MERLIN-Seg: self-supervised despeckling for label-efficient semantic segmentation

Emanuele Dalsasso, Clément Rambour, Nicolas Trouvé, Nicolas Thome

Computer Vision and Image Understanding

journal homepage: www.elsevier.com

MERLIN-Seg: self-supervised despeckling for label-efficient semantic segmentation

Emanuele Dalsasso[a,**], Clément Rambour[a], Nicolas Trouvé[b], Nicolas Thome[a,c]

[a] *Conservatoire National des Arts et Métiers, 75003 Paris, France*
[b] *The French Aerospace Lab, ONERA, 91120 Palaiseau, France*
[c] *Sorbonne Université, CNRS, ISIR, F-75005 Paris, France*

ABSTRACT

This is the pre-acceptance version. To read the final version published in 2024 in *Computer Vision and Image Understanding*, please go to: https://doi.org/10.1016/j.cviu.2024.103940
Remote sensing satellites acquire a continuous stream of data on a daily basis. As most of those data are unlabeled, the development of algorithms requiring weak supervision is of paramount importance. In this paper, we show that the need for annotation for Synthetic Aperture Radar data can be reduced by coupling a despeckling task (self-supervised) and a segmentation task (supervised). The proposed self-supervised learning framework, called MERLIN-Seg, has been trained for building footprint extraction and achieves favorable performances even with 1% of annotated data. We show that conditioning the network on despeckling without labels is beneficial for supervised segmentation. Our experiments demonstrate that the joint training of the two tasks achieves better performances than a vanilla segmentation network in terms of IoU, F1 score, and accuracy on both simulated and real SAR images.

1. Introduction

Earth observation satellites carry onboard different kind of remote sensors collecting information characterizing the Earth. Among these, Synthetic Aperture Radar (SAR) is an active system with imaging capabilities. As an active sensor, it can collect information at any time of the day and in (almost) all weather conditions, providing continuous and global coverage of the Earth's surface. This powerful feature of SAR allows access to cloud-covered areas, such as the tropics and subtropics, subject to long wet seasons and frequent precipitation.

Deep learning algorithms play an essential role in the analysis of remote sensing data (Zhu et al., 2017). Once they are tuned and deployed, they can process data in a quick manner to extract useful information for Earth monitoring. To train such algorithms, one often needs a large set of labeled data. In remote sensing, the great abundance of images is guaranteed by the presence of numerous sensors embarked on satellites in orbit around the Earth and the increased adoption of open data policies. However, these images are generally not annotated, making data labeling one of the big open challenges in remote sensing (Wang et al., 2022b).

In the computer vision community, alternatives to vanilla supervised algorithms requiring labeled data have been proposed: so-called self-supervised learning (SSL) approaches have been shown to learn powerful representations for many tasks (Xian et al., 2017; Chen et al., 2020b; Lehtinen et al., 2018; Finn et al., 2017; Le-Khac et al., 2020). SSL has recently sparked

**Corresponding author:
e-mail: work.emanuele.dalsasso@outlook.com (Emanuele Dalsasso)

great attention in remote sensing (Wang et al., 2022b), but still few works were proposed to deal with SAR data. In an SSL framework, a neural network is pre-trained on a self-supervised pretext task and then fine-tuned on the actual downstream task. This relieves the network of the burden of seeing many labeled data, reducing the need for annotation (Zheng et al., 2021). Contrastive learning approaches comprise another family of SSL methods where a network is trained to force a close latent representation between augmented views of the same image (and possibly push away representations from negative samples). To this aim, one may resort to techniques such as image rotation, mirroring, and image colorization. (Baranchuk et al., 2022; Brempong et al., 2022) have demonstrated that denoising pre-training learns useful semantic representations. As an alternative to this sequential scheme, a neural network can be trained simultaneously on both downstream task and pretext task, such as image denoising. In DenoiSeg (Buchholz et al., 2021), while learning to reduce noise from both unlabeled and labeled images, the network efficiently co-learns to segment even on a few labeled samples.

The application of SSL to SAR data is not straightforward (Wang et al., 2022b). Augmenting SAR data is indeed very sensitive: unlike optical images, SAR images are complex-valued, their pixels represent measurements with physical meaning, and the appearance of the scene is intimately related to the sensor position and acquisition mode. Moreover, SAR images are affected by speckle, a multiplicative perturbation with specific statistics.

In this paper, we present MERLIN-Seg, a general framework to address the shortage of annotations in SAR images, relying on self-supervised despeckling (Dalsasso et al., 2021a). As in (Buchholz et al., 2021), we simultaneously train our model on the self-supervised despeckling and supervised segmentation tasks (Fig.1). We empirically demonstrate that our SSL framework is beneficial for the segmentation of SAR images, especially when the availability of labels is low. We test our approach on the specific task of building footprint segmentation on simulated and real data. The main contributions of our work are the following:

- To the best of our knowledge, we are the first to propose to use despeckling as an SSL technique to extract semantic features from SAR data for label-efficient semantic segmentation. In particular, we propose a generalization of the MERLIN framework (Dalsasso et al., 2021a) for the simultaneous learning of despeckling and segmentation.
- The proposed MERLIN-Seg approach extracts semantically meaningful features from SAR images without the need of designing SAR-specific augmentation. We argue that SAR images can be seen as inherently augmented due to the presence of speckle. Thus, we exploit the nature of SAR data to devise a physically meaningful self-supervised task.
- We show that the simultaneous learning of SAR despeckling and segmentation is more effective than SSL pre-training on SAR despeckling, although both strategies are relevant and outperform a fully supervised learning on the footprint segmentation task when few labels are available.
- We validate MERLIN-Seg on both EMPRISE simulated SAR images and real TerraSAR-X data, highlighting the versatility of the proposed approach to different acquisition modalities and resolutions. Moreover, our approach shows to perform well both in dense and sparse urban environments.

2. Related Work

SAR despeckling SAR is a coherent imaging system. As such, the complex SAR signal z is the result of the coherent summation of many elementary echoes coming from the same resolution cell. For rough surfaces at the wavelength scale and in absence of predominant scatterers, the measured intensity $I = |z^2|$ presents strong fluctuations that severely limits the exploitation of SAR images. Such fluctuations take the name of speckle phenomenon. The SAR intensity I, the underlying reflectivity R and the speckle S are linked by a multiplicative relation:

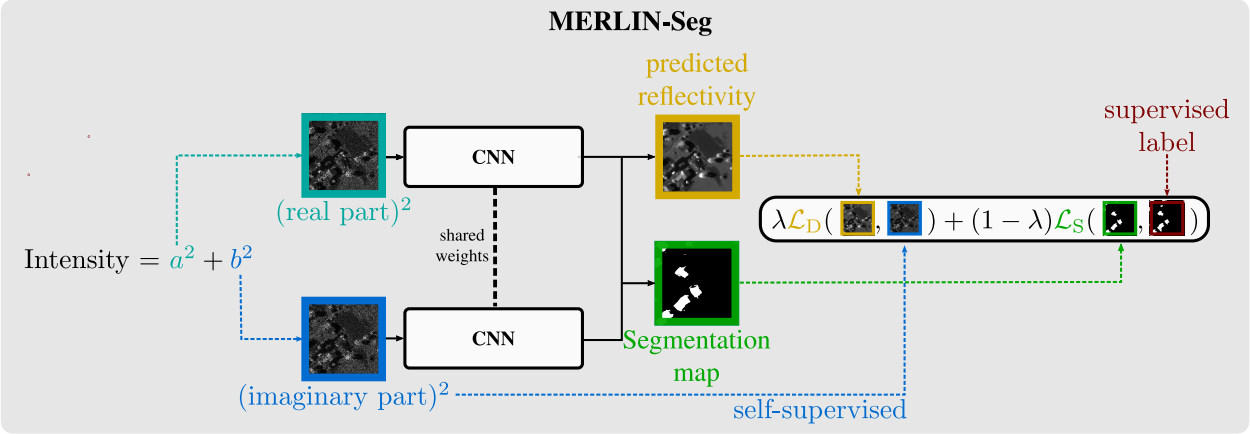


Fig. 1: The training strategy of the proposed MERLIN-Seg approach. A neural network is jointly trained for despeckling (self-supervised) and segmentation (supervised) into a unique weekly-supervised framework.

$I = R \times S$ (Goodman, 2007). The aim of despeckling techniques is to suppress the speckle and restore the underlying reflectivity.

Previously unseen performances has been obtained by SAR despeckling approaches relying on a self-supervised training strategy (Fracastoro et al., 2021; Dalsasso et al., 2022). Indeed, the difficulty in SAR imaging to produce the clean groundtruth signal R limits the development of supervised methods. SSL for SAR despeckling stems from the advances in self-supervised image denoising (Lehtinen et al., 2018; Krull et al., 2019; Buchholz et al., 2021). Given two independent noisy images y_i and y_j representing the same underlying clean signal x, a neural network $f_\theta(.)$ can be trained to maximize the agreement between network's output $f_\theta(y_i)$ and an independent noisy image y_j. If the noise likelihood is known, the network loss can be defined as $\mathcal{L} = -\log \mathrm{p}(y_j|f_\theta(y_i))$. Because y_i and y_j are i.i.d. samples, the network cannot predict the random component of y_j starting from y_i. Instead, it will predict the deterministic common component, i.e. the clean signal x. This way, a neural network can be trained for image denoising by only seeing noisy examples.

Depending on the way of splitting the dataset into two independent sets, one can class self-supervised despeckling approaches into three categories:

- Multi-image self-supervised approaches exploit images acquired at different dates (sufficiently spaced-apart so that the speckle is temporally decorrelated). Thus, one obtains several samples representing the same underlying scene, provided that the changes can be compensated (Dalsasso et al., 2021b).

- In single-image self-supervised approaches, one can rely on specific architectures excluding the central pixel from the network receptive field (so-called blind-spot networks (Laine et al., 2019; Lee and Jeong, 2022)) and keep it apart to supervise network training. The specificity required by such a receptive field limits the choice of architecture and might reduce the network's performance. Moreover, to verify the independence of the hidden pixel with the neighboring ones, it requires the speckle to be spatially independent: such a hypothesis is however not verified in practice, as data providers often deliver spatially-correlated SAR products, and whitening the speckle requires a tedious and touchy pre-processing, especially if sensors parameters are unknown (Lapini et al., 2013).

- Most recently, MERLIN approach (Dalsasso et al., 2021a) has demonstrated that any architecture can be trained on single SAR images by exploiting the complex nature of Single-Look Complex (SLC) SAR data. MERLIN exploits the phase information to split the intensity image into two independent sub-images, namely the real and imaginary parts, creating ideal conditions for self-supervised learning based on noise2noise (Lehtinen et al., 2018).

Unlike the first two families of techniques, MERLIN does not make any hypothesis on temporal coherence (as it is trained on single dates), nor on speckle spatial correlation (as the network implicitly learns it). For these reasons, MERLIN would fit like a glove in an SSL framework for SAR imaging.

Semantic Segmentation and building footprint Semantic segmentation has been a popular topic in computer vision for many years. In the last decade, Convolutional Neural Networks (CNNs) have been shown to be highly effective in learning discriminative features and achieving high accuracy in various image segmentation tasks. U-shaped CNN architectures such as Unet (Ronneberger et al., 2015), UpperNet (Xiao et al., 2018) or DeepLab (Chen et al., 2018) and their variants set new standards in computer vision. Recently, transformer architectures tend to show better performances in semantic segmentation at the cost of strong parameters overhead (Liu et al., 2021; Xie et al., 2021; Themyr et al., 2023).

In the domain of remote sensing, land cover is a challenging task due to the strong contrast dynamics, the lack and/or quality of annotation, and the large size of remote sensing images. Hybrid CNN-Transformers architectures tend to show the best performances for multi-spectral image segmentation (Wang et al., 2022a; Scheibenreif et al., 2022). Concerning SAR data, the lack of large high-resolution annotated datasets, the precence of speckle and geometrical distortions strongly limit the capacity of very large models. Moreover, due to the different sensing techniques large models learned on optical images can not be directly transferred for this task.

Semi and Self Supervised Learning Semi-supervised learning aims to make the most of the available data in cases where only a small portion of the dataset is labeled. We can roughly classify the semi-supervised learning methods into three families: self-training with pseudo labels (Xie et al., 2020b), generative models (Ehsan Abbasnejad et al., 2017; Dai et al., 2017), and consistency regularization (Miyato et al., 2018; Tarvainen and Valpola, 2017; Laine and Aila, 2017; Valpola, 2015). Self-training involves iteratively extending the labeled dataset with high-confidence predictions from the model. Generative models learn the underlying distribution of the data and then use it to generate to unseen samples. Finally, consistency regularization techniques involve enforcing the model's output to be consistent across different perturbations of the input such as additive noise, contrast jittering, or random flip (Xie et al., 2020a; Cubuk et al., 2020). This enforces a smoothness assumption on the input space. Seminal works include Ladder Networks (Valpola, 2015) or Pi-model (Laine and Aila, 2017) where consistency is enforced between features maps or predictions from clean and noisy input. Adversarial training in semi-supervised learning aims to ensure smooth prediction on a ball around the data (Miyato et al., 2018).

Recent approaches in SSL rely on the same strategy where a pretext task is designed and solving it requires learning useful image representations. These techniques achieve state-of-the-art performance over approaches that learn representations from unsupervised data only. Visual pretext tasks include grayscale image colorization (Larsson et al., 2017), image patches localization and orientation prediction (Komodakis and Gidaris, 2018), jigsaw puzzles(Noroozi and Favaro, 2016), or inpainting (Pathak et al., 2016). Similarly to earlier semi-supervised approaches, contrastive methods encourage the model to learn representations invariant to strong augmentations (Chen et al., 2020a; Grill et al., 2020; Chen and He, 2021). Due to the intrinsic lack of annotations in remote sensing data, semi-supervised and self-supervised approaches appear as valuable strategies to train efficient classification and segmentation models. SSL appraoches for SAR image classification and target recognition include contrastive strategies as well as patch rotation prediction (Ren et al., 2021; Xu et al., 2021; Zhang et al., 2019).

However, the design of pretext tasks for SAR images is critical as they need to preserve the physical sense and the geometry of SAR data. (Zhang and Ma, 2022) shows that some augmentations can have a negative effect on certain downstream tasks, as they can lead to undesired invariances suppressing discriminative features or to the loss of fine-grained information. For instance, geometric distortions appearing in SAR images such as shadowing and layover depend on sensor's orientation and

are not rotation invariant.

3. Method

We propose to reduce the need for annotations by coupling a despeckling task (self-supervised) with a segmentation task (supervised) into a unique label-efficient framework: MERLIN-Seg (Fig.2). While segmentation can only rely on labeled data, self-supervised despeckling can benefit from the large amount of raw data available to learn a semantic representation of the data, without the need of augmenting them. The joint training of the two tasks reduces the labeling requirements for the supervised task. In the following, the despeckling task will be first addressed. Then, we will discuss how despeckling and segmentation can simultaneously be trained into a unique semi-supervised framework.

3.1. Self-supervised despeckling: MERLIN

We propose to use in this framework the self-supervised MERLIN speckle reduction strategy (Dalsasso et al., 2021a). Not only it achieves good restoration performances, but it is also very flexible as it is agnostic to the network architecture. Moreover, by training the network on Single-Look Complex (SLC) SAR images, it learns sensor-specific parameters (*e.g.*, speckle spatial correlation) by its own.

Goodman's model of speckle (Goodman, 2007) describes the complex speckle as a circular Gaussian distributed random variable s. The signal $z = a + jb$ measured over an area with reflectivity R is defined as $z = s\sqrt{R}$. Thus, the distribution of z is given by:

$$\mathrm{p}_Z(z) = \frac{1}{\pi R}\exp\left(-\frac{|z|^2}{R}\right) = \frac{1}{\pi R}\exp\left(-\frac{a^2+b^2}{R}\right). \quad (1)$$

Based on the complex speckle model, in (Dalsasso et al., 2021a) it has been shown that a complex SAR image $\boldsymbol{z} = \boldsymbol{a} + j\boldsymbol{b}$ can be decomposed into two i.i.d. components $\boldsymbol{a}$ and $\boldsymbol{b} \sim \mathcal{N}(0, \boldsymbol{R}/2)$ (with $\boldsymbol{R}$ the image of the reflectivity, i.e., the variance of the k-th pixel is $R_k/2$), each one containing half of the information of an intensity image ($\boldsymbol{I} = \boldsymbol{a}^2 + \boldsymbol{b}^2$). This implies that a neural network can be trained in a self-supervised manner for speckle reduction as follows: the network $f_\theta(.)$ takes as input one component (*e.g.*, the real part $\boldsymbol{a}$) and it evaluates the quality of the restored reflectivity image ($\tilde{\boldsymbol{R}} = f_\theta(\boldsymbol{a})$) with respect to the other component (*e.g.*, the imaginary part $\boldsymbol{b}$). This is achieved by maximizing the likelihood of $\boldsymbol{b}$ with respect to the network output $\tilde{\boldsymbol{R}}$:

$$\begin{aligned}\mathcal{L}_{\mathrm{MERLIN}}(\tilde{\boldsymbol{R}}, \boldsymbol{b}) &= \sum_k -\log \mathrm{p}(b_k|\tilde{R}_k) \\ &= \sum_k \frac{1}{2}\log\left(\tilde{R}_k\right) + \frac{b_k^2}{\tilde{R}_k}, \end{aligned} \quad (2)$$

where index k indicates the k-th pixel of the estimated reflectivity image $\tilde{\boldsymbol{R}}$ or of the imaginary part $\boldsymbol{b}$. In practice, because of their independence, during training the real and the imaginary parts are permuted at each iteration.

At inference time, the trained network is applied separately on the real and imaginary part: the two intermediate estimations $f_\theta(\boldsymbol{a})$ and $f_\theta(\boldsymbol{b})$ are finally averaged to obtain the final reflectivity estimation.

3.2. MERLIN-Seg: joint despeckling and segmentation

In a classical segmentation framework through deep learning, a network is trained to produce, from the input image scene, a segmentation map $\tilde{\boldsymbol{l}}$ that has to be as close as possible to the groundtruth map $\boldsymbol{l}$. To optimize network weights for a binary segmentation problem, one can use the weighted Binary Cross-Entropy (BCE) defined as:

$$\mathcal{L}_{\mathrm{BCE}}(\tilde{\boldsymbol{l}}, \boldsymbol{l}) = \sum_k p l_k \log(\tilde{l}_k) + (1 - l_k)\log(1 - \tilde{l}_k) \quad (3)$$

with p = #negative samples/#positive samples being a weight assigned to positive examples, l_k being the class label and $\tilde{l}_k$ the output class label of pixel k.

To implement MERLIN-Seg, we propose a generalization of MERLIN as follows. The SAR image $\boldsymbol{z} = \boldsymbol{s}\sqrt{\boldsymbol{R}}$ can be seen as augmentation of its reflectivity $\boldsymbol{R}$ by the complex speckle $\boldsymbol{s}$. The noisy nature of SAR images is leveraged within MERLIN-Seg to extract semantic features through self-supervised despeckling with MERLIN, without the need of augmenting SAR data. According to (Baranchuk et al., 2022), intermediate feature maps of a U-Net model for image denoising encode semantic representations of the input image. Thus, we upsample

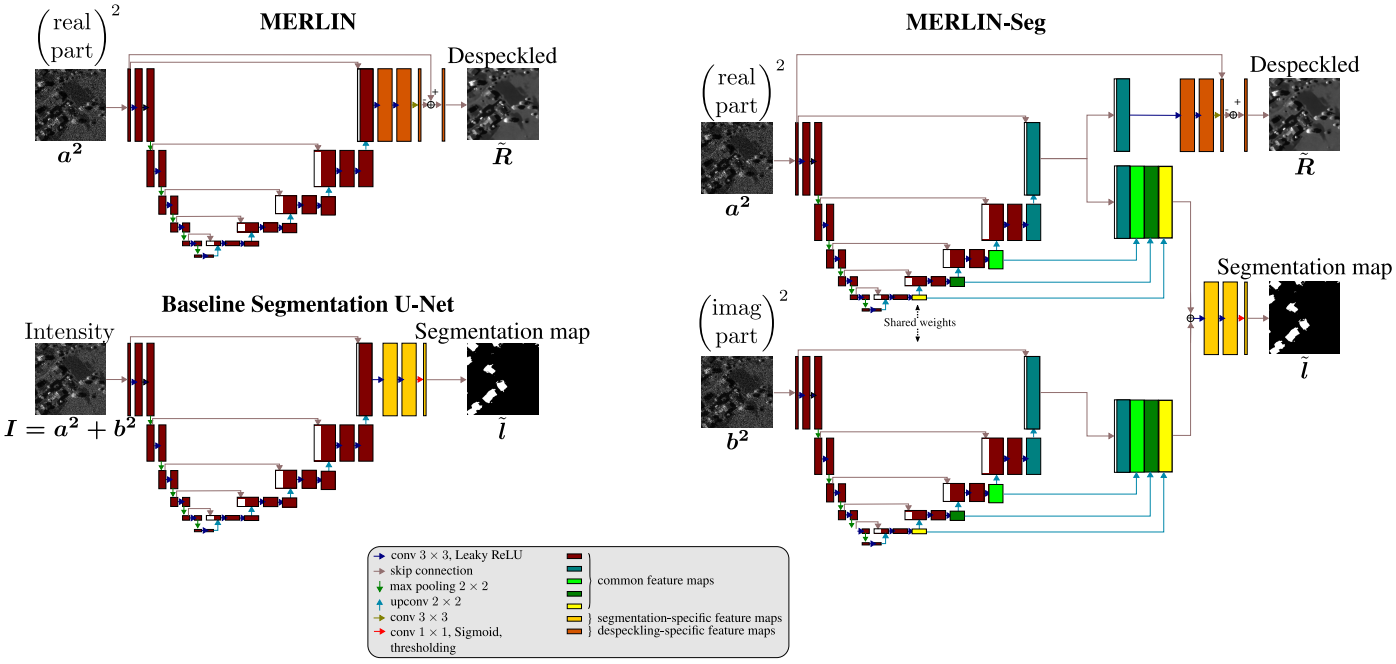


Fig. 2: Illustration of the proposed MERLIN-Seg approach for joint despeckling and segmentation

all feature maps from the upscaling branch of our U-Net model employed in MERLIN and stack a 3 convolutional layers to extract the segmentation map. For the despeckling task, MERLIN processes separately squared real and imaginary parts $\boldsymbol{a}^2$ and $\boldsymbol{b}^2$, characterized by a poorer signal-to-noise ratio (SNR) than the intensity image $\boldsymbol{I} = \boldsymbol{a}^2 + \boldsymbol{b}^2$ by a factor of $\sqrt{2}$. To exploit all the available information for the segmentation task, we compute the feature maps from both real and imaginary parts with a sharing-weights network and we do an average before feeding them to the 3-layers segmentation head, as illustrated in Fig.2. We choose to have a lightweight segmentation head to limit the number of segmentation-specific parameters, and maximize the parameters shared between the tasks. Such a choice is beneficial to facilitate the tuning of the segmentation head when only few labels are available for the downstream task.

The proposed two-heads architecture is trained end-to-end to minimize the following loss function:

$$\mathcal{L}_{\text{MERLIN-Seg}} = \lambda \mathcal{L}_{\text{MERLIN}}(\tilde{\boldsymbol{r}}, \tilde{\boldsymbol{b}}) + (1-\lambda)\mathcal{L}_{\text{BCE}}(\tilde{\boldsymbol{l}}, \boldsymbol{l}) \qquad (4)$$

with $\lambda \in [0, 1]$ an hyperparameter to balance the weight of the two tasks during training. Setting it to 0 corresponds to plain MERLIN despeckling. When no labels are available, the segmentation loss $\mathcal{L}_{\text{BCE}}(\tilde{\boldsymbol{l}}, \boldsymbol{l})$ is set to 0. However, cross-task parameters are still tuned thanks to self-supervised despeckling: while learning to suppress speckle from raw SAR data, the network co-learns to segment on labeled SAR data.

4. Experiments

This section describes the set up of the experiments, the two considered datasets for building footprint segmentation and discusses the obtained results.

4.1. Experimental Settings

Datasets. We evaluated the proposed MERLIN-seg method on two different SAR land cover datasets. The first one is a Single Look Complex (SLC) high resolution simulation of 12600x10800 pixels at X-band provided by the Onera's EMPRISE simulator[1] (Fig.3, top image). The slant-range resolution is 0.71 m and azimuth resolution is 1 m. EMPRISE can generate raw data and realistic SAR images based on the modeling of the physical properties of scatterers present in the 3D scene. In our case, the scene is rendered from cadastral and landcover information, digital elevation models, and fine-grained field surveys of the city of Sainte-Marie in the south of

[1] `http://emprise-em.fr`

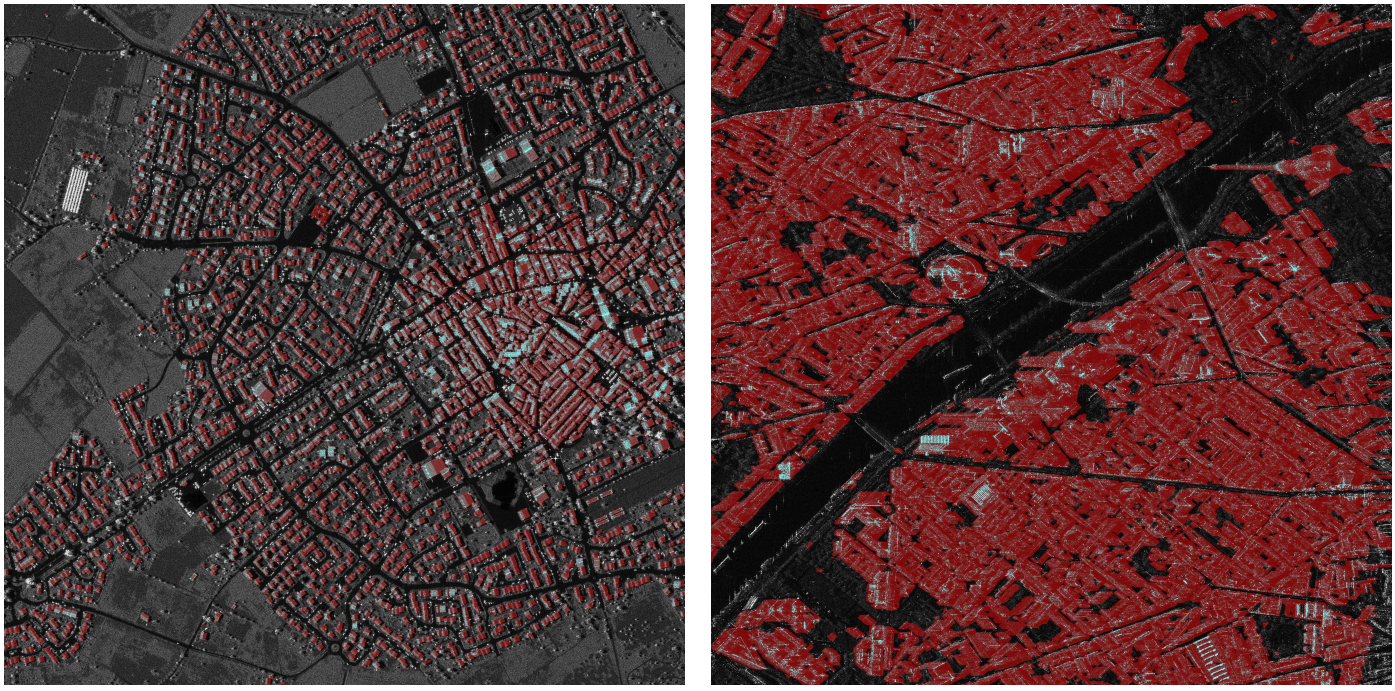


Fig. 3: Left image: example of building footprint superposed to an EMPRISE simulated image of Sainte-Marie. Right image: example of building footprint superposed to a TerraSAR-X SpotLight image acquired over Paris. Both pictures display a 3000×3000 pixels image crop.

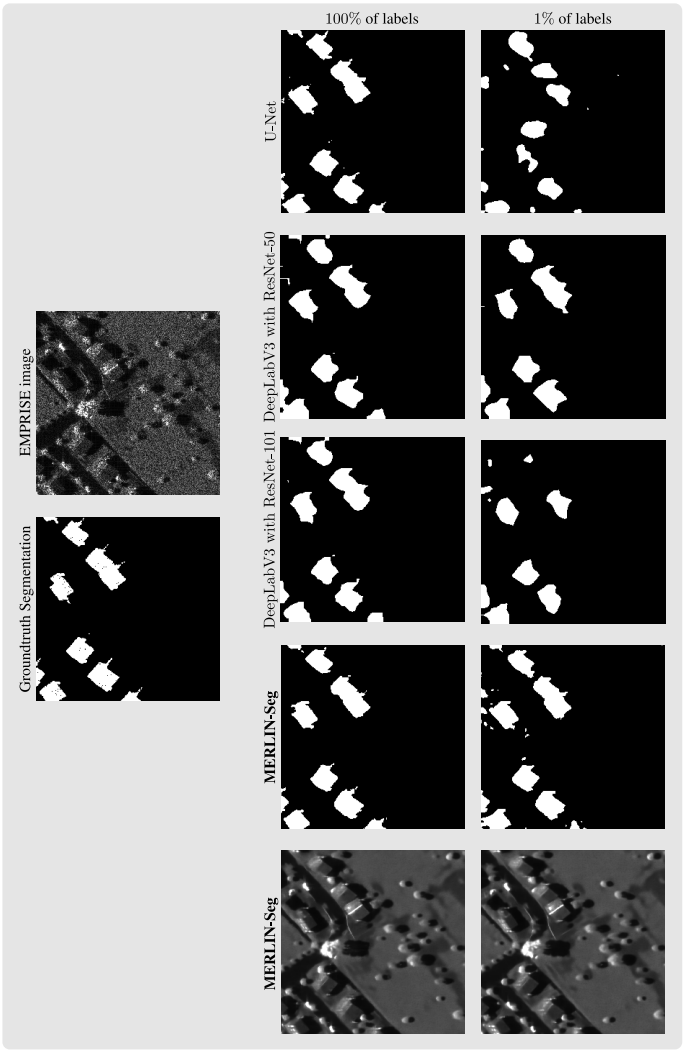
France. All ground cover elements ranging from small bushes to whole buildings are modeled in 3D and ground surface is generated at an average resolution of 5.3cm. The full 3D database cover over 10 by 10 km area. The simulation process allows us to generate perfect labels at the image resolution including SAR effects such as layover or side lobes.

The second dataset is obtained from an SLC TerraSAR-X 6000x10000 SpotLight image acquired in July 2012 over the city of Paris[2] and its surrounding areas. The slant-range resolution is 0.45 m and the azimuth resolution is 0.87 m (Fig.3, bottom image). The labels are extracted from the BDtopo database provided by the french National Geographic Institute (IGN). The BDtopo can be freely downloaded and consists of vector data describing the 2.5D buildings geometry: each georeferenced footprint is associated with the object's height. We performed the projection of the labels from nadir to the true incidence angle from the TerraSAR-X sensor. The scene is composed of dense urban areas and its segmentation is challenging due to the multiple overlays and shadows projected by the buildings.

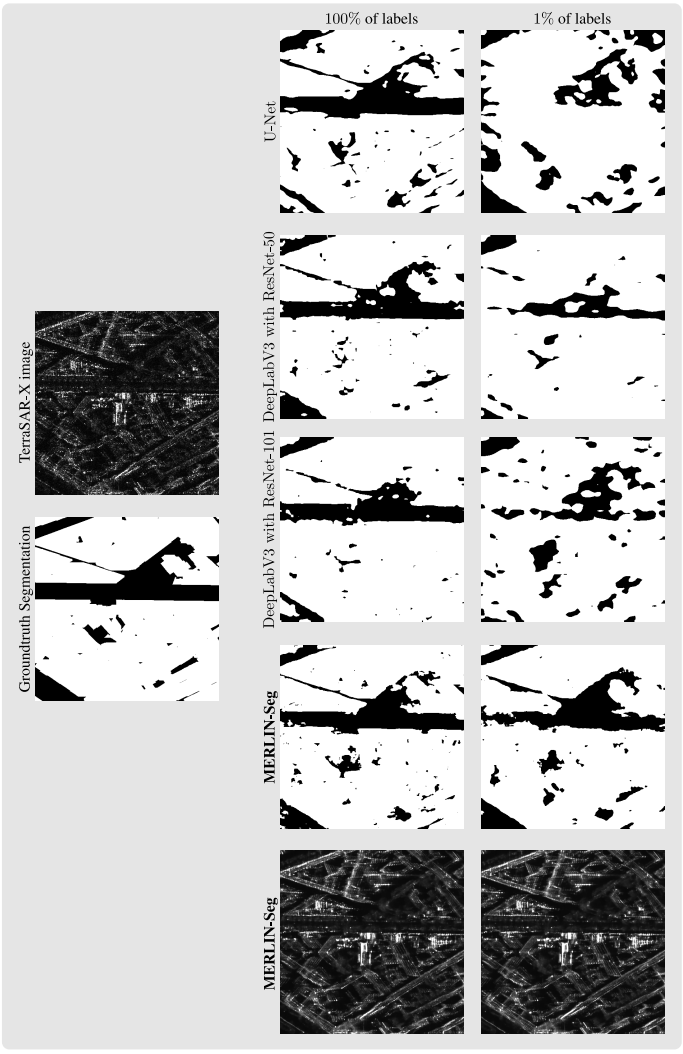
Implementation details. To set up our experiments, we build on the U-Net model as in (Dalsasso et al., 2021a). To generalize the network so that it performs simultaneous despeckling and segmentation, we define an extra 3-layers segmentation head composed of 64 kernels of size 3×3 followed by Leaky ReLU, 32 kernels of size 3×3 followed by Leaky ReLU and a final 1×1 sized kernel followed by sigmoid activation to extract the class labels for each image pixel, see Fig.2. The final binary segmentation mask is obtained by applying a threshold set to 0.5.

The EMPRISE simulated dataset is split into 3 non-overlapping subsets as follows: network is fed with 11161 image patches with a stride of 64, the validation set is composed of 5 non-overlapping image patches, whereas the test is conducted on 70 non-overlapping image patches. All patches are of size 256×256. The U-Net backbone is composed of 5 downsampling layers (compressing the 256×256 image patch down to an 8×8 latent representation) and 5 upscaling layers. The TerraSAR-X dataset is split into 3 independent subsets

[2]These data have been provided by the DLR in the framework of the project LAN1746.



(a) Results on a 256×256 EMPRISE simulated image patch on Sainte-Marie.



(b) Results on a 512×512 TerraSAR-X image patch on Paris.

Fig. 4: Qualitative results highlight that performances of MERLIN-Seg remain stable when reducing the number of labeled data. A segmentation map close to the groundtruth label is obtained even with 1% of labeled images, with an accurate reconstruction of building contours. Whatever the number of available labels, the network restores well the despeckled image. On the contrary, the baseline segmentation collapses: some building are missed and false positive are also detected.

as follows: 1189 patches with a stride of 128 for training, 3 non-overlapping patches for validation and 65 non-overlapping patches for testing. All patches are of 512×512 pixels. The U-Net backbone as one extra downscaling and upscaling layer w.r.t. the network described above. The choice of patch size and network depth is done to enlarge network receptive field.

For the segmentation baseline, we consider the same U-Net model as in MERLIN and replace the last layer with a 1×1 convolutional filter followed by Sigmoid, in order to extract class labels (Fig.2). This allows to disentangle the contribution of the architecture to the advantages brought by the training strategy. Moreover, to compare with a state-of-the-art segmentation model we consider two DeepLabV3 models (Chen et al., 2017), with ResNet-50 and ResNet-101 as backbones. The input of these networks is the intensity SAR image. The networks are trained with $\mathcal{L}_{\text{BCE}}$ loss function on labeled data only. Both the segmentation baselines and MERLIN-Seg are trained for 100 epochs with an initial learning rate of 0.001 and Adam optimizer. For each percentage of labeled data, network weight's are randomly initialized and training is done from scratch. We highlight that when only a limited number of annotated samples are available, the segmentation baseline only sees labeled patches in input. On the contrary, thanks to the self-supervised despeckling task, MERLIN-Seg always sees all the data. In the

Table 1: Results with λ = 0.5 on EMPRISE simulated dataset

	100%			**20%**			**5%**			**1%**		
	mIoU	F1	Acc	mIoU	F1	Acc	mIoU	F1	Acc	mIoU	F1	Acc
U-Net	0.710	**0.803**	0.974	0.674	0.758	0.969	0.551	0.669	0.952	0.195	0.299	0.897
DeepLabV3-50	**0.717**	0.797	0.971	0.698	0.783	0.971	0.630	0.737	0.965	0.545	0.671	0.953
DeepLabV3-101	0.665	0.747	0.967	0.653	0.740	0.965	0.534	0.647	0.953	0.367	0.478	0.933
MERLIN-Seg	0.711	0.799	**0.976**	**0.707**	**0.794**	**0.976**	**0.680**	**0.781**	**0.971**	**0.634**	**0.745**	**0.965**

Table 2: Results with λ = 0.5 on TerraSAR-X dataset

	100%			**20%**			**5%**			**1%**		
	mIoU	F1	Acc	mIoU	F1	Acc	mIoU	F1	Acc	mIoU	F1	Acc
U-Net	0.748	0.832	0.874	0.746	0.831	0.873	0.714	0.806	0.847	0.633	0.750	0.790
DeepLabV3-50	0.735	0.820	0.874	0.753	0.836	0.877	0.725	0.814	0.867	0.718	0.811	0.843
DeepLabV3-101	0.720	0.809	0.864	0.738	0.824	0.869	0.700	0.794	0.852	0.688	0.788	0.831
MERLIN-Seg	**0.765**	**0.844**	**0.886**	**0.760**	**0.840**	**0.882**	**0.760**	**0.840**	**0.883**	**0.756**	**0.838**	**0.882**

latter, batches are sampled randomly, *i.e*. we don't impose any constraint on the number of labeled images composing a batch.

4.2. Results

The baselines for both experiments are a standard U-Net, DeepLabV3 with ResNet-50 and DeepLabV3 with ResNet-101 trained with binary cross-entropy on labeled data. Table 1 summarize our results on the simulated dataset for different proportions of labeled data. On this dataset, MERLIN-Seg is on par with the baselines when 100% of labels are available, with the best accuracy attained by MERLIN-Seg, while U-Net and DeepLabV3 with ResNet-50 obtaining respectively the best F1-score and mIoU. MERLIN-Seg performances are stable even when only 1% labels are used for training. In this case, MERLIN-Seg yields 0.63 mIoU and 0.75 F1 score while the fully supervised baselines performances collapse with significantly slower mIoU and F1 score. On real data, the same behavior can be observed and is summarized in table 2. In this configuration, MERLIN-Seg slightly outperforms the supervised baseline when all the labels are available showing 0.76 against 0.75 mIoU for the U-Net baseline. The model trained with the MERLIN-Seg strategy is particularly stable as the performance drops are limited to 0.01 both for mIoU and F1 score. The fully supervised U-Net model shows a drop up to 0.12 mIoU and 0.08 F1 score and is outperformed by MERLIN-Seg with about 0.12 mIoU and 0.1 F1 score. Quantitative evaluation is confirmed by visual inspection of the results (Fig.4).

Not only our experiments tell that the U-Net model performs better when trained within the MERLIN-Seg framework than in a fully supervised fashion, but it also outperforms the two DeepLabV3 models. DeepLabV3 with ResNet-50 and with ResNet-101 have $\sim$ 40M and $\sim$ 60M parameters. Instead, the U-Net baseline has $\sim$ 1.2M parameters and MERLIN-Seg has $\sim$ 1.5M parameters, *i.e*. only $\sim$ 350k more parameters the U-Net to compute two tasks simultaneously. This allows to efficiently train a U-Net with MERLIN-Seg in a low-data regime, which is common in remote sensing applications.

While the role of despeckling is to learn semantically meaningful features for downstream segmentation and we are not interested in its performances, it is worth to point out that high quality restoration is achieved on both EMPRISE and

TerraSAR-X images. No residual speckle fluctuations seem to be left in the restored image, which is finely restored. We cannot provide a quantitative evaluation on SAR despeckling as the groundtruth is not available.

4.3. Ablation study

In this part, we analyze different aspect of MERLIN-Seg.

Features aggregation MERLIN processes separately the real and the imaginary parts for the despeckling task, due to their independence. A naive extension of MERLIN would consist in providing to the segmentation head only the feature maps estimated from the real part (resp. the imaginary part) and keep the imaginary part (resp. the real part) to supervise the despeckling task. In this configuration, the segmentation task only rely on a partial information and the performances are even below the baseline for all metrics on pure segmentation ($\lambda = 0$): see table 3, second row. To incorporate the information of both real and imaginary parts for downstream segmentation, two strategies are considered. The feature maps are computed for both real and imaginary parts. Then, they can be aggreated either by concatenated them, or by doing features average. While real and imaginary parts remain separated for despeckling, they can be both exploited for segmentation. The two aggregation strategies have proved to be effective as they outperform the baseline. We choose to keep the average as it shows to be slighlty more efficient and it requires less training parameters.

Table 3: Results on EMPRISE dataset when setting $\lambda = 0$. MERLIN-Seg with features mean shows higher performances than the other strategies on building footprint segmentation.

	IoU	**F1**	**Accuracy**
U-Net	0.710	0.803	0.974
MERLIN-Seg (w/o aggregation)	0.671	0.750	0.970
MERLIN-Seg (w/ concatenation)	**0.713**	0.805	**0.975**
MERLIN-Seg (w/ mean)	**0.713**	**0.806**	**0.975**

The impact of λ The value of the λ parameter directly impact the relative importance of the denoising and segmentation tasks. This tuning would depend on the proportion of annotated data but one would expect the performances to be sufficiently stable with respect to λ to avoid painful hand-tuning. Table 4 shows the performances of the segmentation task for different values of λ in the context where only 1% of labels are available. The performances achieved by MERLIN-Seg are quite stable for λ values samples from 0.1 to 0.9 and the best regime is obtained for $\lambda = 0.5$ which gives a practical rull of thumb.

It is fundamental to observe the particular case of $\lambda = 0$. A drastic drop of performances occurs if self-supervised despeckling is not applied. This result strongly supports our thesis: given the same network architecture, the joint learning of segmentation and self-supervised despeckling ($\lambda \neq 0$) is highly beneficial w.r.t. a fully supervised training on segmentation only ($\lambda = 0$).

Table 4: Results on EMPRISE building footprint dataset when varying λ. The label percentage is fixed to 1% for all experiments.

MERLIN-Seg	**IoU**	**F1**	**Accuracy**
$\boldsymbol{\lambda = 0}$	0.263	0.382	0.899
$\boldsymbol{\lambda = 0.1}$	0.600	0.715	0.959
$\boldsymbol{\lambda = 0.2}$	0.589	0.704	0.960
$\boldsymbol{\lambda = 0.5}$	**0.634**	**0.745**	**0.965**
$\boldsymbol{\lambda = 0.7}$	0.628	0.744	0.964
$\boldsymbol{\lambda = 0.9}$	0.607	0.727	0.960

End-to-end *vs.* Fine-tuning We compare the proposed joint learning strategy with a fully SSL pre-training approach: once the network is pre-trained on self-supervised despeckling, its weights are frozen and the feature maps are fed to the segmentation head to be fine-tuned on labeled data. We refer to this strategy as sequential approach. In this training configuration, we employ the same network as in MERLIN-Seg. The quantitative evaluation reported in table 5 allows to say while self-supervised pre-training learns semantically meaningful features, showing good segmentation performances of building footprint segmentation when only 1% of the EMPRISE

dataset is annotated, the sequential approach is outperformed by MERLIN-Seg. We believe that pre-training the backbone for denoising vast quantities of diverse scenes could significantly enhance overall performance and is an interesting lead for future works. However, in the context of data scarcity, end-to-end training brings higher performance. Additionally, we also acknowledge the potential benefits of pre-training the backbone on a semantic segmentation task. However, the high cost of acquiring high-resolution SAR images and the inherent challenges in producing accurately registered ground truth data pose significant obstacles. Furthermore, the distinctive nature of SAR images introduces a notable domain shift, making it impractical to leverage off-the-shelf models or define a suitable segmentation-related pretext task for self-supervised pretraining.

Table 5: Results of sequential approach (fine-tuning) versus MERLIN-Seg (end-to-end) on EMPRISE building footprint dataset. The label percentage is fixed to 1% and for MERLIN-Seg we set $\lambda = 0.5$.

	IoU	**F1**	**Accuracy**
Sequential	0.422	0.556	0.930
MERLIN-Seg	**0.634**	**0.745**	**0.965**

5. Conclusion

We have shown that self-supervised despeckling extract semantically meaningful features of SAR images for downstream segmentation task. The segmentation task benefits from despeckling, especially when few annotated samples are available: the features learned on despeckling guide segmentation, making the learning of the segmentation task easier.

In contrast to classical contrastive learning frameworks, our approach does not require to apply any augmentation to the data. Indeed, the speckle is seen as an intrinsic augmentation of SAR images and our model is trained to recover the speckle-free image in a self-supervised fashion.

While our experiments have been conducted on a U-shaped model, MERLIN-Seg is model agnostic and it can thus profit from the most performing networks for image segmentation. Moreover, its performances can be improved by resorting to additional input information, may it be multi-modal (*e.g.* multispectral images (Bergamasco et al., 2023)) and/or multi-temporal (*e.g.* SAR time series (Meraoumia et al., 2023)): the network would extract richer semantic information from the extra input to better despeckle SAR images, thus facilitating the downstream task.

6. Acknowledgements

This project has been funded by ANR (the French National Research Agency) and DGA (Direction Générale de l'Armement) under ASTRAL project ANR-21-ASTR-0011. We would like to thanks french DGA for supporting the EMPRISE project and Onera's partner Scalian-DS for its major implication in the development of the simulation software.

References

Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A., 2022. Label-efficient semantic segmentation with diffusion models, in: International Conference on Learning Representations.

Bergamasco, L., Bovolo, F., Bruzzone, L., 2023. A dual-branch deep learning architecture for multisensor and multitemporal remote sensing semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16, 2147–2162. doi:10.1109/JSTARS.2023.3243396.

Brempong, E.A., Kornblith, S., Chen, T., Parmar, N., Minderer, M., Norouzi, M., 2022. Denoising pretraining for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4175–4186.

Buchholz, T.O., Prakash, M., Schmidt, D., Krull, A., Jug, F., 2021. Denoiseg: joint denoising and segmentation, in: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I, Springer. pp. 324–337.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: ECCV.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020a. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020b. Big self-supervised models are strong semi-supervised learners. Advances in neural information processing systems 33, 22243–22255.

Chen, X., He, K., 2021. Exploring simple siamese representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15750–15758.

Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V., 2020. Randaugment: Practical automated data augmentation with a reduced search space, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 702–703.

Dai, Z., Yang, Z., Yang, F., Cohen, W.W., Salakhutdinov, R.R., 2017. Good semi-supervised learning that requires a bad gan. Advances in neural information processing systems 30.

Dalsasso, E., Denis, L., Muzeau, M., Tupin, F., 2022. Self-supervised training strategies for sar image despeckling with deep neural networks, in: EUSAR 2022; 14th European Conference on Synthetic Aperture Radar, VDE. pp. 1–6.

Dalsasso, E., Denis, L., Tupin, F., 2021a. As if by magic: self-supervised training of deep despeckling networks with merlin. IEEE Transactions on Geoscience and Remote Sensing 60, 1–13.

Dalsasso, E., Denis, L., Tupin, F., 2021b. Sar2sar: A semi-supervised despeckling algorithm for sar images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 4321–4329.

Ehsan Abbasnejad, M., Dick, A., van den Hengel, A., 2017. Infinite variational autoencoder for semi-supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5888–5897.

Finn, C., Abbeel, P., Levine, S., 2017. Model-agnostic meta-learning for fast adaptation of deep networks, in: International conference on machine learning, PMLR. pp. 1126–1135.

Fracastoro, G., Magli, E., Poggi, G., Scarpa, G., Valsesia, D., Verdoliva, L., 2021. Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives. IEEE Geoscience and Remote Sensing Magazine 9, 29–51.

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

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271–21284.

Komodakis, N., Gidaris, S., 2018. Unsupervised representation learning by predicting image rotations, in: International conference on learning representations (ICLR).

Krull, A., Buchholz, T.O., Jug, F., 2019. Noise2void-learning denoising from single noisy images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2129–2137.

Laine, S., Aila, T., 2017. Temporal ensembling for semi-supervised learning, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=BJ6oOfqge.

Laine, S., Karras, T., Lehtinen, J., Aila, T., 2019. High-quality self-supervised deep image denoising. Advances in Neural Information Processing Systems 32.

Lapini, A., Bianchi, T., Argenti, F., Alparone, L., 2013. Blind speckle decorrelation for sar image despeckling. IEEE transactions on geoscience and remote sensing 52, 1044–1058.

Larsson, G., Maire, M., Shakhnarovich, G., 2017. Colorization as a proxy task for visual understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6874–6883.

Le-Khac, P.H., Healy, G., Smeaton, A.F., 2020. Contrastive representation learning: A framework and review. Ieee Access 8, 193907–193934.

Lee, K., Jeong, W.K., 2022. Noise2kernel: Adaptive self-supervised blind denoising using a dilated convolutional kernel architecture. Sensors 22, 4255.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T., 2018. Noise2noise: Learning image restoration without clean data, in: International Conference on Machine Learning, PMLR. pp. 2965–2974.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. International Conference on Computer Vision (ICCV) .

Meraoumia, I., Dalsasso, E., Denis, L., Abergel, R., Tupin, F., 2023. Multitemporal speckle reduction with self-supervised deep neural networks. IEEE Transactions on Geoscience and Remote Sensing 61, 1–14.

Miyato, T., Maeda, S.i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence 41, 1979–1993.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI, Springer. pp. 69–84.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536–2544.

Ren, B., Zhao, Y., Hou, B., Chanussot, J., Jiao, L., 2021. A mutual information-based self-supervised learning model for polsar land cover classification. IEEE Transactions on Geoscience and Remote Sensing 59, 9224–9237.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer. pp. 234–241.

Scheibenreif, L., Hanna, J., Mommert, M., Borth, D., 2022. Self-supervised vision transformers for land-cover segmentation and classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1422–1431.

Tarvainen, A., Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems 30.

Themyr, L., Rambour, C., Thome, N., Collins, T., Hostettler, A., 2023. Full contextual attention for multi-resolution transformers in semantic segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 3224–3233.

Valpola, H., 2015. From neural pca to deep unsupervised learning, in: Advances in independent component analysis and learning machines. Elsevier, pp. 143–171.

Wang, H., Xing, C., Yin, J., Yang, J., 2022a. Land cover classification for polarimetric sar images based on vision transformer. Remote Sensing 14, 4656.

Wang, Y., Albrecht, C.M., Braham, N.A.A., Mou, L., Zhu, X.X., 2022b. Self-supervised learning in remote sensing: A review. IEEE Geoscience and Remote Sensing Magazine 10, 213–247. doi:10.1109/MGRS.2022.3198244.

Xian, Y., Schiele, B., Akata, Z., 2017. Zero-shot learning-the good, the bad and the ugly, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4582–4591.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J., 2018. Unified perceptual parsing for scene understanding, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 418–434.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. arXiv preprint arXiv:2105.15203 .

Xie, Q., Dai, Z., Hovy, E., Luong, T., Le, Q., 2020a. Unsupervised data augmentation for consistency training. Advances in neural information processing systems 33, 6256–6268.

Xie, Q., Luong, M.T., Hovy, E., Le, Q.V., 2020b. Self-training with noisy student improves imagenet classification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10687–10698.

Xu, Y., Sun, H., Chen, J., Lei, L., Ji, K., Kuang, G., 2021. Adversarial self-supervised learning for robust sar target recognition. Remote Sensing 13, 4158.

Zhang, J., Ma, K., 2022. Rethinking the augmentation module in contrastive learning: Learning hierarchical augmentation invariance with expanded views, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16650–16659.

Zhang, S., Wen, Z., Liu, Z., Pan, Q., 2019. Rotation awareness based self-supervised learning for sar target recognition, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, IEEE. pp. 1378–1381.

Zheng, X., Kellenberger, B., Gong, R., Hajnsek, I., Tuia, D., 2021. Self-supervised pretraining and controlled augmentation improve rare wildlife recognition in uav images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 732–741.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE geoscience and remote sensing magazine 5, 8–36.