On the impact of despeckling for supervised SAR super-resolution

Max Muzeau[⋆,†], Chengfang Ren[⋆], Jeremy Fix[§], Frederic Brigui[⋆,†], and Jean Philippe Ovarlez[⋆]
[⋆]SONDRA, CentraleSupélec, Université Paris-Saclay, 91192 Gif-sur-Yvette, France
[†]DEMR, ONERA, Université Paris-Saclay, 91120 Palaiseau, France
[§]LORIA, CNRS, CentraleSupélec, Université Paris-Saclay, F-57000 Metz, France

Abstract

Enhancement of SAR resolution is essential for various applications in earth observation. Since SAR images are highly corrupted by speckle noise, we propose to help super-resolution neural network learning with a despeckling preprocessing step. Unlike optical images, low-resolution SAR images are extracted from the sub-apertures of the original SAR image. To evaluate the impact of the despeckling, SwinIR, SRCNN, and ESPCN neural networks are trained in three ways: Noisy2Noisy, Noisy2Denoised, and Denoised2Denoised. The ONERA SAR database experiments show the despeckling improvement gap and the slight enhancement of SwinIR over SRCNN and ESPCN according to the visual reconstruction and to L_1, L_2, PSNR, and SSIM metrics.

1 Introduction

Synthetic Aperture Radar (SAR) is a cutting-edge remote sensing system that plays a significant role in earth observation and environmental monitoring. High-resolution SAR imaging provides finer details in images, allowing for detecting and identifying smaller objects and features on the ground. However, the resolution of Side-Looking Airborne Radars (SLAR) is theoretically limited by the radar bandwidth in slant range and the antenna footprint width in azimuth [1] and practically degraded by targets sidelobes [2].

To overcome these issues, the Spatial Variant Apodization (SVA) algorithm and its variant aimed to reduce or cancel sidelobes have been proposed in [2, 3, 4, 5]. These unsupervised algorithms based on the impulse response model are computationally fast and efficiently reduce the sidelobes. However, the main lobe width remains unchanged. The latter issue can be solved using supervised learning approaches based on neural networks by leveraging prior information from a database of paired High Resolution (HR) and Low Resolution (LR) SAR images [6, 7, 8]. To sharp mainlobes, neural networks have to learn to restore HR SAR images from downsampled LR SAR inputs similarly to the setup in the challenge on optical image super-resolution [9].

However, SAR image formation is specific to radar waves which differ from optic. Particularly, the SAR range and azimuth axis are not permutable, and classical augmentations (e.g. rotation and flip) are unrealistic. Additionally, the speckle noise highly corrupts SAR images making the despeckling process decisive for target and anomaly detection [10]. Fortunately, SAR despeckling methods such as [11, 12] are able to reduce the speckle noise with few Single Look Complex (SLC) SAR images.

In this paper, we propose to evaluate the impact of the despeckling process for SAR super-resolution using well-known neural networks, namely Super-Resolution Convolutional Neural Network (SRCNN) [13], Efficient Sub-Pixel Convolutional Neural network (ESPCN) [14] and Shifted WINdows transformer Image Restoration neural network (SwinIR) [15]. Unlike the previous study, the LR SAR database is extracted from the subapertures of the original (HR) SAR image using subband and sublook processing [16, 17]. This process is enabled using the complex-valued SLC SAR data to filter range and azimuth spectrum. Additionally, we perform the MERLIN despeckling algorithm [11] to build a denoised SAR database. We aim to evaluate whether the despeckling process can help the above-mentioned neural networks enhance SAR super-resolution. A quantitative evaluation is performed using L_1, L_2 losses, PSNR and SSIM metrics.

The rest of the paper is organized as follows. Section 2 explains the neural network architectures and training setups used for SAR super-resolution. Section 3 analyzes the network performance over real SAR data acquired from ONERA.

2 Methodology

2.1 Neural networks architecture

To better understand the approach, there is a brief explanation of the three super-resolution networks used. Then, the adaptation for the SAR image is explained.

Super-resolution convolutional neural network

The SRCNN [13] is one of the earliest deep-learning networks used for such a task. It consists of an upsampling with a bicubic interpolation, followed by three convolutions. This network is very shallow, which makes the super-resolution capacity limited. Yet it achieved state-of-the-art restoration quality when it was published.

The SRCNN architecture for SAR experiments is the fol-

lowing. LR SAR input is first upsampled using nearest-neighbor interpolation followed by 3 convolutional layers. The feature extraction layer is performed by a 9×9 kernel and 64 output channels, followed by a ReLu activation function. The non-linear-mapping layer is a 5×5 kernel and 32 output channels, followed by a ReLu activation function. The output reconstruction layer is a 5×5 kernel and 1 output channel. All the convolutional layers are zero-padded to keep the output image size constant.

Efficient sub-pixel convolutional neural network

The main drawback of SRCNN is that the convolution is done in a high-resolution space, which can be time-consuming. The ESPCN [14] proved that the convolutions can be done in the low-resolution space. This led to run time improvement and also to better performances. The first layers are standard convolution layers and the last one is a sub-pixel convolution. This is like a normal convolution but the number of output channels will be r^2C, r being the resolution improvement coefficient and C the number of channels. In this way, the upsampling method is defined by the network, in opposition to SRCNN. To improve the resolution of an image of shape (C, H, W) to an image of shape (C, rH, rW) the last convolution layer will have the shape (r^2C, H, W), the elements of this tensor will then be unfolded to obtain the desired super-resolution image.
Since we experiment only on SAR data single channel $C = 1$, the ESPCN architecture is the following. There are 5 convolutional layers. The first two convolution layers have 32 output channels, the next two have 64 output channels and the last one has $r^2 = 4$ output channels which leads to $r = 2$ the upsampling factor for each spatial axis. The convolutional output of size $(4, H, W)$ is then unfolded according to the process defined in [14] to obtain the output HR image of size $(1, 2H, 2W)$. All the convolutions are done with 3×3 kernels with zero padding, followed by the ReLu activation function except for the last convolution.

Swin transformer image restoration neural network

Recently, most of the state-of-the-art super-resolution methods are based on transformers [18]. As we can see in [9], the Swin transformer [15] is used. It introduces a hierarchical architecture that utilizes a shifted windowing scheme for the computation of representations. The goal is to adapt the architecture of transformers (initially made for NLP tasks) for vision applications. The local attention mechanism allows the computation of high-resolution images at low-level space (each token is extracted from a 4×4 patch). This architecture has been adapted to image restoration in [19] with the SwinIR network. The architecture is decomposed in 3 steps :

- **Shallow Feature Extraction**: A standard 3×3 convolution to get high-level information
- **Deep feature extraction**: A combination of multiple blocks, each combining multiple Swin Transformer layers followed by a convolution. There is a residual connection between each block.

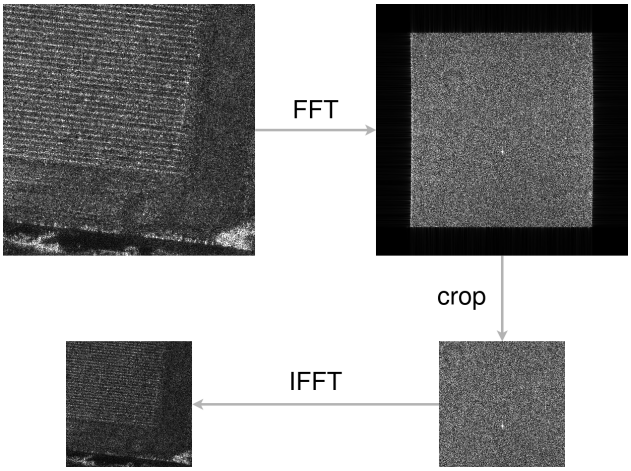

Figure 1 Process to subsample a SAR image. It consists of a crop in the Fourier domain.

- **High quality image reconstruction** : Shallow feature and deep feature are aggregated. They are supposed to assess respectively for the input image's low-frequency and high-frequency information. Then, the sub-pixel convolution defined in the previous model is used to obtain the high-resolution output image.

The SwinIR architecture for SAR experiments is the following. All the convolutions are done with 3×3 kernels. The shallow feature extraction convolution layer has 60 output channels. Then, the deep feature extraction is composed of 3 Residual Swin Transformer Block, each composed of 4 Swin Transformer Layer, which is a window-based multi-head self-attention (W-MSA) and a multilayer perceptron (MLP), each preceded by a layer normalization. The W-MSA has a window size of 8 and is composed of 4 heads. The MLP got a depth of 2 with a hidden feature ratio of 2. The convolution at the end of the feature extraction has 60 output channels. For the upsample, there is also a convolution with 60 output channels followed by an ESPCN type upsample for the features, i.e. a convolution of $60 \times r^2 = 240$ output channels followed by a pixel shuffle of ratio 2. After the upsample, there is a last convolution with 1 output channel for the final reconstruction.
The architecture of the networks can be found in detail in the repository `https://github.com/muzmax/SAR_super_resolution.git`

2.2 SAR images preprocessing

The architectures described above are made for optics images. Because SAR images are different in many ways, they have to be adapted to work efficiently. Notably, we propose in the following to generate LR SAR images from the sub-apertures of HR SAR data and then to despeckle them to obtain a better Clutter to Noise Ratio (CNR).

LR-HR SAR images generation

Contrary to optical images, there is a possibility to extract sub-apertures of SAR images with lower resolution as explained in [20, 16, 17]. SLC SAR data are complex-valued images allowing the computation of range and

azimuth spectrum by applying 2D Fourier transform or wavelet transform. The original SAR spectrum should be first shifted to the center (notably, the Doppler centroid is moved to zero azimuth frequency) as shown in Fig. 1. Then we filter the spectral image in the center to obtain the desired resolution without padding. The LR image is then obtained by 2D inverse Fourier transform.

SAR despeckling

An important aspect of SAR images is the strong noise called speckle [21]. This phenomenon occurs due to the coherent summation of many backscatters in a single pixel, causing constructive or destructive interference. This noise complicates the task of super-resolution because when the image is upsampled by a network, it tends to remove parts of the speckle and replace it with artifacts (see Fig. 4). One solution for this problem is proposed in [22] when a network has to minimize the distance between its output z and a set of observation $(y_1, ..., y_n)$ with a L_2 loss it will make an average of all the observations to obtain the best result such that

$$\underset{\mathbf{z}}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{y}}\left[\, \|\mathbf{z}-\mathbf{y}\|_2^2 \,\right] = \mathbb{E}_{\mathbf{y}}[\mathbf{y}] \tag{1}$$

Even though super-resolution neural networks will despeckle a bit following the noise2noise principle [22], the result won't be good enough. This is why we proposed in this article to incorporate speckle-free images in the pipeline. The advantage of this approach is that the network won't have to learn how to remove the speckle of the image (or to keep it identical depending on the objective) in an unsupervised manner. To remove the speckle from our images, the network Merlin [11] is used. It is a network that uses an image's real and imaginary parts as its input and label. It is easy to train because no pair of images or labels are needed, only SLC images. The model used for this network is a standard U-net. The training phase is done from scratch with the Adam optimizer on ONERA SETHI dataset (see section 3.1) divided into 5264 patches of size 256×256. A batch size of 30 is used with an initial learning rate of 10^{-3} that decreases by a factor of 10 after the 5th and 20th epochs for 30 epochs.

2.3 Training setup

The super-resolution network is trained in three different manners to observe the improvement brought by despeckling. The SAR input image is the LR SAR image downsampled from the high resolution SAR which is the objective/label image. The generation of LR SAR is mentioned in section 2.2. We decide to despeckle or not the LR-HR SAR images according to the following setups:

- **Noisy2Noisy**: Both the input and the label are images with speckle. This baseline will help to see how other methods are improving the performances.
- **Noisy2Denoised**: The input and the label are respectively images with speckle and speckle-free images denoised by MERLIN [11]. This method will check if a network can learn how to make a super-resolution version of a SAR image and remove its speckle at the same time.
- **Denoised2Denoised**: Both the input and the label are speckle-free images denoised by MERLIN [11]. This method will show how far the performance of the super-resolution network can be if we have already removed the speckle of the input image.

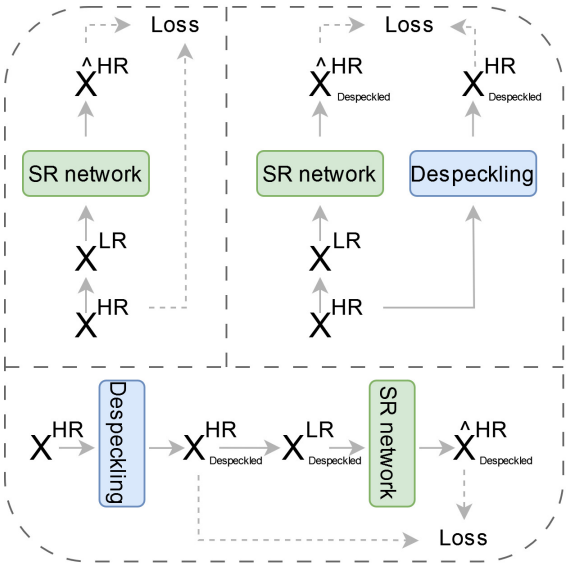

Figure 2 Different training for super-resolution. Upper-left: standard method. Upper-right: the label is the speckle-free HR image. Bottom: Both the LR input and the HR label are speckle-free images. SR, HR, and LR stand respectively for Super Resolution, High Resolution, and Low Resolution.

A summary of this training method is displayed in Fig. 2 After completing the training process for the despeckling network, the three super-resolution architectures are trained, each for the three different methods which makes a total of 9 networks. All networks use the Huber loss [23] in pair with the Adam optimizer. The data described above is sliced in patches of size 512×512 for a total of 1316 patches that are separated into training and validation parts (80% and 20% respectively) and grouped in a batch size of 2 during training. The initial learning rate is 10^{-3}, which decreases by a factor of 10 if the loss does not improve more than 10^{-4} in 30 epochs. The training is done for 100 epochs.

3 Experiments

All the training methods described above will be tested in the following experiences in the case of a $2\times$ super-resolution. To assess the networks' performances, metrics L_1, L_2, SSIM, and PSNR are used on a validation dataset. $L1$ and $L2$ metrics are respectively the mean absolute error and the mean square error (MSE) between original and reconstructed HR images. PSNR is defined by $10 \log_{10} \frac{L^2}{MSE}$

with L the dynamic range. And SSIM metric is defined as:

$$\mathsf{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

where μ_z and σ_z^2 are respectively the empirical mean and variance of image patch z of size 11×11, σ_{xy} the empirical covariance between patches x and y, $c_1 = 0.01^2$ and $c_2 = 0.03^2$. The bicubic interpolation is added to have a baseline of super-resolution methods.

3.1 Dataset

To make the experiments, airborne SAR data from the ONERA instrument called SETHI [24] is used. It consists of X-band high-resolution data (about $25cm$ in azimuth and range) in Hh polarization. The image was acquired over the Nimes-Garons airport area. Multiple types of scatterers are present in the scene such as urban areas, agricultural fields, or small woods. Finally, we have one 37400×9230 pixels image. The full SLC SAR image is shown in Fig. 3.

3.2 Qualitative results

When the image is subsampled, the high frequency content is removed. The speckle contains much information in this frequency range, because of this and its randomness, it is complicated to retrieve its original aspect. As we can see in Fig. 4 first row, the results of the $Noisy2Noisy$ experiments are not convincing. Some parts of the speckle in the crops are replaced with a local averaging. With the SwinIR, the result is more visually pleasing but still, there is a clear difference compared to the high-resolution original SAR image.

The experiment $Noisy2Denoised$ is the most complicated because the network has to learn to do a super-resolution and a despeckling simultaneously. In addition, the image of Fig. 4 makes the task difficult because there is high-frequency information in the crop lanes. The convolution networks $SRCNN$ and $ESPCN$ cannot retrieve the denoised SAR image (Fig. 4 bottom row). The areas that were not well restored in the last experiment are the same that are not well restored in this case. Some parts of the lines are replaced with a global average. It may be because the receptive field of the convolution is too small or because the networks are too shallow to understand the complexity of the tasks. In comparison, the SwinIR network achieved fairly good results. There are some restoration difficulties in the same areas where the convolution networks struggled, but it is largely less significant.

The final experiment $Denoised2Denoised$ is the easiest one. All networks give visually good results, and even the bicubic interpolation can achieve good performance. It highlights the fact that super-resolution for SAR images is hard mainly because of the noise that corrupts the image.

3.3 Quantitative results

As we can see in Table 1, the transformer architecture gives the best results for all cases and this was to be expected because SwinIR is a state-of-the-art network for image restoration, and it also has a lot more parameters than the convolution networks. For the same reasons that are explained in the last subsection, the $Noisy2Noisy$ quantitative evaluation is not convincing. The difference between the simplest convolution network $SRCNN$ and a huge transformer $SwinIR$ is small. It shows that the task in itself, which is estimating a high-resolution speckle, is not feasible with this approach. Because $Noisy2Denoised$ is the most complicated approach, it is useful to have a powerful network. This is why the difference between $SwinIR$ and other methods is essential, it goes along with the qualitative evaluation. Because the $Denoised2Denoised$ approach is fairly simple, all networks give good results. It shows that for this specific task, a huge transformer such as $SwinIR$ is too powerful for the task.

Noisy2Noisy

Loss	Bicubic	SRCNN	ESPCN	SwinIR
L_1	0.6481	0.4952	0.4899	**0.4810**
L_2	0.7441	0.4528	0.4446	**0.4322**
SSIM	0.1929	0.2934	0.3028	**0.3224**
PSNR	47.314	49.476	49.555	**49.677**

Noisy2Denoised

Loss	Bicubic	SRCNN	ESPCN	SwinIR
L_1	-	0.2705	0.2408	**0.1687**
L_2	-	0.1689	0.1428	**0.0790**
SSIM	-	0.2897	0.3475	**0.5465**
PSNR	-	55.607	56.622	**59.591**

Denoised2Denoised

Loss	Bicubic	SRCNN	ESPCN	SwinIR
L_1	0.0509	0.0119	0.0124	**0.0099**
L_2	0.0093	0.0005	0.0004	**0.0003**
SSIM	0.8830	0.9795	0.9815	**0.9844**
PSNR	69.061	81.751	81.893	**83.301**

Table 1 Quantitative evaluation of the training methods.

4 Conclusion

In this paper, we evaluated the impact of the despeckling process for SAR super-resolution tasks. We proposed to generate the paired monochannel LR-HR SAR images using the SAR sub-apertures approach. Additionally, this database is despeckled to obtain better CNR that eases the SAR super-resolution learning by neural networks. Three neural network architectures, SRCNN, ESPCNN, and SwinIR, have been proposed to learn SAR super-resolution. Quantitative metrics are evaluated on the ONERA database, highlighting the importance of the despeckling pre-processing. We also note that SwinIR has a reasonable advantage over SRCNN and ESPCN. Further

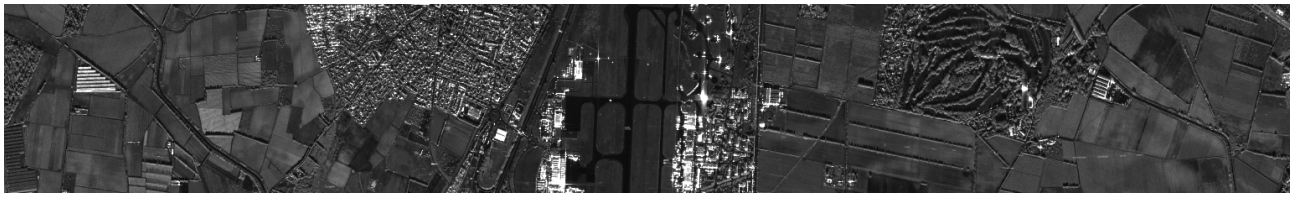

Figure 3 SETHI SLC SAR Image. HH polarization. X-band. Range direction is on the vertical axis and the azimuth direction is on the horizontal axis.

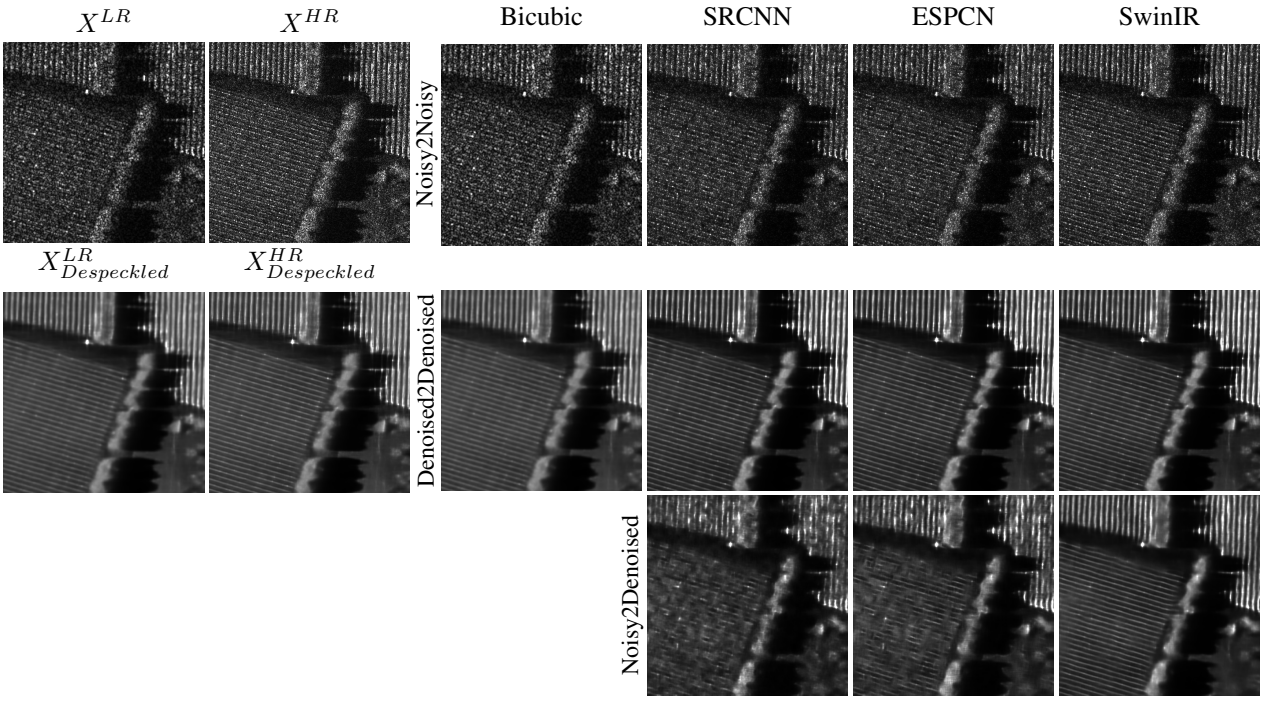

Figure 4 Qualitative evaluation of the different super-resolution methods and networks.

experiments can be conducted for multichannel SAR images such as polarimetric, interferometric, or tomographic SAR.

5 Literature

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

[2] H. C. Stankwitz, R. J. Dallaire, and J. R. Fienup, "Nonlinear apodization for sidelobe control in SAR imagery," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 31, no. 1, pp. 267–279, 1995.

[3] B. H. Smith, "Generalization of spatially variant apodization to noninteger Nyquist sampling rates," *IEEE Transactions on Image Processing*, vol. 9, no. 6, pp. 1088–1093, 2000.

[4] T. Xiong, S. Wang, B. Hou, Y. Wang, and H. Liu, "A resample-based SVA algorithm for sidelobe reduction of SAR/ISAR imagery with noninteger Nyquist sampling rate," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 2, pp. 1016–1028, 2015.

[5] M. Tay, "Unsupervised deep learning parameter estimation for high fidelity synthetic aperture radar super resolution," in *2022 23rd International Radar Symposium (IRS)*, 2022, pp. 241–246.

[6] C. Zheng, X. Jiang, Y. Zhang, X. Liu, B. Yuan, and Z. Li, "Self-Normalizing Generative Adversarial Network for Super-Resolution Reconstruction of SAR Images," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, July 2019, pp. 1911–1914, ISSN: 2153-7003.

[7] H. Shen, L. Lin, J. Li, Q. Yuan, and L. Zhao, "A residual convolutional neural network for polarimetric SAR image super-resolution," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 90–108, Mar. 2020.

[8] L. Lin, J. Li, Q. Yuan, and H. Shen, "Polarimetric SAR Image Super-Resolution VIA Deep Convolutional Neural Network," in *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, July 2019, pp. 3205–3208, ISSN: 2153-7003.

[9] Y. Zhang et al, "NTIRE 2023 Challenge on Image Super-Resolution (×4): Methods and Results," in *2023 IEEE/CVF Conference on Computer Vision and*

Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, June 2023, pp. 1865–1884, IEEE.

[10] M. Muzeau, C. Ren, S. Angelliaume, M. Datcu, and J.-P. Ovarlez, "Self-supervised learning based anomaly detection in synthetic aperture radar imaging," *IEEE Open Journal of Signal Processing*, vol. 3, pp. 440–449, 2022.

[11] E. Dalsasso, L. Denis, and F. Tupin, "As if by magic: self-supervised training of deep despeckling networks with MERLIN," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

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

[13] C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, Feb. 2016, Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[14] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016, pp. 1874–1883, IEEE.

[15] 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," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002, IEEE.

[16] J.-C. Souyris, C. Henry, and F. Adragna, "On the use of complex sar image spectral analysis for target detection: assessment of polarimetry," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2725–2734, 2003.

[17] C. Brekke, S. N. Anfinsen, and Y. Larsen, "Subband extraction strategies in ship detection with the subaperture cross-correlation magnitude," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 786–790, 2013.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[19] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, Oct. 2021, pp. 1833–1844, IEEE.

[20] J.-P. Ovarlez, L. Vignaud, J.-C. Castelli, M. Tria, and M. Benidir, "Analysis of SAR images by multidimensional wavelet transform," *IEE Proceedings, Radar, Sonar and Propagation*, vol. 150, no. 4, pp. 234–241, 2003.

[21] J. W. Goodman, "Some fundamental properties of speckle," *JOSA*, vol. 66, no. 11, pp. 1145–1150, 1976.

[22] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2noise: Learning image restoration without clean data," *arXiv preprint arXiv:1803.04189*, 2018.

[23] K. Gokcesu and H. Gokcesu, "Generalized Huber loss for robust learning and its efficient minimization for a robust statistics," *arXiv preprint arXiv:2108.12627*, 2021.

[24] R. Baqué, P. Dreuillet, and H. Oriot, "Sethi: Review of 10 years of development and experimentation of the remote sensing platform," in *International Radar Conference (RADAR)*, 2019, pp. 1–5.