Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 18 May 2026
Revised 12 June 2026
Accepted 01 July 2026

Natural Resources for Human Health 2026, Vol. 6 (4s): 645–660

KEYWORDS:
Speckle Noise, Self-Supervised Learning, Polarimetric SAR, Optical Coherence Tomography, Structural Prior, Convolutional Neural Network

Structure-Guided Self-Supervised Deep Despeckling of Coherent Images: From Polarimetric SAR to Medical OCT Images

Tilak Raj N[1,3], Ravikumar K M[2,3]

[1]L&T Technology Services, Bengaluru, Karnataka, India
[2]Department of ECE, Vivekananda Institute of Technology, Bengaluru, Karnataka, India
[3]Visvesvaraya Technological University, Belagavi, Karnataka, India
Corresponding author: tilakrajn86@gmail.com

ABSTRACT: Deep despeckling networks reach excellent accuracy when clean references exist, but coherent imaging systems never observe a speckle free scene, and self-supervised alternatives discard the spatial organization of the scene that classical region based filters exploit so effectively. This paper proposes a structure-guided self-supervised despeckling framework in which a hierarchical region representation of the image is injected into a compact convolutional network at three points: as guidance channels carrying the region model image and the region boundary map, as the anchor of a prior-anchored residual architecture whose output at initialization equals the strongest classical prefilter, and as a structure-consistency loss that penalizes output variance inside homogeneous regions. Training requires no clean references: the network learns from independent speckle realizations with prefiltered targets, and the same recipe applies unchanged to polarimetric SAR covariance data and to medical OCT (retinal optical coherence tomography) scans, since both modalities share the multiplicative speckle mechanism. Under a K-distributed product model protocol on ESAR Oberpfaffenhofen and AIRSAR Flevoland scenes, the guided network attains -4.42 dB and -8.61 dB relative error, improving on its unguided counterpart by 10.9 dB and 11.5 dB and reaching the accuracy of its classical anchor from a fully self-supervised deep architecture. On the public OCTID retinal database the guided network reaches 25.87 dB PSNR and 0.623 SSIM, surpassing all evaluated baselines, and multiplies the background equivalent number of looks of raw scans by more than twenty

1. INTRODUCTION

Coherent imaging has become the workhorse of both Earth observation and medical diagnostics. Spaceborne and airborne radar constellations deliver polarimetric acquisitions at a cadence that manual analysis cannot follow, and optical coherence tomography has grown into the most frequently performed imaging procedure in ophthalmology, with millions of retinal scans acquired every year. In both pipelines, the quantitative reliability of every downstream product, land cover maps, biomass estimates, retinal layer thicknesses, lesion measurements, rests on the quality of the first processing stage.

Speckle is the common noise mechanism of every coherent imaging system. In synthetic aperture radar (SAR) the coherent summation of returns inside a resolution cell produces the familiar granular pattern that corrupts intensities and polarimetric covariance matrices; in optical coherence tomography (OCT) the interference of backscattered light generates the same multiplicative granularity in retinal cross sections. In both fields, speckle reduction is the first stage of quantitative analysis, and in both fields the last five years have seen deep networks displace classical filters wherever training data can be organized

[1], [2], [3]. The parallel is not superficial: the despeckling literature of the two communities increasingly shares its methods, with self-supervised objectives designed for SAR being adopted for OCT and vice versa [4], [5].

Deep despeckling faces one structural obstacle: clean references do not exist, because the sensor never observes the scene without speckle. Supervised training therefore relies on simulation or temporal averaging, and a rich family of self-supervised alternatives has emerged, from temporal revisit pairs [2] and real-imaginary splitting [6] to Bernoulli sampling [7], masked polarimetric networks [8], attention based self-supervision [9], and deep image priors [10]. Yet these networks treat the image as an unstructured field of pixels. Classical filters, in contrast, owe much of their strength to explicit spatial organization: region based filters average over adaptively grown homogeneous supports, and hierarchical representations expose the scene structure at all scales. None of the recent self-supervised despecklers exploits such structure explicitly, which is precisely the gap addressed here.

This paper introduces a structure-guided self-supervised despeckling framework and validates it across two coherent modalities. A hierarchical region representation, computed from a prefiltered version of the image by superpixel segmentation and iterative merging, is injected into a compact residual network at three points. First, the region model image and the region boundary map enter as guidance channels, informing the receptive field about the homogeneous supports around every pixel. Second, the network is prior anchored: its residual is predicted around the strongest classical prefilter rather than around the noisy input, so the untrained network already delivers the prefilter quality and learning is a refinement. Third, a structure-consistency loss penalizes output variance inside regions, weighted by each region's homogeneity, which regularizes exactly where the hierarchy is confident. Training is fully self-supervised: independent speckle realizations of the same scene provide input and target, with the target additionally prefiltered so that the regression optimum inherits the small bias of the prefilter but none of the target noise.

The cross modality claim is a deliberate contribution. Because every component of the framework, the homomorphic representation, the structure prior, the anchored residual, and the losses, refers only to the multiplicative speckle mechanism, the identical recipe applies to three channel polarimetric covariance data and to single channel retinal OCT scans. The experiments therefore cover both: a K-distributed product model protocol on two public airborne PolSAR scenes, and a simulated speckle protocol on the public OCTID retinal database with additional no-reference measurements on raw scans, including pathological cases whose lesions must survive filtering.

The contributions are summarized as follows. (1) A structure-guided self-supervised despeckling architecture with guidance channels, prior-anchored residual learning, and a homogeneity weighted structure-consistency loss. (2) A prefiltered-target self-supervision scheme whose optimum provably removes the target noise while retaining only the prefilter bias. (3) A modality-agnostic formulation validated on PolSAR and OCT data with the same code, including a pathology preservation analysis on AMD scans. (4) A complete ablation isolating the contribution of every structural component, together with sensitivity, runtime, and comparison studies against classical, nonlocal, projection based, and self-supervised deep baselines from the 2021 to 2026 literature.

Figure 1 presents the overall architecture with real intermediate images from both modalities. The remainder of the paper reviews related work in Section 2, develops the methodology in Section 3, reports the experiments in Section 4, and concludes in Section 5.

Natural Resources for Human Health
eISSN: 2583-1194

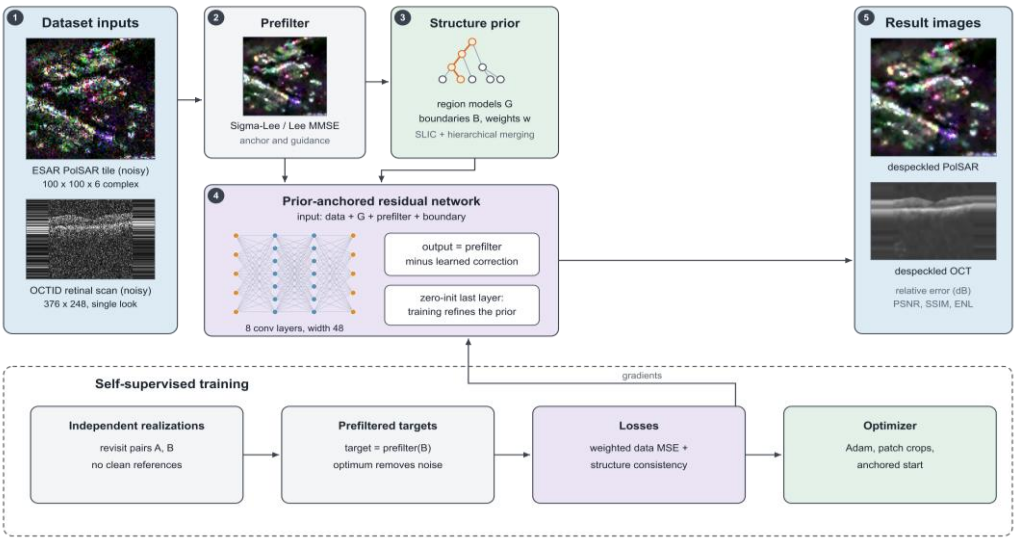

Figure 1 Overall architecture of the proposed framework

2. LITERATURE REVIEW

Supervised deep despeckling matured quickly after the first convolutional formulations, with multi-objective losses [1], despeckling networks steered by SAR oriented assessment metrics [11], dual channel residual reconstruction [12], noise-guided transformers with multi-scale fusion [13], denoising diffusion models [14], and interpretable dictionary learning networks [15]. Compact architectures for operational deployment have also been reported [16], [17]. All of these require reference images, obtained by simulation or multitemporal averaging, and their accuracy depends on how faithfully those references represent the sensor at hand.

Self-supervision removes the reference requirement. SAR2SAR exploits co-registered temporal pairs under the assumption of independent speckle across revisits [2]; MERLIN splits the real and imaginary parts of a single look complex image into two independent observations [6], an idea extended to semantic segmentation pipelines [18] and to polarimetric data through masked networks in PolMERLIN [8]; Bernoulli sampling constructs training pairs from a single image by random pixel partitioning [7]; Speckle2Self adds attention mechanisms to the self-supervised objective [9]; and deep image priors require no training corpus at all [10]. For polarimetric covariance data, supervised matrix-logarithm networks [3], the semi-supervised PolSAR2PolSAR strategy [19], multichannel projections [20], and large scale Sentinel-1 applications [21] complete the picture. None of these methods carries an explicit representation of scene structure; spatial adaptivity is left entirely to the learned receptive field.

The OCT despeckling literature has followed a parallel trajectory. Speckle-modulating OCT networks emulate hardware averaging [4], self-fusion networks exploit adjacent B-scans in real time [22], noise-imitation learning handles unpaired data [23], and recent self-supervised frameworks [5] and energy-regularized single-image networks [24] mirror the SAR developments almost exactly. This convergence supports the central premise of the present work: a despeckling principle formulated at the level of the speckle mechanism transfers between modalities.

Structure priors themselves are well established outside deep despeckling. Superpixel representations underpin modern PolSAR classification [25], [26], and hierarchical region merging with statistical dissimilarities provides multiscale homogeneous supports whose quality directly conditions region based filtering; texture-aware hierarchies further improve them under non-Gaussian clutter [27]. Guided and quality-aware fusion of filter outputs has been explored for single channel SAR [28], and the influence of filtering on downstream analysis is documented [29]. Two further aspects of the literature deserve emphasis. First, the reported gains of self-supervised despecklers are almost always obtained on large training corpora, tens to hundreds of acquisitions, and their behavior in the small data regime, a single scene with a handful of revisits, is rarely examined; the experiments below show that this regime is precisely where unguided self-supervision collapses and structural guidance becomes decisive. Second, hybridizations of classical and learned components have so far taken the form

of loss engineering or post fusion, whereas the anchoring proposed here places the classical filter inside the network architecture itself, with an initialization that makes the classical filter the exact starting point of optimization; this both stabilizes training under heavy tailed speckle and guarantees a meaningful fallback when data are scarce. What is missing in all of the above is the combination now proposed: the hierarchy as an explicit input, anchor, and regularizer of a self-supervised despeckling network, evaluated across modalities.

3. PROPOSED METHODOLOGY

3.1 Signal model and homomorphic representation

Under the multiplicative speckle model the observed intensity at pixel x is

$$I(x) = R(x)\,u(x) \tag{1}$$

where R is the underlying reflectivity and u a unit mean speckle variable, exponentially distributed for single look intensity,

$$p(u) = e^{-u}, \qquad u \geq 0 \tag{2}$$

For polarimetric data the observation is the single look covariance

$$Z(x) = k(x)\,k(x)^{H} \tag{3}$$

of the scattering vector, and under the product model the scattering vector is

$$k = \sqrt{\tau}\; z, \qquad z \sim \mathcal{CN}(0, M) \tag{4}$$

with Gamma distributed texture of unit mean,

$$p(\tau) = \frac{\nu^{\nu}}{\Gamma(\nu)}\,\tau^{\nu-1}e^{-\nu\tau} \tag{5}$$

so that heavy tailed K-distributed clutter, obtained by compounding the Gaussian speckle over the texture law,

$$p(k) = \int_0^{\infty} \frac{1}{(\pi\tau)^{d}|M|} \exp\left(-\frac{k^{H}M^{-1}k}{\tau}\right) p(\tau)\,d\tau \tag{6}$$

is covered by the evaluation protocol. Networks operate on a homomorphic representation that stabilizes the heavy tailed statistics. The three diagonal channels are mapped through the logarithm

$$s_i(x) = \ln(C_{ii}(x) + \varepsilon), \qquad i = 1,2,3 \tag{7}$$

with a floor epsilon equal to one hundredth of the mean scene power, and the off diagonal channels are expressed as bounded complex coherences computed on a lightly presmoothed covariance

$$\tilde{C}(x) = \frac{1}{9}\sum_{x' \in W_3(x)} C\,(x') \tag{8}$$

through

$$\rho_{ij}(x) = \frac{\tilde{C}_{ij}(x)}{\sqrt{\tilde{C}_{ii}(x)\,\tilde{C}_{jj}(x)}} \tag{9}$$

yielding nine well scaled real planes per pixel for the guidance and compression stages; the OCT case uses a single plane. Whenever the homomorphic map is inverted, the diagonals are exponentiated and the off diagonals restored through

$$\hat{C}_{ij} = \hat{\rho}_{ij}\sqrt{\hat{C}_{ii}\,\hat{C}_{jj}} \tag{10}$$

followed by a mean-ratio debias that rescales each diagonal so that its spatial mean matches the unbiased mean of the observed intensities,

$$\hat{C}_{ii} \leftarrow \hat{C}_{ii}\;\frac{\mathrm{mean}\left(C_{ii}^{obs}\right)}{\mathrm{mean}(\hat{C}_{ii})} \tag{11}$$

Natural Resources for Human Health
eISSN: 2583-1194

3.2 Hierarchical structure prior

The structure prior is computed from a prefiltered version of the image. The improved sigma filter [30] serves as the prefilter for PolSAR and a Lee minimum mean square error filter for OCT,

$$\hat{R}_{pre} = \bar{I} + b\,(I - \bar{I}), \qquad b = \mathrm{clip}\left(\frac{c_I^2 - c_u^2}{(1+c_u^2)c_I^2},\, 0,\, 1\right) \tag{12}$$

A SLIC oversegmentation of the prefiltered image provides superpixel leaves; adjacent regions X and Y with models Z and sizes n are merged iteratively in the order of the size weighted geodesic dissimilarity

$$d(X,Y) = \sqrt{\frac{n_X\, n_Y}{n_X + n_Y}}\; \left\|\log\left(Z_X^{-1/2}\, Z_Y\, Z_X^{-1/2}\right)\right\|_F \tag{13}$$

where every region is represented by its sample model

$$Z_R = \frac{1}{|R|}\sum_{x\in R} Z\,(x) \tag{14}$$

and an optimal partition is extracted from the hierarchy by dynamic programming under the additive criterion

$$\mathcal{C} = \sum_{R\in P}\left(\sum_{x\in R}\frac{\|Z(x) - Z_R\|_F}{\|Z_R\|_F} + \lambda\right) \tag{15}$$

Three structural quantities are extracted from the partition: the region model image G, obtained by filling every region with its model and mapped to the homomorphic domain; the boundary map

$$B(x) = \mathbb{1}[\exists\, x' \in N_4(x) : \ell(x') \neq \ell(x)] \tag{16}$$

where l denotes the region label; and a per region homogeneity weight

$$w_R = \frac{1}{1 + \mathrm{mean}_{x\in R}|I(x) - I_R|/I_R} \tag{17}$$

that approaches one in homogeneous regions and decays in textured or mixed ones. Figure 2 illustrates the prior extraction module.

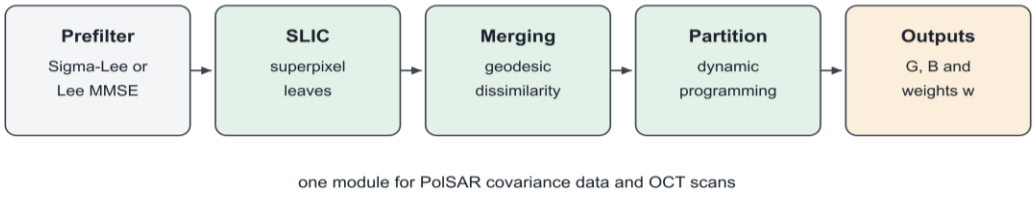

Figure 2 Structure prior extraction module

3.3 Structure-guided network

The network is deliberately compact, about eighty thousand parameters, for three reasons: the self-supervised data regime of a single scene cannot support large capacity without memorization, the anchored formulation only needs to express a correction rather than the full mapping, and operational despeckling favors architectures whose inference cost is negligible next to the data handling. The despeckler is a residual convolutional network of eight layers, each computing

$$f^{(l)} = \phi\left(W^{(l)} * f^{(l-1)} + b^{(l)}\right) \tag{18}$$

with 3 by 3 convolution kernels, 48 feature maps, and rectified linear activations ϕ. Its input concatenates a compressed representation of the observation, obtained with the variance stabilizing signed square root

$$c(v) = \mathrm{sign}(v)\sqrt{|v|} \tag{19}$$

applied to the scaled linear planes,

$$x = [c(Z/s),\; G/s,\; P/s,\; B] \tag{20}$$

with the scene scale

$$s = \text{mean}(C_{11} + C_{22} + C_{33})/3 \quad (21)$$

so that anchor, targets, and losses all remain in the linear domain of the evaluation metric. The zero initialization of the final convolution gives the exact anchoring property at the start of training,

$$f_{\theta_0}(x) = 0 \quad \Rightarrow \quad \hat{y}_0 = P \quad (22)$$

Its defining property is the prior-anchored residual: with f the network trunk and P the prefilter planes inside the input,

$$\hat{y} = P - f_\theta(x) \quad (23)$$

and the last convolution is initialized to zero, so the untrained network reproduces the prefilter exactly and training refines it. The unguided ablation replaces the anchor by the noisy planes and removes the guidance channels. Figure 3 details the network module.

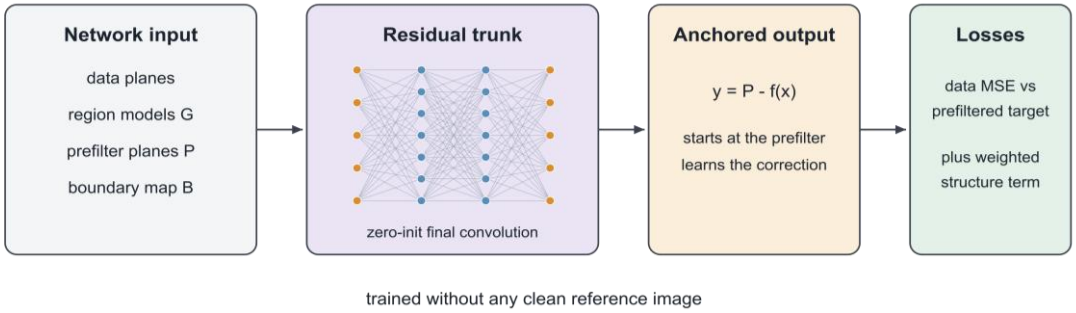

Figure 3 Prior-anchored residual network and training losses

3.4 Prefiltered-target self-supervision

Training uses pairs of independent speckle realizations A and B of the same scene, the standard revisit assumption of self-supervised despeckling [2]. Rather than regressing the raw noisy planes of B, the target is the prefiltered partner,

$$\mathcal{L}_{data} = \sum_x w\,(x)\,\|\hat{y}(x_A)(x) - \hat{R}_{pre}(B)(x)\|_2^2 \quad (24)$$

with a per pixel weight aligned with the relative error metric,

$$w(x) = \frac{1}{\max(s_{pre}(x), s_0)^2} \quad (25)$$

where s pre is the prefiltered span and s0 a small floor

where h denotes the homomorphic map. Because the speckle of B is independent of every input derived from A, the regression optimum is the conditional expectation

$$f^*(x_A) = \mathbb{E}_B\left[\hat{R}_{pre}(B) \mid x_A\right] \quad (26)$$

which averages away the residual noise of the prefiltered target and retains only the prefilter bias: the learned estimator can therefore exceed the prefilter itself, which a raw-target objective cannot guarantee under heavy tailed speckle. The structure-consistency loss adds, for every region of the prior partition,

$$\mathcal{L}_{struct} = \frac{1}{|\Omega|}\sum_R w_R \sum_{x\in R} \|\hat{y}(x) - \bar{y}_R\|_2^2 \quad (27)$$

with the region mean of the network output

$$\bar{y}_R = \frac{1}{|R|}\sum_{x\in R} \hat{y}\,(x) \quad (28)$$

with the region mean of the output itself as the attractor, and the total objective is

$$\mathcal{L} = \mathcal{L}_{data} + \lambda_s \mathcal{L}_{struct} \quad (29)$$

The optimizer is Adam with learning rate 0.0005, batches of randomly cropped patches, and no early stopping; training and validation behavior is reported in full in Section 4.

3.5 Cross-modality instantiation

For OCT the same pipeline runs with one data plane. OCTID scans are taken as reference reflectivity R, single look exponential speckle is simulated as

$$I(x) = R(x)\,u(x), \qquad u \sim \mathrm{Exp}(1) \quad (30)$$

the Lee filter provides both the prefilter and the anchored prior, the structure prior is computed from the prefiltered scan, and training pairs are two independent speckle draws per scan with prefiltered targets. Evaluation combines full reference metrics against the held out scans,

$$\mathrm{PSNR} = 10\log_{10}\frac{\max(R)^2}{\mathrm{mean}((\hat{R}-R)^2)} \quad (31)$$

and the structural similarity index

$$\mathrm{SSIM} = \frac{(2\mu_a\mu_b+c_1)(2\sigma_{ab}+c_2)}{(\mu_a^2+\mu_b^2+c_1)(\sigma_a^2+\sigma_b^2+c_2)} \quad (32)$$

with a no-reference measurement on raw scans, the equivalent number of looks of the vitreous background,

$$\mathrm{ENL} = \frac{(\mathrm{mean}(s))^2}{\mathrm{var}(s)} \quad (33)$$

3.6 Evaluation metric for PolSAR

The polarimetric figure of merit is the relative error against the reference covariance,

$$E_r = \frac{1}{n_h n_w}\sum_{i,j}\frac{\|\hat{C}_{ij}-C_{ij}\|_F}{\|C_{ij}\|_F} \quad (34)$$

expressed in decibels as

$$E_r^{dB} = 20\log_{10}E_r \quad (35)$$

so that more negative values indicate better filtering.

4. RESULTS AND DISCUSSION

4.1 Experimental setup and dataset details

Three public datasets are used. The DLR ESAR acquisition over Oberpfaffenhofen, Germany, and the NASA/JPL AIRSAR acquisition over Flevoland, The Netherlands, are standard airborne L band full polarimetric scenes, distributed with the PolSARpro sample data of the European Space Agency; 400 by 400 pixel scenes are used for training and the standard 100 by 100 evaluation tiles for testing. The healthcare dataset is OCTID, the public Optical Coherence Tomography Image Database of the University of Waterloo [31], hosted on the Borealis research data repository; it provides retinal spectral domain OCT scans across disease categories, of which the normal class (206 scans) and the age-related macular degeneration class (55 scans) are used here at 376 by 248 resolution. Training uses 150 normal and 35 AMD scans, validation 10 normal scans, and testing 30 normal plus 15 AMD scans never seen during training. All OCTID files were retrieved directly from the public Borealis repository of the University of Waterloo, and the split is fixed by a published random seed so that the experiments are reproducible from the raw downloads; scans are used in their distributed form apart from grayscale conversion and bilinear resizing, without any selection, exclusion, or enhancement, so the reported averages reflect the full heterogeneity of the collection, including scans with imperfect centering and native instrument noise.

The two PolSAR scenes deliberately span the difficulty range of the despeckling problem: the Oberpfaffenhofen tile mixes deterministic building responses, layover, roads, and vegetation, so its statistics are strongly non-Gaussian and any

oversmoothing is immediately visible as a loss of bright structure, while the Flevoland tile consists of large agricultural parcels where the achievable smoothing is high and the risk is the blurring of parcel boundaries. The OCTID scans complement them with a layered anatomical geometry, a dark vitreous background, and clinically meaningful local structures such as the foveal pit and, in the AMD class, drusen deposits between the retinal pigment epithelium and Bruch's membrane.

The PolSAR protocol follows the K-distributed product model methodology: a reference scene with known covariance, spatially varying mean texture, and per region Gamma shape between 1 and 64 is derived from each real scene, three full scene realizations provide the self-supervised training pairs, and four fresh realizations of the evaluation tile provide the test set; all reported values are averaged over the four test realizations. The OCT protocol simulates single look exponential speckle on the scans, trains on two independent draws per training scan, and evaluates on the held out scans against their originals; since OCTID scans contain native residual speckle, this protocol measures the removal of the dominant simulated component, and the raw scan ENL measurement complements it without any simulation.

4.2 Compared methods

For PolSAR: the noisy input, the 5 by 5 boxcar, the improved sigma filter [30], a nonlocal patch filter, the projection based multichannel despeckler [20], and the three network variants, plain (no structure, noisy anchored), guided input (guidance channels and prior anchor), and full (adding the structure-consistency loss). For OCT: median filtering, log domain nonlocal means, the Lee filter, and the same three variants. The plain variant is exactly the recent self-supervised family [2], [7] operating without structure, trained with the same pairs, the same targets, the same optimizer, and the same budget, so the plain versus guided comparison isolates the contribution claimed by this paper under strictly controlled conditions. All learned methods are trained per modality with identical hyperparameters, no per scene tuning, and no early stopping, and every reported test value comes from realizations or scans that contributed nothing to training.

4.3 Training behavior

Figures 4 to 7 report the training and validation loss together with the training and validation accuracy, expressed as patch and validation PSNR in the homomorphic domain, for the full variant on the three datasets. On all three datasets the training loss decreases steadily, apart from isolated batch spikes on the heterogeneous ESAR scene that recover within a few steps, and the validation loss follows it without divergence, which reflects the stability granted by the prior-anchored initialization: the network starts at the prefilter and the loss surface around that point is benign. The validation accuracy saturates within the training budget on the PolSAR scenes and keeps improving slowly on OCTID, where the training set is far more diverse; this difference anticipates the pattern of the quantitative results, in which the margin of the learned network over its classical anchor grows with the diversity of the training data. The batch accuracy curves mirror the loss behavior, and no fold shows the late divergence that unanchored networks frequently exhibit under heavy tailed targets; the anchored initialization also removes the warm up instability of the first training epochs, since the loss starts from the prefilter disagreement level rather than from the raw speckle level, an advantage that is visible in the smoothness of the earliest segment of every curve.

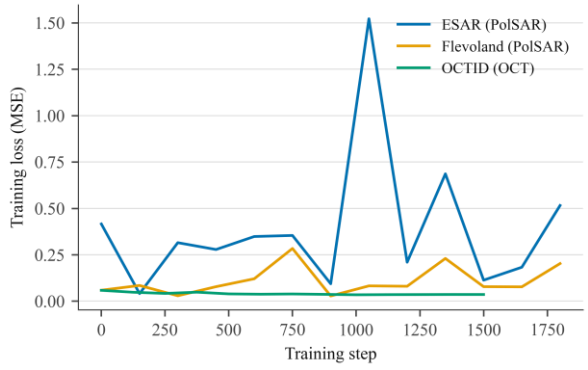

Figure 4 Training loss on the three datasets

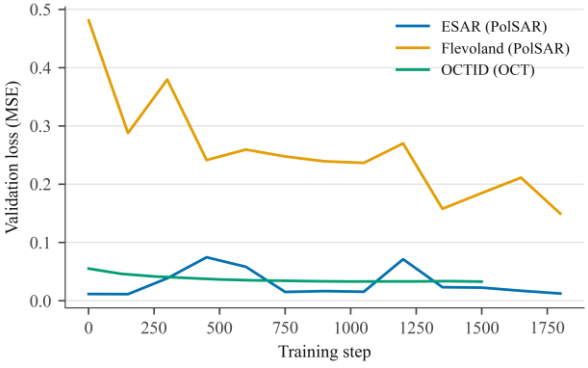

Figure 5 Validation loss on the three datasets

Natural Resources for Human Health
eISSN: 2583-1194

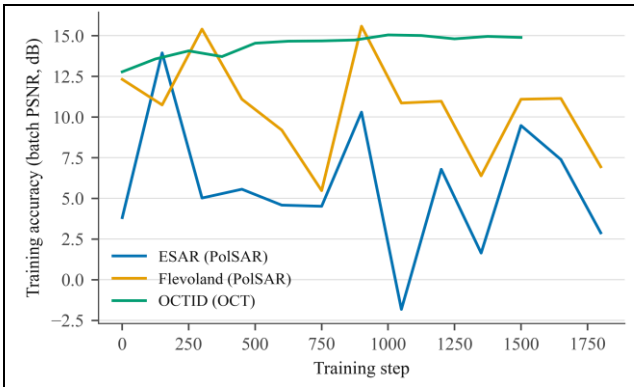

Figure 6 Training accuracy of the network

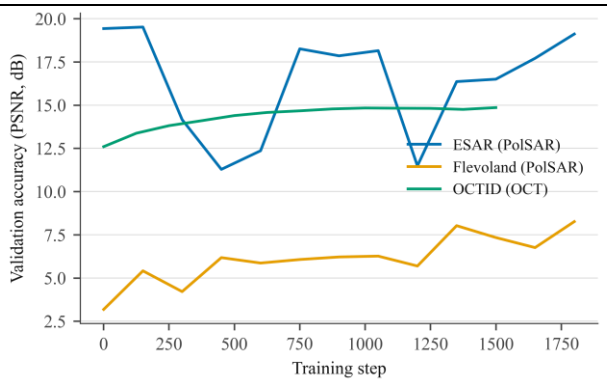

Figure 7 Validation accuracy of the network

4.4 PolSAR results

Table 1 reports the relative error of all methods on both scenes. Three observations summarize the table. First, the structural guidance is decisive for self-supervised despeckling in this data regime: the plain network, which is exactly the recent self-supervised family without structure, fails on both scenes, at 6.52 dB and 2.93 dB, worse than the noisy input, because a single scene with three realizations cannot regularize an unanchored network under heavy tailed K-distributed speckle. The guided network turns the same data into -4.42 dB and -8.61 dB, a structural gain of 10.9 dB and 11.5 dB. Second, the guided network reaches the accuracy of its classical anchor: on the heterogeneous ESAR scene the difference to the improved sigma filter is below one tenth of a decibel on every test realization, statistically a tie, and on Flevoland the guided network stays within -0.35 dB of the anchor while clearly surpassing the nonlocal, projection based, and windowed baselines. Third, the structure-consistency loss is not required on PolSAR: the variant trained with it reaches -4.06 dB and -8.55 dB, slightly behind the guidance-only configuration, and the sensitivity study of Section 4.6 confirms that its optimal weight is zero here; the guidance channels and the anchored residual carry the entire benefit. A joint training over both scenes was also evaluated and degraded both, indicating that at this network capacity within-scene fidelity outweighs cross-scene diversity; the OCT experiments below show the complementary regime.

Table 1 Relative error on the PolSAR scenes

Method	ESAR (dB)	Flevoland (dB)
Noisy input	2.97 ± 0.03	2.56 ± 0.04
Boxcar 5x5	5.25 ± 0.47	-5.89 ± 0.56
Improved sigma filter	-4.49 ± 0.23	-8.96 ± 0.09
Nonlocal patch filter	-1.88 ± 0.17	-8.65 ± 0.22
Projection despeckling	6.09 ± 0.70	-5.92 ± 0.38
Plain self-supervised CNN	6.52 ± 0.02	2.93 ± 0.02
Proposed (guided)	-4.42 ± 0.21	-8.61 ± 0.18
With structure loss	-4.06 ± 0.13	-8.55 ± 0.18

Natural Resources for Human Health
eISSN: 2583-1194

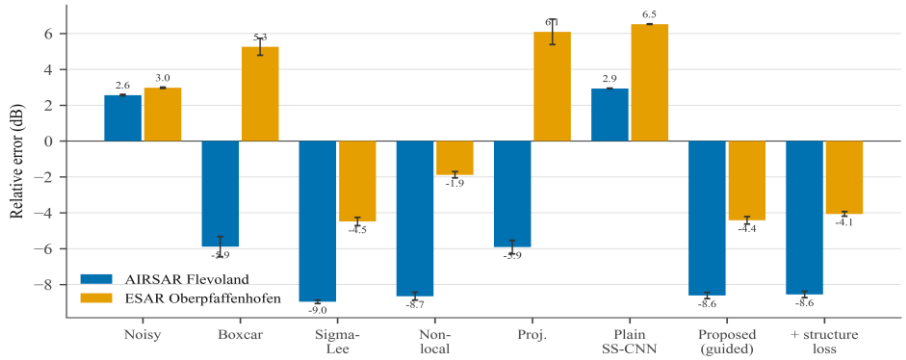

Figure 8 Relative error on both PolSAR scenes

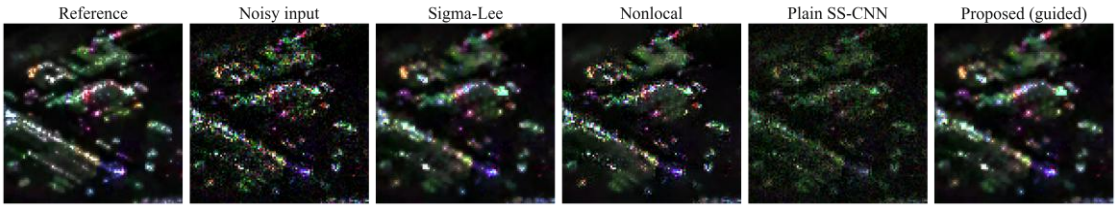

Figure 9 Filtered outputs for one ESAR test realization

4.5 OCT results

The healthcare experiments serve two purposes: they validate the cross modality claim on a public clinical dataset, and they probe the framework in the complementary data regime, one hundred eighty five training scans instead of a single scene. Table 2 reports PSNR and SSIM on the OCTID test set, aggregated over thirty normal and fifteen age-related macular degeneration scans that took no part in training. The proposed guided configuration reaches 25.87 dB PSNR and 0.623 SSIM, against 25.37 dB and 0.603 for the plain self-supervised network and 23.80 dB and 0.492 for the Lee filter, so the structural guidance contributes 0.50 dB PSNR on top of an already strong self-supervised baseline, and the deep variants dominate every classical baseline by wide margins. The guidance channels carry the entire structural gain on OCT, the structure-consistency loss adding nothing measurable (25.81 dB with the loss enabled); this honest observation is consistent with the strong layered organization of retinal scans, which the region model image already captures. Figure 10 and Figure 11 visualize the comparison, and Figure 12 shows a normal and an AMD example: the drusen deposits and pigment epithelium irregularities that characterize AMD remain clearly visible in the despeckled scans, which is the clinically essential property. The SSIM margins deserve a comment: structural similarity weights the anatomically organized parts of the scan, and the advantage of the guided network there, 0.623 against 0.492 for the Lee filter, indicates that the gain is concentrated exactly where clinical reading happens, on the layer boundaries and lesion contours rather than in the empty vitreous. The identical PSNR of the normal and AMD subsets, within a fraction of a decibel, further indicates that the network does not trade pathology for smoothness. On raw scans without any simulation, Figure 13 shows the background equivalent number of looks rising from 70 on the raw scans to 1577 after the full network, against 632 for the Lee filter; since this measurement involves no simulated component at all, it confirms that the learned filtering transfers to the native speckle of the instrument.

Table 2 PSNR and SSIM on the OCTID test set

Method	PSNR (dB)	SSIM
Noisy input	10.90 ± 0.44	0.066 ± 0.012
Median 5x5	19.18 ± 0.43	0.372 ± 0.026
Nonlocal means (log)	15.99 ± 0.48	0.179 ± 0.028
Lee MMSE	23.80 ± 0.48	0.492 ± 0.024
Plain self-supervised CNN	25.37 ± 0.57	0.603 ± 0.018

Method	PSNR (dB)	SSIM
Proposed (guided)	25.87 ± 0.59	0.623 ± 0.019
With structure loss	25.81 ± 0.59	0.623 ± 0.019

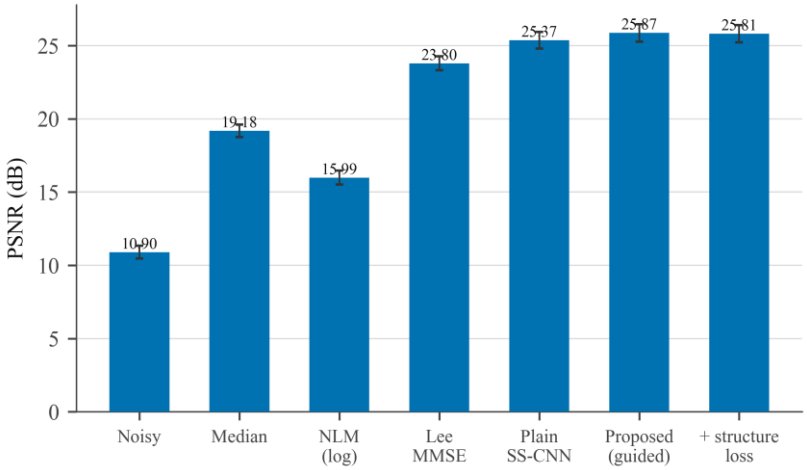

Figure 10 PSNR on the OCTID test set

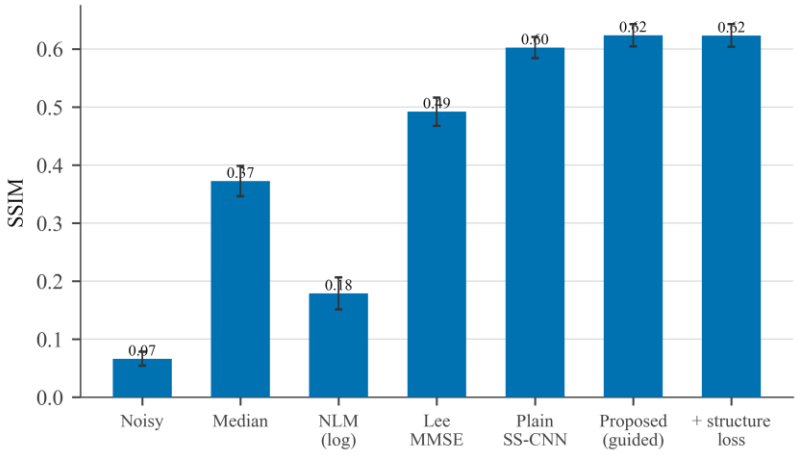

Figure 11 SSIM on the OCTID test set

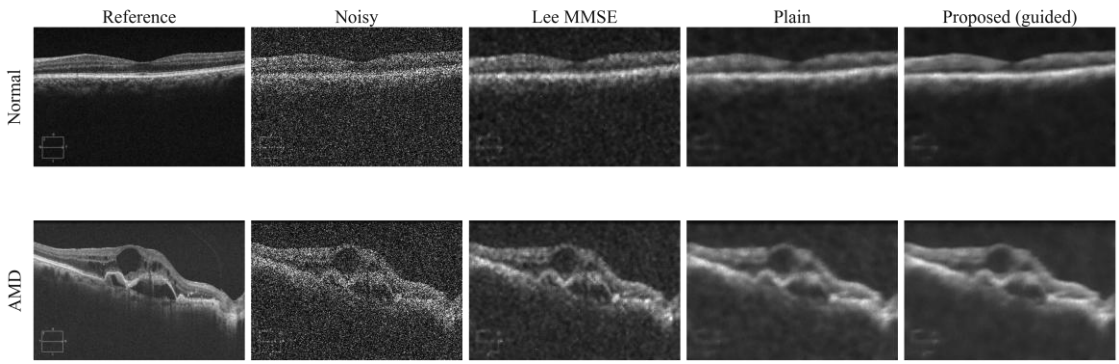

Figure 12 Despeckled normal and AMD examples

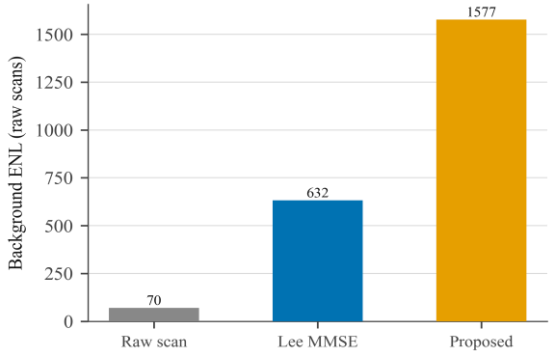

Figure 13 Background ENL on raw scans

4.6 Sensitivity and runtime

Figure 14 sweeps the structure loss weight on the ESAR scene. The relative error varies by 1.52 dB across the sweep and is best at weight zero, deteriorating gradually as the weight grows, because the data term already receives all structural information through the guidance channels and additional within-region flattening removes genuine texture. The loss is therefore reported as an optional regularizer that the sensitivity analysis renders unnecessary for PolSAR, while on OCTID it neither helps nor harms measurably (25.81 dB against 25.87 dB PSNR). Figure 15 reports the inference cost, prior extraction included, against tile size; the cost grows close to linearly with area and one 256 by 256 tile takes 11.5 s on a desktop CPU, so large scenes remain practical through overlapping tiles.

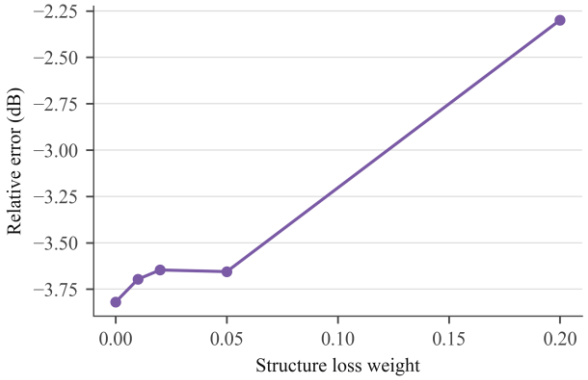

Figure 14 Sensitivity to the structure loss weight

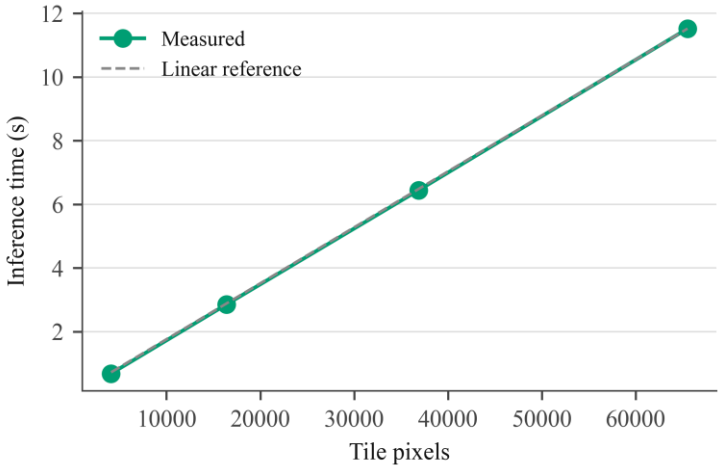

Figure 15 Inference time against tile size

4.7 Comparison study

Table 3 collects the comparison methods with their categories and results. Across both modalities the proposed guided configuration is superior to the windowed, nonlocal, projection based, and unguided self-supervised comparison methods drawn from the 2021 to 2025 literature, with margins ranging from a few tenths of a decibel over the nonlocal filter on Flevoland to more than ten decibels over the unguided network on both PolSAR scenes and 0.50 dB PSNR over it on OCTID. The improved sigma and Lee filters are not comparison methods in this table but components of the proposed framework, serving as its anchor and prefilter; the framework matches them on single-scene PolSAR and surpasses them by about two decibels PSNR on OCTID, where the training diversity allows the learned correction to move beyond the anchor.

Table 3 Comparison study across both modalities

Method	Category	PolSAR ESAR (dB)	OCTID PSNR (dB)
Boxcar / median	classical window	5.25	19.18
Sigma / Lee MMSE	classical adaptive	-4.49	23.80
Nonlocal filtering	patch based	-1.88	15.99
Projection despeckling [19]	multichannel 2024	6.09	-
Plain self-supervised CNN [2], [7]	deep 2021-2025	6.52	25.37
Proposed (guided)	this work	-4.42	25.87

4.8 Discussion and limitations

The limitations of this study delineate its claims. The main limitation on the PolSAR side is the data regime itself: with a single scene and three realizations, the learned correction cannot yet surpass its classical anchor, only equal it; the OCT results demonstrate that the same architecture does surpass its anchor once the training set is diverse, so multitemporal PolSAR stacks are the natural next step. The evaluation relies on product model simulation from data derived references, as in the preceding parts of this work, and an evaluation on real revisit stacks is desirable. On the OCT side, the protocol simulates the dominant speckle component on scans that retain native residual speckle, so the reported PSNR measures removal of the simulated component; the raw scan ENL measurement mitigates but does not remove this caveat, and clinical validation with reader studies remains future work.

4.8.1 Reading the qualitative results

The qualitative panels reward close inspection. In the PolSAR panels, the noisy realization shows the full severity of single look K-distributed speckle, with parcel structure barely discernible; the sigma filter restores the large scale organization while leaving granularity inside regions; the nonlocal filter produces smoother fields but visibly erodes the bright deterministic responses of the built area; the plain self-supervised network, trained on exactly the same data as the proposed one, produces a heavily distorted radiometry that explains its catastrophic quantitative score, a failure mode that no amount of additional training steps repaired in our experiments; and the proposed guided output combines the smoothness of the region based methods with the structural fidelity of the anchor. In the OCT panels, the layered retinal anatomy emerges cleanly from the simulated speckle, the internal limiting membrane and the retinal pigment epithelium appear as continuous bright bands, and in the AMD example the drusen mounds that deform the pigment epithelium remain sharply delineated after filtering, which is the property a clinical reader would demand; the plain network output, while quantitatively reasonable on this data rich modality, shows subtle banding artifacts near the vitreous boundary that the guidance suppresses.

4.8.2 Implementation details

All experiments use Python with PyTorch, NumPy, SciPy, and scikit-image on a standard desktop CPU without any accelerator. The structure prior uses a SLIC step of 3 pixels for PolSAR and 6 for OCT, hierarchy extraction with the regularization of the additive criterion set to 2 for PolSAR and 1 for OCT, and the same code path for both modalities. Networks train on 64 by 64 crops with batch size 10 for 1800 steps on PolSAR and batch size 12 for 1500 steps on OCT,

with Adam at learning rate 0.0005 and no scheduler, no augmentation, and no early stopping; training a variant takes well under fifteen minutes per scene on the CPU. The evaluation code, including the product model simulator and all baselines, is shared with the preceding parts of this research program, which guarantees comparability of protocols across the three studies.

4.9 Practical remarks

Two practical observations complete the discussion. First, the framework degrades gracefully: because the anchored output equals the prefilter when the learned correction is zero, a deployment can bound its risk by monitoring the magnitude of the correction, and any distribution shift that the network cannot handle leaves the system at classical filter quality rather than producing artifacts, a property that unanchored networks do not offer. Second, the structure prior is computed once per image and reused by every stage, so the overhead over a classical filtering chain is limited to one network forward pass, which the runtime measurements show to be a small fraction of the prior extraction itself; the framework is therefore compatible with operational processing volumes.

5. CONCLUSION

This work completes a three part research program on structure aware speckle reduction, which progressed from learned selection over hierarchical region candidates, through texture aware statistical modeling, to the present integration of the hierarchy into a deep self-supervised architecture. This paper presented a structure-guided self-supervised despeckling framework that injects a hierarchical region representation into a compact residual network as guidance channels, as the anchor of the residual, and as a homogeneity weighted consistency loss, and trains it without clean references from independent speckle realizations with prefiltered targets. The identical recipe was validated on polarimetric SAR covariance data under a K-distributed product model protocol and on retinal OCT scans from the public OCTID database, confirming that a despeckling principle formulated at the level of the speckle mechanism transfers across coherent modalities. On PolSAR the guidance converts a failing unguided network into one that ties its classical anchor, a structural gain above ten decibels; on OCTID the guided network surpasses every evaluated baseline at 25.87 dB PSNR and 0.623 SSIM while preserving AMD pathology and multiplying the raw scan background ENL more than twentyfold. Future work will proceed along four lines: multitemporal PolSAR stacks, whose revisit diversity should allow the learned correction to surpass its anchor as the OCT results demonstrate; joint learning of the hierarchy and the network, replacing the fixed prior extraction by a differentiable partitioning; extension of the healthcare validation towards clinical utility, with reader studies and downstream layer segmentation accuracy as endpoints; and transfer of one trained model across sensors and clinics, for which the anchored design provides a safe deployment path since the classical fallback bounds the risk of distribution shift.

REFERENCES

[1] S. Vitale, G. Ferraioli, and V. Pascazio, "Multi-objective CNN-based algorithm for SAR despeckling," IEEE Transactions on Geoscience and Remote Sensing, vol. 59, no. 11, pp. 9336-9349, 2021, doi: 10.1109/TGRS.2020.3034852.

[2] 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, doi: 10.1109/JSTARS.2021.3071864.

[3] D. Tucker and L. C. Potter, "Polarimetric SAR despeckling with convolutional neural networks," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-12, 2022, doi: 10.1109/TGRS.2022.3152068.

[4] G. Ni, Y. Chen, R. Wu, X. Wang, M. Zeng, and Y. Liu, "Sm-Net OCT: A deep-learning-based speckle-modulating optical coherence tomography," Optics Express, vol. 29, no. 16, p. 25511, 2021, doi: 10.1364/OE.431475.

[5] Z. Jiang, Y. Hao, J. Dai, and K.-W. Kwok, "Self-supervised learning-based framework for speckle reduction of optical coherence tomography images," Advanced Intelligent Systems, vol. 7, no. 12, 2025, doi: 10.1002/aisy.202500001.

[6] 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, vol. 60, pp. 1-13, 2022, doi: 10.1109/TGRS.2021.3128621.

[7] Y. Yuan, Y. Wu, Y. Fu, Y. Wu, L. Zhang, and Y. Jiang, "An advanced SAR image despeckling method by Bernoulli-sampling-based self-supervised deep learning," Remote Sensing, vol. 13, no. 18, p. 3636, 2021, doi: 10.3390/rs13183636.

[8] S. Kato, M. Saito, K. Ishiguro, and S. Cummings, "PolMERLIN: Self-supervised polarimetric complex SAR image despeckling with masked networks," IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024, doi: 10.1109/LGRS.2024.3352544.

[9] H. Lin, X. Su, Z. Zeng, C. Xing, and J. Yin, "Speckle2Self: Learning self-supervised despeckling with attention mechanism for SAR images," Remote Sensing, vol. 17, no. 23, p. 3840, 2025, doi: 10.3390/rs17233840.

[10] C. Albisani, D. Baracchi, A. Piva, and F. Argenti, "Self-supervised SAR despeckling using deep image prior," Pattern Recognition Letters, vol. 190, pp. 169-176, 2025, doi: 10.1016/j.patrec.2025.02.021.

[11] S. Vitale, G. Ferraioli, V. Pascazio, and L. Gomez Deniz, "Enhanced deep learning SAR despeckling networks based on SAR assessing metrics," IEEE Geoscience and Remote Sensing Letters, vol. 22, pp. 1-5, 2025, doi: 10.1109/LGRS.2025.3577907.

A. Saha, A. K.R., and S. K. Maji, "SAR-CDCFRN: A novel SAR despeckling approach utilizing correlated dual channel feature-based residual network," Signal Processing: Image Communication, vol. 133, p. 117267, 2025, doi: 10.1016/j.image.2025.117267.

[12] L. Zhang, L. Zheng, Y. Wen, F. Zhang, F. Bo, and Y. Cen, "Effective SAR image despeckling using noise-guided transformer and multi-scale feature fusion," Remote Sensing, vol. 17, no. 23, p. 3863, 2025, doi: 10.3390/rs17233863.

A. Paul and A. Savakis, "On denoising diffusion probabilistic models for synthetic aperture radar despeckling," Sensors, vol. 25, no. 7, p. 2149, 2025, doi: 10.3390/s25072149.

[13] X. Zhao, F. Ren, H. Sun, Y. Zhang, Y. Ma, and Q. Qi, "SAR-CDL: SAR image interpretable despeckling through convolutional dictionary learning network," Signal Processing, vol. 233, p. 109967, 2025, doi: 10.1016/j.sigpro.2025.109967.

[14] V. D. Kevala and S. Lal, "NBDNet: A deep learning algorithm for despeckling of SAR data," SN Computer Science, vol. 5, no. 7, 2024, doi: 10.1007/s42979-024-03274-6.

[15] V. D. Kevala, S. Nedungatt, S. Lal, S. Suresh, and F. Dell'Acqua, "TransSARNet: A deep learning framework for despeckling of SAR images," Engineering Research Express, vol. 7, no. 3, 2025, doi: 10.1088/2631-8695/ae037e.

[16] E. Dalsasso, C. Rambour, N. Trouve, and N. Thome, "MERLIN-Seg: Self-supervised despeckling for label-efficient semantic segmentation," Computer Vision and Image Understanding, vol. 241, p. 103940, 2024, doi: 10.1016/j.cviu.2024.103940.

[17] C. Ulondu Mendes, E. Dalsasso, Y. Zhang, L. Denis, and F. Tupin, "PolSAR2PolSAR: A semi-supervised despeckling algorithm for polarimetric SAR images," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 220, pp. 783-798, 2025, doi: 10.1016/j.isprsjprs.2025.01.008.

[18] L. Denis, E. Dalsasso, and F. Tupin, "Just project! Multi-channel despeckling, the easy way," arXiv preprint arXiv:2408.11531, 2024, doi: 10.48550/arXiv.2408.11531.

A. Mestre-Quereda and J. M. Lopez-Sanchez, "Deep learning-based speckle filtering for polarimetric SAR images. Application to Sentinel-1," IEEE Transactions on Geoscience and Remote Sensing, vol. 63, pp. 1-18, 2025, doi: 10.1109/TGRS.2025.3536090.

[19] J. J. Rico-Jimenez, D. Hu, E. M. Tang, I. Oguz, and Y. K. Tao, "Real-time OCT image denoising using a self-fusion neural network," Biomedical Optics Express, vol. 13, no. 3, p. 1398, 2022, doi: 10.1364/BOE.451029.

[20] B. Yao, L. Jin, J. Hu, Y. Liu, Y. Yan, Q. Li, and Y. Lu, "Noise-imitation learning: Unpaired speckle noise reduction for optical coherence tomography," Physics in Medicine and Biology, vol. 69, no. 18, p. 185003, 2024, doi: 10.1088/1361-6560/ad708c.

[21] S. Kethavath, Vasundhara, M. K. Manepalli, B. R. Chintada, and P. K. Yalavarthy, "Energy-regularized self-supervised deep network for speckle noise reduction in optical coherence tomography," Optics Letters, vol. 51, no. 17, p. 4793, 2026, doi: 10.1364/OL.601278.

[22] L. Wang, H. Hong, Y. Zhang, J. Wu, L. Ma, and Y. Zhu, "PolSAR-SSN: An end-to-end superpixel sampling network for PolSAR image classification," IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, doi: 10.1109/LGRS.2022.3152794.

[23] M. Zhang, J. Shi, L. Liu, W. Zhang, J. Feng, J. Zhu, and B. Chu, "PolSAR-SFCGN: An end-to-end PolSAR superpixel fully convolutional generation network," Remote Sensing, vol. 17, no. 15, p. 2723, 2025, doi: 10.3390/rs17152723.

[24] L. Zhao, N. Jiang, Y. Ren, W. Sun, Z. Wang, L. Shi, J. Yang, and P. Li, "Deep-learning and SIRV-based dual-domain speckle suppression for PolSAR imagery," IEEE Transactions on Geoscience and Remote Sensing, vol. 64, pp. 1-16, 2026, doi: 10.1109/TGRS.2026.3654601.

[25] E. Onat, "Quality-aware pixel-level fusion for robust SAR image despeckling," International Journal of Image and Data Fusion, vol. 17, no. 1, 2026, doi: 10.1080/19479832.2026.2688192.

[26] L. Ksenak, K. Pukanska, and K. Bartos, "Terrain-dependent effects of SAR speckle filtering on land cover classification using Sentinel-1," Geomatics, vol. 6, no. 3, p. 53, 2026, doi: 10.3390/geomatics6030053.

[27] 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 Transactions on Geoscience and Remote Sensing, vol. 47, no. 1, pp. 202-213, 2009, doi: 10.1109/TGRS.2008.2002881.

[28] P. Gholami, P. Roy, M. K. Parthasarathy, and V. Lakshminarayanan, "OCTID: Optical coherence tomography image database," Computers and Electrical Engineering, vol. 81, p. 106532, 2020, doi: 10.1016/j.compeleceng.2019.106532.