Transformer-Based Feature-Preserving Despeckling Framework for Synthetic Aperture Radar Imagery

Bibek Kumar[1], Ajay Kumar[2]
[1] Lincoln University College; [2] IILM University
drbibek.pdf@lincoln.edu.my

Abstract: Synthetic Aperture Radar (SAR) images are regularly used in the remote sensing because these images can be generated in any weather or light condition. However, the SAR images are always exaggerated by speckle noise, which makes the SAR images less clear and makes it difficult to get useful structural and textural data for processing. To make it effective, despeckling methods must minimize noise while retaining the important image features such as edges and fine textures. To solve these issues, this article developed a Transformer Based Despeckling Network (TBDN). The developed framework initially extracts low level features from the noisy SAR image by use of convolutional layers. Then, transformer encoder blocks use extracted features with the help of multi head self-attention mechanisms to achieve long range spatial dependencies as well as contextual data. Hierarchical characteristics are collected in a multi-scale feature fusion module to support preserve the structure while reducing noise. Finally, the despeckled SAR image is enhanced with the help of image reconstruction layer.

Keywords: Peak Signal to Noise Ratio; SAR Imagery; Speckle Noise; SSIM; Transformer.

Introduction

Synthetic Aperture Radar (SAR) imagery has become an important innovation in the area of remote sensing because of its ability to capture high resolution images independent of weather conditions and lighting variations. The SAR utilized unique microwave signals and these can go through smoke, darkness and clouds, for making it useful for watching the environment, military operation, and dealing with disasters. But, due to bounce back of signals, many things in a area and get mixed up and results in form of speckle noise. Speckle noise is responsible for making analysis difficult and it requires interpretation before analysis.

There are so many statistical filtering methods to eliminate speckle noise. The Lee filter is one of the first method to reduce speckle noise while trying to keep clear image with edge information preservation. This method is used to minimize the mean square error in the imagery [1]. The Kuan is another filtering method which try to enhance things by adjusting the image filtering from one spot to other spot [2]. There is another despeckling filtering technique known as Frost filter which uses special set of weights that reduce quickly to smooth out image [3]. These old filtering techniques are little bit fast and too hard to put into action. These despeckling techniques look for similarities in various parts of images not just nearer pixels.

The Non-Local Means (NLM) algorithm were introduced to overcome the statistical filtering techniques that uses redundancy within the image. In NLM, we can compute pixel similarities across distant image

areas with averages similar patches to minimize noise with structural details preservation [4]. The NLM methods improves performance of despeckling as compared to local filters but they are more computational expensive and also need substantial processing time in case of large SAR datasets. The performance may also get affected in highly heterogeneous regions because of difficulties In identifying similar patches.

In recent time, the convolutional neural network (CNN) based approaches comes into picture and achieved a better result but they faced some of the challenges. The CNN methods are basically capturing information from small spatial neighborhoods. Whereas deeper CNN can be used to increase the receptive filed but with the cost of large computational resources and these networks needed a complex architecture. For example, Zhang et al. presented a residual learning-based CNN architecture which enhanced the performance of the despeckling process with mapping the noisy and clean images [5]. Another CNN methodology developed by Chierchia et al. in which they enhanced noise suppression and visual quality of SAR imagery [6].

The transformer-based architectures specially developed for the natural language processing work. They have been adapted for solving computer vision problems just because of its powerful self-attention mechanism. In the transformer every element of the input allows to attend every other element and it enable the capability of capturing global contextual relationship with the information. The different transformer architecture including Visual Transformer (ViT) have demonstrated strong performance in tasks such as image segmentation, restoration and classification. Transformers have capability in modelling long range interactions among image patches and making them particularly perfect for tasks that need comprehensive contextual understanding [7].
Transformer based models have been used in the field of image restoration, denoising, and deblurring in recent days. Such models leverage multi head self-attention process in extracting both global and local features. The use of transformers into image restoration methodologies gives us a great solution for SAR despeckling but retaining structural information is difficult while despeckling is critical.
In this article, we proposed a Transformer Based Despeckling Network (TBDN) designed particularly for improvement in SAR images. In the proposed architecture, we combine the strength of convolutional neural networks with the transformer-based attention techniques. In the extraction of low-level spatial features from the SAR input imagery, we used a convolutional layer block whereas transformer encoder blocks in capturing long range contextual information with the multi head self-attention operations. In this architecture we used multi scale feature fusion model additionally, which help us in integrating feature representation to enhance edge preservation and structural consistency. At the last stage, an image reconstruction module generated the despeckled output image with minimized noise and improve visual quality.

The major objectives we finalized for this study are as follows:

1. To design and develop an improved despeckling method with the capability of redcuing speckle noise in SAR images without losing critical structure and textual data.
2. To improve capability of storing features in the despeckling process by leveraging transformer-based attention model and multi scale fusion techniques.

The goal of the proposed method is to improve qualitative as well as quantitative performance compared to available conventional filtering techniques and available deep learning-based methods/models. The evaluation of these models uses standard performance metrics like peak signal to noise ration (PSNR), Structural Similarity Index (SSIM), and Edge preservation Index (EPI). These metrices are responsible to explain effectiveness of the proposed methodologies in getting high quality SAR image restoration.

The remaining article is organized as given. In section II, reviews related work in SAR despeckling field, which includes traditional filtering approaches, NLM based methodology and deep learning-based methods. In section III, the architecture and methodology of the proposed network have been included. Section IV is used to explain the experimental setup with different evaluation metrics used in this work. In section V, the experimental result as well as comparison with other available state of arts are being discussed. Finally, the conclusion of the article is discussed in section VI.

Related work

SAR imagery has been popular research field in reducing speckle noise since many decades. There are so many numerous methods have been designed and developed ranging from transformer-based filtering to advanced neural network-based frameworks. The main challenge is SAR image despeckling is to maintain balance among the noise reduction as well as useful structural feature preservation like edges, boundaries, and textures. In recent years, development in the machine learning and attention based models have opened new opportunities in enhancing despeckling performance while maintaining image information.

Argenti et al. developed a wavelet domain despeckling method specially for SAR imagery. In this approach speckle noise is redued by using adaptive thresholding in transform area [8]. This method achieve high noise reduction performance when copared with other classical spatial filtering approach.

However, wavelet-based methods may produce artifacts in highly textured areas due to limitations in complex spatial structures representation.

The variational models and optimization based methodologies are used for image restoration. Aubert and Aujol developed a variational model for SAR despeckling that formulates the issue as an energy minimization task by adding fidelity term and a regularization part [9]. These techniques tried to preserve edge information while smoothing similar area. Sparse representation methods have also been utilized for SAR image restoration. Dabov et al. proposed a Block Matching and 3D (BM3D) filtering technique and it groups same image patched to perform collaborative filtering in transform domain [10]. This techniques recognized for it powerful denoising capabilities as well as ability to hold fine image structures. In other developments, BM3D were applied to SAR imagery by incorporating logarithmic transformations to handle multiplicative noise.

In the era of machine learning, researchers began exploring learning-based methods for despeckling. Chen et al. developed a deep neural network architecture which learns the mapping among noisy and clean SAR imagery with the use of supervised learning [11]. The deep neural network model achieve better performance when compared with traditional filters, especially in different regions.

Another impactful contribution was made by Mao et al., who demonstrated a deep residual encoder–decoder network for the use of image restoration [12]. The encoder–decoder structure permitted the model to detention of multi-scale representations, which proved themselves in the reducing noise with

maintaining structural data. Nevertheless, convolution-based framework are still facing challenges in designing long range dependencies across large spatial regions.

To overcome such issues, attention based methods were developed and introduced in image restoration frameworks. Zhang et al. came up with a residual channel attention network (RCAN). The RCAN integrates attention modules to emphasize useful feature channels in the training [13]. This method enhanced feature representation as well restoration quality of degraded imagery.

Now a days, transformer-based architectures have expanded noteworthy attention in the area of computer vision. Liang et al. came up with Swin Transformer, a hierarchical vision transformer architecture and it processes images with shifted window attention process [14]. This technique competently captures both local and global dependencies and simultaneously maintaining computational efficiency. Its success in different vision work has encouraged researchers to investigate transformer-based techniques for image restoration and despeckling. Similarly, a restormer was proposed by Zamir et al. for high resolution image restoration work [15]. This method includes multi head attention with efficient feature aggregation in improving restoration quality. On the basis of achieved result, it can be consider that transformer based frameworks outperform CNN based models in task.

Because of these advancements, The SAR image despeckling area becomes an emerging research area in the transformer architecture applications. The SAR imagery are affected by a unique structural pattern as well as speckle noise and this is the reason it requires some special modeling techniques. The available state of arts is not sufficient to handle noise reduction and edge preservation simultaneously. These challenges make transformer-based attention methods useful in multi scale feature extraction which may give a proper direction in enhancing SAR image analysis.

In the Transformer based despeckling network (TBDN), three blocks are introduced; Convolutional feature extraction, multi scale feature fusion, and transformer encoder block. The main goal of this combination of all three block is to control the things in both convolutional network and transformer based attention to reduce speckle noise with preservation of edge information in the image.

Proposed Methodology

To handle the issue of speckle noise reduction with edge information preservation in SAR imagery, we are proposing a new architecture TBDN. The proposed architecture uses a combination of convolutional feature extraction and transformer based attention techniques which is able to capture spatial data and contextual relationship globally.

The ability of capturing long rang dependencies in an image is restricted in the traditional convolution neural network. In this contrast, the TBDN model employs self attention technique that can help in the analyzing interaction among distant pixel by enabling network. It is much more capable in allowing the system to differentiate speckle noise with useful image structure like edges, boundaries and tectures.

The proposed TBDN network uses four important components and represented by Figure 1:

I. Feature Extraction Module
II. Transformer Encoder Blocks
III. Multi-Scale Feature Fusion Module
IV. Image Reconstruction Layer

The complete architecture is developed to reduce speckle noise with retaining image characters and boundaries information.

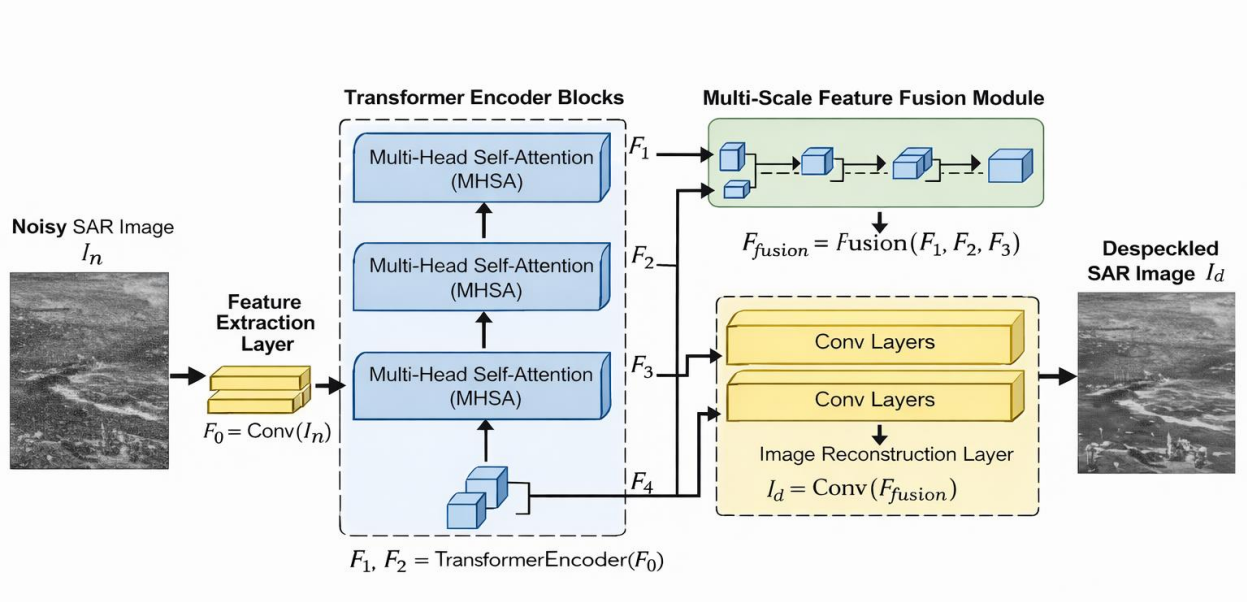


Figure 1. The flow diagram of Transformer Based Despeckling Network (TBDN) Architecture.

Feature Extraction Model

The despeckling process starts from the feature extraction step. In this step the noisy SAR image is use to transform it into high dimensional feature representation. The noisy SAR image can be represented by given Equation(1):

$$\mathrm{F_0} = \mathrm{Conv}(\mathrm{I_n}) \qquad (1)$$

In this Equation (1), F_0 is used to represent the extracted feature map and it contains image edges and texture information. The extracted features can be used as initial representation and can be process by the transformer blocks later.

Transformer Encoder Blocks

After feature extraction, the obtained feature maps are processed through a sequence of transformer encoder blocks. These blocks can be used for modelling long range dependencies in the image. Every transformer encoder will carry following parts.

i. Multi-Head Self-Attention (MHSA)
ii. Layer Normalization
iii. Feed-Forward Network (FFN)
iv. Residual Connections

The self-attention mechanism computes relationships among all pixel positions in the feature map. The attention operation can be expressed as:

$$Attention(Q, K, V) = Softmax(QK^T/\sqrt{d^k})$$

where:

Q represents query vectors

K represents key vectors

V represents value vectors

d^k is the dimension of the key vectors

With the above process, the network will assign weights to various areas of the image, and it will differentiate between speckle noise artifacts and structural feature. There are several transformer blocks those are stacked to improve model's ability for capturing complex spatial relationship in the complete image.

Multi-Scale Feature Fusion Module

The speckle noise reduction needs the model to consider data at diverse spatial scales. With this consideration, the proposed model uses a multi scale feature fusion module that use to combine extracted feature from different transformer layers. Suppose F1, F2, and F3 represents feature maps received from various stages of the transformer encoder. Then the received feature maps are combine with the help of fusion operation.

$$F_{fusion} = Fusion(F_1, F_2, F_3$$

The fusion process helps the model in retaining high level contextual data as well as fine grained information. With the help of features integration at multiple levels, the network enhances the ability of edges and texture data preservation during despeckling.

Image Reconstruction Layer

The output received from feature fusion model; the received features is passed through the image reconstruction block for converting the processed feature maps into a speckle noise free module.

The regenerated SAR mage I_d is achieved and represented :

$$I_d = Conv(F_{fusion}). \quad (2)$$

The image construction layer uses convolutional fillter to map the fused featured back to the actual and original image space. The generated output image represented the decpeckle image with enhanced quality and structural data.

Loss Function

To effectively train the proposed model, a combination of reconstruction and structural similarity losses is used. The total loss function is defined as:

$$L_{total} = \alpha L_{MSE} + \beta L_{SSIM}$$

where:

L_{MSE} is used to calculate pixel-wise reconstruction error

L_{SSIM} is used to evaluates structural similarity among the reconstructed and reference images

$\alpha \; and \; \beta$ are representing weighting parameters.

Advantages of the Proposed Model

The proposed TBDN network is very useful with many advantages such as :

Global Context Awareness – Self-attention captures long-range spatial relationships.

Improved Feature Preservation – Multi-scale fusion can be used to maintains edges and textures information.
Robust Noise Suppression – Transformer blocks effectively differentiate speckle patterns from meaningful features.
Better Reconstruction Quality – Combined loss functions ensure accurate image restoration.
The above advantages use to enable the model in achieving superior performance as compared with traditional filters and other available CNN based despeckling methodologies.

Result and Analysis

Experimental Setup
The proposed Transformer-Based Despeckling Network (TBDN) was evaluated using SAR images affected by speckle noise. The experiments were conducted using simulated speckle noise with different noise variances to assess the robustness of the proposed model. The performance of the proposed approach was compared with widely used despeckling techniques including Lee Filter, Frost Filter, Non-Local Means (NLM), and a CNN-based denoising model.

The evaluation was performed using two widely accepted image quality metrics:

- Peak Signal-to-Noise Ratio (PSNR)
- Structural Similarity Index Measure (SSIM)

These metrics quantify the restoration quality of despeckled images in terms of pixel-level accuracy and structural similarity.
Peak Signal-to-Noise Ratio (PSNR)
PSNR measures the similarity between the despeckled image and the reference image by comparing pixel intensities. A higher PSNR value indicates better noise suppression and image restoration.

$$PSNR = 10log_{10}\left({MAX^2}/{MSE}\right)$$

where:
MAX represents the maximum pixel value and
MSE represents the mean squared error between the restored and reference images.

Structural Similarity Index (SSIM)
SSIM evaluates image similarity by comparing luminance, contrast, and structural information between the restored and reference images.

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

where:
μ_x, μ_y represent the mean intensity values
σ_x^2, σ_y^2 represent the variances
σ_{xy} represents covariance between images

Table 1. Compares this work with the related work or previous research by other researchers

Method	PSNR (dB)	SSIM
Lee Filter	24.5	0.71
Frost Filter	25.2	0.73
Non-Local Means (NLM)	27.1	0.80
CNN-Based Model	29.3	0.85

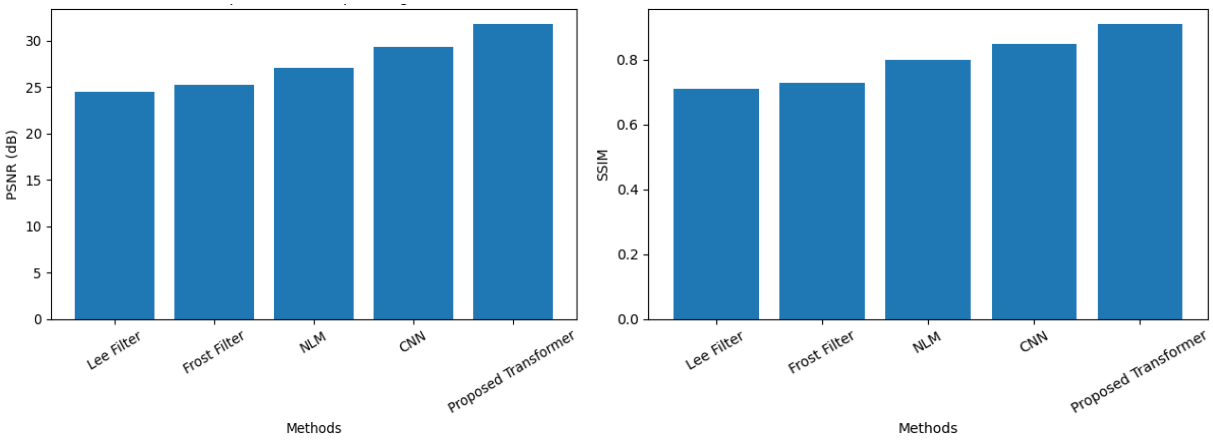
Table 1 presents a quantitative comparison of different despeckling approaches using PSNR and SSIM metrics. The Table shows that the traditional filtering approaches show limited performance because of tendency to smoot structural information. The NLM method use to enhance noise reduction with the help of utilizing image redundancy.

Deep learning-based CNN models provide better performance by learning complex feature representations. However, the proposed Transformer-Based Despeckling Network achieves the highest PSNR and SSIM values among all evaluated methods. The PSNR improvement shows more perfect noida reduction whereas the high SSIM value represent the better edge preservation. This performance gain can be attributed to the self-attention mechanism of the transformer architecture, which effectively captures global contextual relationships in SAR imagery.

To further evaluate the robustness of the proposed model, experiments were conducted with different speckle noise levels. Table 2 represented PSNR vs Noise Variance graph which illustrates that the proposed model consistently outperforms competing methods across different noise levels.

Table 2. Compares this work with the related work or previous research by other researchers

Noise Variance	Lee	Frost	NLM	CNN	Proposed
0.1	26.1	26.8	28.7	30.4	**32.2**
0.2	25.4	26.0	27.9	29.8	**31.5**
0.3	24.5	25.2	27.1	29.3	**31.0**
Noise Variance	**Lee**	**Frost**	**NLM**	**CNN**	**Proposed**



(a) (b)

Figure 2. (a) Comparison of PSNR value among different traditional filters, CNN-based methods, and the proposed TBDM model (b) Comparison between structural preservation capability of different despeckling methods.

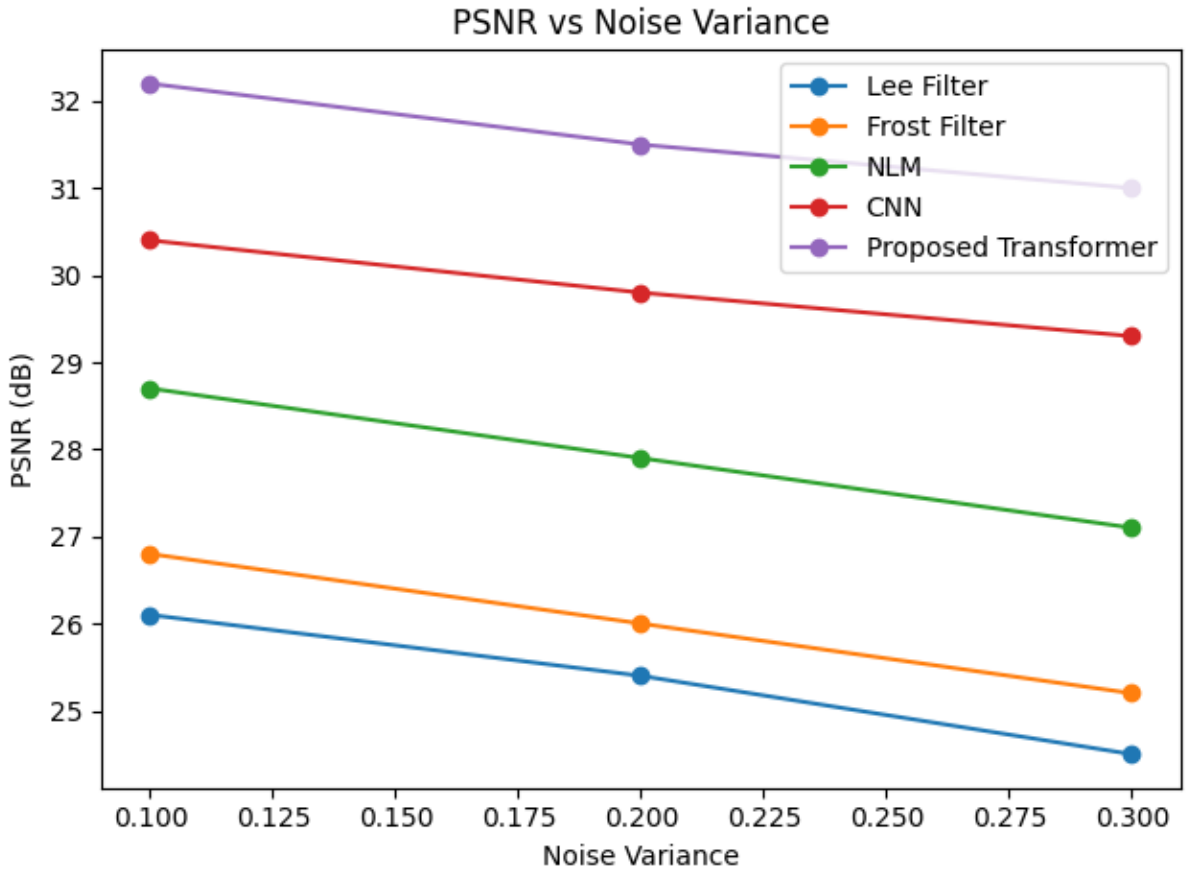


Figure 3. The performance of PSNR under various speckle noise variances demonstrating the proposed model.

The graphical analysis further confirms the superiority of the proposed Transformer-Based Despeckling Network. As illustrated in Figure 2(a) , the proposed method achieves the highest PSNR value compared

with traditional and CNN-based methods, indicating improved noise suppression capability. Similarly, Figure 2(b) shows that the proposed model achieves the highest SSIM value, demonstrating better preservation of structural and textural information in SAR images.

Figure 3 illustrates the PSNR performance under varying noise levels. The proposed model consistently outperforms competing methods across all noise variances, confirming its robustness in handling different speckle noise intensities. This improvement can be attributed to the global contextual modeling ability of the transformer architecture combined with multi-scale feature fusion.

Conclusions

This research proposed and demonstrated a transfer-based framework for SAR despeckling with goal of speckle noise reduction and structural and textural data preservation. The available traditional methods typically reduce speckle noise but at the cost of losing important image information. To remove this drawback, the proposed model combines convolution feature extraction with transformer based attention methodology in capturing local image patterns as well as global contextual relationships.

Experimental result achieve by this architecture represents that the proposed architecture performs better than other available traditional filtering methods and CNN based methods. The proposed architecture evaluated on the basis of PSNR and SSMI parameters. The higher PSNR and SSIM value highlight the potential of transformer-based architecture in improving image restoration tasks.

In summary, the proposed architecture approach provides a very effective and reliable solution for SAR image despeckling. The proposed model can enhance the quality of remote sensing imagery used in different applications.

Future research may focus on optimizing the computational efficiency of the model, exploring lightweight transformer variants, and applying the framework to other types of remote sensing imagery and real-time processing systems.

References

1. J. S. Lee, "Digital image enhancement and noise filtering by use of local statistics", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, no. 2, pp. 165–168, 1980. https://doi.org/10.1109/TPAMI.1980.4766994
2. D. T. Kuan, A. A. Sawchuk, T. C. Strand and P. Chavel, "Adaptive noise smoothing filter for images with signal-dependent noise", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 7, no. 2, pp. 165–177, 1985. https://doi.org/10.1109/TPAMI.1985.4767641
3. V. S. Frost, J. A. Stiles, K. S. Shanmugan and J. C. Holtzman, "A model for radar images and its application to adaptive digital filtering of multiplicative noise", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 4, no. 2, pp. 157–166, 1982. https://doi.org/10.1109/TPAMI.1982.4767223
4. A. Buades, B. Coll and J. M. Morel, "A non-local algorithm for image denoising", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 60–65, 2005. https://doi.org/10.1109/CVPR.2005.38

5. K. Zhang, W. Zuo, Y. Chen, D. Meng and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising", IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142–3155, 2017. https://doi.org/10.1109/TIP.2017.2662206
6. G. Chierchia, D. Cozzolino, G. Poggi and L. Verdoliva, "SAR image despeckling through convolutional neural networks", IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 11, pp. 1936–1940, 2017. https://doi.org/10.1109/LGRS.2017.2739351
7. A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale", International Conference on Learning Representations (ICLR), 2021. https://doi.org/10.48550/arXiv.2010.11929
8. F. Argenti, L. Alparone and G. Benelli, "Speckle removal from SAR images in the undecimated wavelet domain", IEEE Transactions on Geoscience and Remote Sensing, vol. 40, no. 11, pp. 2363–2374, 2002. https://doi.org/10.1109/TGRS.2002.804721
9. G. Aubert and J. F. Aujol, "A variational approach to removing multiplicative noise", SIAM Journal on Applied Mathematics, vol. 68, no. 4, pp. 925–946, 2008. https://doi.org/10.1137/060671814
10. K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering", IEEE Transactions on Image Processing, vol. 16, no. 8, pp. 2080–2095, 2007. https://doi.org/10.1109/TIP.2007.901238
11. Y. Chen, W. Yu and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5261–5269, 2015. https://doi.org/10.1109/CVPR.2015.7299141
12. X. Mao, C. Shen and Y. Yang, "Image restoration using very deep convolutional encoder–decoder networks with symmetric skip connections", Advances in Neural Information Processing Systems, vol. 29, pp. 2802–2810, 2016.
13. Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong and Y. Fu, "Image super-resolution using very deep residual channel attention networks", Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301, 2018. https://doi.org/10.1007/978-3-030-01234-2_18
14. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows", Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 10012–10022, 2021. https://doi.org/10.1109/ICCV48922.2021.00986
15. S. Zamir, A. Arora, S. Khan, M. Hayat, F. Khan and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5728–5739, 2022. https://doi.org/10.1109/CVPR52688.2022.00567