

Despeckling of Synthetic Aperture Radar Images using Linear–Angular Attention Transformer

Souraja Kundu, M. K. Bhuyan, *Senior Member, IEEE*, Karl F. MacDorman, *Senior Member, IEEE*, Neeraj Kumar Sharma, *Member, IEEE*, and Manish Bhatt, *Member, IEEE*

Abstract—Synthetic aperture radar (SAR) images are often contaminated by speckle noise, a type of multiplicative noise resulting from the imaging process. SAR image despeckling is a crucial preprocessing step for satellite imaging, enhancing image visualization and facilitating downstream analysis. In this study, we propose a linear–angular attention transformer network for SAR despeckling. The approach efficiently captures both local and global context in linear time within a multiscale transformer–convolutional neural network architecture. Our transformer integrates nonlocal denoising and multiscale feature extraction in a single model. Using a smoothing loss and fast nonlocal post-processing, the model achieved a 17% improvement in structural similarity index, a 61% enhancement in contrast-to-noise ratio, and an increase above 100% in the equivalent number of looks metric across three datasets compared to multiple state-of-the-art baselines, even when trained on a small dataset and for only 13 epochs. Comparison with different multi-head self-attention mechanisms revealed the effectiveness of linear–angular attention as a step towards green AI, showing both quantitative and qualitative performance improvements. Unlike models that rely on optical images for training and lack domain-specific features for real SAR despeckling, the proposed network is trained directly on SAR images in a self-supervised manner.

Index Terms—Deep learning, despeckling, image/signal processing, linear–angular attention, natural disasters and hazards, synthetic aperture radar (SAR).

I. INTRODUCTION

SPECKLE is a form of multiplicative noise arising from the reflection of radar signals off electromagnetically rough surfaces. Its presence can lead to several challenges, including degraded visualization, reduced analysis accuracy, difficulties in image translation, and interpretation errors. Despeckling of synthetic aperture radar (SAR) images is a necessary preprocessing step for subsequent satellite image analysis tasks such as segmentation, object detection, or image fusion [1], [2], [3]. Despeckling enhances the interpretability of SAR images and improves the performance of downstream algorithms, including classification, segmentation, and object detection. Potential real-world applications include disaster monitoring and management, agriculture and forestry, urban infrastructure mapping, and climate research [4].

S. Kundu, M. K. Bhuyan, and M. Bhatt are with the Department of Electronics and Electrical Engineering, and N. K. Sharma is with the Mehta Family School of Data Science and Artificial Intelligence at Indian Institute of Technology Guwahati, Assam 781039, India, e-mails: k.souraja@gmail.com, {mkb, manishb, neerajs}@iitg.ac.in.

K. F. MacDorman is with the Luddy School of Informatics, Computing and Engineering, Indiana University, Indianapolis, IN 46202 USA. (e-mail: kmacdorm@indiana.edu).

Manuscript received March, 2025; revised October 2025. Corresponding author: Manish Bhatt

To despeckle SAR images fully is challenging due to the multiplicative nature of speckle noise and the absence of noise-free ground truth. For a SAR image with the average number of looks L (radar pulses transmitted and received), the speckled SAR image can be expressed as

$$y = x \cdot n, \quad (1)$$

where x is the clean image, y is the speckled image, and n is the speckle noise distributed as

$$p(n) = \frac{1}{\Gamma(k)} \theta^{-k} n^{k-1} e^{-n/\theta}, \quad (2)$$

where $k = L$ and $\theta = 1/L$, so that $E[n] = 1$, and θ and k are the scale and shape parameters of the Gamma distribution [5].

Recent SAR despeckling approaches have explored entropy-guided dual wavelet shrinkage and intelligent Bayesian wavelet thresholding [4], [6]. Supervised models are typically trained on optical datasets [7], [8], [9] using synthetic speckle noise (as in Eq. 2) but suffer from domain mismatch on real SAR images. While techniques such as pixel-shuffle downsampling [10] are helpful, training on real SAR images remains essential for domain-specific feature learning. Therefore, this study presents a model trained on synthetically speckled SAR images, rather than optical images, for SAR despeckling.

The utility of deep learning in SAR despeckling began with convolutional neural networks (CNNs) [1], later progressing to vision transformers [11]. A multitask framework for jointly performing despeckling and change detection on dual-polarization SAR images was proposed in [12] by integrating polarization decomposition, spatiotemporal attention, and a transformer–CNN change detection branch. Vision transformers were effective owing to their global modeling capability, albeit at the cost of quadratic complexity. Subsequently, Shifted Windowed Multi-Head Self-Attention (WMSA or Swin) Transformers [13] gained popularity [14], [2] due to their windowed self-attention mechanism with cross-window connections, which achieves linear computational complexity with respect to image size. You et al. [15] introduced a castling vision transformer with kernel-based linear–angular multi-head self-attention (LAMSA) to capture global context more effectively. This design also mitigated the accuracy drop of kernel-based linear attention compared with vanilla softmax-based attention while maintaining lower complexity. In this study, we extend this concept to SAR despeckling by integrating nonlocal means (NLM) filtering within an LAMSA

transformer and enhancing detail preservation through a novel dilated hierarchical regional feature extraction block.

The denoising diffusion probabilistic model is another popular method for despeckling [16], [17]. However, training such a model is computationally expensive and time-consuming because of the diffusion process involved. Among generative models, generative adversarial networks have also been used for SAR despeckling [18], [19], [20]; however, they suffer from the mode collapse problem. Apart from supervised models, SAR2SAR [21] employs the Noise2Noise framework for self-supervised despeckling, addressing temporal variations in SAR images. MERLIN [22] improves training by using the real and imaginary components of single-look complex images as speckle pairs, but this does not fully capture the spatial correlation of speckle noise. Speckle2Void [23], based on the Noise2Void framework, leverages blind-spot CNNs to enable single-image SAR denoising. However, blind-spot CNNs inherently restrict the model's access to complete contextual information, resulting in a loss of detail in highly textured or structured regions, typically found in SAR images. In contrast, the proposed model leverages the complementary strengths of CNNs for local feature extraction and transformer-based attention for modeling long-range dependencies. The inclusion of linear-angular attention further enhances its ability to capture directional patterns in speckle, while the U-Net architecture ensures effective multiscale representation, together resulting in improved despeckling performance.

Despeckling aims to smooth uniform regions while preserving edges. We achieve this by incorporating mean-squared gradients of the despeckled image (smoothing loss) alongside L1 loss in the objective function. Additionally, a fast nonlocal filtering post-processing step further enhances performance, mitigating artifacts that arise from training on real SAR images. The application of LAMSA offers computational benefits consistent with the principles of green AI by reducing resource usage while maintaining performance.

The work makes the following contributions:

- We propose a multiscale neural network with nonlocal means transformers that efficiently capture global and local context through linear-angular attention for SAR image despeckling. Unlike prior approaches, the model is trained on synthetically speckled SAR images, eliminating the need for clean ground truth.
- A dilated hierarchical regional feature extraction block operates in parallel with the transformer to enhance fine features in the despeckled SAR image and compensate for the sparse regularization required by LAMSA.
- A fast nonlocal filtering-based post-processing strategy is applied to remove artifacts caused by direct training on real SAR images. The model outperforms state-of-the-art methods in contrast preservation and speckle suppression, achieving a 17% improvement in structural similarity index, a 61% enhancement in contrast-to-noise ratio, and over a 100% increase in the equivalent number of looks across three datasets.
- The use of LAMSA is a step towards green AI, as it requires less training and inference time, fewer training

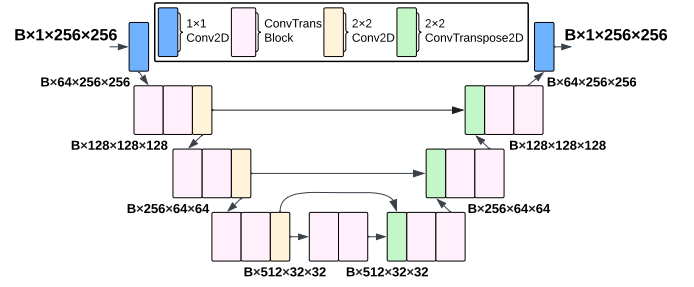


Fig. 1. The proposed model. The ConvTrans block is shown in detail in Fig. 2.

parameters, and $O(N)$ self-attention computation comparable to most other multi-head self-attention mechanisms.

II. METHODOLOGY

A. Preliminaries of Linear-Angular Attention

In transformers, self-attention computes correlations among input tokens using query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) vectors, obtained from linear projections of the tokens with three learnable weight matrices ($\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$):

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (3)$$

where d_k is the feature dimension, and $\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$ represents token similarity.

Computing the pairwise correlations of N tokens requires $O(N^2)$ complexity. Linear attention decomposes this softmax similarity function between $\mathbf{Q}$ and $\mathbf{K}$ into separate kernel embeddings, reducing the computational cost from quadratic in N to quadratic in d_k , using the associative property of matrix multiplication:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\phi(\mathbf{Q}) \sum_{j=1}^N \phi(\mathbf{K}_j)^T \mathbf{V}_j}{\phi(\mathbf{Q}) \sum_{j=1}^N \phi(\mathbf{K}_j)^T}, \quad (4)$$

where $\phi(\cdot)$ is a projection function used to approximate different kernels.

Polynomial, exponential, or RBF kernels measure spatial similarity. Yorsh and Kovalenko [24] used a learnable feedforward network as $\phi(\cdot)$. However, the angular kernel measures spectral angle distance-based similarity, defined as [15]:

$$\text{Sim}(\mathbf{Q}_i, \mathbf{K}_j) = 1 - \frac{1}{\pi} \arccos\left(\frac{\langle \mathbf{Q}_i, \mathbf{K}_j \rangle}{\|\mathbf{Q}_i\| \|\mathbf{K}_j\|}\right). \quad (5)$$

The angular kernel implicitly maps input data to a high (potentially infinite) dimensional feature space. This similarity can be expanded using trigonometric identities and written as a sum of linear-angular terms and higher-order nonlinear residual kernels. The linear-angular terms, $\frac{1}{2} + \frac{1}{\pi} \langle \mathbf{Q}_i, \mathbf{K}_j^T \rangle$, can be computed in $O(N)$ time, while the higher-order residual terms are approximated using a learnable depthwise convolution (DWConv) module to capture neighboring-token dependencies. This simplified similarity score between query and key is used later in Eq. 9 in Section II-B to derive the linear-angular attention.

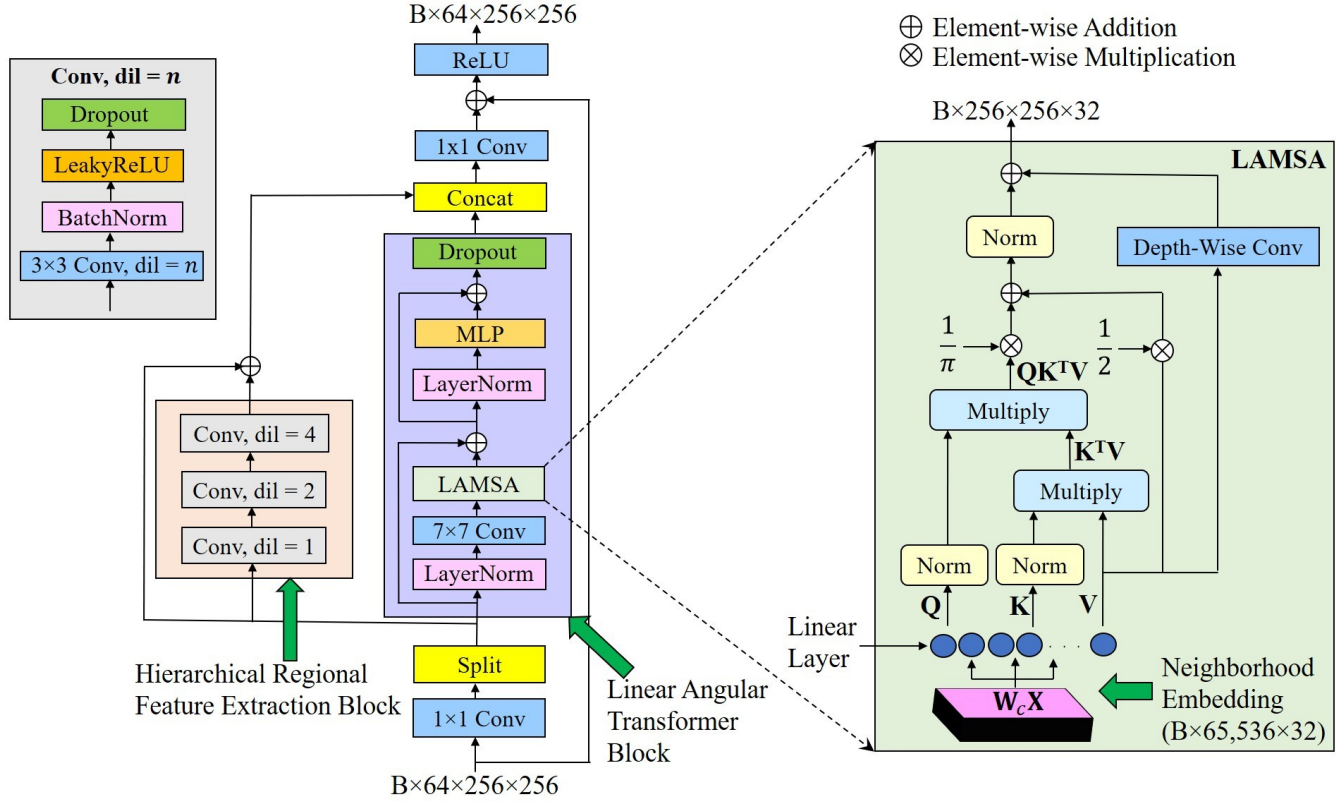


Fig. 2. ConvTrans block containing LAMSA transformer and hierarchical regional feature extraction block.

Nonlocal Mean Filtering: Xiao et al. [3] noted that a transformer's multi-head attention mechanism resembles the NLM filter, where the query $\mathbf{Q}$ and key $\mathbf{K}$ represent neighborhood matrices, and the value $\mathbf{V}$ represents pixel intensities. For this analogy, $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ must be linearly projected from neighborhood vectors, with softmax-normalized similarity serving as attention weights. Instead of this direct formulation, we use a CNN layer to extract local neighborhood features before projecting them into $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, followed by LAMSA computation. In this way, we integrate convolutional NLM filtering into the transformer's embedding process.

B. Model architecture

The schematic block diagram of the proposed network is presented in Fig. 1, and the corresponding details of the ConvTrans block and hierarchical regional feature extraction block are presented in Fig. 2. The model's backbone is similar to the SCUNet [14], which integrates Swin-Conv blocks within a multiscale U-Net model. WMSA transformers alternate self-attention between regular and cyclically shifted window partitioning, thereby reducing complexity from quadratic to linear. However, their fixed window size limits global feature capture. Therefore, we have replaced WMSA attention with linear-angular multi-head self-attention (LAMSA) to improve despeckling by efficiently capturing both local and global features. A dilated hierarchical regional feature extraction

block is included to further enhance multiscale processing. An ablation study is presented in Section III-A to justify this selection.

The network comprises an encoder with three strided convolution-based downsampling modules and a decoder with three transposed convolution-based upsampling modules, each with residual connections. Each module contains two ConvTrans blocks, with two additional blocks in the U-Net body. In a ConvTrans block, input feature map $\mathbf{X}$ undergoes a 1×1 convolution and splits into $\mathbf{X}_1$ and $\mathbf{X}_2$ for transformer and convolutional processing, respectively. It then concatenates, passes through another 1×1 convolution with a residual connection to $\mathbf{X}$, and concludes with a ReLU layer.

Linear-Angular Transformer Module: The input feature map $\mathbf{X}$ first passes through a 7×7 convolution incorporating NLM filtering. This convolutional operation effectively performs spatial aggregation over a neighborhood in a learnable manner, thereby emulating the averaging behavior of traditional NLMs. The convolutional NLM filtering layer replaces each pixel with a weighted combination of its neighborhood pixels, followed by the transformer's multi-head attention computation. This CNN layer produces embeddings $\mathbf{W}_c\mathbf{X}$ that capture neighborhood information as a linear combination of pixel values with the convolutional weight matrix $\mathbf{W}_c$.

The embeddings $\mathbf{W}_c\mathbf{X}$ are subsequently passed through three linear layers with weights $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ to

generate the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ feature maps as defined below:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{W}_c \mathbf{X}, \quad (6)$$

$$\mathbf{K} = \mathbf{W}_K \mathbf{W}_c \mathbf{X}, \quad (7)$$

$$\mathbf{V} = \mathbf{W}_V \mathbf{W}_c \mathbf{X}, \quad (8)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are weight matrices. Query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are labeled within the LAMSA block in Fig. 2.

The LAMSA is computed on normalized $\mathbf{Q}, \mathbf{K}$ as

$$\begin{aligned} \text{Att}_{\text{LA}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Sim}(\mathbf{Q}, \mathbf{K}) \cdot \mathbf{V} \\ &= \frac{1}{2} \mathbf{V} + \frac{1}{\pi} \mathbf{Q} \mathbf{K}^T \mathbf{V} + \mathbf{W}_{\text{DWC}} \mathbf{V}, \end{aligned} \quad (9)$$

where the first two terms correspond to the linear and angular similarity components computed through tensor multiplications between $\mathbf{Q}$ and $\mathbf{K}$, followed by weighting of $\mathbf{V}$. The higher-order residual term $\mathbf{W}_{\text{DWC}} \mathbf{V}$ is implemented using a depthwise convolution layer, as shown in Fig. 2. The simplified $\text{Sim}(\mathbf{Q}, \mathbf{K})$ is used as described in Section II-A. The complexity to compute Eq. 9 is $O(N)$.

Hierarchical Regional Feature Extraction Block: This block comprises a multiscale CNN with three convolutional layers, each with a 3×3 convolution, batch normalization, leaky ReLU, and dropout, applied with 1, 2, and 4 dilation rates.

Objective Function: Speckle noise in SAR images causes high-intensity fluctuations. Therefore, L_1 loss, which is robust to outliers, is used here. Additionally, a smoothing loss based on the mean squared gradient of the despeckled image helps smooth homogeneous regions. The smoothing loss can be written as

$$L_1 = \frac{1}{N} \sum_{i,j} |\hat{x}_{i,j} - x_{i,j}| \quad (10)$$

$$L_{\text{smooth}} = \frac{1}{2N} \sum_{i,j} ((\hat{x}_{i+1,j} - \hat{x}_{i,j})^2 + (\hat{x}_{i,j+1} - \hat{x}_{i,j})^2), \quad (11)$$

where $\hat{x}_{i,j}$ is the despeckled pixel intensity at (i, j) and $x_{i,j}$ is the real SAR pixel intensity at (i, j) . Therefore, the total loss is

$$L_{\text{total}} = \lambda_1 L_1 + \lambda_2 L_{\text{smooth}}, \quad (12)$$

where $\lambda_1 = 1$ and $\lambda_2 = 0.001$ are determined heuristically to prevent excessive smoothing and edge loss. We performed experiments with different combinations of loss functions, and this was our final objective function. The performance of other loss functions is summarized in Section III-A.

Finally, as a post-processing step, the model's despeckled output was refined using the Fast Non-Local Means (FastNLM) filter [25] to remove artifacts from direct SAR image training. The implementation utilized Python's `findpeaks` library in conjunction with OpenCV's `fastNlMeansDenoising` and was tested with various window sizes. The final window size was set to 5 after empirical optimization.

Algorithm 1 Proposed SAR Image Despeckling Algorithm

- 1: **Input:** Speckled SAR images
 - 2: **Output:** Despeckled SAR images
 - 3: **Step 1: Preprocessing**
 - 4: Convert images to grayscale.
 - 5: Normalize pixel intensities to $[0, 1]$.
 - 6: Square intensities to enhance differences.
 - 7: Resize images to 256×256 .
 - 8: **Step 2: Neural Network**
 - 9: Pass each image through the model (Fig. 1).
 - 10: At each resolution level, apply two ConvTrans blocks:
 - 11: Apply 1×1 convolution to input feature map X .
 - 12: Split input into two branches:
 - 13: Transformer branch (X_1): NLM filtering + LAMSA (Fig. 2).
 - 14: Convolutional branch (X_2): hierarchical feature extraction.
 - 15: Concatenate X_1 and X_2 .
 - 16: Apply 1×1 convolution on concatenated features.
 - 17: Add residual connection with X .
 - 18: Apply ReLU activation.
 - 19: **Step 3: Post-processing**
 - 20: Apply FastNLM filtering to the output.
 - 21: Return the final despeckled image.
-

C. Competing Methods

We compared our results with the following two state-of-the-art models, which have demonstrated superior despeckling performance.

1) *SAR transformer* [11]: Perera et al. introduced a novel transformer-based network for despeckling SAR images. The network incorporated a transformer-based encoder. The network was trained end-to-end using synthetically generated speckled images with a composite L_2 and total variation loss function. This SAR transformer model outperformed several nonlocal filters and CNN-based methods [11], such as probabilistic patch-based denoising [26], block-matching 3D algorithm, wavelet-domain shrinkage-based SAR denoising [27], SAR-CNN [28], and Image Despeckling CNN [1].

2) *SCUNet* [14]: Zhang et al. proposed the Swin-ConvUNet architecture, which integrates residual convolutional layers for local feature modeling with Swin transformer blocks for nonlocal context representation. Wang et al. employed Swin Transformers and residual CNNs to improve despeckling of SAR images [2]. The Swin Conv block consisted of a residual convolutional block for extracting local features and a WMSA block for capturing long-range dependencies. A pixel-shuffle downsampling post-processing strategy was used to address spatially correlated real SAR speckle. The original SCUNet employed an L_1 loss only and outperformed multiple strong baselines, such as neural nearest neighbors networks [29], nonlocal recurrent network-based image restoration [30], and Restormer [31], in practical blind image denoising.

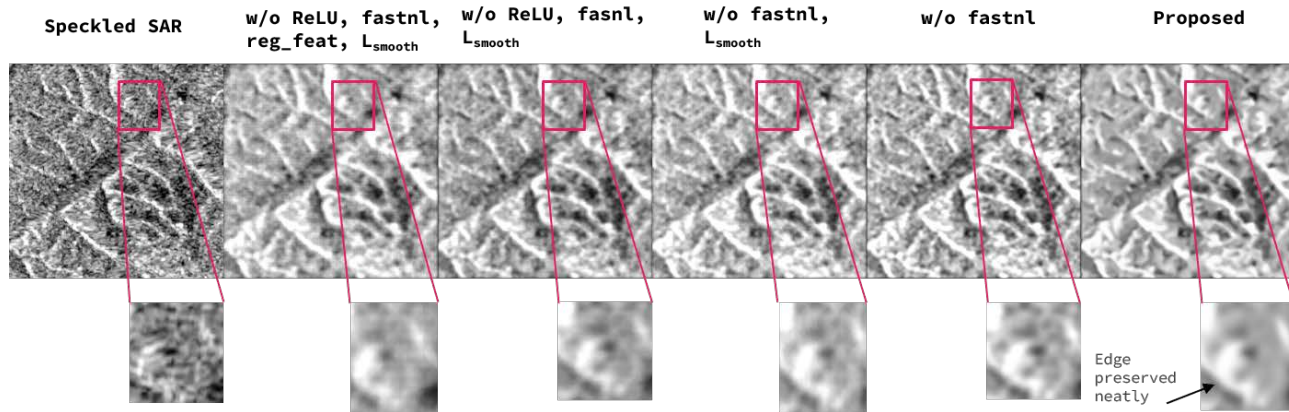


Fig. 3. Ablation study results with different modules deactivated. The region inside the red box is zoomed in below each image.

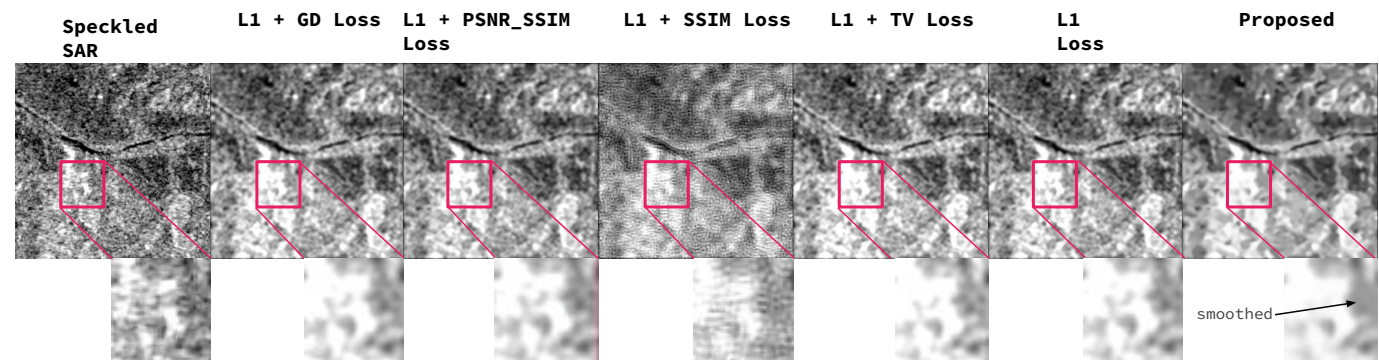


Fig. 4. Ablation study results with different loss function combinations. Zoomed-in portions show smoothing effect in different loss functions.

D. Datasets and Modeling Details

Datasets: We used three public datasets in this study. (1) Sentinel-1 SAR images across the globe in spring from the Technical University of Munich [32], (2) Sentinel-1 SAR images specifically in the Himalayas throughout the year from 2014 to 2024 collected from Google Earth Engine, and (3) PALSAR-2 ScanSAR HH polarized SAR images across the globe from 2014 to 2024 from Google Earth Engine. The model was trained on a dataset consisting of 162 training, 50 validation, and 144 test images.

Data Preprocessing: For fair comparison across models, all images were converted to single-channel grayscale, normalized (pixel intensity in [0,1]), squared (to enhance intensity differences), and resized to 256×256 . Speckled images were generated by multiplying the clean image with simulated Gamma noise, as per Equation 2.

Modeling Parameters: We experimented with learning rates (10^{-5} to 10^{-3}) and epochs (5 to 50), selecting 1×10^{-4} and 13 epochs via cross-validation. The Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) was used with a constant learning rate because scheduling (e.g., applying a 0.5 decay every 5 epochs) yielded inferior results. The model has 20.5M parameters, requiring 10 m 40 s to train 13 epochs on

162 images using PyTorch on an Nvidia RTX A6000 GPU (batch size = 1).

Evaluation Metrics: For quantitative comparison of the results, we used peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), contrast-to-noise ratio (CNR), and equivalent number of looks (ENL). The PSNR can be defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{I_{\max}^2}{\text{MSE}} \right)$$

where $I_{\max}$ is the maximum pixel intensity (e.g., 255 for 8-bit images), and MSE is the mean squared error between the reference and the despeckled images. Higher PSNR indicates better image quality with lower distortion.

The SSIM is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (13)$$

where μ and σ represent the mean and standard deviation of images x and y , and C_1, C_2 are stability constants. Higher SSIM indicates better structural similarity and perceptual quality.

The CNR can be defined as

$$\text{CNR} = \frac{|\mu_A - \mu_B|}{\sigma_B},$$

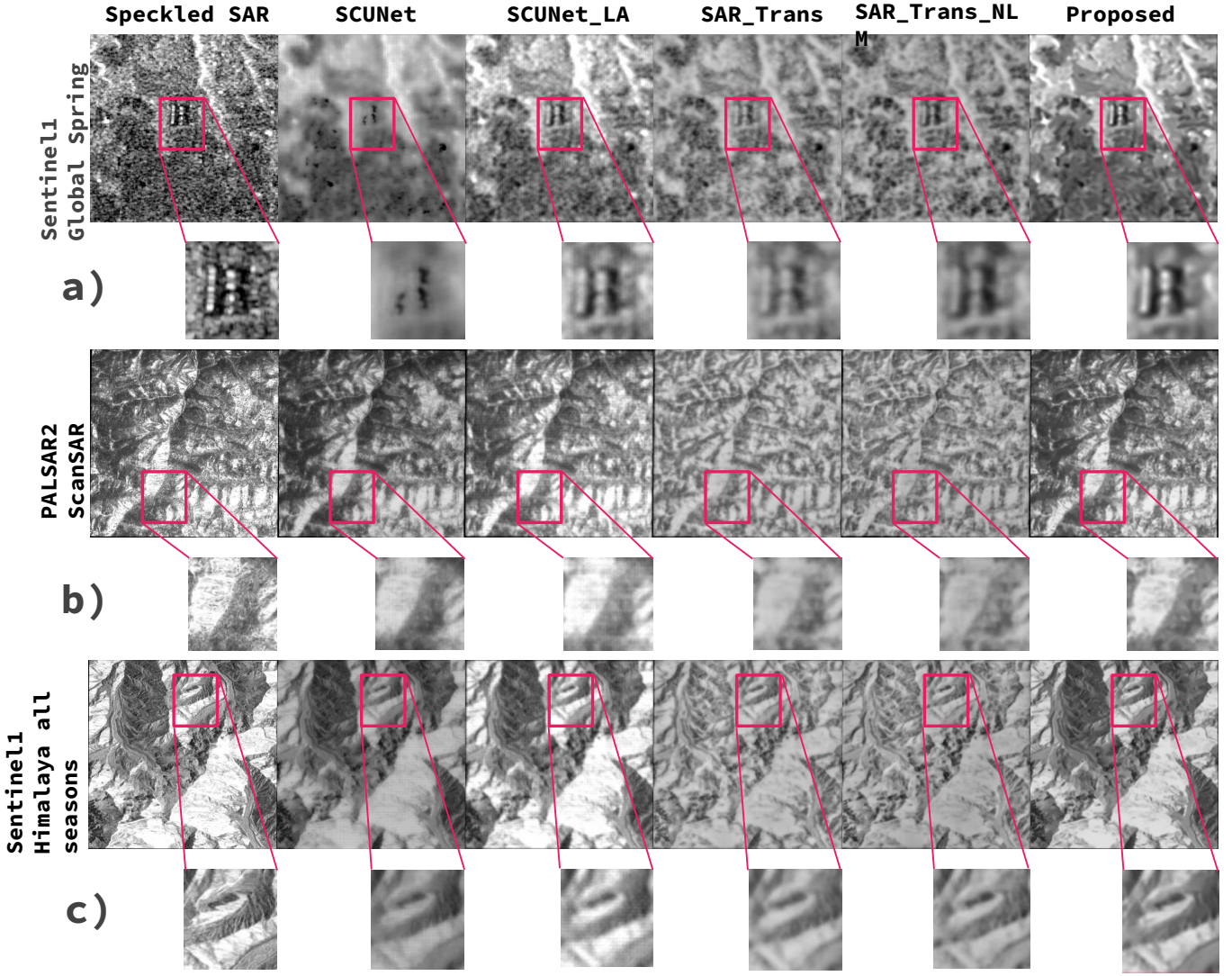


Fig. 5. Comparison of the despeckled images obtained from state-of-the-art models on a) Sentinel1 Global Spring dataset, b) PALSAR2 ScanSAR dataset, and c) Sentinel1 Himalaya all seasons dataset. The corresponding contrast-to-noise ratio appears in Table III.

where A and B are the foreground and background regions of interest, respectively, μ is the mean, and σ is the standard deviation of pixel intensity in the boxes. A higher CNR implies better edge contrast preservation.

The ENL in a region of interest R in the despeckled SAR can be defined as

$$\text{ENL} = \frac{\mu_R^2}{\sigma_R^2},$$

where higher ENL indicates better noise reduction in homogeneous regions.

III. EXPERIMENTAL RESULTS

We validated our model through an ablation study and by comparing its performance with two strong SOTA baselines that surpass other CNN and transformer models (refer to Section II-C and two of their improved versions). Prioritizing computational efficiency, we excluded denoising diffusion

probabilistic models, which offer marginal gains at significantly higher computational costs [17].

A. Ablation Results

The proposed model's key modules were systematically deactivated to assess each component's impact. First, only the linear-angular attention was retained in the ConvTrans blocks, and only the L_1 loss was used. This setup excluded the regional feature extractor (denoted as reg_feat), smoothing loss (denoted as L_{smooth}), the ReLU activation at the end of each ConvTrans block (denoted as ReLU), and the dropout layer at the output of LAMSA transformer block—the characteristic features differentiating our model from the SCUNet. Subsequently, the hierarchical feature extractor was added, and the relevance of each batch normalization layer was tested by selective deactivation. The ReLU activation, smoothing loss, and FastNL post-processing strategies were then reintroduced sequentially. As shown in Table I and Figure

TABLE I
ABLATION STUDY ON THE PALSAR-2 SCANSAR DATASET REPORTING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH WEIGHTING PARAMETERS OPTIMAL FOR EACH LOSS AND BEST VALUES BOLDED.

Experiment 1	Different Model Architectures			
Models	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)
Proposed	27.96 $\pm$ 0.24	0.70 $\pm$ 0.06	1.57 $\pm$ 2.39	87.33 $\pm$ 129.70
w/o reg_feat, ReLU, L_{smooth} , FastNL	28.47 $\pm$ 0.26	0.65 $\pm$ 0.08	1.29 $\pm$ 1.77	69.80 $\pm$ 86.86
w/o ReLU, L_{smooth} , FastNL	27.92 $\pm$ 0.15	0.66 $\pm$ 0.07	0.92 $\pm$ 1.15	38.34 $\pm$ 35.34
w/o L_{smooth} , FastNL	28.14 $\pm$ 0.28	0.69 $\pm$ 0.06	1.24 $\pm$ 1.64	53.66 $\pm$ 67.25
w/o FastNL	27.97 $\pm$ 0.23	0.70 $\pm$ 0.06	1.35 $\pm$ 1.84	66.02 $\pm$ 83.97
Experiment 2	Different Loss Functions			
Objective Functions	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)
Proposed ($L_1 + 0.001 \times L_{smooth}$)	27.96 $\pm$ 0.24	0.70 $\pm$ 0.06	1.26 $\pm$ 2.02	98.08 $\pm$ 259.52
$L_1 + 0.00001 \times L_{GD}$	27.55 $\pm$ 0.13	0.63 $\pm$ 0.06	1.19 $\pm$ 1.90	85.23 $\pm$ 226.00
$L_1 + 0.001 \times L_{PSNR\ SSIM}$	27.81 $\pm$ 0.18	0.68 $\pm$ 0.06	1.17 $\pm$ 1.72	91.16 $\pm$ 248.46
$L_1 + 0.1 \times L_{SSIM}$	28.03 $\pm$ 0.25	0.70 $\pm$ 0.06	1.19 $\pm$ 1.82	85.96 $\pm$ 212.90
$L_1 + 0.0000005 \times L_{TV}$	27.82 $\pm$ 0.23	0.66 $\pm$ 0.06	1.08 $\pm$ 1.56	82.61 $\pm$ 228.22
L_1	28.15 $\pm$ 0.28	0.69 $\pm$ 0.06	1.07 $\pm$ 1.46	69.20 $\pm$ 167.22

TABLE II
COMPARISON OF THE PROPOSED MODEL WITH SELF-ATTENTION BASELINES (NO POST-PROCESSING) ON THE SENTINEL-1 HIMALAYA ALL-SEASON DATASET, REPORTING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH TRAINING ON 162 SAMPLES, PER-IMAGE INFERENCE TIME, AND BEST VALUES BOLDED.

Model	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)	Training Time	Inference Time	Parameters
Proposed	27.61 $\pm$ 0.06	0.81 $\pm$ 0.02	1.97 $\pm$ 1.57	22.99 $\pm$ 27.27	10 m 40 s	0.244 s	20.5 M
WMSA	27.39 $\pm$ 0.05	0.78 $\pm$ 0.02	0.79 $\pm$ 0.02	18.42 $\pm$ 19.00	7 m 36 s	0.110 s	21.4 M
FlashMSA	27.41 $\pm$ 0.05	0.76 $\pm$ 0.02	1.89 $\pm$ 1.44	20.58 $\pm$ 24.02	70 m 48 s	8.196 s	21.4 M
LRMSA	27.59 $\pm$ 0.04	0.82 $\pm$ 0.02	1.88 $\pm$ 1.56	19.09 $\pm$ 20.89	13 m 59 s	0.091 s	20.5 M
RoPEMSA	27.54 $\pm$ 0.06	0.79 $\pm$ 0.02	1.90 $\pm$ 1.47	20.01 $\pm$ 21.44	81 m 55 s	8.260 s	20.5 M

TABLE III
COMPARISON OF MODELS USING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH BEST VALUES BOLDED.

Dataset 1	Sentinel-1 Himalaya All Seasons				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.41 $\pm$ 0.07	28.43 $\pm$ 0.22	28.28 $\pm$ 0.16	28.22 $\pm$ 0.14	27.61 $\pm$ 0.05
SSIM ($\uparrow$)	0.67 $\pm$ 0.03	0.73 $\pm$ 0.03	0.73 $\pm$ 0.02	0.73 $\pm$ 0.02	0.79 $\pm$ 0.02
CNR ($\uparrow$)	0.99 $\pm$ 1.82	0.99 $\pm$ 1.77	0.37 $\pm$ 0.56	0.38 $\pm$ 0.46	1.06 $\pm$ 1.84
ENL ($\uparrow$)	620.99 $\pm$ 831.87	262.04 $\pm$ 323.37	993.99 $\pm$ 1595.43	859.82 $\pm$ 1415.07	2721.45 $\pm$ 6125.83
Dataset 2	PALSAR2 ScanSAR HH				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.74 $\pm$ 0.25	28.47 $\pm$ 0.26	28.03 $\pm$ 0.27	28.06 $\pm$ 0.28	27.96 $\pm$ 0.24
SSIM ($\uparrow$)	0.63 $\pm$ 0.06	0.65 $\pm$ 0.08	0.62 $\pm$ 0.08	0.62 $\pm$ 0.08	0.70 $\pm$ 0.06
CNR ($\uparrow$)	0.50 $\pm$ 0.45	0.58 $\pm$ 0.44	0.34 $\pm$ 0.40	0.36 $\pm$ 0.41	0.61 $\pm$ 0.47
ENL ($\uparrow$)	207.91 $\pm$ 368.12	128.37 $\pm$ 220.03	206.27 $\pm$ 373.05	241.52 $\pm$ 436.17	300.08 $\pm$ 523.54
Dataset 3	Sentinel-1 Global Spring				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.91 $\pm$ 0.10	28.07 $\pm$ 0.12	28.25 $\pm$ 0.05	28.21 $\pm$ 0.04	28.16 $\pm$ 0.10
SSIM ($\uparrow$)	0.43 $\pm$ 0.09	0.59 $\pm$ 0.02	0.52 $\pm$ 0.02	0.53 $\pm$ 0.02	0.66 $\pm$ 0.01
CNR ($\uparrow$)	0.38 $\pm$ 0.53	0.54 $\pm$ 0.59	0.43 $\pm$ 0.60	0.38 $\pm$ 0.45	0.61 $\pm$ 0.63
ENL ($\uparrow$)	587.96 $\pm$ 1148.71	2201.45 $\pm$ 6145.01	189.44 $\pm$ 474.42	130.10 $\pm$ 315.65	20994.17 $\pm$ 48338.10

3, our proposed model outperformed all other ablated variants across evaluation metrics.

As part of the ablation study, we also evaluated other loss functions and learning rate scheduling strategies, as described in Section II-D. The tested losses are L_1 loss with multiscale SSIM loss (L_{SSIM}) [33], total variation loss (L_{TV}) [34], gradient difference loss (L_{GD}) [35], loss with a linear combination of PSNR and SSIM called PSNR-SSIM loss ($L_{PSNR\ SSIM}$), and L_1 loss. The results appear in Figure 4 and Table I. Figure 4's zoomed-in portions show that due to the use of a smoothing loss, the despeckled images are much smoother than those coming from other loss functions. Moreover, the SSIM loss noticeably worsens performance.

To justify the use of LAMSA for SAR despeckling, we compared it with other commonly used multi-head self-attention (MSA) mechanisms, including WMSA (as in SCUNet), FlashMSA [36], low-rank factorization MSA (LRMSA) [37], and rotary positional encoding-based MSA (RoPEMSA) [38]. The quantitative results are summarized in Table II. Although WMSA trained faster, its limited global context made it less effective for despeckling. LRMSA achieved the second-best performance and the fastest inference owing to improved token interactions over a larger context. FlashMSA and LRMSA primarily optimize the self-attention computation. LAMSA was selected based on its superior overall performance on the metrics.

B. Model Comparison Results

The model comparison results with the SAR transformer and SCUNet model are shown in Table III and Figure 5. The qualitative results in Fig. 5 (a) highlight that the proposed model more effectively restores sharp crucial details, such as the two building structures, than other competitive models. One can observe in (b) and (c) that the proposed model smooths out the homogeneous regions far better than other models and also keeps clear boundaries between the two different slopes of the hilly region without smudging the boundaries.

The quantitative results confirm the model's performance, as shown in Table III. On average, compared to the four other competitive models, the proposed model achieves a 10.6% improvement in SSIM, a 94.9% enhancement in CNR, and a 416.7% increase in ENL on the Sentinel-1 Himalaya all-season dataset. For the PALSAR-2 ScanSAR dataset, it shows an 11.2% improvement in SSIM, a 44.0% enhancement in CNR, and a 62.0% increase in ENL. Additionally, the proposed approach delivers a 29.2% improvement in SSIM, a 44.0% enhancement in CNR, and a staggering 7835.9% increase in ENL on the Sentinel-1 Global Spring dataset. Based on a combined qualitative and quantitative evaluation, the proposed model outperformed the strong baselines, with linear computational complexity.

We also experimented with improved versions of the SCUNet and the SAR transformer. SCUNet's WMSA was replaced with linear-angular attention to retain global contexts with fast training. (This variant is called SCUNet LA.) However, the SAR transformer was improved with the introduction of FastNL-based preprocessing and NLM-based embedding generation before being fed to the transformer. Even if the SCUNet LA model's despeckled SARs mostly yield the best or second-best quantitative results, it is evident from the qualitative evaluation of the images in Figure 5 that SCUNet and SCUNet LA are over-smoothing. The reason may be that SCUNet was designed for general denoising purposes, for which multiplicative noise is a special case.

Some statistical tests were conducted to determine whether the proposed despeckling method significantly outperformed the four other methods (SCUNet, SCUNet LA, SAR Trans, and SAR Trans NLM). The Kruskal-Wallis one-way analysis of variance was used to test if the five despeckling methods significantly differed from each other on the four evaluation metrics (PSNR, SSIM, CNR, and ENL) for the three datasets. A statistically significant ($p < 0.05$) Kruskal-Wallis one-way analysis was achieved. Further, multiple comparisons using the Mann-Whitney U-test were performed with Holm-Bonferroni-corrected p -values at a significance level of $p < 0.012$ for four comparisons.

On the Sentinel-1 Himalaya All Seasons dataset, the proposed model provided statistically significant improvements in SSIM over all other models with a $p < 0.00001$. On CNR, the proposed model significantly outperformed SAR Trans and SAR Trans NLM with $p < 0.00001$; and on ENL, the proposed model significantly improved despeckling compared to SCUNet and SCUNet LA with p -values of 0.00233 and

0.00056, respectively.

On the PALSAR dataset, the proposed method significantly outperformed the four other methods on SSIM with $p < 0.00001$. On CNR, the proposed method significantly outperformed SCUNet with $p < 0.00001$, SAR Trans with $p = 0.00005$, and SAR Trans NLM with $p = 0.00015$. On ENL, the proposed method significantly outperformed SCUNet with $p < 0.00001$ and SAR Trans with $p = 0.00042$.

On the Sentinel-1 Global Spring dataset, the proposed method provided statistically significant improvements in SSIM compared with all other models with a $p < 0.00001$. On CNR, the proposed method exceeded SCUNet with $p < 0.00001$, SAR Trans with $p = 0.00005$, and SAR Trans NLM with $p = 0.00015$. On ENL, the proposed method outperformed SCUNet with $p < 0.00001$ and SAR Trans NLM with $p = 0.00042$. Overall, the proposed method significantly improved performance on SSIM, CNR, and ENL compared with the other methods.

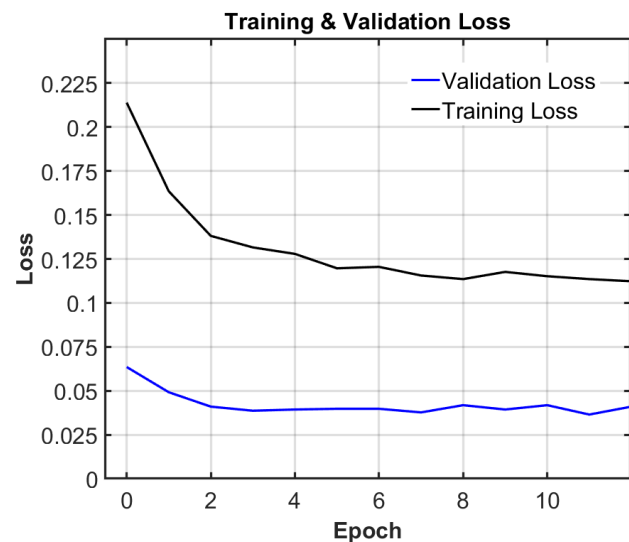


Fig. 6. Training and validation loss vs. epoch curves for the proposed model.

IV. DISCUSSION

Our proposed model's benefits are (1) self-supervised training without the need for pristine SAR images (which are difficult to acquire); (2) training the model with inherently noisy SAR images to preserve domain-specific features, unlike any other SAR despeckling models. This is possible because of the efficient local-global feature extraction capability of LAMSA, and (3) lower computational complexity, aiming for greener AI.

The cost of an AI model grows with the cost of executing the model on a single sample (for example, the number of model parameters), the size of the training dataset, and the number of hyperparameter experiments [39]. The proposed model could achieve better performance on SSIM, CNR, and ENL metrics when trained with a much smaller dataset (about 160 images) for only 13 epochs. Under similar conditions, SCUNet produces blurry, despeckled images, and the SAR

transformer produces poor, deformed images, as shown in Figure 5. The study did not perform an extensive hyperparameter search for the loss function weights (instead adopting conventionally used loss term weights), as that was neither our primary goal nor sustainable for green AI [39]. The number of model parameters of the proposed model (20.5 M) is substantially less compared with the SAR transformer (25.3 M). It takes about three days to train SCUNet on four NVIDIA RTX 2080 Ti GPUs. The proposed model can be trained using a single NVIDIA RTX A6000 GPU in 10 m 40 s, yielding perceptually satisfying despeckling results. Figure 6 shows parallel decreasing trends without divergence, indicating that the model generalizes well with no evidence of overfitting within these epochs. The lower validation loss relative to training loss is due to regularization (e.g., weight decay, dropout) and differences in dataset characteristics [40].

Evaluating model performance posed some challenges. Due to the unavailability of clean ground truth SAR images, the training dataset was created by pairing real SAR images with noise-added images. The PSNR and SSIM metrics require a reference or ground-truth image. These metrics were calculated between the real SAR image and the despeckled image. Due to the inherent noise present in the reference SAR image, the PSNR metric is not a reliable indicator. Although PSNR values are presented in the tables, the proposed method has not outperformed the others on PSNR for the same reason.

The SSIM metric is more reliable because it reflects the structural and perceptual similarity. Non-reference metrics like CNR and ENL are better as they are calculated solely from the despeckled image. CNR and ENL values are highly dependent on the choice of position and size of foreground and background boxes. For CNR, the background box must contain a homogeneous region, and the foreground box should contain high-frequency details. For ENL, the region of interest must be homogeneous, as ENL concerns the smoothness of the image. For each experimental result reported in the ablation study and the model comparison tables, different background and foreground boxes for CNR and ENL are chosen heuristically to present the superior values of CNR and ENL metrics.

Apart from quantitative performance evaluation, we also rely on qualitative visualization. The zoomed-in areas in all the figures show that the despeckled image output from the proposed model not only gets a smooth texture, but also preserves edges and finer details, unlike the other methods.

The despeckled image output from the model was further filtered using the FastNLM filter [25] to eliminate artifacts generated by direct training on SAR images. This algorithm replaces the window similarity with pixel intensity similarity at each level in the multiresolution pyramid structure, filtering out non-similar pixels within a neighborhood. We experimented with different window sizes for the FastNLM filtering step (3 to 15, with a step size of 1) while applying it as a post-processing step. We also directly applied FastNLM filtering on noisy images to check whether a deep learning model is even necessary.

One interesting observation is that the PSNR and SSIM values between the ground truth and noisy images are the same as those between the ground truth and FastNLM filtered

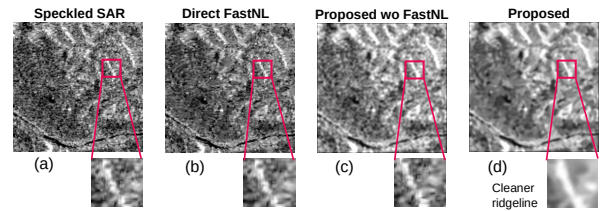


Fig. 7. (a) A reference image with speckle noise is presented. The corresponding despeckled images obtained for qualitative comparisons using (b) direct FastNLM filtering, (c) the proposed model without FastNLM filtering, and (d) the proposed model with FastNLM filtering. Zoomed-in regions compare despeckling quality and structural detail preservation.

images of different window sizes. In contrast, when FastNLM is applied as a post-processing step of our proposed model, the PSNR increases (indicating more denoising) and the SSIM decreases (indicating more smoothing, which results in the loss of details). Also, the appearance of the directly FastNLM filtered image is smudged at the edges, whereas it is continuous for the proposed model. Thus, applying FastNLM as post-processing is more effective than using it standalone. Based on a qualitative comparison and SSIM, CNR, and ENL values, we selected a window size of 5 for application in the FastNLM post-processing step. For the direct application of FastNLM to the noisy image, a filter size of 15 yields superior performance; however, it still falls short of the proposed model's results.

Figure 7 compares image quality when using only FastNLM filtering versus applying it after the proposed despeckling algorithm. The NLM-filtered image reduces granular noise while preserving fine structural details such as edges and texture boundaries. The filtering leads to cleaner homogeneous areas, well-maintained linear features, and fewer speckle artifacts, producing a visually coherent image and clearly demonstrating an improvement in human-perceived image quality. For example, in the zoomed-in region of Fig. 7(d), the high-altitude mountain ridgeline is clearly visible with substantially reduced speckle noise. Homogeneous regions, such as slopes and flat terrain, appear smooth due to effective noise suppression, while crucial linear details like rivers are preserved.

In the linear-angular attention block, as described in [15], a sparse masked softmax-based attention can capture connections to nonadjacent tokens, but it quickly converges to all zeros as the training progresses, leaving it unused for a considerable training duration. Additionally, we aimed to avoid the costly softmax-based attention calculation; therefore, we eliminated this term. The experimental results supported this decision, yielding better results without the sparse regularization term.

The hierarchical regional feature extraction block was added to capture features at multiple scales. This integration provided an alternative to computationally intensive sparse regularization. The ReLU at the end of the ConvTrans block creates sparse activation by suppressing any negative contribution through the residual connection, thus acting as an implicit regularizer and improving overall performance.

Among the various loss functions tested, only the combination of $L_1 + L_{\text{smooth}}$ performed well for despeckling be-

cause this balances noise suppression and detail preservation. L_{SSIM} focuses on perceptual similarity but does not explicitly suppress noise, often retaining noise-like high-frequency details. L_{TV} enforces smoothness but can excessively blur fine textures, degrading important details, like those observed in the SAR transformer. L_{GD} , which enforces similarity in image gradients, does not sufficiently reduce speckle noise in SAR images. $L_{PSNR\ SSIM}$ does not directly optimize for noise suppression, as PSNR prioritizes intensity matching while SSIM emphasizes structural similarity rather than despeckling. Using only L_1 loss ensures pixel reconstruction but does not enforce smoothness, allowing residual noise to persist.

V. CONCLUSION

SAR images require despeckling as a preprocessing step for various satellite imaging tasks. This study presents an efficient transformer-based SAR despeckling algorithm that integrates LAMSA with nonlocal means, capturing both local and global context in linear time, thereby overcoming the WMSA transformer's local-only limitation. The model outperforms four strong baselines on SSIM (perceptual and structural similarity retention), CNR (contrast preservation), and ENL (speckle suppression) for three datasets without requiring ground-truth references. The results show that LAMSA performs the best for SAR despeckling because of its sensitivity to angular orientation. We envision that the proposed model will be a greener approach to performing SAR despeckling with multiple use cases.

ACKNOWLEDGMENT

The authors would like to thank the late Prof. Yuji Iwahori (Chubu University, Japan) for his initial discussions on this work. Prof. Iwahori passed away in December 2024. This work acknowledges support from the ISRO-IIT Guwahati Space Technology Cell.

REFERENCES

- [1] Puyang Wang, He Zhang, and Vishal M Patel. SAR image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24(12):1763–1767, 2017.
- [2] Can Wang, Rongyao Zheng, Jingzhen Zhu, Wentao Xu, and Xiwen Li. A practical SAR despeckling method combining SWIN transformer and residual CNN. *IEEE Geoscience and Remote Sensing Letters*, 2023.
- [3] Siyao Xiao, Shunsheng Zhang, Libing Huang, and Wen-Qin Wang. Trans-NLM network for SAR image despeckling. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- [4] Prabhishek Singh, Manoj Diwakar, Achyut Shankar, Raj Shree, and Manoj Kumar. A review on SAR image and its despeckling. *Archives of Computational Methods in Engineering*, 28(7):4633–4653, 2021.
- [5] Vishal M. Patel, Glenn R. Easley, Rama Chellappa, and Nasser M. Nasrabadi. Separated component-based restoration of speckled SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(2):1019–1029, 2014.
- [6] Prabhishek Singh, Achyut Shankar, Manoj Diwakar, and Mohammad R Khosravi. MSPB: Intelligent SAR despeckling using wavelet thresholding and bilateral filter for big visual radar data restoration and provisioning quality of experience in real-time remote sensing. *Environment, Development and Sustainability*, pages 1–31, 2022.
- [7] David Martin, Charles Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001.
- [8] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2016.
- [9] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017.
- [10] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang, Honghui Shi, and Thomas Huang. When AWGN-based denoiser meets real noises. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13074–13081, 2020.
- [11] Malsha V Perera, Wele Gedara Chaminda Bandara, Jeya Maria Jose Valanarasu, and Vishal M Patel. Transformer-based SAR image despeckling. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 751–754. IEEE, 2022.
- [12] Jie Li, Shaowei Shi, Liupeng Lin, Qiangqiang Yuan, Huanfeng Shen, and Liangpei Zhang. A multi-task learning framework for dual-polarization SAR imagery despeckling in temporal change detection scenarios. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:155–178, 2025.
- [13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.
- [14] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical blind image denoising via swin-conv-unet and data synthesis. *Machine Intelligence Research*, 20(6):822–836, 2023.
- [15] Haoran You, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peizhao Zhang, Haoqi Fan, Peter Vajda, and Yingyan Celine Lin. Castling-ViT: Compressing self-attention via switching towards linear-angular attention at vision transformer inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14431–14442, 2023.
- [16] Malsha V Perera, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, and Vishal M Patel. SAR despeckling using a denoising diffusion probabilistic model. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.
- [17] Xuran Hu, Ziqiang Xu, Zhihan Chen, Zhenpeng Feng, Mingzhe Zhu, and Ljubiša Stanković. SAR despeckling via regional denoising diffusion probabilistic model. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 7226–7230. IEEE, 2024.
- [18] Puyang Wang, He Zhang, and Vishal M Patel. Generative adversarial network-based restoration of speckled SAR images. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, 2017.
- [19] Feng Gu, Hong Zhang, and Chao Wang. A GAN-based method for SAR image despeckling. In *2019 SAR in Big Data Era (BIGSAR DATA)*, pages 1–5. IEEE, 2019.
- [20] Ruijiao Liu, Yangyang Li, and Licheng Jiao. SAR image speckle reduction based on a generative adversarial network. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2020.
- [21] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. SAR2SAR: A semi-supervised despeckling algorithm for SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4321–4329, 2021.
- [22] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. As if by magic: Self-supervised training of deep despeckling networks with merlin. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2021.
- [23] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Speckle2void: Deep self-supervised SAR despeckling with blind-spot convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2021.
- [24] Uladzislau Yorsh and Alexander Kovalenko. Linear self-attention approximation via trainable feedforward kernel. In *International Conference on Artificial Neural Networks*, pages 807–810. Springer, 2022.
- [25] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.
- [26] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. Iterative weighted maximum likelihood denoising with probabilistic patch-based weights. *IEEE Transactions on Image Processing*, 18(12):2661–2672, 2009.
- [27] SARa Parrilli, Mariana Poderico, CeSARio Vincenzo Angelino, and Luisa Verdoliva. A nonlocal SAR image denoising algorithm based

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

on lmmse wavelet shrinkage. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):606–616, 2011.

[28] Giovanni Chierchia, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. SAR image despeckling through convolutional neural networks. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5438–5441. IEEE, 2017.

[29] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[30] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in neural Information Processing Systems*, 31, 2018.

[31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.

[32] Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in SAR-optical data fusion. *arXiv preprint arXiv:1807.01569*, 2018.

[33] Zhou Wang, Eero Simoncelli, and Alan Bovik. Multi-scale structural similarity for image quality assessment. *Proceedings of the IEEE Asilomar Conference Signals, Systems and Computers*, 02 2004.

[34] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[35] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[36] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[37] Sneha Mehta, Huzefa Rangwala, and Naren Ramakrishnan. Low rank factorization for compact multi-head self-attention. *arXiv preprint arXiv:1912.00835*, 2019.

[38] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.

[39] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[40] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning*, pages 10181–10192. PMLR, 2020.



Souraja Kundu was born in Kolkata, West Bengal, India. She completed her B.Tech. in Electronics and Communication Engineering from the Indian Institute of Technology (IIT) Guwahati in 2025, where she was awarded the Governor of Assam Gold Medal. She is currently pursuing a Ph.D. in the Machine Learning Department at Carnegie Mellon University’s School of Computer Science in the United States. She has previously worked as a Deep Learning and Computer Vision Research Intern at Chubu University, Japan (2023).



M.K. Bhuyan (Senior Member, IEEE) received a Ph.D. degree in electronics and communication engineering from the India Institute of Technology (IIT) Guwahati, India. He was a postdoctoral researcher at the School of Information Technology and Electrical Engineering at the University of Queensland, Australia. He was an Assistant Professor in the Department of Electrical Engineering at IIT Roorkee, India, and Jorhat Engineering College, Assam, India. He also worked in the Indian Engineering Services. In 2014, he was a Visiting Professor at Indiana University and Purdue University, Indiana, USA. He received the National Award for Best Applied Research/Technological Innovation from the President of India in 2012. He is an IEEE senior member. He is currently a professor with the Department of Electronics and Electrical Engineering at IIT Guwahati. He is also a visiting professor in the Department of Computer Science at Chubu University, Japan. His research interests include image and video processing, computer vision, machine learning, human–computer interaction (HCI), virtual reality, augmented reality, and biomedical signal processing.



Karl Fredric MacDorman (Senior Member, IEEE) received the Ph.D. degree in computer science from Cambridge University, Cambridge, U.K., in 1997. He is an Associate Professor in the Human–Computer Interaction Program with the Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA. He is also the Associate Dean of Academic Affairs. His research interests include cognitive science, human–computer interaction, machine learning, and robotics.



Neeraj Kumar Sharma (Member, IEEE) is an Assistant Professor at the School of Data Science and Artificial Intelligence, in the Indian Institute of Technology, Guwahati (India). He received his PhD from the Indian Institute of Science, Bangalore, India, in 2018. He was a BrainHub Postdoctoral Fellow at Carnegie Mellon University (CMU) from 2017 to 2018. He was a Postdoctoral Researcher at the Neuroscience Institute and Department of Psychology, CMU, Pittsburgh, from 2019 to 2020. In 2020, he was a Postdoctoral Researcher at the Society of Innovation and Development (SiD), Indian Institute of Science, Bengaluru, and in 2021, he was a CV Raman Postdoctoral Researcher at the Indian Institute of Science, Bengaluru (India). From 2021 to 2022, he was a Scientist at the Fraunhofer Institute for Integrated Circuits, IIS in Erlangen (Germany). His research interests include deep learning, speech and audio signal processing, and multimodal signal processing.



Manish Bhatt (Member, IEEE) is an Assistant Professor at the Department of Electronics and Electrical Engineering at the Indian Institute of Technology (IIT), Guwahati, India. He received his PhD degree from the Indian Institute of Science, Bangalore, India, in 2017. He was a postdoctoral fellow at the University of Montreal, Canada, from 2017 to 2020. His research interests include image processing, remote sensing, medical imaging, inverse problems, machine learning, and deep learning.

Despeckling of Synthetic Aperture Radar Images using Linear–Angular Attention Transformer

Souraja Kundu, M. K. Bhuyan, *Senior Member, IEEE*, Karl F. MacDorman, *Senior Member, IEEE*, Neeraj Kumar Sharma, *Member, IEEE*, and Manish Bhatt, *Member, IEEE*

Abstract—Synthetic aperture radar (SAR) images are often contaminated by speckle noise, a type of multiplicative noise resulting from the imaging process. SAR image despeckling is a crucial preprocessing step for satellite imaging, enhancing image visualization and facilitating downstream analysis. In this study, we propose a linear–angular attention transformer network for SAR despeckling. The approach efficiently captures both local and global context in linear time within a multiscale transformer–convolutional neural network architecture. Our transformer integrates nonlocal denoising and multiscale feature extraction in a single model. Using a smoothing loss and fast nonlocal post-processing, the model achieved a 17% improvement in structural similarity index, a 61% enhancement in contrast-to-noise ratio, and an increase above 100% in the equivalent number of looks metric across three datasets compared to multiple state-of-the-art baselines, even when trained on a small dataset and for only 13 epochs. Comparison with different multi-head self-attention mechanisms revealed the effectiveness of linear–angular attention as a step towards green AI, showing both quantitative and qualitative performance improvements. Unlike models that rely on optical images for training and lack domain-specific features for real SAR despeckling, the proposed network is trained directly on SAR images in a self-supervised manner.

Index Terms—Deep learning, despeckling, image/signal processing, linear–angular attention, natural disasters and hazards, synthetic aperture radar (SAR).

I. INTRODUCTION

SPECKLE is a form of multiplicative noise arising from the reflection of radar signals off electromagnetically rough surfaces. Its presence can lead to several challenges, including degraded visualization, reduced analysis accuracy, difficulties in image translation, and interpretation errors. Despeckling of synthetic aperture radar (SAR) images is a necessary preprocessing step for subsequent satellite image analysis tasks such as segmentation, object detection, or image fusion [1], [2], [3]. Despeckling enhances the interpretability of SAR images and improves the performance of downstream algorithms, including classification, segmentation, and object detection. Potential real-world applications include disaster monitoring and management, agriculture and forestry, urban infrastructure mapping, and climate research [4].

S. Kundu, M. K. Bhuyan, and M. Bhatt are with the Department of Electronics and Electrical Engineering, and N. K. Sharma is with the Mehta Family School of Data Science and Artificial Intelligence at Indian Institute of Technology Guwahati, Assam 781039, India, e-mails: k.souraja@gmail.com, {mkb, manishb, neerajs}@iitg.ac.in.

K. F. MacDorman is with the Luddy School of Informatics, Computing and Engineering, Indiana University, Indianapolis, IN 46202 USA. (e-mail: kmacdorm@indiana.edu).

Manuscript received March, 2025; revised October 2025. Corresponding author: Manish Bhatt

To despeckle SAR images fully is challenging due to the multiplicative nature of speckle noise and the absence of noise-free ground truth. For a SAR image with the average number of looks L (radar pulses transmitted and received), the speckled SAR image can be expressed as

$$y = x \cdot n, \quad (1)$$

where x is the clean image, y is the speckled image, and n is the speckle noise distributed as

$$p(n) = \frac{1}{\Gamma(k)} \theta^{-k} n^{k-1} e^{-n/\theta}, \quad (2)$$

where $k = L$ and $\theta = 1/L$, so that $E[n] = 1$, and θ and k are the scale and shape parameters of the Gamma distribution [5].

Recent SAR despeckling approaches have explored entropy-guided dual wavelet shrinkage and intelligent Bayesian wavelet thresholding [4], [6]. Supervised models are typically trained on optical datasets [7], [8], [9] using synthetic speckle noise (as in Eq. 2) but suffer from domain mismatch on real SAR images. While techniques such as pixel-shuffle downsampling [10] are helpful, training on real SAR images remains essential for domain-specific feature learning. Therefore, this study presents a model trained on synthetically speckled SAR images, rather than optical images, for SAR despeckling.

The utility of deep learning in SAR despeckling began with convolutional neural networks (CNNs) [1], later progressing to vision transformers [11]. A multitask framework for jointly performing despeckling and change detection on dual-polarization SAR images was proposed in [12] by integrating polarization decomposition, spatiotemporal attention, and a transformer–CNN change detection branch. Vision transformers were effective owing to their global modeling capability, albeit at the cost of quadratic complexity. Subsequently, Shifted Windowed Multi-Head Self-Attention (WMSA or Swin) Transformers [13] gained popularity [14], [2] due to their windowed self-attention mechanism with cross-window connections, which achieves linear computational complexity with respect to image size. You et al. [15] introduced a castling vision transformer with kernel-based linear–angular multi-head self-attention (LAMSA) to capture global context more effectively. This design also mitigated the accuracy drop of kernel-based linear attention compared with vanilla softmax-based attention while maintaining lower complexity. In this study, we extend this concept to SAR despeckling by integrating nonlocal means (NLM) filtering within an LAMSA

transformer and enhancing detail preservation through a novel dilated hierarchical regional feature extraction block.

The denoising diffusion probabilistic model is another popular method for despeckling [16], [17]. However, training such a model is computationally expensive and time-consuming because of the diffusion process involved. Among generative models, generative adversarial networks have also been used for SAR despeckling [18], [19], [20]; however, they suffer from the mode collapse problem. Apart from supervised models, SAR2SAR [21] employs the Noise2Noise framework for self-supervised despeckling, addressing temporal variations in SAR images. MERLIN [22] improves training by using the real and imaginary components of single-look complex images as speckle pairs, but this does not fully capture the spatial correlation of speckle noise. Speckle2Void [23], based on the Noise2Void framework, leverages blind-spot CNNs to enable single-image SAR denoising. However, blind-spot CNNs inherently restrict the model's access to complete contextual information, resulting in a loss of detail in highly textured or structured regions, typically found in SAR images. In contrast, the proposed model leverages the complementary strengths of CNNs for local feature extraction and transformer-based attention for modeling long-range dependencies. The inclusion of linear-angular attention further enhances its ability to capture directional patterns in speckle, while the U-Net architecture ensures effective multiscale representation, together resulting in improved despeckling performance.

Despeckling aims to smooth uniform regions while preserving edges. We achieve this by incorporating mean-squared gradients of the despeckled image (smoothing loss) alongside L1 loss in the objective function. Additionally, a fast nonlocal filtering post-processing step further enhances performance, mitigating artifacts that arise from training on real SAR images. The application of LAMSA offers computational benefits consistent with the principles of green AI by reducing resource usage while maintaining performance.

The work makes the following contributions:

- We propose a multiscale neural network with nonlocal means transformers that efficiently capture global and local context through linear-angular attention for SAR image despeckling. Unlike prior approaches, the model is trained on synthetically speckled SAR images, eliminating the need for clean ground truth.
- A dilated hierarchical regional feature extraction block operates in parallel with the transformer to enhance fine features in the despeckled SAR image and compensate for the sparse regularization required by LAMSA.
- A fast nonlocal filtering-based post-processing strategy is applied to remove artifacts caused by direct training on real SAR images. The model outperforms state-of-the-art methods in contrast preservation and speckle suppression, achieving a 17% improvement in structural similarity index, a 61% enhancement in contrast-to-noise ratio, and over a 100% increase in the equivalent number of looks across three datasets.
- The use of LAMSA is a step towards green AI, as it requires less training and inference time, fewer training

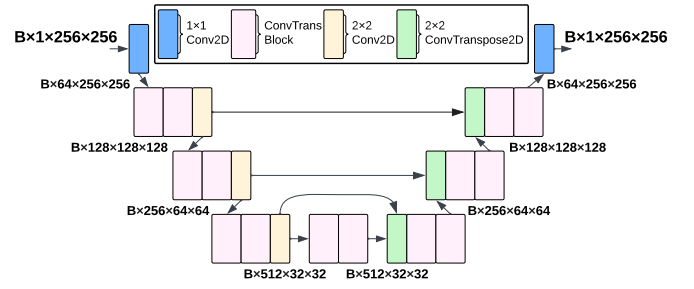


Fig. 1. The proposed model. The ConvTrans block is shown in detail in Fig. 2.

parameters, and $O(N)$ self-attention computation comparable to most other multi-head self-attention mechanisms.

II. METHODOLOGY

A. Preliminaries of Linear-Angular Attention

In transformers, self-attention computes correlations among input tokens using query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) vectors, obtained from linear projections of the tokens with three learnable weight matrices ($\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$):

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (3)$$

where d_k is the feature dimension, and $\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)$ represents token similarity.

Computing the pairwise correlations of N tokens requires $O(N^2)$ complexity. Linear attention decomposes this softmax similarity function between $\mathbf{Q}$ and $\mathbf{K}$ into separate kernel embeddings, reducing the computational cost from quadratic in N to quadratic in d_k , using the associative property of matrix multiplication:

$$\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \frac{\phi(\mathbf{Q}) \sum_{j=1}^N \phi(\mathbf{K}_j)^T \mathbf{V}_j}{\phi(\mathbf{Q}) \sum_{j=1}^N \phi(\mathbf{K}_j)^T}, \quad (4)$$

where $\phi(\cdot)$ is a projection function used to approximate different kernels.

Polynomial, exponential, or RBF kernels measure spatial similarity. Yorsh and Kovalenko [24] used a learnable feedforward network as $\phi(\cdot)$. However, the angular kernel measures spectral angle distance-based similarity, defined as [15]:

$$\text{Sim}(\mathbf{Q}_i, \mathbf{K}_j) = 1 - \frac{1}{\pi} \arccos\left(\frac{\langle \mathbf{Q}_i, \mathbf{K}_j \rangle}{\|\mathbf{Q}_i\| \|\mathbf{K}_j\|}\right). \quad (5)$$

The angular kernel implicitly maps input data to a high (potentially infinite) dimensional feature space. This similarity can be expanded using trigonometric identities and written as a sum of linear-angular terms and higher-order nonlinear residual kernels. The linear-angular terms, $\frac{1}{2} + \frac{1}{\pi} \langle \mathbf{Q}_i, \mathbf{K}_j^T \rangle$, can be computed in $O(N)$ time, while the higher-order residual terms are approximated using a learnable depthwise convolution (DWConv) module to capture neighboring-token dependencies. This simplified similarity score between query and key is used later in Eq. 9 in Section II-B to derive the linear-angular attention.



The embeddings $\mathbf{W}_c\mathbf{X}$ are subsequently passed through three linear layers with weights $\mathbf{W}_Q$, $\mathbf{W}_K$, and $\mathbf{W}_V$ to

generate the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ feature maps as defined below:

$$\mathbf{Q} = \mathbf{W}_Q \mathbf{W}_c \mathbf{X}, \quad (6)$$

$$\mathbf{K} = \mathbf{W}_K \mathbf{W}_c \mathbf{X}, \quad (7)$$

$$\mathbf{V} = \mathbf{W}_V \mathbf{W}_c \mathbf{X}, \quad (8)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are weight matrices. Query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ are labeled within the LAMSA block in Fig. 2.

The LAMSA is computed on normalized $\mathbf{Q}, \mathbf{K}$ as

$$\begin{aligned} \text{Att}_{\text{LA}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{Sim}(\mathbf{Q}, \mathbf{K}) \cdot \mathbf{V} \\ &= \frac{1}{2} \mathbf{V} + \frac{1}{\pi} \mathbf{Q} \mathbf{K}^T \mathbf{V} + \mathbf{W}_{\text{DWC}} \mathbf{V}, \end{aligned} \quad (9)$$

where the first two terms correspond to the linear and angular similarity components computed through tensor multiplications between $\mathbf{Q}$ and $\mathbf{K}$, followed by weighting of $\mathbf{V}$. The higher-order residual term $\mathbf{W}_{\text{DWC}} \mathbf{V}$ is implemented using a depthwise convolution layer, as shown in Fig. 2. The simplified $\text{Sim}(\mathbf{Q}, \mathbf{K})$ is used as described in Section II-A. The complexity to compute Eq. 9 is $O(N)$.

Hierarchical Regional Feature Extraction Block: This block comprises a multiscale CNN with three convolutional layers, each with a 3×3 convolution, batch normalization, leaky ReLU, and dropout, applied with 1, 2, and 4 dilation rates.

Objective Function: Speckle noise in SAR images causes high-intensity fluctuations. Therefore, L_1 loss, which is robust to outliers, is used here. Additionally, a smoothing loss based on the mean squared gradient of the despeckled image helps smooth homogeneous regions. The smoothing loss can be written as

$$L_1 = \frac{1}{N} \sum_{i,j} |\hat{x}_{i,j} - x_{i,j}| \quad (10)$$

$$L_{\text{smooth}} = \frac{1}{2N} \sum_{i,j} ((\hat{x}_{i+1,j} - \hat{x}_{i,j})^2 + (\hat{x}_{i,j+1} - \hat{x}_{i,j})^2), \quad (11)$$

where $\hat{x}_{i,j}$ is the despeckled pixel intensity at (i, j) and $x_{i,j}$ is the real SAR pixel intensity at (i, j) . Therefore, the total loss is

$$L_{\text{total}} = \lambda_1 L_1 + \lambda_2 L_{\text{smooth}}, \quad (12)$$

where $\lambda_1 = 1$ and $\lambda_2 = 0.001$ are determined heuristically to prevent excessive smoothing and edge loss. We performed experiments with different combinations of loss functions, and this was our final objective function. The performance of other loss functions is summarized in Section III-A.

Finally, as a post-processing step, the model's despeckled output was refined using the Fast Non-Local Means (FastNLM) filter [25] to remove artifacts from direct SAR image training. The implementation utilized Python's `findpeaks` library in conjunction with OpenCV's `fastNlMeansDenoising` and was tested with various window sizes. The final window size was set to 5 after empirical optimization.

Algorithm 1 Proposed SAR Image Despeckling Algorithm

- 1: **Input:** Speckled SAR images
 - 2: **Output:** Despeckled SAR images
 - 3: **Step 1: Preprocessing**
 - 4: Convert images to grayscale.
 - 5: Normalize pixel intensities to $[0, 1]$.
 - 6: Square intensities to enhance differences.
 - 7: Resize images to 256×256 .
 - 8: **Step 2: Neural Network**
 - 9: Pass each image through the model (Fig. 1).
 - 10: At each resolution level, apply two ConvTrans blocks:
 - 11: Apply 1×1 convolution to input feature map X .
 - 12: Split input into two branches:
 - 13: Transformer branch (X_1): NLM filtering + LAMSA (Fig. 2).
 - 14: Convolutional branch (X_2): hierarchical feature extraction.
 - 15: Concatenate X_1 and X_2 .
 - 16: Apply 1×1 convolution on concatenated features.
 - 17: Add residual connection with X .
 - 18: Apply ReLU activation.
 - 19: **Step 3: Post-processing**
 - 20: Apply FastNLM filtering to the output.
 - 21: Return the final despeckled image.
-

C. Competing Methods

We compared our results with the following two state-of-the-art models, which have demonstrated superior despeckling performance.

1) *SAR transformer* [11]: Perera et al. introduced a novel transformer-based network for despeckling SAR images. The network incorporated a transformer-based encoder. The network was trained end-to-end using synthetically generated speckled images with a composite L_2 and total variation loss function. This SAR transformer model outperformed several nonlocal filters and CNN-based methods [11], such as probabilistic patch-based denoising [26], block-matching 3D algorithm, wavelet-domain shrinkage-based SAR denoising [27], SAR-CNN [28], and Image Despeckling CNN [1].

2) *SCUNet* [14]: Zhang et al. proposed the Swin-ConvUNet architecture, which integrates residual convolutional layers for local feature modeling with Swin transformer blocks for nonlocal context representation. Wang et al. employed Swin Transformers and residual CNNs to improve despeckling of SAR images [2]. The Swin Conv block consisted of a residual convolutional block for extracting local features and a WMSA block for capturing long-range dependencies. A pixel-shuffle downsampling post-processing strategy was used to address spatially correlated real SAR speckle. The original SCUNet employed an L_1 loss only and outperformed multiple strong baselines, such as neural nearest neighbors networks [29], nonlocal recurrent network-based image restoration [30], and Restormer [31], in practical blind image denoising.

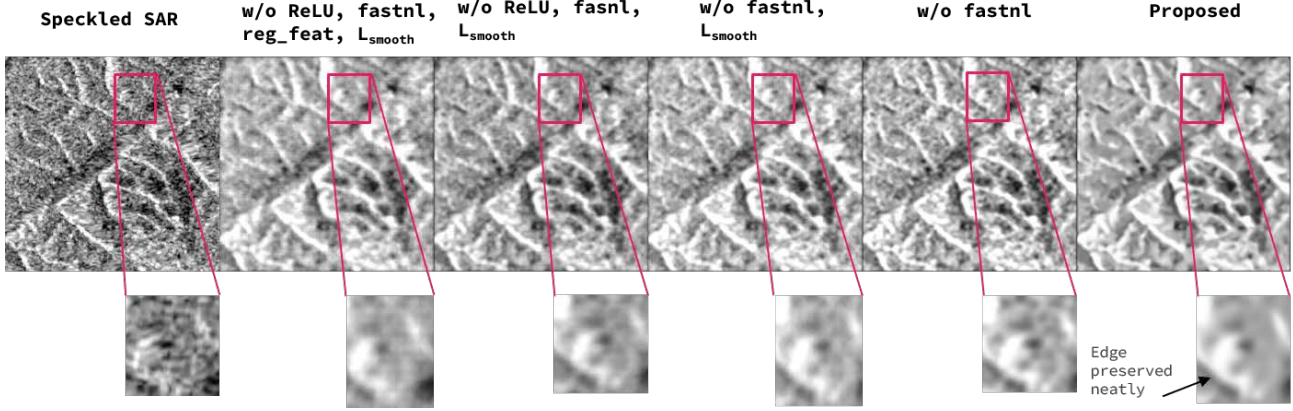


Fig. 3. Ablation study results with different modules deactivated. The region inside the red box is zoomed in below each image.

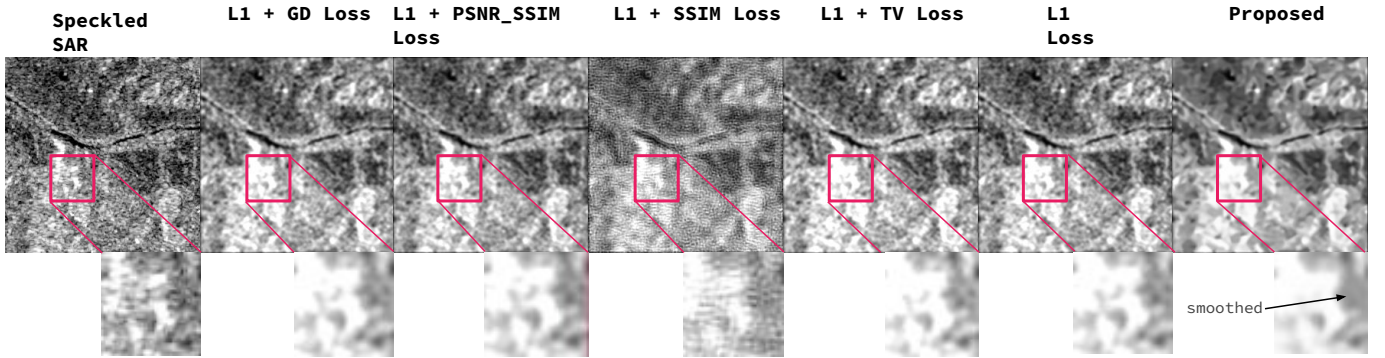


Fig. 4. Ablation study results with different loss function combinations. Zoomed-in portions show smoothing effect in different loss functions.

D. Datasets and Modeling Details

Datasets: We used three public datasets in this study. (1) Sentinel-1 SAR images across the globe in spring from the Technical University of Munich [32], (2) Sentinel-1 SAR images specifically in the Himalayas throughout the year from 2014 to 2024 collected from Google Earth Engine, and (3) PALSAR-2 ScanSAR HH polarized SAR images across the globe from 2014 to 2024 from Google Earth Engine. The model was trained on a dataset consisting of 162 training, 50 validation, and 144 test images.

Data Preprocessing: For fair comparison across models, all images were converted to single-channel grayscale, normalized (pixel intensity in [0,1]), squared (to enhance intensity differences), and resized to 256×256 . Speckled images were generated by multiplying the clean image with simulated Gamma noise, as per Equation 2.

Modeling Parameters: We experimented with learning rates (10^{-5} to 10^{-3}) and epochs (5 to 50), selecting 1×10^{-4} and 13 epochs via cross-validation. The Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$) was used with a constant learning rate because scheduling (e.g., applying a 0.5 decay every 5 epochs) yielded inferior results. The model has 20.5M parameters, requiring 10 m 40 s to train 13 epochs on

162 images using PyTorch on an Nvidia RTX A6000 GPU (batch size = 1).

Evaluation Metrics: For quantitative comparison of the results, we used peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), contrast-to-noise ratio (CNR), and equivalent number of looks (ENL). The PSNR can be defined as

$$\text{PSNR} = 10 \log_{10} \left(\frac{I_{\max}^2}{\text{MSE}} \right)$$

where $I_{\max}$ is the maximum pixel intensity (e.g., 255 for 8-bit images), and MSE is the mean squared error between the reference and the despeckled images. Higher PSNR indicates better image quality with lower distortion.

The SSIM is defined as

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (13)$$

where μ and σ represent the mean and standard deviation of images x and y , and C_1, C_2 are stability constants. Higher SSIM indicates better structural similarity and perceptual quality.

The CNR can be defined as

$$\text{CNR} = \frac{|\mu_A - \mu_B|}{\sigma_B},$$

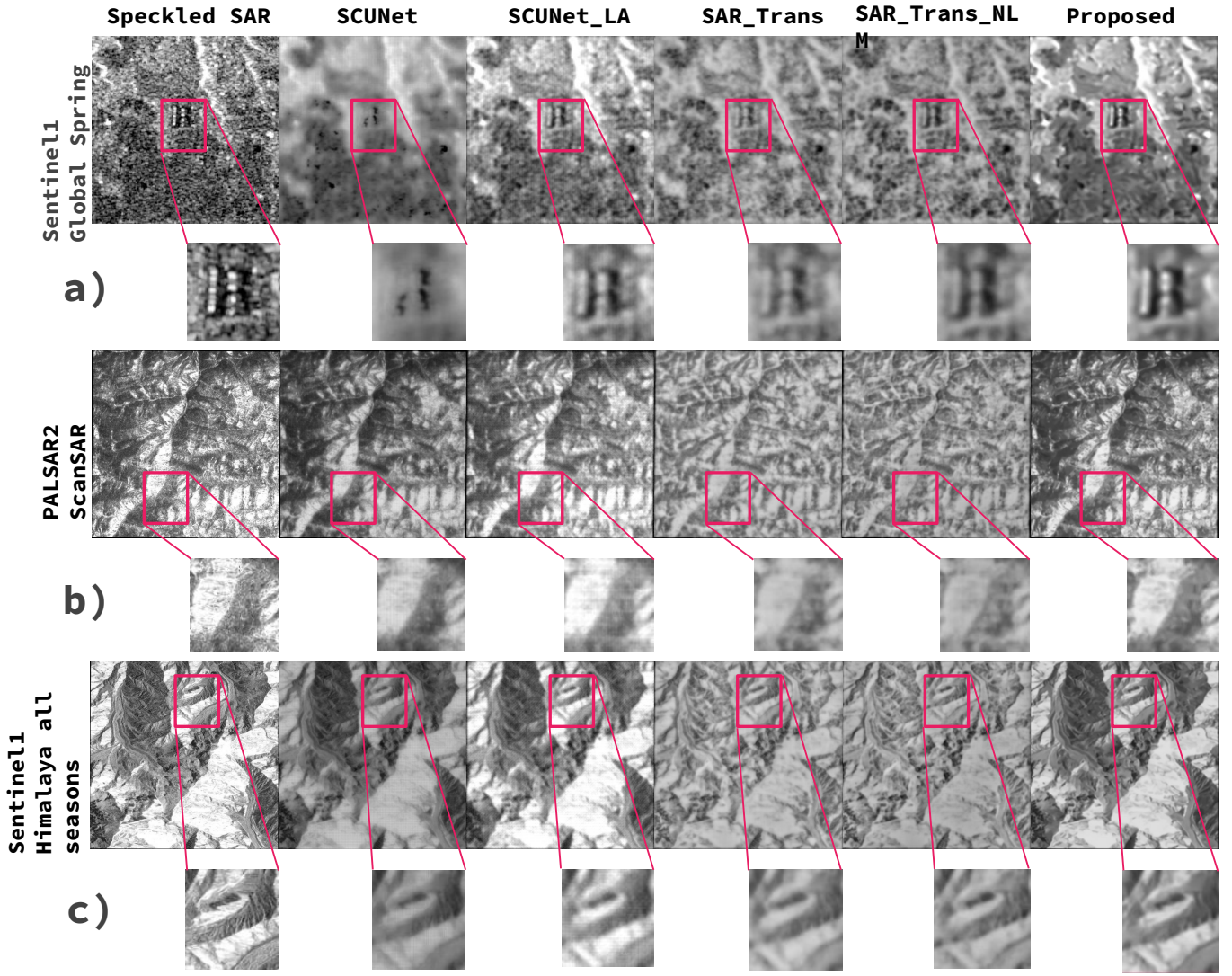


Fig. 5. Comparison of the despeckled images obtained from state-of-the-art models on a) Sentinel1 Global Spring dataset, b) PALSAR2 ScanSAR dataset, and c) Sentinel1 Himalaya all seasons dataset. The corresponding contrast-to-noise ratio appears in Table III.

where A and B are the foreground and background regions of interest, respectively, μ is the mean, and σ is the standard deviation of pixel intensity in the boxes. A higher CNR implies better edge contrast preservation.

The ENL in a region of interest R in the despeckled SAR can be defined as

$$\text{ENL} = \frac{\mu_R^2}{\sigma_R^2},$$

where higher ENL indicates better noise reduction in homogeneous regions.

III. EXPERIMENTAL RESULTS

We validated our model through an ablation study and by comparing its performance with two strong SOTA baselines that surpass other CNN and transformer models (refer to Section II-C and two of their improved versions). Prioritizing computational efficiency, we excluded denoising diffusion

probabilistic models, which offer marginal gains at significantly higher computational costs [17].

A. Ablation Results

The proposed model's key modules were systematically deactivated to assess each component's impact. First, only the linear-angular attention was retained in the ConvTrans blocks, and only the L_1 loss was used. This setup excluded the regional feature extractor (denoted as reg_feat), smoothing loss (denoted as L_{smooth}), the ReLU activation at the end of each ConvTrans block (denoted as ReLU), and the dropout layer at the output of LAMSA transformer block—the characteristic features differentiating our model from the SCUNet. Subsequently, the hierarchical feature extractor was added, and the relevance of each batch normalization layer was tested by selective deactivation. The ReLU activation, smoothing loss, and FastNL post-processing strategies were then reintroduced sequentially. As shown in Table I and Figure

TABLE I
ABLATION STUDY ON THE PALSAR-2 SCANSAR DATASET REPORTING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH WEIGHTING PARAMETERS OPTIMAL FOR EACH LOSS AND BEST VALUES BOLDED.

Experiment 1	Different Model Architectures			
Models	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)
Proposed	27.96 $\pm$ 0.24	0.70 $\pm$ 0.06	1.57 $\pm$ 2.39	87.33 $\pm$ 129.70
w/o reg_feat, ReLU, L_{smooth} , FastNL	28.47 $\pm$ 0.26	0.65 $\pm$ 0.08	1.29 $\pm$ 1.77	69.80 $\pm$ 86.86
w/o ReLU, L_{smooth} , FastNL	27.92 $\pm$ 0.15	0.66 $\pm$ 0.07	0.92 $\pm$ 1.15	38.34 $\pm$ 35.34
w/o L_{smooth} , FastNL	28.14 $\pm$ 0.28	0.69 $\pm$ 0.06	1.24 $\pm$ 1.64	53.66 $\pm$ 67.25
w/o FastNL	27.97 $\pm$ 0.23	0.70 $\pm$ 0.06	1.35 $\pm$ 1.84	66.02 $\pm$ 83.97
Experiment 2	Different Loss Functions			
Objective Functions	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)
Proposed ($L_1 + 0.001 \times L_{smooth}$)	27.96 $\pm$ 0.24	0.70 $\pm$ 0.06	1.26 $\pm$ 2.02	98.08 $\pm$ 259.52
$L_1 + 0.00001 \times L_{GD}$	27.55 $\pm$ 0.13	0.63 $\pm$ 0.06	1.19 $\pm$ 1.90	85.23 $\pm$ 226.00
$L_1 + 0.001 \times L_{PSNR\ SSIM}$	27.81 $\pm$ 0.18	0.68 $\pm$ 0.06	1.17 $\pm$ 1.72	91.16 $\pm$ 248.46
$L_1 + 0.1 \times L_{SSIM}$	28.03 $\pm$ 0.25	0.70 $\pm$ 0.06	1.19 $\pm$ 1.82	85.96 $\pm$ 212.90
$L_1 + 0.0000005 \times L_{TV}$	27.82 $\pm$ 0.23	0.66 $\pm$ 0.06	1.08 $\pm$ 1.56	82.61 $\pm$ 228.22
L_1	28.15 $\pm$ 0.28	0.69 $\pm$ 0.06	1.07 $\pm$ 1.46	69.20 $\pm$ 167.22

TABLE II
COMPARISON OF THE PROPOSED MODEL WITH SELF-ATTENTION BASELINES (NO POST-PROCESSING) ON THE SENTINEL-1 HIMALAYA ALL-SEASON DATASET, REPORTING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH TRAINING ON 162 SAMPLES, PER-IMAGE INFERENCE TIME, AND BEST VALUES BOLDED.

Model	PSNR ($\uparrow$)	SSIM ($\uparrow$)	CNR ($\uparrow$)	ENL ($\uparrow$)	Training Time	Inference Time	Parameters
Proposed	27.61 $\pm$ 0.06	0.81 $\pm$ 0.02	1.97 $\pm$ 1.57	22.99 $\pm$ 27.27	10 m 40 s	0.244 s	20.5 M
WMSA	27.39 $\pm$ 0.05	0.78 $\pm$ 0.02	0.79 $\pm$ 0.02	18.42 $\pm$ 19.00	7 m 36 s	0.110 s	21.4 M
FlashMSA	27.41 $\pm$ 0.05	0.76 $\pm$ 0.02	1.89 $\pm$ 1.44	20.58 $\pm$ 24.02	70 m 48 s	8.196 s	21.4 M
LRMSA	27.59 $\pm$ 0.04	0.82 $\pm$ 0.02	1.88 $\pm$ 1.56	19.09 $\pm$ 20.89	13 m 59 s	0.091 s	20.5 M
RoPEMSA	27.54 $\pm$ 0.06	0.79 $\pm$ 0.02	1.90 $\pm$ 1.47	20.01 $\pm$ 21.44	81 m 55 s	8.260 s	20.5 M

TABLE III
COMPARISON OF MODELS USING PSNR, SSIM, CNR, AND ENL (MEAN $\pm$ STD), WITH BEST VALUES BOLDED.

Dataset 1	Sentinel-1 Himalaya All Seasons				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.41 $\pm$ 0.07	28.43 $\pm$ 0.22	28.28 $\pm$ 0.16	28.22 $\pm$ 0.14	27.61 $\pm$ 0.05
SSIM ($\uparrow$)	0.67 $\pm$ 0.03	0.73 $\pm$ 0.03	0.73 $\pm$ 0.02	0.73 $\pm$ 0.02	0.79 $\pm$ 0.02
CNR ($\uparrow$)	0.99 $\pm$ 1.82	0.99 $\pm$ 1.77	0.37 $\pm$ 0.56	0.38 $\pm$ 0.46	1.06 $\pm$ 1.84
ENL ($\uparrow$)	620.99 $\pm$ 831.87	262.04 $\pm$ 323.37	993.99 $\pm$ 1595.43	859.82 $\pm$ 1415.07	2721.45 $\pm$ 6125.83
Dataset 2	PALSAR2 ScanSAR HH				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.74 $\pm$ 0.25	28.47 $\pm$ 0.26	28.03 $\pm$ 0.27	28.06 $\pm$ 0.28	27.96 $\pm$ 0.24
SSIM ($\uparrow$)	0.63 $\pm$ 0.06	0.65 $\pm$ 0.08	0.62 $\pm$ 0.08	0.62 $\pm$ 0.08	0.70 $\pm$ 0.06
CNR ($\uparrow$)	0.50 $\pm$ 0.45	0.58 $\pm$ 0.44	0.34 $\pm$ 0.40	0.36 $\pm$ 0.41	0.61 $\pm$ 0.47
ENL ($\uparrow$)	207.91 $\pm$ 368.12	128.37 $\pm$ 220.03	206.27 $\pm$ 373.05	241.52 $\pm$ 436.17	300.08 $\pm$ 523.54
Dataset 3	Sentinel-1 Global Spring				
Models	SCUNet	SCUNet LA	SAR Trans	SAR Trans NLM	Proposed
PSNR (dB) ($\uparrow$)	27.91 $\pm$ 0.10	28.07 $\pm$ 0.12	28.25 $\pm$ 0.05	28.21 $\pm$ 0.04	28.16 $\pm$ 0.10
SSIM ($\uparrow$)	0.43 $\pm$ 0.09	0.59 $\pm$ 0.02	0.52 $\pm$ 0.02	0.53 $\pm$ 0.02	0.66 $\pm$ 0.01
CNR ($\uparrow$)	0.38 $\pm$ 0.53	0.54 $\pm$ 0.59	0.43 $\pm$ 0.60	0.38 $\pm$ 0.45	0.61 $\pm$ 0.63
ENL ($\uparrow$)	587.96 $\pm$ 1148.71	2201.45 $\pm$ 6145.01	189.44 $\pm$ 474.42	130.10 $\pm$ 315.65	20994.17 $\pm$ 48338.10

3, our proposed model outperformed all other ablated variants across evaluation metrics.

As part of the ablation study, we also evaluated other loss functions and learning rate scheduling strategies, as described in Section II-D. The tested losses are L_1 loss with multiscale SSIM loss (L_{SSIM}) [33], total variation loss (L_{TV}) [34], gradient difference loss (L_{GD}) [35], loss with a linear combination of PSNR and SSIM called PSNR-SSIM loss ($L_{PSNR\ SSIM}$), and L_1 loss. The results appear in Figure 4 and Table I. Figure 4's zoomed-in portions show that due to the use of a smoothing loss, the despeckled images are much smoother than those coming from other loss functions. Moreover, the SSIM loss noticeably worsens performance.

To justify the use of LAMSA for SAR despeckling, we compared it with other commonly used multi-head self-attention (MSA) mechanisms, including WMSA (as in SCUNet), FlashMSA [36], low-rank factorization MSA (LRMSA) [37], and rotary positional encoding-based MSA (RoPEMSA) [38]. The quantitative results are summarized in Table II. Although WMSA trained faster, its limited global context made it less effective for despeckling. LRMSA achieved the second-best performance and the fastest inference owing to improved token interactions over a larger context. FlashMSA and LRMSA primarily optimize the self-attention computation. LAMSA was selected based on its superior overall performance on the metrics.

B. Model Comparison Results

The model comparison results with the SAR transformer and SCUNet model are shown in Table III and Figure 5. The qualitative results in Fig. 5 (a) highlight that the proposed model more effectively restores sharp crucial details, such as the two building structures, than other competitive models. One can observe in (b) and (c) that the proposed model smooths out the homogeneous regions far better than other models and also keeps clear boundaries between the two different slopes of the hilly region without smudging the boundaries.

The quantitative results confirm the model's performance, as shown in Table III. On average, compared to the four other competitive models, the proposed model achieves a 10.6% improvement in SSIM, a 94.9% enhancement in CNR, and a 416.7% increase in ENL on the Sentinel-1 Himalaya all-season dataset. For the PALSAR-2 ScanSAR dataset, it shows an 11.2% improvement in SSIM, a 44.0% enhancement in CNR, and a 62.0% increase in ENL. Additionally, the proposed approach delivers a 29.2% improvement in SSIM, a 44.0% enhancement in CNR, and a staggering 7835.9% increase in ENL on the Sentinel-1 Global Spring dataset. Based on a combined qualitative and quantitative evaluation, the proposed model outperformed the strong baselines, with linear computational complexity.

We also experimented with improved versions of the SCUNet and the SAR transformer. SCUNet's WMSA was replaced with linear-angular attention to retain global contexts with fast training. (This variant is called SCUNet LA.) However, the SAR transformer was improved with the introduction of FastNL-based preprocessing and NLM-based embedding generation before being fed to the transformer. Even if the SCUNet LA model's despeckled SARs mostly yield the best or second-best quantitative results, it is evident from the qualitative evaluation of the images in Figure 5 that SCUNet and SCUNet LA are over-smoothing. The reason may be that SCUNet was designed for general denoising purposes, for which multiplicative noise is a special case.

Some statistical tests were conducted to determine whether the proposed despeckling method significantly outperformed the four other methods (SCUNet, SCUNet LA, SAR Trans, and SAR Trans NLM). The Kruskal-Wallis one-way analysis of variance was used to test if the five despeckling methods significantly differed from each other on the four evaluation metrics (PSNR, SSIM, CNR, and ENL) for the three datasets. A statistically significant ($p < 0.05$) Kruskal-Wallis one-way analysis was achieved. Further, multiple comparisons using the Mann-Whitney U-test were performed with Holm-Bonferroni-corrected p -values at a significance level of $p < 0.012$ for four comparisons.

On the Sentinel-1 Himalaya All Seasons dataset, the proposed model provided statistically significant improvements in SSIM over all other models with a $p < 0.00001$. On CNR, the proposed model significantly outperformed SAR Trans and SAR Trans NLM with $p < 0.00001$; and on ENL, the proposed model significantly improved despeckling compared to SCUNet and SCUNet LA with p -values of 0.00233 and

0.00056, respectively.

On the PALSAR dataset, the proposed method significantly outperformed the four other methods on SSIM with $p < 0.00001$. On CNR, the proposed method significantly outperformed SCUNet with $p < 0.00001$, SAR Trans with $p = 0.00005$, and SAR Trans NLM with $p = 0.00015$. On ENL, the proposed method significantly outperformed SCUNet with $p < 0.00001$ and SAR Trans with $p = 0.00042$.

On the Sentinel-1 Global Spring dataset, the proposed method provided statistically significant improvements in SSIM compared with all other models with a $p < 0.00001$. On CNR, the proposed method exceeded SCUNet with $p < 0.00001$, SAR Trans with $p = 0.00005$, and SAR Trans NLM with $p = 0.00015$. On ENL, the proposed method outperformed SCUNet with $p < 0.00001$ and SAR Trans NLM with $p = 0.00042$. Overall, the proposed method significantly improved performance on SSIM, CNR, and ENL compared with the other methods.

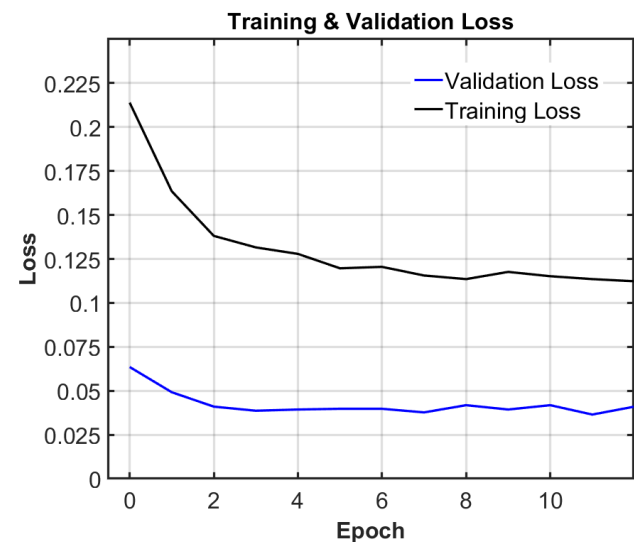


Fig. 6. Training and validation loss vs. epoch curves for the proposed model.

IV. DISCUSSION

Our proposed model's benefits are (1) self-supervised training without the need for pristine SAR images (which are difficult to acquire); (2) training the model with inherently noisy SAR images to preserve domain-specific features, unlike any other SAR despeckling models. This is possible because of the efficient local-global feature extraction capability of LAMSA, and (3) lower computational complexity, aiming for greener AI.

The cost of an AI model grows with the cost of executing the model on a single sample (for example, the number of model parameters), the size of the training dataset, and the number of hyperparameter experiments [39]. The proposed model could achieve better performance on SSIM, CNR, and ENL metrics when trained with a much smaller dataset (about 160 images) for only 13 epochs. Under similar conditions, SCUNet produces blurry, despeckled images, and the SAR

transformer produces poor, deformed images, as shown in Figure 5. The study did not perform an extensive hyperparameter search for the loss function weights (instead adopting conventionally used loss term weights), as that was neither our primary goal nor sustainable for green AI [39]. The number of model parameters of the proposed model (20.5 M) is substantially less compared with the SAR transformer (25.3 M). It takes about three days to train SCUNet on four NVIDIA RTX 2080 Ti GPUs. The proposed model can be trained using a single NVIDIA RTX A6000 GPU in 10 m 40 s, yielding perceptually satisfying despeckling results. Figure 6 shows parallel decreasing trends without divergence, indicating that the model generalizes well with no evidence of overfitting within these epochs. The lower validation loss relative to training loss is due to regularization (e.g., weight decay, dropout) and differences in dataset characteristics [40].

Evaluating model performance posed some challenges. Due to the unavailability of clean ground truth SAR images, the training dataset was created by pairing real SAR images with noise-added images. The PSNR and SSIM metrics require a reference or ground-truth image. These metrics were calculated between the real SAR image and the despeckled image. Due to the inherent noise present in the reference SAR image, the PSNR metric is not a reliable indicator. Although PSNR values are presented in the tables, the proposed method has not outperformed the others on PSNR for the same reason.

The SSIM metric is more reliable because it reflects the structural and perceptual similarity. Non-reference metrics like CNR and ENL are better as they are calculated solely from the despeckled image. CNR and ENL values are highly dependent on the choice of position and size of foreground and background boxes. For CNR, the background box must contain a homogeneous region, and the foreground box should contain high-frequency details. For ENL, the region of interest must be homogeneous, as ENL concerns the smoothness of the image. For each experimental result reported in the ablation study and the model comparison tables, different background and foreground boxes for CNR and ENL are chosen heuristically to present the superior values of CNR and ENL metrics.

Apart from quantitative performance evaluation, we also rely on qualitative visualization. The zoomed-in areas in all the figures show that the despeckled image output from the proposed model not only gets a smooth texture, but also preserves edges and finer details, unlike the other methods.

The despeckled image output from the model was further filtered using the FastNLM filter [25] to eliminate artifacts generated by direct training on SAR images. This algorithm replaces the window similarity with pixel intensity similarity at each level in the multiresolution pyramid structure, filtering out non-similar pixels within a neighborhood. We experimented with different window sizes for the FastNLM filtering step (3 to 15, with a step size of 1) while applying it as a post-processing step. We also directly applied FastNLM filtering on noisy images to check whether a deep learning model is even necessary.

One interesting observation is that the PSNR and SSIM values between the ground truth and noisy images are the same as those between the ground truth and FastNLM filtered

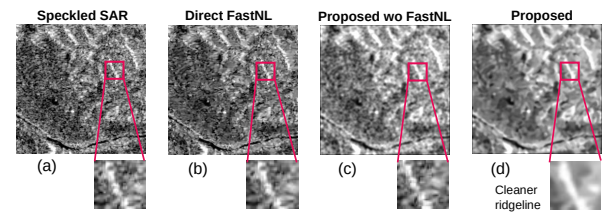


Fig. 7. (a) A reference image with speckle noise is presented. The corresponding despeckled images obtained for qualitative comparisons using (b) direct FastNLM filtering, (c) the proposed model without FastNLM filtering, and (d) the proposed model with FastNLM filtering. Zoomed-in regions compare despeckling quality and structural detail preservation.

images of different window sizes. In contrast, when FastNLM is applied as a post-processing step of our proposed model, the PSNR increases (indicating more denoising) and the SSIM decreases (indicating more smoothing, which results in the loss of details). Also, the appearance of the directly FastNLM filtered image is smudged at the edges, whereas it is continuous for the proposed model. Thus, applying FastNLM as post-processing is more effective than using it standalone. Based on a qualitative comparison and SSIM, CNR, and ENL values, we selected a window size of 5 for application in the FastNLM post-processing step. For the direct application of FastNLM to the noisy image, a filter size of 15 yields superior performance; however, it still falls short of the proposed model's results.

Figure 7 compares image quality when using only FastNLM filtering versus applying it after the proposed despeckling algorithm. The NLM-filtered image reduces granular noise while preserving fine structural details such as edges and texture boundaries. The filtering leads to cleaner homogeneous areas, well-maintained linear features, and fewer speckle artifacts, producing a visually coherent image and clearly demonstrating an improvement in human-perceived image quality. For example, in the zoomed-in region of Fig. 7(d), the high-altitude mountain ridgeline is clearly visible with substantially reduced speckle noise. Homogeneous regions, such as slopes and flat terrain, appear smooth due to effective noise suppression, while crucial linear details like rivers are preserved.

In the linear-angular attention block, as described in [15], a sparse masked softmax-based attention can capture connections to nonadjacent tokens, but it quickly converges to all zeros as the training progresses, leaving it unused for a considerable training duration. Additionally, we aimed to avoid the costly softmax-based attention calculation; therefore, we eliminated this term. The experimental results supported this decision, yielding better results without the sparse regularization term.

The hierarchical regional feature extraction block was added to capture features at multiple scales. This integration provided an alternative to computationally intensive sparse regularization. The ReLU at the end of the ConvTrans block creates sparse activation by suppressing any negative contribution through the residual connection, thus acting as an implicit regularizer and improving overall performance.

Among the various loss functions tested, only the combination of $L_1 + L_{\text{smooth}}$ performed well for despeckling be-

1 cause this balances noise suppression and detail preservation.
2 L_{SSIM} focuses on perceptual similarity but does not explic-
3 itly suppress noise, often retaining noise-like high-frequency
4 details. L_{TV} enforces smoothness but can excessively blur
5 fine textures, degrading important details, like those observed
6 in the SAR transformer. L_{GD} , which enforces similarity in
7 image gradients, does not sufficiently reduce speckle noise in
8 SAR images. $L_{PSNR\ SSIM}$ does not directly optimize for noise
9 suppression, as PSNR prioritizes intensity matching while
10 SSIM emphasizes structural similarity rather than despeckling.
11 Using only L_1 loss ensures pixel reconstruction but does not
12 enforce smoothness, allowing residual noise to persist.

13
14
15 V. CONCLUSION

16 SAR images require despeckling as a preprocessing step for
17 various satellite imaging tasks. This study presents an efficient
18 transformer-based SAR despeckling algorithm that integrates
19 LAMSA with nonlocal means, capturing both local and global
20 context in linear time, thereby overcoming the WMSA trans-
21 former’s local-only limitation. The model outperforms four
22 strong baselines on SSIM (perceptual and structural similarity
23 retention), CNR (contrast preservation), and ENL (speckle
24 suppression) for three datasets without requiring ground-truth
25 references. The results show that LAMSA performs the best
26 for SAR despeckling because of its sensitivity to angular
27 orientation. We envision that the proposed model will be
28 a greener approach to performing SAR despeckling with
29 multiple use cases.

30
31 ACKNOWLEDGMENT

32 The authors would like to thank the late Prof. Yuji Iwahori
33 (Chubu University, Japan) for his initial discussions on this
34 work. Prof. Iwahori passed away in December 2024. This work
35 acknowledges support from the ISRO-IIT Guwahati Space
36 Technology Cell.

37
38 REFERENCES

39
40 [1] Puyang Wang, He Zhang, and Vishal M Patel. SAR image despeckling
41 using a convolutional neural network. *IEEE Signal Processing Letters*,
42 24(12):1763–1767, 2017.
43 [2] Can Wang, Rongyao Zheng, Jingzhen Zhu, Wentao Xu, and Xiwen Li.
44 A practical SAR despeckling method combining SWIN transformer and
45 residual CNN. *IEEE Geoscience and Remote Sensing Letters*, 2023.
46 [3] Siyao Xiao, Shunsheng Zhang, Libing Huang, and Wen-Qin Wang.
47 Trans-NLM network for SAR image despeckling. *IEEE Transactions*
48 *on Geoscience and Remote Sensing*, 2024.
49 [4] Prabhishek Singh, Manoj Diwakar, Achyut Shankar, Raj Shree, and
50 Manoj Kumar. A review on SAR image and its despeckling. *Archives*
51 *of Computational Methods in Engineering*, 28(7):4633–4653, 2021.
52 [5] Vishal M. Patel, Glenn R. Easley, Rama Chellappa, and Nasser M.
53 Nasrabadi. Separated component-based restoration of speckled SAR
54 images. *IEEE Transactions on Geoscience and Remote Sensing*,
55 52(2):1019–1029, 2014.
56 [6] Prabhishek Singh, Achyut Shankar, Manoj Diwakar, and Mohammad R
57 Khosravi. MSPB: Intelligent SAR despeckling using wavelet threshold-
58 ing and bilateral filter for big visual radar data restoration and provi-
59 sioning quality of experience in real-time remote sensing. *Environment,*
60 *Development and Sustainability*, pages 1–31, 2022.
[7] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A
database of human segmented natural images and its application to
evaluating segmentation algorithms and measuring ecological statistics.
In *Proceedings Eighth IEEE International Conference on Computer*
Vision. ICCV 2001, volume 2, pages 416–423. IEEE, 2001.

[8] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong,
Hongliang Li, and Lei Zhang. Waterloo exploration database: New
challenges for image quality assessment models. *IEEE Transactions*
on Image Processing, 26(2):1004–1016, 2016.
[9] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single
image super-resolution: Dataset and study. In *Proceedings of the IEEE*
Conference on Computer Vision and Pattern Recognition Workshops,
pages 126–135, 2017.
[10] Yuqian Zhou, Jianbo Jiao, Haibin Huang, Yang Wang, Jue Wang,
Honghui Shi, and Thomas Huang. When AWGN-based denoiser meets
real noises. In *Proceedings of the AAAI Conference on Artificial*
Intelligence, volume 34, pages 13074–13081, 2020.
[11] Malsha V Perera, Wele Gedara Chaminda Bandara, Jeya Maria Jose
Valanarasu, and Vishal M Patel. Transformer-based SAR image de-
speckling. In *IGARSS 2022-2022 IEEE International Geoscience and*
Remote Sensing Symposium, pages 751–754. IEEE, 2022.
[12] Jie Li, Shaowei Shi, Liupeng Lin, Qiangqiang Yuan, Huanfeng Shen, and
Liangpei Zhang. A multi-task learning framework for dual-polarization
SAR imagery despeckling in temporal change detection scenarios.
ISPRS Journal of Photogrammetry and Remote Sensing, 221:155–178,
2025.
[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang,
Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision
transformer using shifted windows. In *Proceedings of the IEEE/CVF*
International Conference on Computer Vision, pages 10012–10022,
2021.
[14] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao
Tang, Deng-Ping Fan, Radu Timofte, and Luc Van Gool. Practical
blind image denoising via swin-conv-unet and data synthesis. *Machine*
Intelligence Research, 20(6):822–836, 2023.
[15] Haoran You, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peizhao
Zhang, Haoqi Fan, Peter Vajda, and Yingyan Celine Lin. Castling-ViT:
Compressing self-attention via switching towards linear-angular atten-
tion at vision transformer inference. In *Proceedings of the IEEE/CVF*
Conference on Computer Vision and Pattern Recognition, pages 14431–
14442, 2023.
[16] Malsha V Perera, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda
Bandara, and Vishal M Patel. SAR despeckling using a denoising
diffusion probabilistic model. *IEEE Geoscience and Remote Sensing*
Letters, 20:1–5, 2023.
[17] Xuran Hu, Ziqiang Xu, Zhihan Chen, Zhenpeng Feng, Mingzhe Zhu,
and Ljubiša Stanković. SAR despeckling via regional denoising dif-
fusion probabilistic model. In *IGARSS 2024-2024 IEEE International*
Geoscience and Remote Sensing Symposium, pages 7226–7230. IEEE,
2024.
[18] Puyang Wang, He Zhang, and Vishal M Patel. Generative adversarial
network-based restoration of speckled SAR images. In *2017 IEEE*
7th International Workshop on Computational Advances in Multi-sensor
Adaptive Processing (CAMSAP), pages 1–5. IEEE, 2017.
[19] Feng Gu, Hong Zhang, and Chao Wang. A GAN-based method for SAR
image despeckling. In *2019 SAR in Big Data Era (BIGSAR DATA)*, pages
1–5. IEEE, 2019.
[20] Ruijiao Liu, Yangyang Li, and Licheng Jiao. SAR image speckle reduc-
tion based on a generative adversarial network. In *2020 International*
Joint Conference on Neural Networks (IJCNN), pages 1–6. IEEE, 2020.
[21] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. SAR2SAR: A
semi-supervised despeckling algorithm for SAR images. *IEEE Journal*
of Selected Topics in Applied Earth Observations and Remote Sensing,
14:4321–4329, 2021.
[22] Emanuele Dalsasso, Loïc Denis, and Florence Tupin. As if by magic:
Self-supervised training of deep despeckling networks with merlin. *IEEE*
Transactions on Geoscience and Remote Sensing, 60:1–13, 2021.
[23] Andrea Bordone Molini, Diego Valsesia, Giulia Fracastoro, and Enrico
Magli. Speckle2void: Deep self-supervised SAR despeckling with blind-
spot convolutional neural networks. *IEEE Transactions on Geoscience*
and Remote Sensing, 60:1–17, 2021.
[24] Uladzislau Yorsh and Alexander Kovalenko. Linear self-attention
approximation via trainable feedforward kernel. In *International Con-*
ference on Artificial Neural Networks, pages 807–810. Springer, 2022.
[25] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local
means denoising. *Image Processing On Line*, 1:208–212, 2011.
[26] Charles-Alban Deledalle, Loïc Denis, and Florence Tupin. Iterative
weighted maximum likelihood denoising with probabilistic patch-based
weights. *IEEE Transactions on Image Processing*, 18(12):2661–2672,
2009.
[27] SARa Parrilli, Mariana Poderico, CeSARio Vincenzo Angelino, and
Luisa Verdoliva. A nonlocal SAR image denoising algorithm based

on lmmse wavelet shrinkage. *IEEE Transactions on Geoscience and Remote Sensing*, 50(2):606–616, 2011.

- [28] Giovanni Chierchia, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. SAR image despeckling through convolutional neural networks. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 5438–5441. IEEE, 2017.
- [29] Tobias Plötz and Stefan Roth. Neural nearest neighbors networks. *Advances in Neural Information Processing Systems*, 31, 2018.
- [30] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *Advances in neural Information Processing Systems*, 31, 2018.
- [31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022.
- [32] Michael Schmitt, Lloyd Haydn Hughes, and Xiao Xiang Zhu. The sen1-2 dataset for deep learning in SAR-optical data fusion. *arXiv preprint arXiv:1807.01569*, 2018.
- [33] Zhou Wang, Eero Simoncelli, and Alan Bovik. Multi-scale structural similarity for image quality assessment. *Proceedings of the IEEE Asilomar Conference Signals, Systems and Computers*, 02 2004.
- [34] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.
- [35] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.
- [36] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.
- [37] Sneha Mehta, Huzefa Rangwala, and Naren Ramakrishnan. Low rank factorization for compact multi-head self-attention. *arXiv preprint arXiv:1912.00835*, 2019.
- [38] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024.
- [39] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.
- [40] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning*, pages 10181–10192. PMLR, 2020.



Souraja Kundu was born in Kolkata, West Bengal, India. She completed her B.Tech. in Electronics and Communication Engineering from the Indian Institute of Technology (IIT) Guwahati in 2025, where she was awarded the Governor of Assam Gold Medal. She is currently pursuing a Ph.D. in the Machine Learning Department at Carnegie Mellon University's School of Computer Science in the United States. She has previously worked as a Deep Learning and Computer Vision Research Intern at Chubu University, Japan (2023).



M.K. Bhuyan (Senior Member, IEEE) received a Ph.D. degree in electronics and communication engineering from the Indian Institute of Technology (IIT) Guwahati, India. He was a postdoctoral researcher at the School of Information Technology and Electrical Engineering at the University of Queensland, Australia. He was an Assistant Professor in the Department of Electrical Engineering at IIT Roorkee, India, and Jorhat Engineering College, Assam, India. He also worked in the Indian Engineering Services. In 2014, he was a Visiting Professor at Indiana University and Purdue University, Indiana, USA. He received the National Award for Best Applied Research/Technological Innovation from the President of India in 2012. He is an IEEE senior member. He is currently a professor with the Department of Electronics and Electrical Engineering at IIT Guwahati. He is also a visiting professor in the Department of Computer Science at Chubu University, Japan. His research interests include image and video processing, computer vision, machine learning, human-computer interaction (HCI), virtual reality, augmented reality, and biomedical signal processing.



Karl Fredric MacDorman (Senior Member, IEEE) received the Ph.D. degree in computer science from Cambridge University, Cambridge, U.K., in 1997. He is an Associate Professor in the Human-Computer Interaction Program with the Luddy School of Informatics, Computing, and Engineering, Indiana University, Indianapolis, IN, USA. He is also the Associate Dean of Academic Affairs. His research interests include cognitive science, human-computer interaction, machine learning, and robotics.



Neeraj Kumar Sharma (Member, IEEE) is an Assistant Professor at the School of Data Science and Artificial Intelligence, in the Indian Institute of Technology, Guwahati (India). He received his PhD from the Indian Institute of Science, Bangalore, India, in 2018. He was a BrainHub Postdoctoral Fellow at Carnegie Mellon University (CMU) from 2017 to 2018. He was a Postdoctoral Researcher at the Neuroscience Institute and Department of Psychology, CMU, Pittsburgh, from 2019 to 2020. In 2020, he was a Postdoctoral Researcher at the Society of Innovation and Development (SiD), Indian Institute of Science, Bengaluru, and in 2021, he was a CV Raman Postdoctoral Researcher at the Indian Institute of Science, Bengaluru (India). From 2021 to 2022, he was a Scientist at the Fraunhofer Institute for Integrated Circuits, IIS in Erlangen (Germany). His research interests include deep learning, speech and audio signal processing, and multimodal signal processing.



Manish Bhatt (Member, IEEE) is an Assistant Professor at the Department of Electronics and Electrical Engineering at the Indian Institute of Technology (IIT), Guwahati, India. He received his PhD degree from the Indian Institute of Science, Bangalore, India, in 2017. He was a postdoctoral fellow at the University of Montreal, Canada, from 2017 to 2020. His research interests include image processing, remote sensing, medical imaging, inverse problems, machine learning, and deep learning.