Unsupervised deep learning for ship detection in SAR images

João Moura
joao.reis.moura@tecnico.ulisboa.pt

Instituto Superior Técnico, University of Lisbon, Portugal

October 2022

Abstract

From illegal fishing to drug smuggling, environmental protection, and threat prevention, it is evident that maritime surveillance is of extreme importance. One of the key aspects of maritime surveillance is having knowledge of the location of the ships. Since the ocean covers such a wide area, automatic algorithms are necessary to monitor them. Recent advances in deep learning have substantially facilitated the development of ship detection methods for synthetic aperture radar (SAR) images. However, most of the solutions are supervised object detection methods, which require large amounts of labelled data. Labelling the images is an extremely time-consuming process. To take advantage of the huge and increasing amount of SAR data, we propose two unsupervised deep learning frameworks for SAR ship segmentation. The first framework is based on an image-to-image translation model, the CycleGAN, in which we exploit the model's unpaired image style transfer capabilities to learn the mapping from the SAR image domain to a segmentation domain. The second approach, the UDSEP (U-net Detect-Select-Erase-Paste) is a self-supervised segmentation framework, in which we train a segmentation network with data from a novel algorithm that generates synthetic labelled images from the original SAR unlabelled images. Experiments on the SAR-Ship-Dataset and on SSDD reveal promising results but still inferior to those of the supervised methods.
Keywords: Deep learning, Unsupervised, Synthetic Aperture Radar, Ship Semantic Segmentation

1. Introduction

Maritime surveillance has attracted a lot of attention in recent decades, particularly in vessel detection, as knowledge of vessel placements is required to attain complete maritime domain awareness [1]. From threat prevention to national security, safety, and environmental protection, it is crucial to provide relevant organizations, governments, and agencies with real-time data on vessel localizations to assist decision-making processes. Synthetic aperture radar (SAR) is the most suitable type of radar to provide data for ship detection since its resolution is constant even when far from the observed targets, it can image wide areas at constant resolution, and works regardless of daylight and cloud cover [2].

Several ship detection methods have arisen since the first SAR satellite was launched in 1978. Most traditional methods are based on a constant false alarm rate (CFAR). These methods are frequently not robust enough and have detection speeds incompatible to suit the needs of real-time applications. Recently, with the growth of artificial intelligence, various elegant deep learning solutions have obtained state-of-the-art in the SAR ship detection task. However, most of these solutions are object detection methods, which are supervised and, therefore, require large amounts of labelled data. Labelling the images is a process that requires SAR specialists and is extremely time-consuming and expensive. Unsupervised methods, which do not require the labelling of training images for feature extraction, can be a suitable alternative for ship detection, especially given the extensive and expanding amount of available SAR data.

This type of work has not been extensively explored. Ferreira et al. [3] proposed an unsupervised framework for SAR ship detection based on anomaly detection. They start by learning the data representations with a convolutional Variational Autoencoder (VAE) and then perform anomaly detection based on those representations with a clustering algorithm. Dias et al. [4] also proposed an unsupervised anomaly detection framework for ship detection. They train a Bidirectional Generative Adversarial Network (BiGAN) with non-ship images and then use its inability to reconstruct images with ships to detect anomalies. In fact, although referred to as unsupervised, both the mentioned works rely on a supervised preselection of non-ship ocean images to train the models. Furthermore, some weakly-supervised approaches have also been proposed but for ship segmentation, where the authors train the models with two global labels, ship or non-ship [5], or with missing target level annotations [6]. In fact, to our best knowledge, no fully deep learning unsupervised work has been proposed for either SAR ship detection or segmentation.

The goal of this work is to develop unsupervised deep learning methods for ship detection in SAR images, which will be approached as a semantic segmentation problem. Two distinct novel deep learning frameworks are presented. The proposed work should contribute to filling the void in the state-of-the-art of SAR ship detection with unsupervised techniques.

2. Background

In this section, the deep learning models relevant to this thesis are revised.

2.1. U-net

Originally designed for biomedical segmentation, the U-net [7] is a supervised state-of-the-art segmentation network.

The basic architecture of the U-net is composed of two paths: a contracting path and an expansion path. Also known as the encoder, the contracting path follows a typical convolutional network architecture, consisting of repeated convolutions followed by rectified linear unit (ReLU) activations and max-pooling, that allows for high-level feature extraction. Throughout the contracting path, the number of feature channels increases while the image size decreases. The expansion path, or decoder, consists of up-convolutions followed by convolutions and ReLU, and concatenations with features that have been captured in the encoder. Due to convolution, there is a loss of border pixels, therefore, cropping is necessary. Thus, the pixel features near the edges are removed, since they have the least amount of contextual information. The full model architecture resembles a u-shape that is able to propagate contextual information throughout the network, allowing to segment objects in an area using information from a larger overlapping area. Several energy functions have been proposed to optimise the U-net. For binary classification, one of the most commonly used energy functions is the Binary Cross Entropy (BCE),

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) + \left(1-y_i\right)\log\left(1-\hat{y}_i\right), \quad (1)$$

where N is the number of pixels of the training image, $y_i \in \{0,1\}$ is the target value of pixel i, and $\hat{y_i} \in [0,1]$ is the predicted probability for the pixel i.

2.2. CycleGAN

The CycleGAN [8] is an image-to-image translation model that, unlike other GAN-based approaches such as Pix2Pix [9], does not require paired examples to be trained.

The goal of the method is to learn the mapping between the domains X and Y, and vice-versa, given the training data $x_i \in X$ with $i = 1, \ldots, N$ and $y_j \in Y$ with $j = 1, \ldots, M$[1] with distributions $x \sim p_{\text{data}}(x)$ and $y \sim p_{\text{data}}(y)$, respectively. The model includes two generators $G : X \to Y$ and $F : Y \to X$ and two adversarial discriminators, D_X and D_Y, where D_X aims to distinguish between images $\{x\}$ from the X domain and translated images $\{F(y)\}$, and D_Y between images $\{y\}$ from the Y domain and $\{G(x)\}$. To be able to correctly translate domains, the authors proposed three types of terms for the objective function: adversarial loss, cycle consistency loss, and identity loss.

The adversarial loss is responsible for approximating the distribution of the generated images to the target distribution. For the mapping function $G : X \to Y$, G attempts to generate images $G(x)$ that match the Y domain. Then D_Y tries to distinguish the generated image from real samples, $y \in Y$. Therefore, the objective is given by

$$\begin{aligned}\mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) = \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\log D_Y(y)\right] \\ + \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\log\left(1 - D_Y(G(x))\right)\right],\end{aligned}$$

where G attempts to minimise it and D_Y aims to maximise it, i.e., $\min_G \max_{D_Y} \mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right)$. For the mapping function $F : Y \to X$, the process is similar, hence the goal is to solve $\min_F \max_{D_X} \mathcal{L}_{\text{GAN}}\left(F, D_X, Y, X\right)$. With the adversarial loss, the generators are expected to generate plausible images in the target domain, indistinguishable from the real ones. However, it does not guarantee an individual translation from input to the desired output, thus the need to introduce the cycle consistency loss.

The cycle consistency loss is able to further reduce the space of possible mapping functions by attempting to make the mapping functions cycle-consistent via an L1-norm reconstruction loss for a real image. For the X domain, the cycle ought to be able to reconstruct x, that is, $x \to G(x) \to F(G(x)) \approx x$. For the Y domain, the principle remains, thus, G and F should be updated during the training to ensure that y can be correctly reconstructed, i.e., $y \to F(y) \to G(F(y)) \approx y$. To ensure this procedure, the cycle consistency loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\|F(G(x)) - x\|_1\right] \\ + \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\|G(F(y)) - y\|_1\right],\end{aligned}$$

where $\|\cdot\|_1$ represents the L1-norm.

Furthermore, the authors suggested the regularisation of the generators to force an identity mapping when real samples of the target domain are provided as input. Therefore, they defined the identity loss as

$$\begin{aligned}\mathcal{L}_{\text{idty}}(G, F) = \mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\|G(y) - y\|_1\right] \\ + \mathbb{E}_{x\sim p_{\text{data}}(x)}\left[\|F(x) - x\|_1\right].\end{aligned}$$

Although this loss is not fully required to successfully learn the mapping between the domains, it can improve the results depending on the translation task. The intuition behind the loss should be for the CycleGAN to only change parts of the image if required. Therefore, if something already looks like the target domain, the model should learn that it does not need to be changed.

The full objective is given by the weighted sum of the objectives referenced above:

$$\begin{aligned}\mathcal{L}\left(G, F, D_X, D_Y\right) = \mathcal{L}_{\text{GAN}}\left(G, D_Y, X, Y\right) \\ + \mathcal{L}_{\text{GAN}}\left(F, D_X, Y, X\right) + \lambda\mathcal{L}_{\text{cyc}}(G, F) + \alpha\mathcal{L}_{\text{idty}}(G, F),\end{aligned} \quad (2)$$

where λ and α are parameters that determine the importance of each objective. Moreover, the goal is to obtain the generators that solve

$$G^*, F^* = \arg\min_{G,F}\max_{D_x, D_Y} \mathcal{L}\left(G, F, D_X, D_Y\right).$$

[1] Subscripts i and j will be omitted for simplicity.

3. Proposed Approach

3.1. Dataset

Two well-known datasets from the SAR ship research community were separately used to train and evaluate the models proposed in this thesis: the SAR-Ship-Dataset and the SSDD (SAR Ship Detection Dataset).

1) SAR-Ship-Dataset: The SAR-Ship-Dataset [10] consists of 102 Chinese Gaofen-3 and 108 Sentinel-1 images that were processed and cropped to build a dataset with 39729 256x256 ship images with a total of 50885 ships. The ships vary considerably in size and have distinct and complex backgrounds. One of the models used in this work, the CycleGAN, has a complex training architecture. Thus, due to extremely long running times, it is unfeasible to train it with the full dataset. Therefore, we were compelled to create a more concise version of the dataset. In order to capture the original dataset diversity in a balanced manner, we propose to create it with images that are equally distributed in their Shannon entropy [11] value. Therefore, we first compute the entropy of each image in the original dataset, and then create the new dataset with a total of 7000 entropy-distributed images. We call this dataset the concise-SAR-Ship-Dataset. The images of this dataset and the test set, which was created by randomly selecting 1000 of the remaining images from the original dataset, were annotated with their ship segmentation through a threshold-based segmentation method supervised by us. This was done for more accurate test results and to be able to train a supervised comparison segmentation method. In addition, a binary label of the level of complexity (simple or complex) is given to each test image. Typically, images that are considered simple contain offshore ships of average size in calm to moderate sea conditions. Images are deemed complex if the ships are excessively large, or in inshore conditions, or with a significant amount of spectral noise.

2) SSDD: The SSDD [12] consist of RadarSat-2, TerraSAR-X, and Sentinel-1 images that were cropped to build a dataset with 2456 ships in 1160 images, which are labelled with their polygon segmentation. This dataset has a diverse ship population, such as small-sized ships, complex backgrounds, and dense arrangements near harbors. We use the standards provided by the authors. Thus, the train and the test set are made up of 928 and 232 images, respectively. The authors also provided an offshore-inshore separation of the images. Moreover, given that the image sizes vary and in order to maintain consistency throughout the models, each image was resized to 256x256 pixels.

3.2. Methods

3.2.1 CycleGAN

The first proposed approach is based on the CycleGAN. In this framework, we aim to explore the image translation capabilities of the CycleGAN to provide semantic segmentation. To this end, we propose for the CycleGAN to learn the mapping between normal images and binary segmentation masks, and vice-versa. In the context of this thesis, and since we are only interested in the segmentation, the main goal is for the CycleGAN to learn the mapping between the SAR image domain and a binary domain, corresponding to the ships' semantic segmentation. Moreover, one independent CycleGAN model is trained for each of the datasets.

The CycleGAN is based on the original paper implementation [8]. Its simplified architecture is depicted in Figure 1. The model consists of two discriminators, D_L and D_{SAR} and two generators, $G_{L\_to\_SAR}$ and $G_{SAR\_to\_L}$. The generator $G_{L\_to\_SAR}$ translates images from the label domain to the SAR domain and $G_{SAR\_to\_L}$ translates images from the SAR domain to the label domain. The discriminators for the label domain and the SAR domain are D_L and D_{SAR}, respectively.

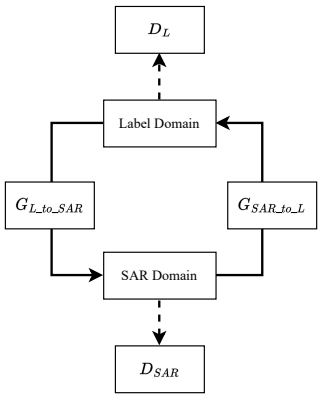

Figure 1: Simplified architecture of the CycleGAN.

The discriminators are deep convolutional neural networks that receive an image as input and compute the likelihood that it came from the training data rather than being generated by the generators. Per the paper's specifications, the discriminators are 70x70 PatchGAN classifiers. The architecture of the PatchGAN discriminator is depicted in Figure 2.

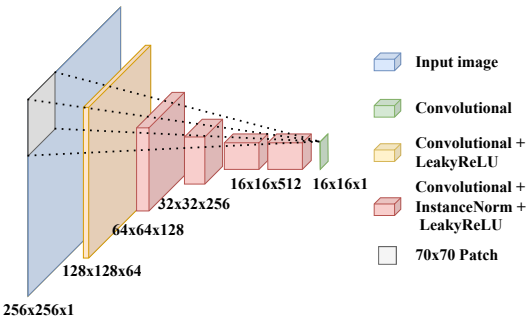

Figure 2: Schematic representation of the architecture of the PatchGAN discriminators. The LeakyRelu activation function has a 0.2 slope for the negative values. The numbers below the blocks represent the image size at its output.

The generators receive an image from one of the domains as input and translate it to the other domain. To accomplish this, the generators have an encoder-decoder architecture, where the input image is initially down-sampled to a latent space that goes through a series of ResNet blocks [13], followed by an upsampling

to the size of the output image. The architecture of the generators is depicted in Figure 3.

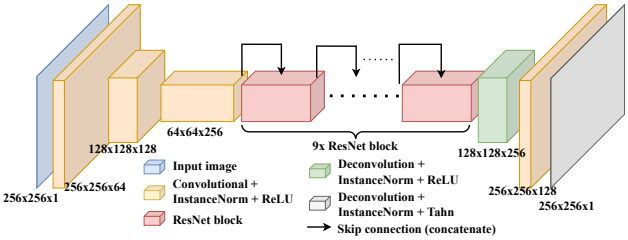

Figure 3: Schematic representation of the architecture of the generators. The numbers below the boxes represent the image size at its output. The ResNet blocks are made of a Convolution-InstanceNorm-ReLU block followed by a Convolution-InstanceNorm block. The convolutions of the ResNet block have a kernel size of 3x3 and a 1x1 stride. The input of each ResNet block is concatenated to its output.

To train the model, it is necessary to provide data from both domains. Since the data does not need to be paired, we just need to ensure that it is representative of the domain. For the SAR image domain, the concise-SAR-Ship-Dataset and the complete SSDD training set are used for each of the models. For the binary label domain, it is required to produce two new datasets, $\mathcal{D}_{label\_SSD}$ and $\mathcal{D}_{label\_SSDD}$, with binary ship segmentation images. To obtain these images, we apply the spectral residual approach for saliency detection [14] followed by Otsu's thresholding to low entropy SAR images, given that we are guaranteed to obtain finer and more accurate segmentation images from this method for these simpler images. For the SAR-Ship-Dataset, we apply this method to the 7000 images with the lowest entropy. For the SSDD, we apply the method to the 300 lowest entropy images, and then augment the results until they match the number of SAR images. A combination of scaling, translating, rotating and elastic deformation [15] is used for the data augmentation. Additionally, the obtained binary images are post-processed using flood fill.

3.2.2 UDSEP

The second proposed framework, the UDSEP (U-net Detect-Select-Erase-Paste), is a self-supervised segmentation method. In this approach, we first generate synthetic labelled images from the original SAR unlabelled images. Then we use the synthetic images and the corresponding masks as training set for semantic segmentation with the U-net. Since the U-net is a supervised method, and in order to avoid generating the segmentation masks manually, we propose a novel algorithm to generate the new SAR images with the corresponding masks, the DSEP (Detect-Select-Erase-Paste) method. The overall framework of the UDSEP method is depicted in Figure 4. The framework is divided into a training phase (a) and an inference phase (b). The training starts with the generation of the synthetic labelled images where the DSEP method takes as input a SAR image x, and outputs a pair of images x_1 and m_1 where x_1 is a new SAR image and m_1 its

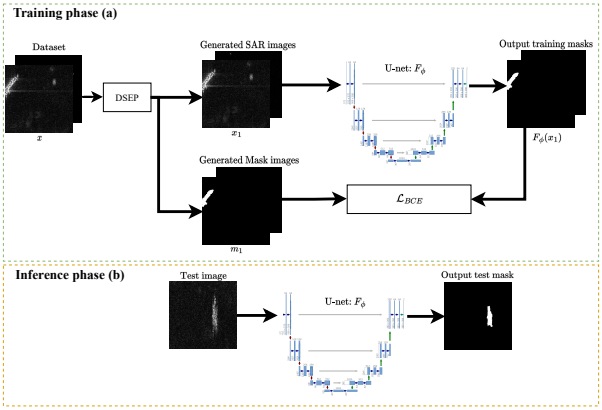

Figure 4: Framework for the UDSEP method.

segmentation mask. This process is repeated for each image in the training set. Then, the generated image pairs are used to train the U-net, F_ϕ with parameters ϕ, which is optimised to minimise the BCE loss (equation 1) between the output training masks and the target generated masks. After training is complete, the optimised U-net is used directly to obtain segmentations of the test set. The detailed architecture of the U-net is shown in Figure 5. Similar to the CycleGAN approach, a separate UDSEP model is trained for each dataset.

To generate the labelled SAR images from the original unlabelled data we introduced the DSEP method. Named after its four main steps: Detect, Select, Erase, and Paste, the DSEP is an unsupervised augmentation process that receives as input an image with objects of the same type and transforms it to a new image, obtaining the corresponding binary segmentation mask with the location of the objects. The overall pipeline of the method is depicted in Figure 6. In a concise manner, the DSEP method consists of the following steps:

- **Detect** the objects in an image.
- **Select** which of the detected objects to keep in the image and add their segmentation to the mask.
- **Erase** the objects that were chosen not to keep in the original image, by covering them with background.
- Optionally **Paste** augmented versions of the detected objects randomly in the image where the objects were erased, adding their segmentation to the mask.

For the detection step, we use the previously trained SAR to label generator from the CycleGAN model. To select which objects to keep in the image and which ones to cover, we start by rejecting objects whose size deviates heavily from the mean, i.e., extremely large or small objects. Then, we randomly keep up to N_{keep} objects. The rest of the objects are erased. To erase the objects, we search the original image to find regions big enough for cutting that were not detected by the generator, i.e., regions of background, and paste them at the locations of the objects to cover. The Select and Erase steps were introduced to mitigate the damage that would come from poor initial detection. Therefore, if the generator performs poorly and unduly

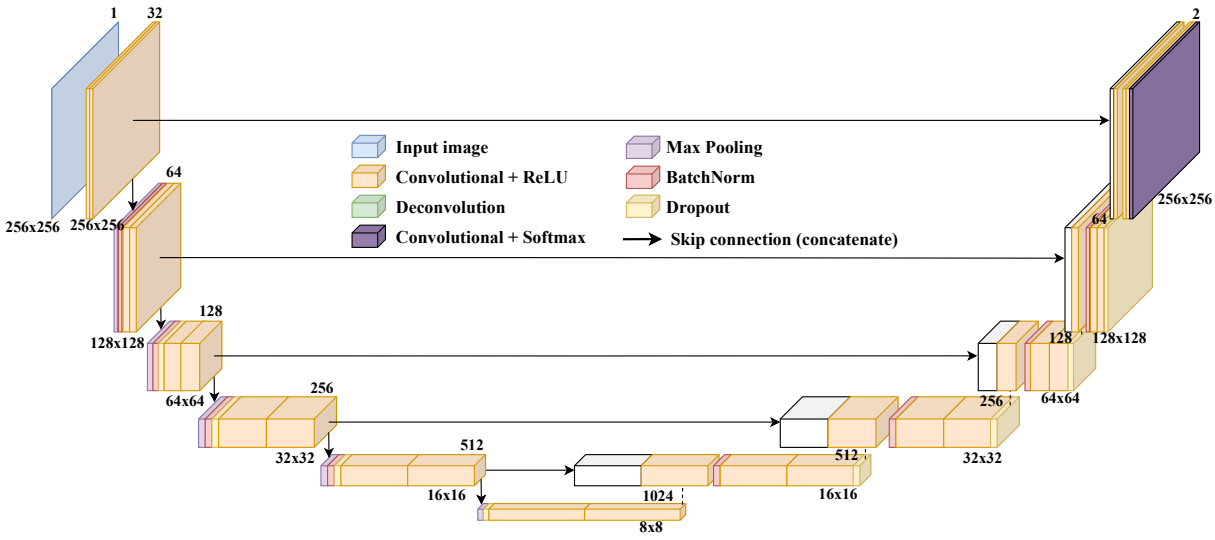

Figure 5: Schematic representation of the architecture of the U-net. The number of channels at the output of the boxes is shown on top of them, and the size of the feature maps at the bottom.

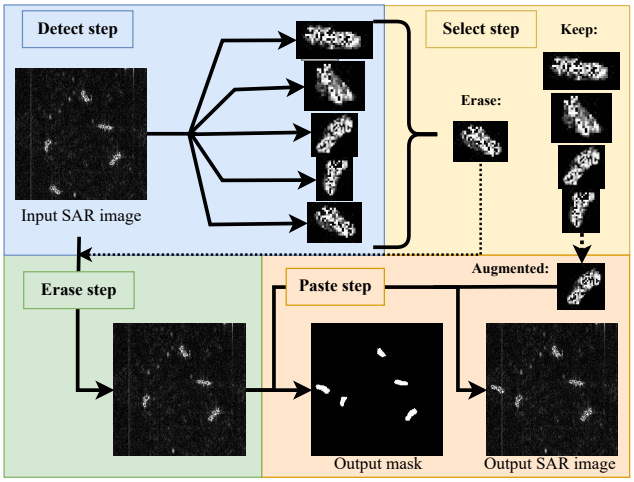

Figure 6: Schematic representation of the DSEP method.

detects a lot of objects, these steps make sure to mitigate the damage that would come from considering all those objects in the segmentation mask. Lastly, in the Paste step, there is a p_a chance of augmenting each of the kept objects and pasting them at a random location in the image that resulted from the Erase step and subsequently on the mask. The augmentation consists of rotation and soft intensity jitter. Therefore, the Paste step was introduced to attempt to increase the amount of information contained in the images without significantly altering them.

It is important to state that we aim to, as far as possible, maintain the structure of the original image. For instance, if the detected objects are all kept and if the Paste step is not employed, the new transformed image will in fact be equal to the original input image. Thus, in practice, the method will work only as a mask generator. Evidently, the quality of the images produced by the DSEP approach is highly dependent on the robustness of the initial object detector.

4. Results

This chapter presents the evaluation metrics and results for the proposed methods as well as comparison methods. First, the data generated as a training set for the models will be analysed. Then, the segmentation and object detection results of the proposed methods will be compared to those of two conventional unsupervised methods, Saliency and CFAR, as well as a supervised U-net. The Saliency approach consists of a spectral residual approach for saliency detection followed by Otsu's thresholding. The CFAR method is a two-parameter CFAR algorithm based on Rayleigh distribution and morphological processing [16]. Both the mentioned methods are traditional unsupervised approaches that do not require any training and, thus, are directly applied to the test set. The Supervised U-net has the architecture of Figure 5 and is trained directly with the supervised ship segmentations masks that were provided for the SSDD and obtained for the concise-SAR-Ship-Dataset. For the CycleGAN, and following the paper recommendation [8], the cycle loss is given ten times the weight of the adversarial loss and double the weight of the identity loss, thus $\lambda = 10$ and $\alpha = 5$ (equation 2). For the UDSEP, p_a is set to 0.2 to have a considerable chance of augmenting the ships, and N_{keep} is set to 4 since we have a moderate level of confidence in the CycleGAN's generator detections.

In addition, a discussion and analysis of the results, an ablation study of the UDSEP method, and a computation evaluation are provided. All the deep learning algorithms were implemented with Python 3.9.10, Tensorflow 2.6.2, Cuda 11.6, Intel(R) Core(TM) i5-7600K CPU @ 3.80GHz, and NVIDIA GeForce GTX 1070 GPU.

4.1. Evaluation Metrics

The methods are evaluated with segmentation metrics and object detection metrics. For the segmentation

evaluation, the IoU (Intersection over Union) and the F1-score for the ship class are computed pixel-wise. The IoU (equation 3) is the area of overlap between the ships' ground truth masks and their prediction masks divided by the area of union between the ships' ground truth masks and their prediction masks. The F1-score (equation 4) is a single evaluation metric that combines precision and recall by taking their harmonic mean. Precision (equation 5) refers to the proportion of correctly assigned ship pixels across all segmentation results, while recall (equation 6) refers to the proportion of correctly segmented ship pixels across all ground truth ship pixels. TP, FP, and FN stand for the pixel number of true positives, false positives, and false negatives, respectively.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of Union}} \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

For the detection evaluation, the F1-score is calculated for the 0.5 IoU threshold. We define this metric as $\text{F1}_{0.5}$. The equations are similar to those used for segmentation evaluation, with the exception that TP, FN, and FN are no longer defined by pixels but by objects. A ship detection is considered a true positive if the IoU value between the ground truth mask and its prediction is greater or equal to 0.5. If the IoU value is lower than 0.5, the detection is considered a false negative. A false positive occurs if there is a detection with no corresponding object in the ground truth.

4.2. Experimental Results and Analysis

4.2.1 Generated Data Analysis

$\mathcal{D}_{label\_SSD}$ & $\mathcal{D}_{label\_SSDD}$: Figure 7 shows several examples of the obtained binary ship segmentation images that make up $\mathcal{D}_{label\_SSD}$ and $\mathcal{D}_{label\_SSDD}$. The proposed saliency-threshold method was able to successfully generate binary segmentation masks from simple SAR images where the sea is calm and there is good contrast between ships and sea. Furthermore, the augmentations implemented for the SSDD also revealed valid results.

DSEP: Figure 8 shows a series of examples of the DSEP transformations. As can be seen from the first two rows of Figure 8 (a) and (b), the method performs extremely well for offshore input images, even if they have bright, noisy backgrounds. In these cases, the initial object detector usually has fewer than 4 detections. Therefore, the method simply works as a mask generator, and given the success of the initial object detector, the obtained image pairs are comparable to those of the desired supervised method. Furthermore, sporadic augmentations that seem to improve the amount of information in the images can also be observed. For inshore images, the quality of the image pairs generated considerably deteriorates. This was already expected given the limitations of the initial CycleGAN based object detector. Nonetheless, it is on these images where we can better see the impact of the Select and Erase steps. For example, in the third row from Figure 8 (a), due to the complexity of the input image, the initial object detector identified 6 objects when the ground truth only indicates the existence of one ship. Since we defined 4 as the maximum number of objects to keep, 2 of those detected objects were covered, avoiding training the U-net with that presumably inadequate labelled data. Although we did not cover all the non-ship objects, we managed to minimise the impact of the poor CycleGAN detection. Moreover, there is always a chance to cover a ship, but we believe that unduly covering a ship from an image should have less negative impact on the network than training it with unlabelled ships. It is important to state that the overall effectiveness of the DSEP method owes a great deal to the CycleGAN generator's robustness.

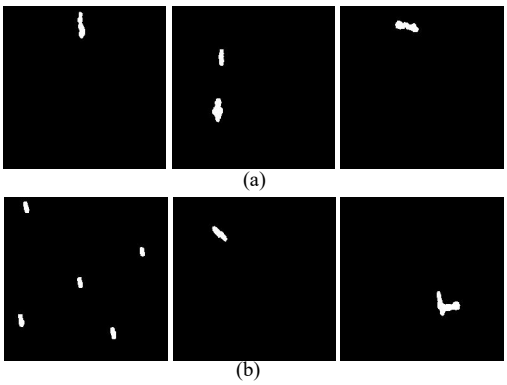

Figure 7: Binary ship segmentation masks obtained with the saliency-threshold method. (a) SAR-Ship-Dataset, (b) SSDD.

4.2.2 Results on SAR-Ship-Dataset

Table 1 presents the pixel-wise IoU and F1-score for the methods described above.

Table 1: Segmentation results for the SAR-Ship-Dataset.

Method	IoU	F1-score
Supervised:		
U-net	0.773	0.841
Unsupervised:		
Saliency	0.551	0.663
CFAR	0.441	0.564
CycleGAN	0.627	0.734
UDSEP	**0.630**	**0.737**

To provide a more in-depth understanding for these results, several segmentation results for images from the test set is represented in Figure 9. Table 2 presents the detection $\text{F1}_{0.5}$ for the complete test set, and computed separately for simple and complex images.

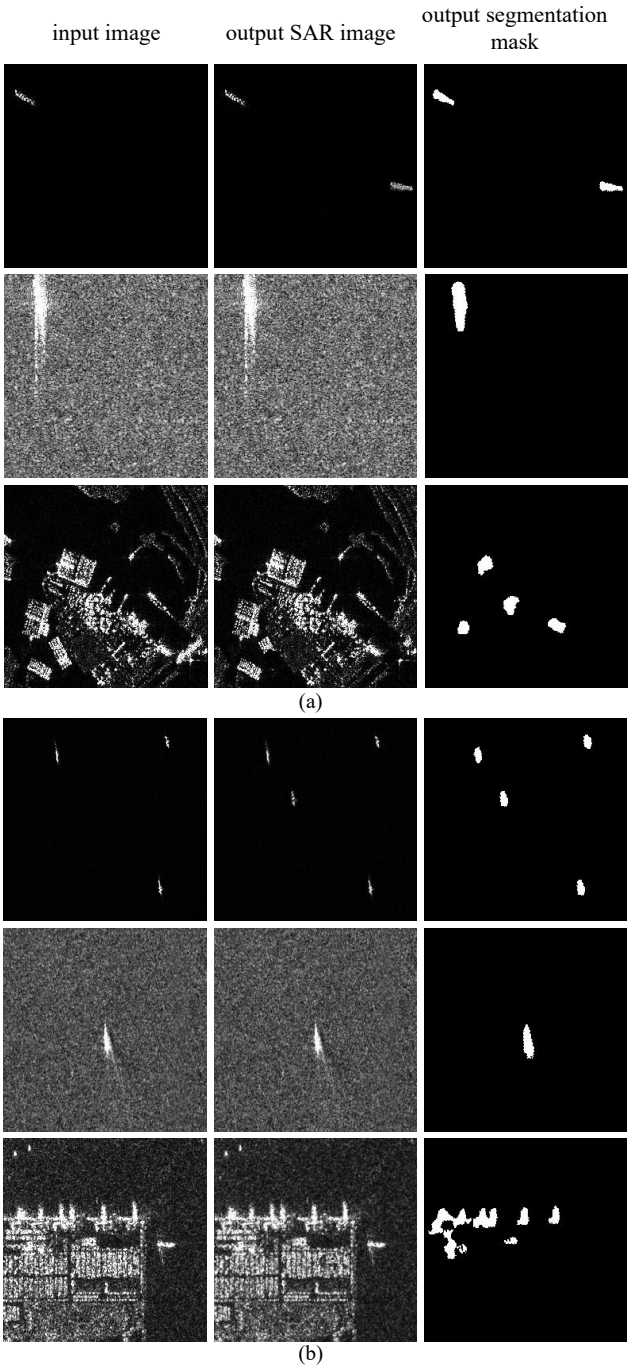

Figure 8: Original input SAR images and the result of the DSEP method: SAR image and its segmentation mask. (a) SAR-Ship-Dataset, (b) SSDD.

Table 2: Detection results for the SAR-Ship-Dataset.

Method	$F1_{0.5}$	$F1_{0.5}$ for simple images	$F1_{0.5}$ for complex images
Supervised:			
U-net	0.859	0.947	0.740
Unsupervised:			
Saliency	0.187	0.715	0.0758
CFAR	0.574	0.704	0.420
CycleGAN	0.706	0.912	0.461
UDSEP	**0.730**	**0.930**	**0.478**

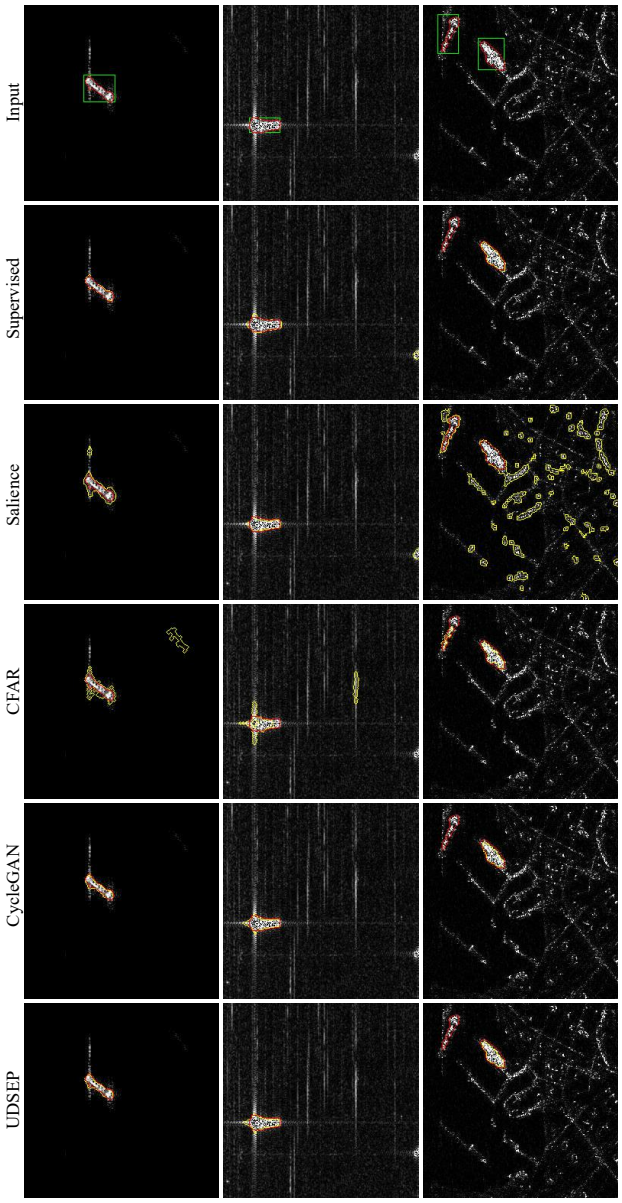

Figure 9: Segmentation results for test images. The original bounding box is represented in green, the ground truth segmentation in red, and the predictions in yellow.

First off, it should come as no surprise that the supervised method outperformed all the remaining methods, which are all unsupervised. Moreover, the good results for the supervised method validate the U-net as the feature extractor for SAR ship semantic segmentation. Furthermore, deep learning techniques performed significantly better than conventional techniques. Among the proposed methods, the UDSEP marginally outperformed the CycleGAN. Given that the CycleGAN generator serves as the foundation for the UDSEP, the similarity of these results between the proposed methods is not surprising. Nonetheless, the integration of the DSEP transformation method with the U-net improved the results of the original CycleGAN. The ablation study provided later in this section will help to better understand the impact of these additions. Moreover, the results of the proposed methods are still considerably lower than those of the supervised method.

When analysing the results separately for simple and complex images, several conclusions can be made. First, it is possible to notice that all methods perform reasonably well for simple images. Traditional methods usually have bigger segmentation masks than ground truth masks and, occasionally, have false detections. Deep learning methods detect the vast majority of ships, with few to no false and missed detections. In fact, for simple images, the results between the proposed unsupervised methods and the supervised method are very comparable.

Furthermore, not only for the proposed methods but also for the comparison methods, there is a significant decline in results for the complex images. This was already expected, given the high degree of similarity between the ships and the background, which may include islands, harbors, noise, etc. This inevitably leads to more false positives. Nonetheless, there is a significantly higher gap in performance between the proposed and the supervised method for the complex images. There are some factors that can account for the lower performance of the proposed methods. First, when training the CycleGAN, even with the efforts of obtaining a concise training set with high image diversity, there is still a class imbalance in the training data. This is due to the fact that there are considerably more simple-to-medium-complex images than complex ones. For this reason, the CycleGAN will likely learn the mapping between simple images and the segmentation domain more effectively. Moreover, being an unsupervised method in which we simply fed images from the two domains, other than the shape of the provided segmentation images, there is nothing that is forcing the CycleGAN to learn to differentiate between shore and ships. For these reasons, and since the CycleGAN generator is directly related to the performance of both proposed methods, it is normal for the results to be worse for complex images. Nevertheless, the proposed methods still obtained satisfactory results for numerous complex images. Moreover, it is possible to observe that the saliency method differs from other approaches in that it produces good results for simple images but terrible for complex ones. This is the reason that allowed the good extraction of the segmentation masks for $\mathcal{D}_{label\_SSD}$ and $\mathcal{D}_{label\_SSDD}$, but only after a preselection of low entropy images.

Ablation study: To further understand the impact of the components of the DSEP method, we conducted an ablation study on the UDSEP. Table 3 presents the conditions and the object detection results of the carried experiments. Case 0 represents the original UDSEP method. Case 5 corresponds to the UDSEP method but using the Saliency as the initial object detector instead of the CycleGAN generator. In case 1, we directly use the CycleGAN predictions to train the U-net, skipping the DSEP transformations. The same is done in case 6, but using the Saliency predictions instead. In the remaining cases, some components of the DSEP transformation method are ignored.

By analysing the outcomes of cases 0 through 4, it is extremely difficult to draw conclusions about the utility of the DSEP approach. Despite the fact that applying all the stages (case 0) results in a slightly better model, the performance of the remaining models is still extremely similar to the original. This can be explained for two reasons. First, given that the CycleGAN generator is already a highly robust object extractor, there are typically less than 4 object detections in the majority of the images. Therefore, the Select and Erase steps, which were introduced as insurance, are unnecessary. Then, the SAR-Ship-dataset is very big, hence the augmentations provided by the Paste step are not that crucial. Therefore, although in our case the DSEP is not that relevant, we aimed at proposing a generic approach that could be used for a variety of scenarios. To validate the relevance of the DSEP, we implemented it with an initial less robust object detector, the Saliency (case 5). In this case, the DSEP method managed to increase the F1-score at the 0.5 threshold by 0.23 points when compared with the model where it was not implemented (case 6), and by 0.38 when compared to the original model where the U-net was not used (Saliency in Table 2), indicating substantial improvements.

Table 3: Conditions of the UDSEP ablation experiment.

Case	Detect	Select	Erase	Paste	$F1_{0.5}$
Case 0	$G_{SAR\_to\_L}$	✓	✓	✓	**0.730**
Case 1	$G_{SAR\_to\_L}$	✗	✗	✗	0.714
Case 2	$G_{SAR\_to\_L}$	✓	✓	✗	0.710
Case 3	$G_{SAR\_to\_L}$	✗	✗	✓	0.715
Case 4	$G_{SAR\_to\_L}$	✓	✗	✓	0.710
Case 5	Saliency	✓	✓	✓	0.563
Case 6	Saliency	✗	✗	✗	0.332

Computation Evaluation: Table 4 presents training and test time for all methods.

Table 4: Training time and inference time per image. *The time taken to generate the images to train the models is accounted for in the training time.

Method	Training time (hours)	Avg. inference time p/ image (ms)
Supervised U-net	1.3	34.9
Saliency	-	1.7
CFAR	-	1.08×10^3
UDSEP*	9	34.9
CycleGAN*	60	82.1

Considering the deep learning models, the CycleGAN model took considerably longer to train, which was expected, given its increased model and training complexity. However, training time is not the most crucial factor. In fact, the inference time is more important given the goal of processing the SAR data in real-time. The U-net-based methods have significantly faster detection speeds than the CycleGAN. This is due to the large model of the CycleGAN generators, which includes a series of ResNet blocks, inherently leading to more expensive computations. Furthermore, the detection speed of the proposed models is lower than the state-of-the-art, especially for the Cy-

cleGAN. Nonetheless, it is important to state that the UDSEP not only managed to slightly increase the detection performance of the CycleGAN but also managed to transfer its knowledge to a U-net which has 2.35 faster detection times.

4.2.3 Results on SSDD

Table 5 presents the pixel-wise IoU and F1-score for the methods described. Several segmentation results are represented in Figure 10. Moreover, Table 6 presents the detection $F1_{0.5}$ for the complete test set, and computed separately for offshore and inshore images.

Table 5: Segmentation results for the SSDD.

Method	IoU	F1-score
Supervised:		
U-net	0.763	0.857
Unsupervised:		
Saliency	0.466	0.585
CFAR	0.483	0.624
CycleGAN	0.554	0.676
UDSEP	**0.571**	**0.693**

Table 6: Detection results for the SSDD.

Method	$F1_{0.5}$	Offshore $F1_{0.5}$	Inshore $F1_{0.5}$
Supervised:			
U-net	0.889	0.952	0.723
Unsupervised:			
Saliency	0.233	0.282	0.109
CFAR	0.367	0.409	**0.277**
CycleGAN	0.578	0.801	0.167
UDSEP	**0.625**	**0.833**	0.257

The results are consistent with the previous data set. The supervised method clearly outperformed the remaining unsupervised methods, and the deep learning methods outperformed the traditional methods. Of the proposed methods, the UDSEP obtained better segmentation and detection results. In light of the fact that the CycleGAN generator has lower performance for this dataset, the improvements of the DSEP method are more noticeable. In addition, the SSDD is a considerably small dataset, thus the augmentations provided by the Paste step likely had a bigger impact than on the SAR-Ship-Dataset. Moreover, the UDSEP managed to clearly outperform the CycleGAN model for inshore images. This is likely due to mitigations endorsed by the Select and Erase steps, which encouraged the U-net to not have as many false detections as it otherwise would have.

Overall, the results for the proposed methods on the SSDD dataset were worse than on the SAR-Ship-Dataset. Several factors support this conclusion. First, the SSDD has a significantly smaller size, which will unavoidably result in models being more vulnerable to overfitting. Then, the dataset is mainly composed of ships of small size. Since the IoU becomes more sensitive as the area of the object decreases, slight discrepancies between the ground truth and the prediction might result in low IoU values for small objects.

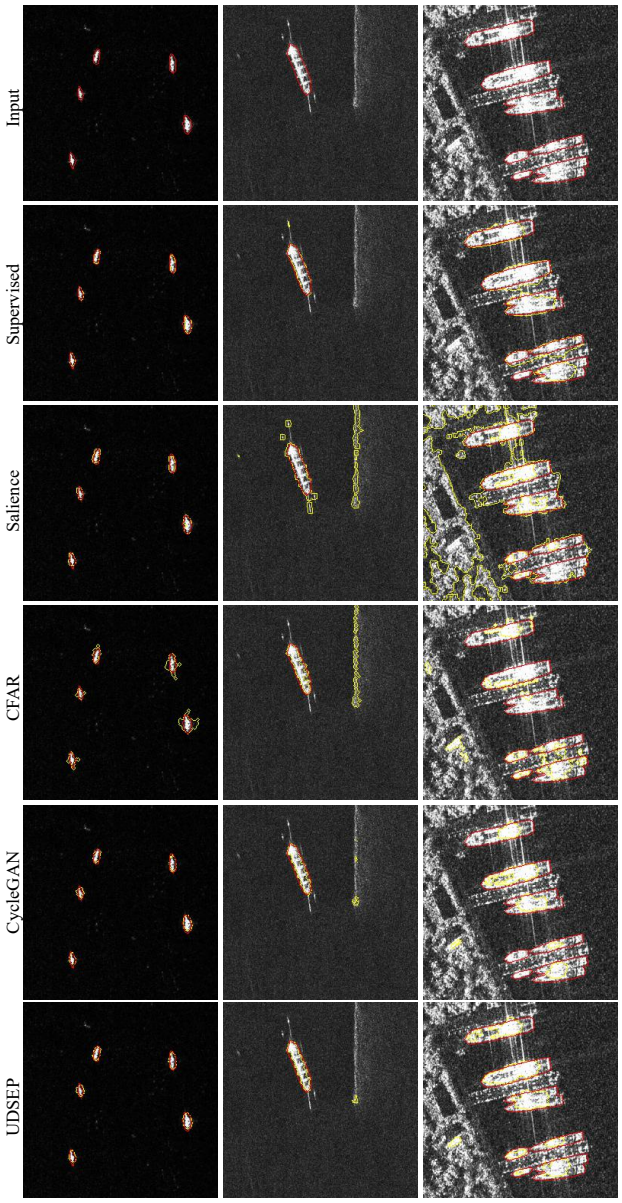

Figure 10: Segmentation results for test images. The ground truth segmentation is represented in red and the predictions in yellow.

Moreover, smaller objects are naturally harder to detect, given that their features may disappear in deeper layers. Nevertheless, the models performed reasonably well for the SSDD, validating the generalisability of the presented methodologies.

5. Conclusions

The main goal of this thesis was to develop unsupervised deep learning techniques for ship detection in SAR images. For this purpose, two fully unsupervised frameworks were proposed for ship segmentation: the CycleGAN, an image-to-image translation model which was explored for segmentation, and the UDSEP, a U-net trained on synthetic generated data from a novel augmentation process. Although still inferior to those of the supervised method, the results obtained for the two proposed methods were extremely promising, especially given the full unsupervised nature of the approaches. Evaluation on simple/offshore images revealed overall competitiveness with the su-

pervised method. Evaluation on complex/inshore images proved that the proposed methods are still insufficiently robust for this type of image. However, it is important to state that ship detection in this type of image is an active challenge, even for supervised research. Given their essentially inferior robustness, the struggle for unsupervised approaches is not surprising.

Moreover, the CycleGAN approach revealed to be effective and robust for domain translation between the SAR domain and the ship segmentation domain. Consequently, the UDSEP managed to enhance the CycleGAN model in two aspects. First, there was a slight improvement in detection quality. Then, there was a severe reduction in detection time, with a decrease of over 57%.

Furthermore, the author believes that the developed work should inspire fellow researchers to develop unsupervised frameworks for SAR ship detection, which can fully exploit the amount of available raw SAR data and the increasing GPU performance.

Future improvements could be made to attempt to improve the accuracy of the models. First, further studies could be employed to attempt to increase the robustness of the models. For instance, the CycleGAN could benefit from a distinct architecture for each of the generators, which could be more task-specific to the domain translation. Moreover, the DSEP method could benefit from improvements in the Select step. Currently, after a size-dependent preliminary removal, the selection of the objects to keep and to erase is essentially random. Several unsupervised strategies could be employed to attempt to address this. For example, a binary cluster could be conducted by K-means to try to classify each object as a ship or non-ship. Then, objects classified as ships would be kept, and objects classified as non-ships would be marked to be erased.

Second, the low accuracy of the inshore scene should be addressed. For the CycleGAN, resolving the class imbalance between the offshore and inshore images could be a good start. The strategy introduced by [17], which used GAN and K-means to create a scene binary cluster and then augmented the inshore scene images, would be a good approach to augment these images in an unsupervised manner. The UDSEP method would indirectly benefit from this improvement.

Lastly, in an effort to make models suitable for real-time detection, the original backbone structures of the models could be replaced with lightweight versions. To this end, we propose not to change the CycleGAN model but experiment with lightweight versions of the U-net to check if it would still be possible to fully transfer the knowledge obtained with the current CycleGAN model. Furthermore, to further test and refine the proposed methods, it would be of interest to evaluate them in other segmentation tasks using datasets other than SAR ship.

References

[1] Geospatial Intelligence Pty Ltd, Maritime surveillance in focus, `https://storymaps.arcgis.com/stories/f03366c92d14470c982bf17c46e21308`, accessed: 2022-06-15 (2021).

[2] C. Oliver, S. Quegan, Understanding Synthetic Aperture Radar Images, EngineeringPro collection, SciTech Publ., 2004.

[3] N. Ferreira, M. Silveira, Ship Detection in SAR Images Using Convolutional Variational Autoencoders, in: IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 2503–2506.

[4] P. Dias, Unsupervised Ship Detection in SAR Images using Generative Adversarial Networks, Master's thesis, Instituto Superior Técnico, University of Lisbon (2022).

[5] J. Wang, Z. Wen, Y. Lu, X. Wang, Q. Pan, Weakly Supervised SAR Ship Segmentation Based on Variational Gaussian G(A)(0) Mixture Model A Learning, in: 2020 Chinese Automation Congress (CAC), 2020, pp. 6072–6077.

[6] F. Gu, H. Zhang, C. Wang, B. Zhang, Weakly supervised ship detection from SAR images based on a three-component CNN-CAM-CRF model, Journal of Applied Remote Sensing 14 (2) (2020) 026506.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W. M. Wells, A. F. Frangi (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.

[8] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242–2251.

[9] P. Isola, J. Zhu, T. Zhou, A. A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, CoRR abs/1611.07004 (2016).

[10] Y. Wang, C. Wang, H. Zhang, Y. Dong, S. Wei, A SAR Dataset of Ship Detection for Deep Learning under Complex Backgrounds, Remote Sensing 11 (7) (2019).

[11] C. E. Shannon, A Mathematical Theory of Communication, The Bell System Technical Journal 27 (1948) 379–423.

[12] T. Zhang, X. Zhang, J. Li, X. Xu, B. Wang, X. Zhan, Y. Xu, X. Ke, T. Zeng, H. Su, I. Ahmad, D. Pan, C. Liu, Y. Zhou, J. Shi, S. Wei, SAR Ship Detection Dataset (SSDD): Official Release and Comprehensive Data Analysis, Remote Sensing 13 (18) (2021).

[13] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[14] X. Hou, L. Zhang, Saliency Detection: A Spectral Residual Approach, IEEE Conference in Computer Vision and Pattern Recognition (2007).

[15] E. Castro, J. S. Cardoso, J. C. Pereira, Elastic deformations for data augmentation in breast cancer mass detection, in: 2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 2018, pp. 230–234.

[16] R. Wu, Two-Parameter CFAR Ship Detection Algorithm Based on Rayleigh Distribution in SAR Images, Preprints (2021).

[17] T. Zhang, X. Zhang, J. Shi, S. Wei, J. Wang, J. Li, H. Su, Y. Zhou, Balance Scene Learning Mechanism for Offshore and Inshore Ship Detection in SAR Images, IEEE Geoscience and Remote Sensing Letters 19 (2022) 1–5.