimated from the power spectrum of image $\tilde{\boldsymbol{z}}(t_{\mathrm{ref}},\cdot)$. This leads to the following simplified expression at t_{ref}:

$$\dot{\boldsymbol{z}}(t_{\mathrm{ref}},\cdot) = \mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_{t_{\mathrm{ref}}} + j\boldsymbol{\psi}_{t_{\mathrm{ref}}}))\boldsymbol{z}_{t_{\mathrm{ref}}}\,. \tag{9}$$

Second, a whitening step may be necessary to address the correlations along the temporal axis, depending both on the coherence matrices $\boldsymbol{\Gamma}_k$ (modeling how temporal decorrelations affect the scene) and the shifts induced by the phases $\boldsymbol{\psi}_t$ (modeling geometric decorrelation according to the interferometric baselines). In the context of multi-temporal speckle filtering, the stronger the correlations along the temporal dimension, the less useful the additional images. It is therefore recommended to consider time series with sufficient temporal speckle decorrelation for which no whitening step is necessary, as illustrated by our results in section III. If images are in interferometric configuration with a large coherence, a whitening step is required. We describe a specific procedure in Appendix A and denote by $\mathring{\boldsymbol{z}}$ the stack after this preprocessing step, i.e., with minimal correlations along the temporal dimension ($\mathring{\boldsymbol{z}} = \dot{\boldsymbol{z}}$ in the absence of whitening step). Assumption 1 summarizes that temporal correlations have been suppressed by the preprocessing step:

[1]as discussed in [15], the Hermitian symmetry of the SAR transfer function must be ensured. This may require additional steps (e.g., demodulation, truncation of the spectrum).

Assumption 1. *The preprocessed image $\mathring{\boldsymbol{z}}_{ref}$ at date t_{ref} is statistically independent of the images $\mathring{\boldsymbol{z}}_t$ for all dates $t \neq t_{ref}$.*

In our MISO framework, we will consider two sets of inputs (noted $\mathscr{E}_a$ and $\mathscr{E}_b$) that contain all images $\mathring{\boldsymbol{z}}_t$ except for the imaginary part $\mathring{\boldsymbol{b}}_{\text{ref}} \in \mathbb{R}^N$ (respectively the real part $\mathring{\boldsymbol{a}}_{\text{ref}} \in \mathbb{R}^N$) of $\mathring{\boldsymbol{z}}_{\text{ref}}$:

$$\mathscr{E}_a = \{\mathring{\boldsymbol{a}}_{\text{ref}}\} \cup \{\mathring{\boldsymbol{z}}_t | t \neq t_{\text{ref}}\} \text{ and}$$
$$\mathscr{E}_b = \{\mathring{\boldsymbol{b}}_{\text{ref}}\} \cup \{\mathring{\boldsymbol{z}}_t | t \neq t_{\text{ref}}\}.$$

In the following proposition, we show that these inputs are independent from the component set aside. This independence will be key to train a network fed with the input set $\mathscr{E}_a$ (or $\mathscr{E}_b$) under the supervision of loss function involving the component $\mathring{\boldsymbol{b}}_{\text{ref}}$ (resp. $\mathring{\boldsymbol{a}}_{\text{ref}}$).

Proposition 1. *Under assumption 1, the input set $\mathscr{E}_a$ is statistically independent from the imaginary part $\mathring{\boldsymbol{b}}_{ref}$ at date t_{ref}, and similarly the input set $\mathscr{E}_b$ is statistically independent from the real part $\mathring{\boldsymbol{a}}_{ref}$.*

Proof. Under assumption 1, the image $\mathring{\boldsymbol{z}}_{\text{ref}}$ is independent from all other images $\mathring{\boldsymbol{z}}_t$ with $t \neq t_{\text{ref}}$. It remains to prove that the real and imaginary parts at time t_{ref} are independent. According to our generative model of Sec.II-A, they can be expressed in terms of the speckle $\boldsymbol{\epsilon}_{\text{ref}}$ and the dominant scatterers $\boldsymbol{d}_{\text{ref}}$

$$\begin{pmatrix}\mathring{\boldsymbol{a}}_{\text{ref}}\\ \mathring{\boldsymbol{b}}_{\text{ref}}\end{pmatrix} = \begin{pmatrix}\Re(\dot{\boldsymbol{d}}_{\text{ref}})\\ \Im(\dot{\boldsymbol{d}}_{\text{ref}})\end{pmatrix} + \mathbf{M}\begin{pmatrix}\Re(\boldsymbol{\epsilon}_{\text{ref}})\\ \Im(\boldsymbol{\epsilon}_{\text{ref}})\end{pmatrix}, \tag{10}$$

where

$$\dot{\boldsymbol{d}}_{\text{ref}} = \mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_{\text{ref}} + j\boldsymbol{\psi}_{\text{ref}}))\boldsymbol{d}_{\text{ref}} \tag{11}$$

and

$$\mathbf{M} = \begin{pmatrix}\mathbf{Q}\mathrm{diag}\big(\cos(\boldsymbol{\alpha}_{\text{ref}})\sqrt{\boldsymbol{r}_{\text{ref}}}\big) & -\mathbf{Q}\mathrm{diag}\big(\sin(\boldsymbol{\alpha}_{\text{ref}})\sqrt{\boldsymbol{r}_{\text{ref}}}\big)\\ \mathbf{Q}\mathrm{diag}\big(\sin(\boldsymbol{\alpha}_{\text{ref}})\sqrt{\boldsymbol{r}_{\text{ref}}}\big) & \mathbf{Q}\mathrm{diag}\big(\cos(\boldsymbol{\alpha}_{\text{ref}})\sqrt{\boldsymbol{r}_{\text{ref}}}\big)\end{pmatrix}$$

with $\boldsymbol{\alpha}_{\text{ref}} = \boldsymbol{\varphi}_{\text{ref}} + \boldsymbol{\psi}_{\text{ref}}$ and where the square root as well as the multiplications between vector $\sqrt{\boldsymbol{r}_{\text{ref}}}$ and the cosine and sine are all applied entry-wise.

Given that $\Re(\boldsymbol{\epsilon}_{\text{ref}})$ and $\Im(\boldsymbol{\epsilon}_{\text{ref}})$ are independent and identically distributed according to a Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{2}\mathbf{I})$, the real-valued vector formed by the real and imaginary components is also distributed according to a Gaussian distribution:

$$\begin{pmatrix}\mathring{\boldsymbol{a}}_{\text{ref}}\\ \mathring{\boldsymbol{b}}_{\text{ref}}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\Re(\dot{\boldsymbol{d}}_{\text{ref}})\\ \Im(\dot{\boldsymbol{d}}_{\text{ref}})\end{pmatrix}, \tfrac{1}{2}\mathbf{M}\mathbf{M}^\dagger\right) \tag{12}$$

with

$$\mathbf{M}\mathbf{M}^\dagger = \begin{pmatrix}\mathbf{Q}\mathrm{diag}(\boldsymbol{r}_{\text{ref}})\mathbf{Q}^\dagger & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}\mathrm{diag}(\boldsymbol{r}_{\text{ref}})\mathbf{Q}^\dagger\end{pmatrix}. \tag{13}$$

This shows that $\mathring{\boldsymbol{a}}_{\text{ref}}$ and $\mathring{\boldsymbol{b}}_{\text{ref}}$ are both jointly Gaussian and decorrelated, and thus, independent. □

C. Self-supervised training strategy

In [15], the following single-date loss function has been introduced:

$$\mathcal{L}_{\text{MERLIN}}(\boldsymbol{a}, \boldsymbol{u}) = \sum_k \frac{1}{2}\log u_k + \frac{a_k^2}{u_k}. \tag{14}$$

It was applied to train a network fed with the imaginary part $\boldsymbol{b}$ of a single SLC image and supervised by the corresponding real part $\boldsymbol{a}$ through $\mathcal{L}_{\text{MERLIN}}(\boldsymbol{a}, \boldsymbol{u})$ (where $\boldsymbol{u}$ represents the output of the network), or conversely by providing $\boldsymbol{a}$ to the network and supervising with $\mathcal{L}_{\text{MERLIN}}(\boldsymbol{b}, \boldsymbol{u})$. Assuming that $\boldsymbol{a}$ and $\boldsymbol{b}$ are statistically independent, the network was shown to learn how to estimate the reflectivities.

We extend this loss to our multi-temporal MISO framework by replacing $\boldsymbol{a}$ with $\mathring{\boldsymbol{a}}_{\text{ref}}$ and $\boldsymbol{b}$ with $\mathring{\boldsymbol{b}}_{\text{ref}}$. The parameters $\boldsymbol{\theta}$ of our regression model $f_{\boldsymbol{\theta}}$ (i.e., the deep neural network) can be learned by minimizing the following multi-temporal extension of the MERLIN loss function:

$$\begin{aligned}\arg\min_{\boldsymbol{\theta}}\ & \mathbb{E}_{\substack{\mathring{\boldsymbol{b}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}\\ \mathscr{E}_a|\boldsymbol{r},\boldsymbol{d}}}\left[\mathcal{L}_{\text{MERLIN}}\left(\mathring{\boldsymbol{b}}_{\text{ref}}, f_{\boldsymbol{\theta}}(\mathscr{E}_a)\right)\right]\\ &+ \mathbb{E}_{\substack{\mathring{\boldsymbol{a}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}\\ \mathscr{E}_b|\boldsymbol{r},\boldsymbol{d}}}\left[\mathcal{L}_{\text{MERLIN}}\left(\mathring{\boldsymbol{a}}_{\text{ref}}, f_{\boldsymbol{\theta}}(\mathscr{E}_b)\right)\right]. \end{aligned}\tag{15}$$

According to Proposition 1, the inputs of the network $\mathscr{E}_a$ or $\mathscr{E}_b$ are independent from the images $\mathring{\boldsymbol{b}}_{\text{ref}}$ and $\mathring{\boldsymbol{a}}_{\text{ref}}$ used in the loss. It is thus impossible for the network to predict the stochastic component in these images (the output $\boldsymbol{u} = \boldsymbol{a}$ would minimize equation (14) but cannot be obtained from the inputs).

In the following proposition, we consider the family of all possible models $f_{\boldsymbol{\theta}}$ that map the input images to a single output image. We then discuss in the proof of Prop.3 the special case of a sub-family of models corresponding to a given parameterization of the regression model $f_{\boldsymbol{\theta}}$ (for example, a fixed neural network architecture).

Proposition 2. *The expectation of the multi-temporal MERLIN loss function (15) is minimal with respect to the predictions $f_{\boldsymbol{\theta}}(\mathscr{E}_a)$ and $f_{\boldsymbol{\theta}}(\mathscr{E}_b)$ if and only if $f_{\boldsymbol{\theta}}(\mathscr{E}_a) = \tilde{\boldsymbol{r}}_{ref} + 2\Im(\dot{\boldsymbol{d}}_{ref})^2$ and $f_{\boldsymbol{\theta}}(\mathscr{E}_b) = \tilde{\boldsymbol{r}}_{ref} + 2\Re(\dot{\boldsymbol{d}}_{ref})^2$, where $\tilde{\boldsymbol{r}}_{ref}$ is the diagonal of covariance matrix $\mathbf{Q}\mathrm{diag}(\boldsymbol{r}_{ref})\mathbf{Q}^\dagger$ and $\dot{\boldsymbol{d}}_{ref} = \mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_{ref} + j\boldsymbol{\psi}_{ref}))\boldsymbol{d}_{ref}$.*

$$\mathrm{Cov}[\tilde{\boldsymbol{z}}] = \begin{pmatrix}\mathrm{diag}(\exp(-j\boldsymbol{\psi}_1))\mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_1 + j\boldsymbol{\psi}_1)) & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathrm{diag}(\exp(-j\boldsymbol{\psi}_T))\mathbf{Q}\mathrm{diag}(\exp(j\boldsymbol{\varphi}_T + j\boldsymbol{\psi}_T))\end{pmatrix}\mathrm{diag}(\sqrt{\boldsymbol{r}})\mathbf{\Pi}^{-1}$$
$$\begin{pmatrix}\mathbf{\Gamma}_1 & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathbf{\Gamma}_N\end{pmatrix}\mathbf{\Pi}\mathrm{diag}(\sqrt{\boldsymbol{r}})\begin{pmatrix}\mathrm{diag}(\exp(-j\boldsymbol{\varphi}_1 - j\boldsymbol{\psi}_1))\mathbf{Q}^\dagger\mathrm{diag}(\exp(j\boldsymbol{\psi}_1)) & & \mathbf{0}\\ & \ddots & \\ \mathbf{0} & & \mathrm{diag}(\exp(-j\boldsymbol{\varphi}_T - j\boldsymbol{\psi}_T))\mathbf{Q}^\dagger\mathrm{diag}(\exp(j\boldsymbol{\psi}_T))\end{pmatrix}.$$

Proof. We start by expressing the values of the two expectations that appear in equation (15). They involve terms of the form $\mathbb{E}[\sum_k \mathring{\boldsymbol{a}}_{\text{ref}}(k)^2/\boldsymbol{u}(k)]$ and $\mathbb{E}[\sum_k \mathring{\boldsymbol{b}}_{\text{ref}}(k)^2/\boldsymbol{v}(k)]$, where $\boldsymbol{u} = f_{\boldsymbol{\theta}}(\mathscr{E}_b)$ and $\boldsymbol{v} = f_{\boldsymbol{\theta}}(\mathscr{E}_a)$. They can be rewritten $\mathbb{E}[\mathring{\boldsymbol{a}}_{\text{ref}}^{\dagger}\text{diag}(1/\boldsymbol{u})\mathring{\boldsymbol{a}}_{\text{ref}}] = \text{Tr}\{\text{diag}(1/\boldsymbol{u})\mathbb{E}[\mathring{\boldsymbol{a}}_{\text{ref}}\mathring{\boldsymbol{a}}_{\text{ref}}^{\dagger}]\}$ where $1/\boldsymbol{u}$ denotes an entry-wise inversion. By marginalization of the Gaussian distribution defined in (12), we obtain $\mathbb{E}[\mathring{\boldsymbol{a}}_{\text{ref}}\mathring{\boldsymbol{a}}_{\text{ref}}^{\dagger}] = \Re(\dot{\boldsymbol{d}}_{\text{ref}})\Re(\dot{\boldsymbol{d}}_{\text{ref}})^{\dagger} + \frac{1}{2}\mathbf{Q}\text{diag}(\boldsymbol{r}_{\text{ref}})\mathbf{Q}^{\dagger}$. Similarly, $\mathbb{E}[\mathring{\boldsymbol{b}}_{\text{ref}}^{\dagger}\text{diag}(1/\boldsymbol{u})\mathring{\boldsymbol{b}}_{\text{ref}}] = \text{Tr}\{\text{diag}(1/\boldsymbol{u})\mathbb{E}[\mathring{\boldsymbol{b}}_{\text{ref}}\mathring{\boldsymbol{b}}_{\text{ref}}^{\dagger}]\}$ with $\mathbb{E}[\mathring{\boldsymbol{b}}_{\text{ref}}\mathring{\boldsymbol{b}}_{\text{ref}}^{\dagger}] = \Im(\dot{\boldsymbol{d}}_{\text{ref}})\Im(\dot{\boldsymbol{d}}_{\text{ref}})^{\dagger} + \frac{1}{2}\mathbf{Q}\text{diag}(\boldsymbol{r}_{\text{ref}})\mathbf{Q}^{\dagger}$. This leads to:

$$\mathbb{E}_{\mathring{\boldsymbol{b}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{a}}_{\text{ref}},\boldsymbol{u})\right] = \sum_k \frac{1}{2}\log \boldsymbol{u}(k) + \frac{\Re(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2 + \frac{1}{2}\tilde{\boldsymbol{r}}_{\text{ref}}(k)}{\boldsymbol{u}(k)} \tag{16}$$

$$\mathbb{E}_{\mathring{\boldsymbol{a}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{b}}_{\text{ref}},\boldsymbol{v})\right] = \sum_k \frac{1}{2}\log \boldsymbol{v}(k) + \frac{\Im(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2 + \frac{1}{2}\tilde{\boldsymbol{r}}_{\text{ref}}(k)}{\boldsymbol{v}(k)}. \tag{17}$$

A necessary condition for the expectations to be minimal is:

$$\frac{\partial}{\partial \boldsymbol{u}(k)}\mathbb{E}_{\mathring{\boldsymbol{a}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{a}}_{\text{ref}},\boldsymbol{u})\right] = 0 \Rightarrow \boldsymbol{u}(k) = \tilde{\boldsymbol{r}}_{\text{ref}}(k) + 2\Re(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2 \tag{18}$$

$$\frac{\partial}{\partial \boldsymbol{v}(k)}\mathbb{E}_{\mathring{\boldsymbol{b}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{b}}_{\text{ref}},\boldsymbol{v})\right] = 0 \Rightarrow \boldsymbol{v}(k) = \tilde{\boldsymbol{r}}_{\text{ref}}(k) + 2\Im(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2. \tag{19}$$

The second-order derivatives for the values of $\boldsymbol{u}(k)$ and $\boldsymbol{v}(k)$ given by equations (18) and (19)

$$\left.\frac{\partial^2\mathbb{E}_{\mathring{\boldsymbol{a}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{a}}_{\text{ref}},\boldsymbol{u})\right]}{\partial \boldsymbol{u}(k)^2}\right|_{\boldsymbol{u}(k)=\tilde{\boldsymbol{r}}_{\text{ref}}(k)+2\Re(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2} = \frac{1}{2(\tilde{\boldsymbol{r}}_{\text{ref}}(k) + 2\Re(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2)^2} \tag{20}$$

$$\left.\frac{\partial^2\mathbb{E}_{\mathring{\boldsymbol{b}}_{\text{ref}}|\boldsymbol{r},\boldsymbol{d}}\left[\mathcal{L}_{\text{MERLIN}}(\mathring{\boldsymbol{b}}_{\text{ref}},\boldsymbol{v})\right]}{\partial \boldsymbol{v}(k)^2}\right|_{\boldsymbol{v}(k)=\tilde{\boldsymbol{r}}_{\text{ref}}(k)+2\Im(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2} = \frac{1}{2(\tilde{\boldsymbol{r}}_{\text{ref}}(k) + 2\Im(\dot{\boldsymbol{d}}_{\text{ref}}(k))^2)^2} \tag{21}$$

are both strictly positive, which shows that the values of $\boldsymbol{u}(k)$ and $\boldsymbol{v}(k)$ correspond to a minimum. Since the solution to equations (18) and (19) is unique, we have identified the only minimum of the objective function. □

Proposition 3. *Minimization of the expectation of the multi-temporal MERLIN loss function leads to an unbiased estimator* $[f_{\boldsymbol{\theta}}(\mathscr{E}_a) + f_{\boldsymbol{\theta}}(\mathscr{E}_b)]/2$ *of the sum of the low-pass filtered reflectivities* $\tilde{\boldsymbol{r}}_{ref}$ *and of the intensity of the low-pass filtered dominant scatterers* $|\dot{\boldsymbol{d}}_{ref}|^2$ *at date* t_{ref}*, provided that* $f_{\boldsymbol{\theta}}$ *is sufficiently expressive (e.g., a deep neural network with sufficient width).*

Proof. Under the Universal Approximation Theorem for width-bounded ReLU networks [26], a network with sufficient width can be built to approximate an arbitrary (Lebesgue-integrable) function $f_{\boldsymbol{\theta}}$. If less expressive estimators $f_{\boldsymbol{\theta}}$ are considered (smaller networks, not fully-connected architectures, other estimators than deep neural networks), a bias may appear due to the reduced ability of the estimator to match the optimal output given in Proposition 2.

For a sufficiently expressive estimator producing the optimal output, according to Proposition 2, the minimum of the expectation of the multi-temporal MERLIN loss function is reached for $f_{\boldsymbol{\theta}}(\mathscr{E}_a) = \tilde{\boldsymbol{r}}_{\text{ref}} + 2\Im(\dot{\boldsymbol{d}}_{\text{ref}})^2$ and $f_{\boldsymbol{\theta}}(\mathscr{E}_b) = \tilde{\boldsymbol{r}}_{\text{ref}} + 2\Re(\dot{\boldsymbol{d}}_{\text{ref}})^2$. The computation of the average concludes the proof:

$$\forall k,\ \frac{f_{\boldsymbol{\theta}}(\mathscr{E}_a)(k) + f_{\boldsymbol{\theta}}(\mathscr{E}_b)(k)}{2} = \tilde{\boldsymbol{r}}_{\text{ref}}(k) + |\dot{\boldsymbol{d}}_{\text{ref}}(k)|^2. \tag{22}$$

□

Figure 2 illustrates the principle of the proposed self-supervised training introduced in Propositions 2 and 3: during training, we minimize MERLIN loss with the sets $\mathscr{E}_a$ or $\mathscr{E}_b$ as input and the images $\mathring{\boldsymbol{b}}_{\text{ref}}$ or $\mathring{\boldsymbol{a}}_{\text{ref}}$ in the supervision. This leads to optimal weights $\boldsymbol{\theta}^*$ at the end of the training phase. At test time, the estimates $f_{\boldsymbol{\theta}^*}(\mathscr{E}_a)$ and $f_{\boldsymbol{\theta}^*}(\mathscr{E}_b)$ are averaged to produce the final estimate.

For practical reasons, we use a convolutional U-Net architecture [27] (also used in the MERLIN method [15]) with a small number of parameters, we consider a limited number of images in the training phase and an approximate minimization based on stochastic gradient computed over mini-batches. The estimator $f_{\boldsymbol{\theta}^*}$ obtained is then only sub-optimal.

III. Experiments

The performance of the proposed multi-temporal MERLIN strategy is first studied on images with simulated speckle in paragraph III-A. Results on Single Look Complex TerraSAR-X images are then presented in paragraph III-B. In both cases, we compare multi-temporal MERLIN networks trained for an increasing number of additional inputs to study the quality improvement brought by these additional dates.

A. Quantitative analysis on simulated speckle

The unsupervised learning strategy presented in Section II is motivated by the lack of speckle-free ground-truth images associated to each speckled SAR image. Yet, in order to perform a quantitative assessment of multi-temporal filtering, we first consider a simulated speckle framework in which both speckle-free and speckle-corrupted images are available. We build high-quality speckle-free stacks by multi-temporal filtering with RABASAR-SAR2SAR [22]. We then generate corrupted versions with simulated speckle corresponding to an ideal SAR transfer function, i.e., speckle with no spatial correlation in the simulated images. This reference data set is composed of 5 multi-temporal stacks of despeckled Sentinel-1 images, each stack containing from 25 to 69 images. Since the

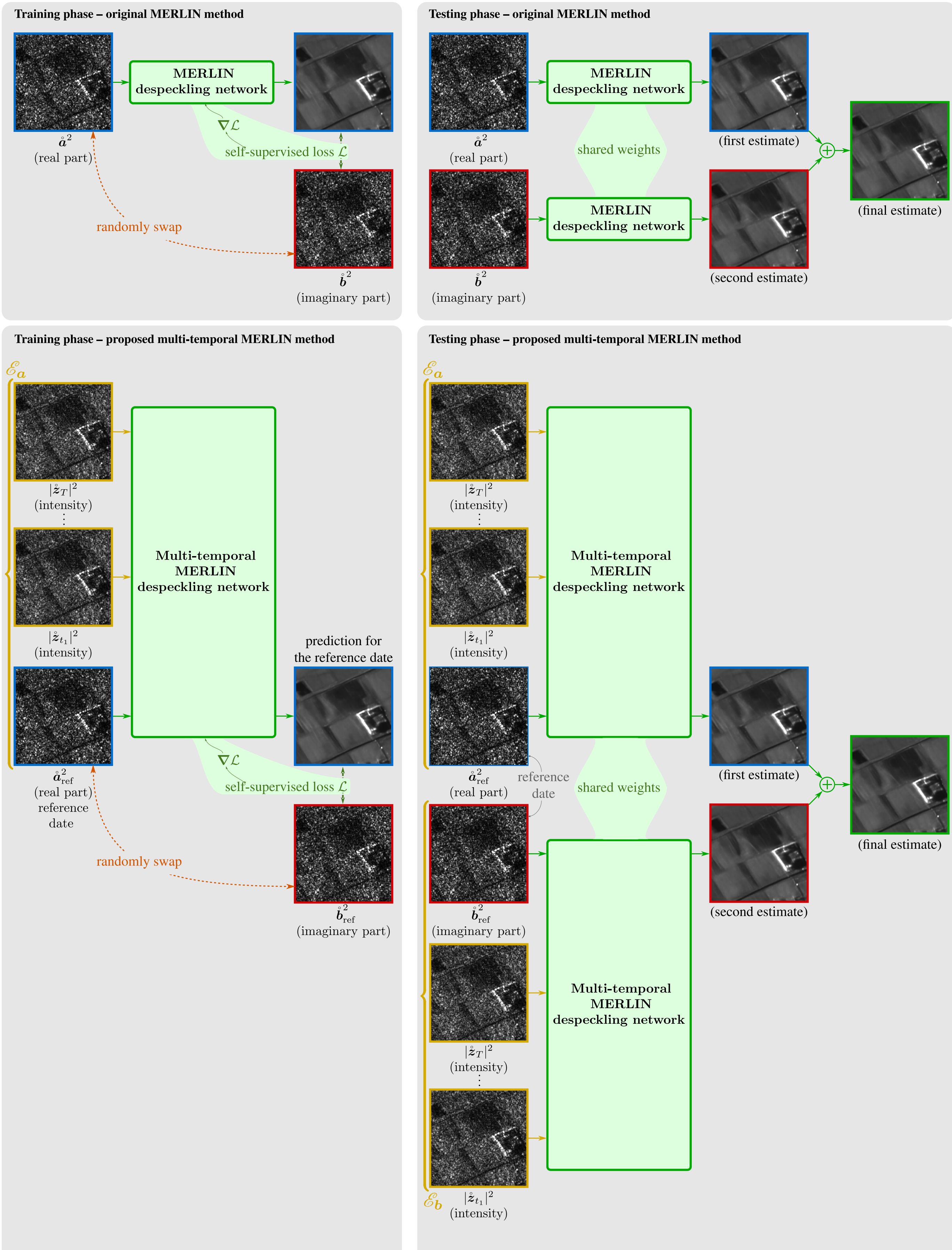

Figure 2. Principle of the self-supervised method MERLIN: original approach [15] (top row) and proposed multi-temporal extension (bottom row). For better visualization, the displayed images are amplitude images.

Table II
TRAINING PARAMETERS OF THE MULTI-TEMPORAL MERLIN NETWORKS (NUMBER OF INPUT CHANNELS FROM 2 TO 20)

	Synthetic speckle Sentinel-1	Actual speckle TerraSAR-X [Sentinel-1]
# stacks	7	2 [1]
# images	237	52 [12]
avg images/stack	33.9	26 [12]
patch size	256×256	256×256
batch size	8	8
# patches	1616	576 [1568]
# batches	202	72 [196]
# epochs	1000	1000
learning rate	10^{-3} 10^{-4} after 10 epochs 10^{-5} after 910 epochs	10^{-3} 10^{-4} after 10 epochs 10^{-5} after 910 [860] epochs

stacks are obtained from actual SAR images, realistic changes can be observed throughout the time series (e.g., evolution of the reflectivities in the fields). To simplify the simulations, we assume fully-developed speckle (the ground-truth images correspond to the reflectivities $\boldsymbol{r}$ and no dominant scatterer is considered: $\boldsymbol{d} = \boldsymbol{0}$). Information on the training sets and the hyperparameters used in all our network trainings are gathered in table II. The hyperparameters are kept unchanged whatever the number of additional inputs.

1) Impact of the number of additional channels: We first evaluate the gain brought by the additional dates on the quality of the estimated speckle-free image. Depending on the presence or absence of change, including an additional input image may disturb or help the despeckling process. When comparing the performances of two networks, a network with fewer inputs that underwent less changes might be favored over a network with more inputs which were all impacted by larger changes. We mitigate the impact of this phenomenon on our analysis by evaluating the performance of our networks on combinations of additional dates forming nested sets, i.e., a network with j additional inputs, $j > i$, shares the same i additional dates as a smaller network with i additional inputs, but also benefits from $j - i$ supplementary inputs.

Figure 3 shows boxplots of the Peak Signal-to-Noise Ratio (PSNR) values computed on the log-reflectivities, for an increasing number of additional input images. The boxplots give for each configuration the minimum PSNR value; first, second, and third quartile PSNR values; and the maximum PSNR value. These statistics are computed over 88400 patches of 256×256 pixels, corresponding to different spatial locations, choices of dates included as input, or speckle realizations. The restoration quality, measured by the PSNR values, improves with the number of images. This improvement is largest when the first additional dates are included, including a few more dates to an already large number of inputs produces a marginal improvement: unsurprisingly, multi-temporal filtering follows a law of diminishing returns with respect to the number of input dates.

Note that the dispersion of PSNR values for the mono-date filtering (leftmost boxplot of Figure 3) is very limited compared to the dispersion of PSNR values obtained with multi-temporal filtering. This is due to the variability of changes present in the additional channels: in multi-temporal filtering, situations with limited changes are more favorable to filtering and lead to better PSNR values while drawing a set of dates with larger changes inevitably gives a worse PSNR value (the variable luck in how similar the additional dates were explains the PSNR fluctuations).

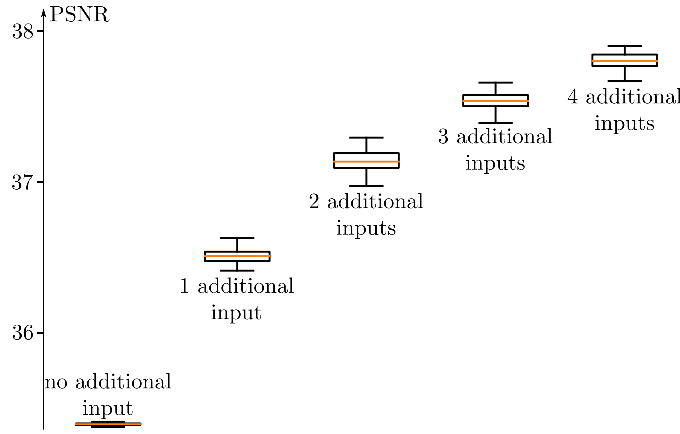

Figure 3. Boxplots of PSNR values obtained for different draws of additional dates and various speckle realizations (each box plot indicates the minimum value, first quartile, in orange: median value, third quartile, and maximum). Our multi-temporal MERLIN method outperforms the baseline methods in terms of PSNR: with 3 additional inputs, the median PSNR with MSAR-BM3D is 36.04 dB (-1.50 dB) and with 2SPPB it is 30.06 dB (-7.48dB).

As illustrated by Figure 4, PSNR values improve when increasing the number of additional input images due to the joint reduction of the estimation bias and of the estimation variance. Additional channels help preserve the spatial resolution, reducing the blur around sharp structures (such as points, lines, edges), as illustrated by the bias term. By not only combining spatial samples but also temporal samples, the estimation variance is reduced by multi-temporal filtering.

The line profiles shown in Figure 5 confirm the improved ability to restore fine structures with multi-temporal filtering (spatial resolution gain): processing a single date (green line) makes it difficult to retrieve the contrast of thin lines (edges at the border of fields); with an additional date, or even better, with 4 additional dates, these structures are much better restored.

2) Impact of temporal correlations: Images acquired in interferometric configuration may suffer from correlations along the temporal axis, as discussed in Section II. This is not ideal in the context of multi-temporal filtering as it reduces the potential benefit of temporal speckle averaging. Beyond this limitation, we illustrate here that, if neglected (i.e., if the temporal decorrelation step presented in Appendix A is omitted), this type of correlations impacts the despeckling performance of networks trained with the multi-temporal MERLIN loss function (the independence assumption between the inputs and the component used for self-supervision is no longer valid).

We repeat the previous experiment with simulated speckle, this time introducing temporal correlations with a coherence matrix $\boldsymbol{\Gamma}_k$ identical for all pixels k, and following a simple

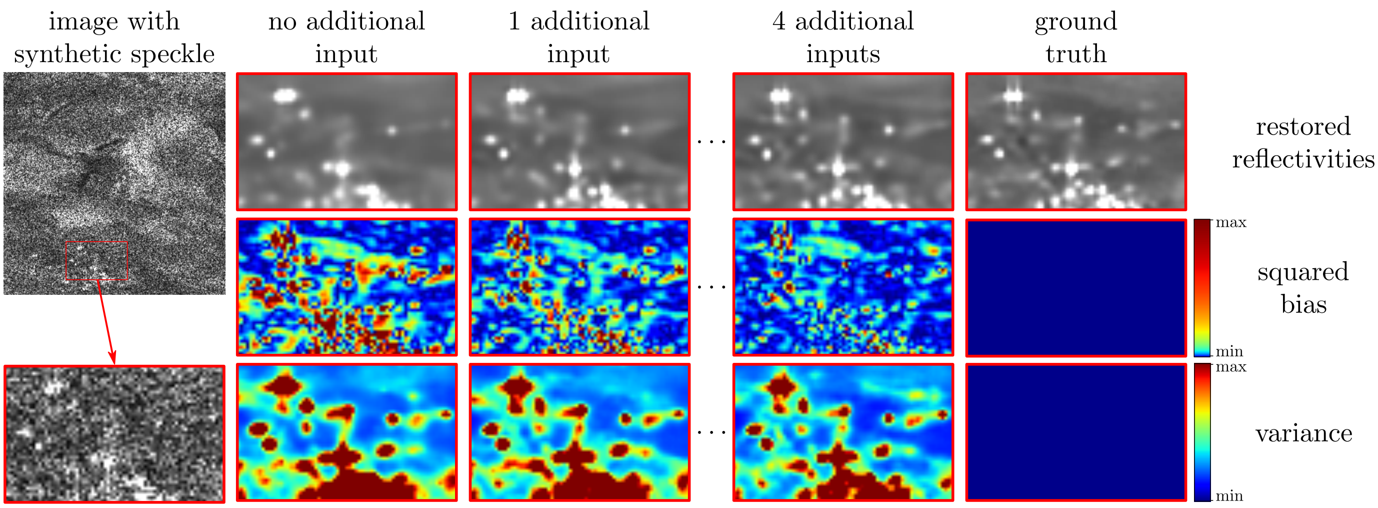

Figure 4. Squared bias and variance averaged over 100 Multi-temporal MERLIN estimations of the reflectivities of a Sentinel-1 stack of Limagne (France). The speckle is simulated based on the method described in [22], and details of the test set are given in III-A1.

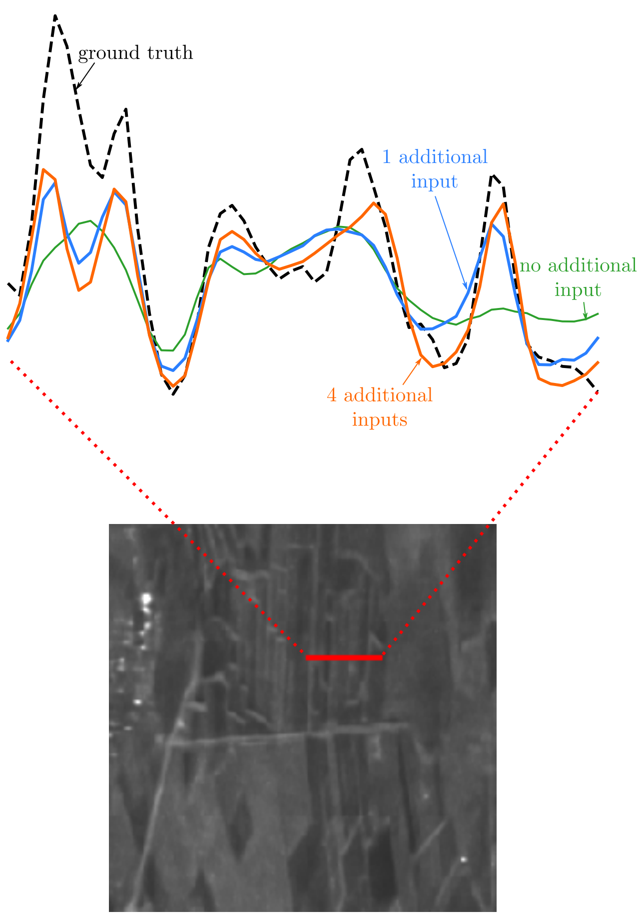

Figure 5. Reflectivites profile along the red line, Marais1, date 14. The profile associated to MERLIN network estimation (green line) is blunt, meaning that the edges of the small observed structures are blurred. The more additional inputs there are, the sharper the profile lines, leading to a better retrieving of small structures.

temporal decorrelation model

$$\forall k,\ \mathbf{\Gamma}_k(t_i, t_j) = \exp\left(-\frac{|t_i - t_j|}{\tau}\right), \tag{23}$$

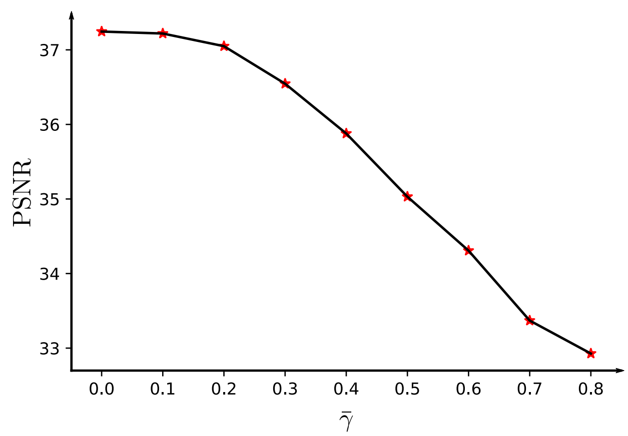

Figure 6. Evolution of the restoration performance (PSNR values computed on log reflectivities) as a function of the average coherence $\bar{\gamma}$ of the multi-temporal stack.

where τ is a characteristic decorrelation time. Rather than reporting how the despeckling performance degrades as a function of parameter τ, we use the more intuitive average coherence $\bar{\gamma}$ defined by

$$\bar{\gamma} = \frac{1}{T^2} \sum_{1 \leq i,j \leq T} \mathbf{\Gamma}_k(t_i, t_j)\,. \tag{24}$$

Figure 6 reports the evolution of the PSNR of restored images (computed on log reflectivities) as a function of the average coherence $\bar{\gamma}$ for a network that uses two additional inputs. Up to $\bar{\gamma} \approx 0.2$ the performance is almost unchanged, then it degrades significantly. At $\bar{\gamma} \approx 0.45$, the PSNR value is no better than that reached by a network with no additional input (mono-date filtering). Beyond $\bar{\gamma} \approx 0.45$, it is worse to include additional dates. The reason is that the temporal correlations of speckle lead the network to "cheat" and to partially guess the speckled component in the images used to supervise the training. Once trained, the network systematically leaves a large fraction of the speckle fluctuations unchanged.

B. Qualitative analysis of networks trained on actual SAR time series

After the successful validation of our approach on time series with simulated speckle, we now turn to real speckle. First, we illustrate our (optional) preprocessing step that performs a temporal decorrelation with respect to the reference date. We recall here how this decorrelation is achieved, more details are given in A:

1) each SLC image of the stack is decomposed into a dominant scatterers component and a background component;
2) interferograms with respect to the reference date are computed on the background components;
3) at each pixel, a temporal whitening is performed based on the local coherence matrix estimated at the previous step;
4) the contribution of dominant scatterers is reintroduced.

Figure 7 illustrates these different steps. We perform step 1) with the method described in [28]: the low-pass filtering effect introduced by the SAR system response (step d of the generative model of Figure 1) is first compensated by resampling and spectral equalization, then an *a contrario* framework is applied to detect the cardinal sines of the dominant scatterers. The contribution of the dominant scatterers is then subtracted from the image and the original spectral apodization is reapplied. In step 2), we estimate interferograms between all pairs of images drawn from the stack of background components $\dot{\boldsymbol{z}} - \dot{\boldsymbol{d}}$. In our experiments, we use the MuLoG algorithm to compute these interferograms. This step is computationally intensive since forming all possible interferograms (in order to maximize the number of training samples to train our despeckling network) requires $\mathcal{O}(T^2)$ interferogram estimations for a multi-temporal stack with T dates. Step 3) is much faster since it only requires applying pixelwise the simple whitening transform of equation (30). Finally, the reintroduction of dominant scatterers in step 4) leads to the temporally whitened stack $\mathring{\boldsymbol{z}}$.

In order to assess the impact of this temporal decorrelation step, we compared the performance of the same network trained in one case directly on a stack of 26 TerraSAR-X images $\dot{\boldsymbol{z}}$ (i.e., the spectra have been shifted to center the spectrum of the reference date, but no temporal decorrelation step has been carried out), and in the other case using a pre-processed stack with our spectrum centering plus the temporal decorrelation technique. Coherences between the first two images of the original and the pre-processed stacks computed with the MuLoG algorithm are presented in Figure 8. It shows that the proposed whitening step strongly reduces the coherence. Despeckling results are presented in Figure 9 and very few differences can be observed (slight changes may be noted around some scatterers). The average coherence on this stack of TerraSAR-X images is equal to 0.23, which corresponds to a mild level of correlation with a negligible impact on the despeckling performance, as shown in our experiments with simulated speckle reported in Figure 6. This illustrates that, even in the case of a satellite with interferometric capabilities, the computationally heavy preprocessing step of temporal decorrelation can be skipped when the coherence level is moderate.

Given the limited impact of this temporal whitening step for the multi-temporal stack we considered, we chose to skip this step and compare the performance of our network trained directly on multi-temporal stacks with other reference methods. Parameters used for our training are recalled in Table II, last column. Figure 10 shows two excerpts taken from the TerraSAR-X stacks used for training. Note that, given our self-supervised training strategy, our network can be tested on the same dataset as used for training. When applying the network to other datasets, the performance might drop if the type of area differs significantly (e.g, training on urban areas and testing on mountainous regions) due to a poor generalization. A fine-tuning step on the data of interest using the self-supervised loss is then preferable. The figure 10 contains two panels with the same numbering, each corresponding to a different stack. The single-look amplitude is shown in (a). In order to identify low-contrasted structures and fine details, the temporal average computed over the whole stack is shown in (b). Due to the changes that occur throughout the time series, this image is not directly comparable to image (a) but is still useful to analyze temporally-stable structures present in the scene given that speckle is strongly reduced by temporal averaging. Areas with fluctuating reflectivities lead to an average value that differs from the reflectivity at the date of interest. Restoration results obtained with several speckle reduction methods are shown in each panel: (c) the mono-date MERLIN network, (d and g) the proposed multi-temporal MERLIN networks, and two baseline patch-based methods: (e and h) MSAR-BM3D introduced in [18] and (f and i) 2SPPB proposed in [19]. Multi-temporal methods are applied to a subset of 4 dates (the reference date + 3 additional dates) in the second row of the figure, or 16 dates (the reference date + 15 additional dates) on the last row. Temporal leakages can be observed in the results of MSAR-BM3D and 2SPPB: spurious information from the other dates contaminate the reference date, this is especially visible by the attenuation of the dark area (almost vertical rectangular field, in the center left of the image on the left pannel). In that respect, multi-temporal MERLIN offers much better results with restored reflectivities in good match with the noisy observation shown in Figure 10(a). Edges are sharper and low-contrast structures are better preserved in the case with a limited amount of dates (3 additional inputs): Figure 10(d-f) left and right panels.

To illustrate that our self-supervised strategy requires only a modest amount of data and that it can be applied to another satellite, we also trained from scratch the networks with a single stack of 12 Sentinel-1 images in Stripmap mode with 2000×2000 pixels. We compare in Figure 11 our despeckling results to the images obtained with the same baseline methods as in Figure 10 (namely, MSAR-BM3D and 2SPPB). Similar observations can be made: fine structures are better restored by the multi-temporal MERLIN and fewer artifacts can be noticed. Increasing the number of input images systematically leads to an improvement of the output image. Note that the use of MERLIN loss to train a network on Sentinel-1 in TOPS mode still requires some additional work in order to find the adequate pre-processing that would lead to statistically independent real and imaginary parts [15].

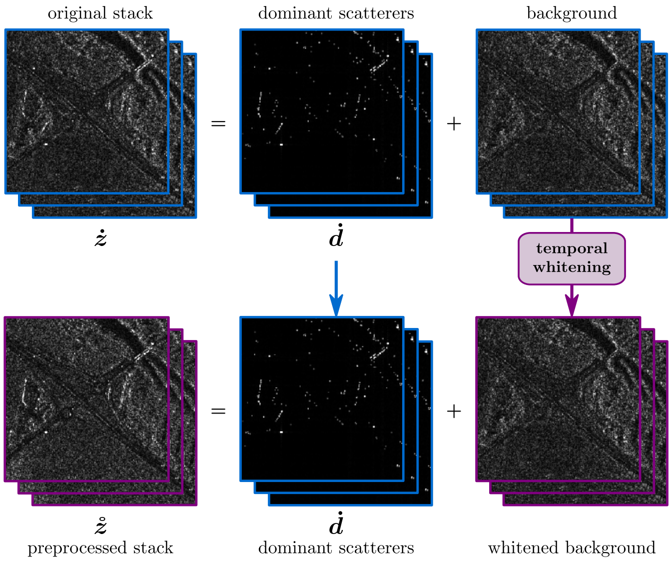

Figure 7. Illustration of the preprocessing step to reduce temporal correlations of the speckle: (first row) the multi-temporal stack is decomposed into dominant scatterers and background using the method in [28]; (second row) the background component is then whitened and the dominant scatterers are added back to produce the preprocessed stack.

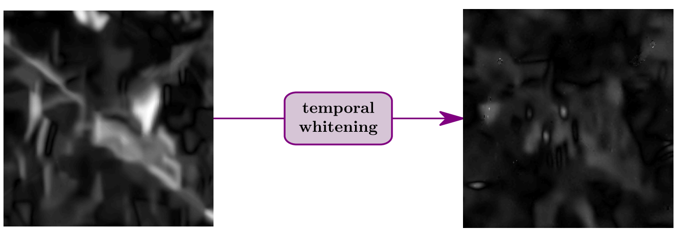

Figure 8. Coherences computed with the MuLoG algorithm on the 2 first dates (2009/05/31 and 2009/06/11) of the Domancy TerraSAR-X stack ©DLR. The studied area is introduced in 7; left: estimated coherence before the whitening step; right: estimated coherence after the whitening step.

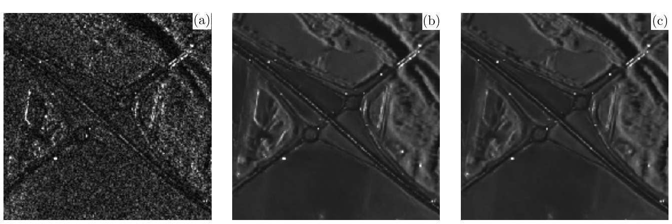

Figure 9. Impact of the temporal whitening step on multi-temporal denoising of TerraSAR-X images ©DLR, city of Domancy (France). (a) noisy image; (b) multi-temporal MERLIN with 2 additional inputs trained on one TerraSAR-X stack without the temporal whitening step; (c) multi-temporal MERLIN with 2 additional inputs trained on one TerraSAR-X stack with the temporal whitening step.

IV. Conclusion

A generative model based on the decomposition of the SLC images into a speckle component and a dominant scatterers component has been introduced in this work. It breaks down the different sources of statistical correlation between spatial, temporal, and real/imaginary components of the complex amplitudes of SAR images. It shows that, under some assumptions like a low coherence or an adequate preprocessing step, the self-supervised training strategy MERLIN can be extended to stacks of SLC images.

This strategy improves the despeckling performance achieved by mono-date networks by exploiting temporal redundancies of the scene and temporal fluctuations of speckle. Our quantitative analysis shows an improvement of the restored reflectivities, a refined spatial resolution, and very few temporal contamination by possible changes in the additional dates provided in input. Networks trained directly on SAR images, without groundtruth, produce restored images of higher quality compared to state-of-the-art techniques.

Deep neural networks trained with our self-supervised strategy seemingly bring a significant improvement to multi-temporal filtering in cases with limited or modest amounts of available dates. If numerous images are available, training a network to process all images becomes heavy, in particular regarding memory issues. Other approaches like ratio-based filtering [22] or a different strategy to combine images from the stack may then be preferable.

The problem of image super-resolution in multi-spectral imaging has led to several multi-image fusion approaches based on deep learning [29], [30]. Future work may study whether the specific network architectures proposed in these methods would benefit the multi-temporal SAR despeckling problem.

Beyond multi-temporal filtering, our framework can straightforwardly be extended to multi-sensor or multi-modality fusion by including as additional input channels some images of the scene acquired by other sources.

Acknowledgments

This project has been funded by the Futur & Ruptures PhD program of the Fondation Mines-Telecom, and partially funded by ASTRAL project (ANR-21-ASTR-0011) and by the grant n°R-S19/OT-0003-086 of the French space agency (CNES).

TerraSAR-X images were provided, as part of the project DLR-MTH0232 and DLR-LAN1746, by the German Space Agency DLR.

Appendix A
Pairwise temporal whitening

Correlations along the temporal axis of the speckle depend both on the coherence matrices $\boldsymbol{\Gamma}_k$ (capturing the temporal decorrelation of the scene) and on the shifts induced by the phases $\boldsymbol{\psi}_t$ (accounting for the geometrical decorrelation due to the change of incidence angles introduced by the interferometric baseline). To reduce these correlations, a whitening process can be designed based on the covariance values $\mathrm{Cov}[\dot{\boldsymbol{z}}(t_{\text{ref}},k);\, \dot{\boldsymbol{z}}(t_i,k)] = \mathbb{E}[(\dot{\boldsymbol{z}}_{\text{ref}}(k) - \dot{\boldsymbol{d}}_{\text{ref}}(k))(\dot{\boldsymbol{z}}_i(k) - \dot{\boldsymbol{d}}_i(k))^*]$.

The dominant scatterer component $\dot{\boldsymbol{d}}$ can be extracted from the images using an iterative algorithm [28]. The 2×2 interferometric covariance matrices $\begin{pmatrix} \mathrm{Cov}[\dot{\boldsymbol{z}}(t_{\text{ref}},k);\, \dot{\boldsymbol{z}}(t_{\text{ref}},k)] & \mathrm{Cov}[\dot{\boldsymbol{z}}(t_{\text{ref}},k);\, \dot{\boldsymbol{z}}(t_i,k)] \\ \mathrm{Cov}[\dot{\boldsymbol{z}}(t_i,k);\, \dot{\boldsymbol{z}}(t_{\text{ref}},k)] & \mathrm{Cov}[\dot{\boldsymbol{z}}(t_i,k);\, \dot{\boldsymbol{z}}(t_i,k)] \end{pmatrix}$ at each pixel k can then be estimated by using an algorithm such

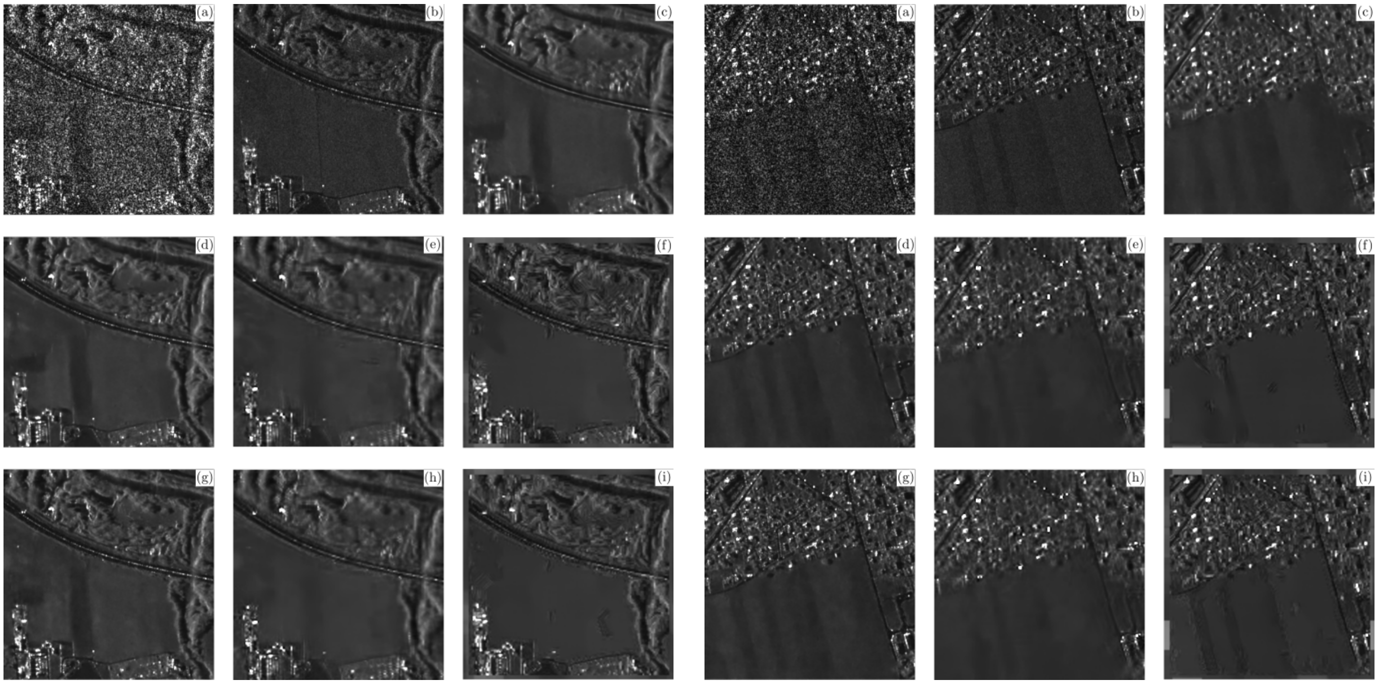

Figure 10. Multi-temporal denoising of TerraSAR-X images ©DLR (SLC images with actual speckle): left panel, city of Saint-Gervais (France); right panel, city of Domancy (France). Each panel shows (a) the noisy image; (b) the temporal average of all 26 images of the stack; (c) mono-date MERLIN filtering [15]; (d) multi-temporal MERLIN with 3 additional inputs; (e) MSAR-BM3D [18] with 3 additional inputs, (f) 2SPPB with 3 additional inputs [19]; (g) multi-temporal MERLIN with 15 additional inputs; (h) MSAR-BM3D [18] with 15 additional inputs, (i) 2SPPB with 15 additional inputs [19].

as MuLoG [31]. We propose to use these estimations to approximate the $2N \times 2N$ covariance matrix

$$\mathrm{Cov}\left[\begin{pmatrix} \dot{\boldsymbol{z}}_i \\ \dot{\boldsymbol{z}}_{\mathrm{ref}} \end{pmatrix}\right] \approx \begin{pmatrix} \mathbf{D}_{i\,i} & \mathbf{D}^{\dagger}_{\mathrm{ref}\,i} \\ \mathbf{D}_{\mathrm{ref}\,i} & \mathbf{D}_{\mathrm{ref\,ref}} \end{pmatrix}, \tag{25}$$

where the four $N \times N$ blocks are diagonal. Neglecting off-diagonal values of the matrices $\mathbf{D}_{i\,i}$ and $\mathbf{D}_{\mathrm{ref\,ref}}$ amounts to considering a limited spatial correlation length (SAR impulse response is close to a Dirac). Neglecting off-diagonal values of the matrices $\mathbf{D}_{i\,\mathrm{ref}}$ and $\mathbf{D}_{\mathrm{ref}\,i}$ is justified when the multi-temporal stack is in interferometric configuration: a shift by one or more pixels of the image $\dot{\boldsymbol{z}}_i$ with respect to image $\dot{\boldsymbol{z}}_{\mathrm{ref}}$ drastically reduces the interferometric coherence (i.e., the diagonal of $\mathbf{D}_{i\,\mathrm{ref}}$ is dominant).

From the expression of the covariance matrix $\mathrm{Cov}[\tilde{z}]$ given at the bottom of page 5 and the definition of $\dot{\boldsymbol{z}}$ in equation (9), we can derive the exact covariances $\mathrm{Cov}[\dot{\boldsymbol{z}}_i]$ and $\mathrm{Cov}[\dot{\boldsymbol{z}}_{\mathrm{ref}}]$ of the centered complex amplitudes of the considered pair of SAR images: $\mathrm{Cov}[\dot{\boldsymbol{z}}_i] = \mathbf{Q}\mathrm{diag}(\boldsymbol{r}_i)\mathbf{Q}^{\dagger}$ (and $\mathrm{Cov}[\dot{\boldsymbol{z}}_{\mathrm{ref}}] = \mathbf{Q}\mathrm{diag}(\boldsymbol{r}_{\mathrm{ref}})\mathbf{Q}^{\dagger}$, respectively). We approximate this covariance matrix by its diagonal: $\mathrm{Cov}[\dot{\boldsymbol{z}}_i] \approx \mathbf{D}_{i\,i}$ (and $\mathrm{Cov}[\dot{\boldsymbol{z}}_{\mathrm{ref}}] \approx \mathbf{D}_{\mathrm{refref}}$) with $\tilde{\boldsymbol{r}}_i$ the diagonal of matrix $\mathbf{Q}\mathrm{diag}(\boldsymbol{r}_i)\mathbf{Q}^{\dagger}$ (and $\tilde{\boldsymbol{r}}_{\mathrm{ref}}$ the diagonal of matrix $\mathbf{Q}\mathrm{diag}(\boldsymbol{r}_{\mathrm{ref}})\mathbf{Q}^{\dagger}$ respectively). These vectors correspond to a low-pass filtered version of the reflectivity maps, according to the SAR response $\mathbf{Q}$.

The anti-diagonal blocks are approximated by $\mathbf{D}_{\mathrm{ref}\,i} = \mathrm{diag}(\tilde{\gamma}_{i\,\mathrm{ref}}\sqrt{\tilde{\boldsymbol{r}}_i\tilde{\boldsymbol{r}}_{\mathrm{ref}}})$ where products between vectors are applied entry-wise, and $\tilde{\boldsymbol{\gamma}}_{i\,\mathrm{ref}} \in \mathbb{C}^N$ is the vector of complex-valued coherences between dates t_i and t_{ref} ($\forall k,\ 0 \leq |\tilde{\gamma}_{i\,\mathrm{ref}}(k)| \leq 1$).

The covariance matrix of a pair of complex amplitudes at a pixel k is finally given by:

$$\begin{aligned}\mathrm{Cov}\left[\begin{pmatrix} \dot{\boldsymbol{z}}(t_i,k) - \dot{\boldsymbol{d}}(t_i,k) \\ \dot{\boldsymbol{z}}(t_{\mathrm{ref}},k) - \dot{\boldsymbol{d}}(t_{\mathrm{ref}},k) \end{pmatrix}\right] &= \mathrm{Cov}\left[\begin{pmatrix} \dot{\boldsymbol{z}}(t_i,k) \\ \dot{\boldsymbol{z}}(t_{\mathrm{ref}},k) \end{pmatrix}\right] \\ &= \begin{pmatrix} \tilde{\boldsymbol{r}}(t_i,k) & \tilde{\gamma}_{i\,\mathrm{ref}}(k)\sqrt{\tilde{\boldsymbol{r}}(t_i,k)\tilde{\boldsymbol{r}}(t_{\mathrm{ref}},k)} \\ \tilde{\gamma}^{*}_{i\,\mathrm{ref}}(k)\sqrt{\tilde{\boldsymbol{r}}(t_i,k)\tilde{\boldsymbol{r}}(t_{\mathrm{ref}},k)} & \tilde{\boldsymbol{r}}(t_{\mathrm{ref}},k) \end{pmatrix}.\end{aligned} \tag{26}$$

The covariance along the temporal dimension between the image of reference $\dot{\boldsymbol{z}}_{\mathrm{ref}}$ and the image at date t_i $\dot{\boldsymbol{z}}_i$ modeled by (25) can be suppressed by multiplying each $2N$ vector $(\dot{\boldsymbol{z}}_i - \dot{\boldsymbol{d}}_i, \dot{\boldsymbol{z}}_{\mathrm{ref}} - \dot{\boldsymbol{d}}_{\mathrm{ref}})$ by a whitening matrix $\mathbf{W}$, leading to the whitened pair of images $(\mathring{\boldsymbol{z}}_i, \mathring{\boldsymbol{z}}_{\mathrm{ref}})$:

$$\begin{pmatrix} \mathring{\boldsymbol{z}}_i \\ \mathring{\boldsymbol{z}}_{\mathrm{ref}} \end{pmatrix} = \mathbf{W}\begin{pmatrix} \mathring{\boldsymbol{z}}_i - \dot{\boldsymbol{d}}_i \\ \mathring{\boldsymbol{z}}_{\mathrm{ref}} - \dot{\boldsymbol{d}}_{\mathrm{ref}} \end{pmatrix} + \begin{pmatrix} \dot{\boldsymbol{d}}_i \\ \dot{\boldsymbol{d}}_{\mathrm{ref}} \end{pmatrix} \tag{27}$$

with

$$\mathbf{W} = \mathbf{\Pi}^{-1}\begin{pmatrix} \mathbf{W}_1^{\dagger} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{W}_N^{\dagger} \end{pmatrix}\mathbf{\Pi} \tag{28}$$

and

$$\mathbf{W}_k\mathbf{W}_k^{\dagger} = \mathrm{Cov}\left[\begin{pmatrix} \dot{\boldsymbol{z}}(t_i,k) - \dot{\boldsymbol{d}}(t_i,k) \\ \dot{\boldsymbol{z}}(t_{\mathrm{ref}},k) - \dot{\boldsymbol{d}}(t_{\mathrm{ref}},k) \end{pmatrix}\right]^{-1}. \tag{29}$$

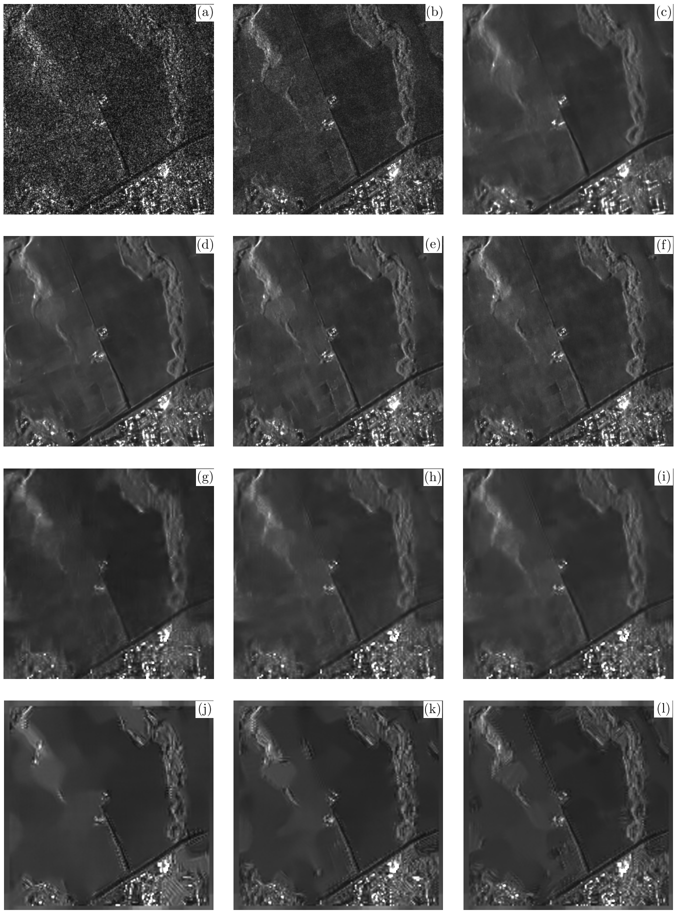

Figure 11. Multi-temporal denoising of Sentinel-1 Stripmap images of the Reunion island (France). Each panel shows (a) the noisy image; (b) the temporal average of all 12 images of the stack; (c) mono-date MERLIN filtering [15]; (d) multi-temporal MERLIN with 1 additional input; (e) multi-temporal MERLIN with 3 additional inputs; (f) multi-temporal MERLIN with 7 additional inputs; (g) MSAR-BM3D [18] with 1 additional input; (h) MSAR-BM3D with 3 additional inputs; (i) MSAR-BM3D with 7 additional inputs; (j) 2SPPB [19] with 1 additional input; (k) 2SPPB with 3 additional inputs; (l) 2SPPB with 7 additional inputs.

The matrices $\mathbf{W}_k$ can be obtained by Cholesky factorization of the inverse of the covariance matrix given in equation (29).

The closed-form expression of the Cholesky factorization in equation (29) leads to a simple definition of the whitened pair

$$\begin{cases} \mathring{\boldsymbol{z}}(t_i,k) = \tau\dot{\boldsymbol{z}}(t_i,k) + (1-\tau)\dot{\boldsymbol{d}}(t_i,k) \\ \qquad\qquad -\sqrt{\frac{\tilde{\boldsymbol{r}}(t_i,k)}{\tilde{\boldsymbol{r}}(t_{\text{ref}},k)}}\tau\tilde{\gamma}^*_{i\,\text{ref}}(k)(\dot{\boldsymbol{z}}(t_{\text{ref}},k) - \dot{\boldsymbol{d}}(t_{\text{ref}},k)) \\ \mathring{\boldsymbol{z}}(t_{\text{ref}},k) = \dot{\boldsymbol{z}}(t_{\text{ref}},k)\,, \end{cases} \tag{30}$$

where $\tau = 1/\sqrt{1-|\tilde{\gamma}_{i\,\text{ref}}(k)|^2}$. Note that only the complex amplitude $\mathring{\boldsymbol{z}}_{t_i}$ is modified while $\mathring{\boldsymbol{z}}_{\text{ref}}$ is left unchanged. This whitening procedure can thus be repeated for all pairs (t_i, t_{ref}), with $1 \leq t_i \leq T$ and $t_i \neq t_{\text{ref}}$, thereby producing a pre-processed stack in which images are all decorrelated with respect to the reference date t_{ref} (used in the subsequent processing as the target date for the despeckling task) and the decorrelated images provide information for the self-supervised training. Only the statistical independence with respect to this target date matters for the validity of the self-supervision used in section II-C.

We prove here that the whitened pair $(\mathring{\boldsymbol{z}}(t_i,k), \mathring{\boldsymbol{z}}(t_{\text{ref}},k))$ has indeed a diagonal covariance matrix.

We can rewrite the whitened pair as follows:

$$\begin{pmatrix} \mathring{\boldsymbol{z}}(t_i,k) \\ \mathring{\boldsymbol{z}}(t_{\text{ref}},k) \end{pmatrix} = \begin{pmatrix} \tau & -\sqrt{\frac{\tilde{\boldsymbol{r}}(t_i,k)}{\tilde{\boldsymbol{r}}(t_{\text{ref}},k)}}\tau\tilde{\gamma}^*_{i\,\text{ref}}(k) \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \dot{\boldsymbol{z}}(t_i,k) - \dot{\boldsymbol{d}}(t_i,k) \\ \dot{\boldsymbol{z}}(t_{\text{ref}},k) - \dot{\boldsymbol{d}}(t_{\text{ref}},k) \end{pmatrix} + \begin{pmatrix} \dot{\boldsymbol{d}}(t_i,k) \\ \dot{\boldsymbol{d}}(t_{\text{ref}},k) \end{pmatrix}. \tag{31}$$

Since the centered dominant component is deterministic, it follows from equations (31) and (26) that

$$\begin{aligned} \operatorname{Cov}\left[\begin{pmatrix} \mathring{\boldsymbol{z}}(t_i,k) \\ \mathring{\boldsymbol{z}}(t_{\text{ref}},k) \end{pmatrix}\right] &= \operatorname{Cov}\left[\begin{pmatrix} \mathring{\boldsymbol{z}}(t_i,k) - \dot{\boldsymbol{d}}(t_i,k) \\ \mathring{\boldsymbol{z}}(t_{\text{ref}},k) - \dot{\boldsymbol{d}}(t_{\text{ref}},k) \end{pmatrix}\right] \\ &= \begin{pmatrix} \tau & -\sqrt{\frac{\tilde{\boldsymbol{r}}(t_i,k)}{\tilde{\boldsymbol{r}}(t_{\text{ref}},k)}}\tau\tilde{\gamma}^*_{i\,\text{ref}}(k) \\ 0 & 1 \end{pmatrix} \\ &\qquad \operatorname{Cov}\left[\begin{pmatrix} \dot{\boldsymbol{z}}(t_i,k) \\ \dot{\boldsymbol{z}}(t_{\text{ref}},k) \end{pmatrix}\right] \begin{pmatrix} \tau & 0 \\ -\sqrt{\frac{\tilde{\boldsymbol{r}}(t_i,k)}{\tilde{\boldsymbol{r}}(t_{\text{ref}},k)}}\tau\tilde{\gamma}_{i\,\text{ref}}(k) & 1 \end{pmatrix} \\ &= \begin{pmatrix} \tilde{\boldsymbol{r}}(t_i,k) & 0 \\ 0 & \tilde{\boldsymbol{r}}(t_{\text{ref}},k) \end{pmatrix}. \end{aligned} \tag{32}$$

This proves that, for each pixel k, the two complex amplitudes are decorrelated. Since they are jointly Gaussian and decorrelated, they are statistically independent.

References

[1] A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 1, pp. 6–43, 2013.

[2] J. Lee, "Digital image smoothing and the sigma filter," *Computer vision, graphics, and image processing*, vol. 24, no. 2, pp. 255–269, 1983.

[3] F. Argenti, A. Lapini, T. Bianchi, and L. Alparone, "A tutorial on speckle reduction in synthetic aperture radar images," *IEEE Geoscience and remote sensing magazine*, vol. 1, no. 3, pp. 6–35, 2013.

[4] C.-A. Deledalle, L. Denis, G. Poggi, F. Tupin, and L. Verdoliva, "Exploiting patch similarity for SAR image processing: the nonlocal paradigm," *IEEE Sig. Proc. Mag.*, vol. 31, no. 4, pp. 69–78, 2014.

[5] X. Zhu, S. Montazeri, M. Ali, Y. Hua, Y. Wang, L. Mou, Y. Shi, F. Xu, and R. Bamler, "Deep learning meets SAR: concepts, models, pitfalls, and perspectives," *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, 2021.

[6] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva, "Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 29–51, 2021.

[7] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, and P. Ghamisi, "Image restoration for remote sensing: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–31, 2021.

[8] S. Vitale, D. Cozzolino, G. Scarpa, L. Verdoliva, and G. Poggi, "Guided patchwise nonlocal sar despeckling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6484–6498, 2019.

[9] S. Vitale, G. Ferraioli, and V. Pascazio, "Analysis on the building of training dataset for deep learning sar despeckling," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[10] E. Dalsasso, L. Denis, and F. Tupin, "SAR2SAR: a semi-supervised despeckling algorithm for SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4321–4329, 2021.

[11] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning Image Restoration without Clean Data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2965–2974.

[12] X. Ma, C. Wang, Z. Yin, and P. Wu, "Sar image despeckling by noisy reference-based deep learning method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8807–8818, 2020.

[13] A. B. Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Speckle2void: Deep self-supervised sar despeckling with blind-spot convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.

[14] S. Laine, T. Karras, J. Lehtinen, and T. Aila, "High-quality self-supervised deep image denoising," in *Advances in Neural Information Processing Systems*, 2019, pp. 6970–6980.

[15] E. Dalsasso, L. Denis, and F. Tupin, "As if by magic: self-supervised training of deep despeckling networks with merlin," *IEEE Transactions on Geoscience and Remote Sensing*, p. early access, 2021.

[16] S. Quegan and J. J. Yu, "Filtering of multichannel sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 11, pp. 2373–2379, 2001.

[17] S. Parrilli, M. Poderico, C. V. Angelino, and L. Verdoliva, "A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 2, pp. 606–616, 2011.

[18] G. Chierchia, M. El Gheche, G. Scarpa, and L. Verdoliva, "Multitemporal sar image despeckling based on block-matching and collaborative filtering," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5467–5480, 2017.

[19] X. Su, C.-A. Deledalle, F. Tupin, and H. Sun, "Two-step multitemporal nonlocal means for synthetic aperture radar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 10, pp. 6181–6196, 2014.

[20] C.-A. Deledalle, L. Denis, and F. Tupin, "Iterative weighted maximum likelihood denoising with probabilistic patch-based weights," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2661–2672, 2009.

[21] W. Zhao, C.-A. Deledalle, L. Denis, H. Maître, J.-M. Nicolas, and F. Tupin, "Ratio-Based Multitemporal SAR Images Denoising: RABASAR," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

[22] E. Dalsasso, I. Meraoumia, L. Denis, and F. Tupin, "Exploiting multi-temporal information for improved speckle reduction of Sentinel-1 SAR images by deep learning," in *IGARSS 2021*, Bruxelles (virtual), Belgium, Jul. 2021.

[23] J. W. Goodman, *Speckle phenomena in optics: theory and applications*. Roberts and Company Publishers, 2007.

[24] R. Bamler and P. Hartl, "Synthetic aperture radar interferometry," *Inverse problems*, vol. 14, no. 4, p. R1, 1998.

[25] M. Havasi, R. Jenatton, S. Fort, J. Z. Liu, J. Snoek, B. Lakshminarayanan, A. M. Dai, and D. Tran, "Training independent subnetworks for robust prediction," in *ICLR*, 2021.

[26] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: a view from the width," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6232–6240.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] R. Abergel, L. Denis, S. Ladjal, and F. Tupin, "Subpixellic methods for sidelobes suppression and strong targets extraction in single look complex SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 759–776, 2018.

[29] A. Bordone Molini, D. Valsesia, G. Fracastoro, and E. Magli, "Deepsum: Deep neural network for super-resolution of unregistered multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3644–3656, 2020.

[30] M. Deudon, A. Kalaitzis, I. Goytom, M. R. Arefin, Z. Lin, K. Sankaran, V. Michalski, S. E. Kahou, J. Cornebise, and Y. Bengio, "Highres-net: Recursive fusion for multi-frame super-resolution of satellite imagery," 2020.

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