Guided Sonar-to-Satellite Translation (Accepted in JINT)

Preprint · January 2021

CITATIONS
0

READS
91

4 authors:

Giovanni Gatti De Giacomo
Universidade Federal do Rio Grande (FURG)
9 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Matheus dos Santos
Universidade Federal do Rio Grande (FURG)
19 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Paulo Drews-Jr
Universidade Federal do Rio Grande (FURG)
170 PUBLICATIONS **1,315** CITATIONS
SEE PROFILE

Silvia Silva da Costa Botelho
Universidade Federal do Rio Grande (FURG)
365 PUBLICATIONS **1,871** CITATIONS
SEE PROFILE

JINT manuscript No.
(will be inserted by the editor)

Guided Sonar-to-Satellite Translation

Giovanni G. De Giacomo · Matheus M. dos Santos · Paulo L. J. Drews-Jr · Silvia S. C. Botelho

Received: date / Accepted: date

Abstract Underwater navigation and localization are greatly enhanced by the use of acoustic images. However, such images are of difficult interpretation. Contrarily, aerial images are easier to interpret, but require Global Positioning System (GPS) sensors. Due to absorption phenomena, GPS sensors are unavailable in underwater environments. Thus, we propose a method to translate sonar images acquired underwater to an aerial counterpart. This process is called sonar-to-satellite translation. To perform the conversion, a U-Net based neural network is proposed, enhanced with state-of-the-art techniques, such as dilated convolutions and guided filters. Afterwards, our approach is validated on two datasets containing sonar images and their satellite analogue. Qualitative experimental results indicate that the proposed method can transfer features from acoustic images to aerial images, generating satellite images that are easier to interpret and visualize.

Keywords Deep Learning · Neural Networks · Robotics

PACS 68 Computer Science

Mathematics Subject Classification (2010) MSC 68T45 · MSC 68T40 · MSC 68T30

Giovanni G. De Giacomo (Corresponding Author)
Federal University of Rio Grande (FURG)
Computer Science Center
Av. Itália Km 8
96203-900
Rio Grande, RS, Brazil
Tel.: +55 (53) 99951-9892
ORCID: 0000-0003-1670-2536
E-mail: ggiacomo@furg.br

Matheus M. dos Santos, Paulo L. J. Drews-Jr, Silvia S. C. Botelho
Federal University of Rio Grande (FURG)
Computer Science Center
ORCID: 0000-0002-6036-8670, 0000-0002-7519-0502, 0000-0002-8857-0221
E-mail: matheusmachado@furg.br, paulodrews@furg.br, silviacb@furg.br

1 Introduction

Converting images into other images, even when dealing with different domains, is an exciting problem that lies at the core of many Machine Learning applications. Performing that conversion is known as image-to-image translation. A general U-Net based Convolutional Neural Network (CNN) was proposed by Isola et al. (2017) to solve it.

Water has physical properties that make it difficult to work with underwater robots, such as Autonomous Underwater Vehicles (AUVs). Mainly, the malfunction of light based sensors, *e.g.* cameras and lasers, and Global Positioning System (GPS) sensors. These malfunctions happen because of the rapid attenuation that electromagnetic waves undergo below water. Therefore, underwater robots traditionally use sonar images as their preferred source of reliable input. However, acoustic images require intensive processing to extract information, due to phenomena such as noise. As a consequence, underwater robot localization and navigation, such as attempted in Dos Santos et al. (2019b,a), is an area that is requiring new and effective methods.

This paper proposes facilitating the interpretation of sonar images through aerial images when operating in underwater environments. This proposal consists in the translation of an acoustic image into a satellite counterpart. Giacomo et al. defined sonar-to-satellite translation as the task of converting an acoustic image into an aerial one. The objective is that satellite images generated by such methods could be used to perform matching with authentic satellite images and locate a robot without needing a GPS sensor. Figure 1 shows a schematic of the sonar-to-satellite translation problem when solved by a CNN.

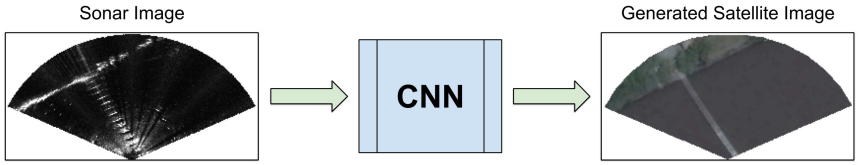


Fig. 1: Sonar-To-Satellite Translation is defined as the conversion of the acoustic image in the left to the satellite image in the right. This diagram also shows a CNN as a solution to the problem.

This work is an extension of Giacomo et al. (2018), that defined sonar-to-satellite translation and presented results with a pix2pix CNN. The method was extended by using a new architecture that includes dilated convolutions from Yu and Koltun (2015); Steffens et al. (2019, 2020) and guided filters from He et al. (2013) in the generator, as well as using more sophisticated loss functions, such as the style reconstruction loss by Johnson et al. (2016). Also, a whole new dataset was included for testing our model. These extensions will be discussed further in later sections.

In the paper, a myriad of experiments are conducted with two local datasets, using a U-Net based neural network architecture augmented with guided layers

from Gonçalves et al. (2018) and with a conditional Generative Adversarial Network (cGAN). After that, the visual quality of the results is verified by doing side comparisons of them with the actual ground truth satellite images, as well as calculating image quality metrics. These ground truth images were obtained by using the data from the GPS and magnetometer sensors of the underwater robot.

The goal of the research is verifying if it is possible to diminish the difficulty in acquiring aerial images when navigating underwater. This difficulty is caused by the unreliability of GPS sensors in said environments. The proposed method attempts to solve the issue by translating sonar images into satellite ones, using a CNN.

This paper is organized in the following way: in the next section, we will discuss the related works; Then, we will introduce the two datasets used in the experiments and talk about their particularities; Afterwards, we will present the methodology used to attack the sonar-to-satellite problem; Subsequently, we will present the experimental results in the two datasets and discuss them. Finally, we will conclude and summarize our contributions.

2 Related Works

Other than Giacomo et al., using a CNN to extract an aerial image from a sonar one is an unprecedented concept. On the other hand, there are various related works upon which this paper is based. Therefore, in this section, papers on CNN, image filtering and general neural network techniques that inspired our model and research will be described. Articles about locating vehicles on land using satellite images will also be discussed, since a similar idea is being proposed on this paper, but for underwater domains.

Image-to-image translation is the task of translating one possible representation of a scene into another, as defined in Isola et al. (2017). To solve this problem, Isola et al. created a CNN architecture called pix2pix. This architecture is based on the U-Net network for medical segmentation that was proposed in Ronneberger et al. (2015). By adding a cGAN component to the network architecture and providing general hyperparameters, Isola et al. provided a standardized approach to solving image translation problems. In addition, Isola et al. performed various experiments using his proposed methodology, working with varied datasets that referred to several problems within the domain of Computer Vision (CV). Among the problems, there was one that aimed to convert aerial images into charts and that inspired the present work.

As one of the basis of our CNN architecture, the U-Net network, described in Ronneberger et al. (2015), is an important related work. In his paper, Ronneberger et al. first proposed the idea of skip connections. A skip connection is a step in a neural network where you feed a feature map from a backward layer into a forward layer in encoder-decoder architectures. Due to their proven capacity of improving learning by returning structures previously discarded by the network, skip connections have been largely used in many applications and neural network architectures, including our own. However, it is also important to mention that skip connections come with a large memory footprint, since they need the required feature maps from previous layers to be stored.

Yu and Koltun introduced dilated convolutions that are useful for aggregating contextual information without losing coverage or resolution. By making use of dilated convolutions, one is able to expand the receptive field dramatically. Therefore, the network is able to capture significantly higher amounts of context than it would with traditional convolutional layers.

He et al. proposed the guided filter, an image filter that outputs a locally linear transform of the guidance image. As detailed in its original paper, the guided filter has good edge-preserving properties. However, it does not suffer from gradient reversal artifacts, like the bilateral filter from Tomasi and Manduchi (1998). Also, unlike the bilateral and simple linear translation-invariant (LTI) filters, the guided filter can be used to transfer structure from the guidance image to the output. Due to the structure-transferring property of the guided filter, He et al. (2013) envisioned that it could be used in applications such as feathering, dehazing and high-quality stereo matching methods. Wu et al. (2018) published and made available an implementation for a fast end-to-end trainable guided filter.

Gonçalves et al. introduced GuidedNet, a model that used a new neural network layer, called guided layer. Guided layers use guided filters as a way to transfer structural information, which are partially lost due to convolutions, back into the output of the neural network. Although the model was initially proposed for image dehazing, it also works well for other tasks involving image generation or restoration.

Viswanathan et al. localizes a ground vehicle by using satellite images as a map. The method creates a feature database by splitting the satellite image in a grid and describing each cell. First, the ground-based panoramic images are warped into a top-down view of the scene. Then, the view is described and used as a query on the satellite database. Finally, a particle filter framework is integrated on the solution to estimate the vehicle position and orientation during the navigation. In Viswanathan et al. (2014), the proposal is validated with experimental tests that often shows better position estimates than the GPS.

Kim and Walter proposed a ground localization method using a learning embedding strategy. A CNN based on the Siamese architecture is used to extract a 4096-dimensional feature vector able to match ground-level imagery with their respective satellite view. Then, these matches serve as a noise observation of the position and orientation of the vehicle. These observations are then used into a particle filter that maintains a distribution on the pose during navigation.

Deng et al. proposed a method to generate ground level images from aerial images. Combined with the method proposed in this paper, it would be possible to produce an aerial image from an underwater acoustic image. Thereafter, a ground level image can be generated, using the method by Deng et al., and used for appropriate applications.

3 Datasets

To evaluate our model under different conditions and locations, two real-world datasets were used: datasets ARACATI 2014 and ARACATI 2017. These datasets were both captured in the Yacht Club of Rio Grande, Brazil. However, they were obtained in different places inside the Yacht Club, as well as in different years.

Therefore, the acoustic images contained in these datasets are substantially different, a fact you can verify in Figure 2.

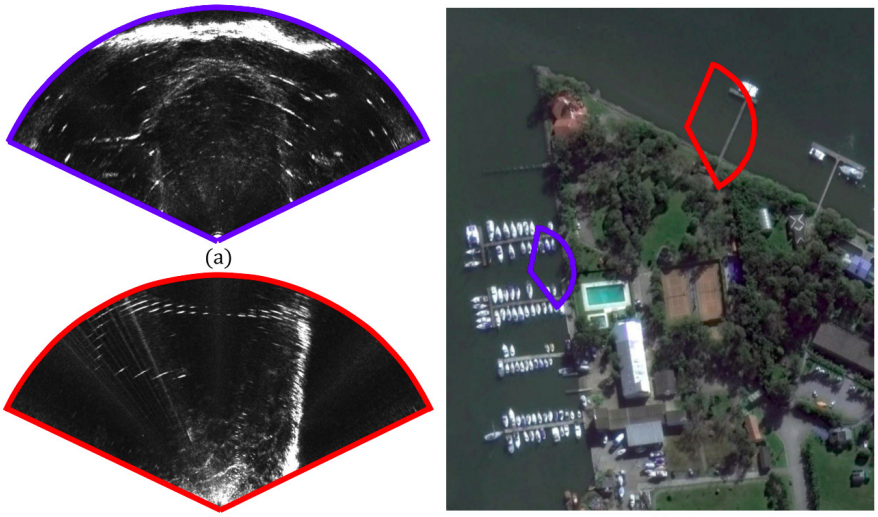


Fig. 2: 2a shows an example of acoustic image for the ARACATI 2014 dataset. 2b presents a sonar image from the ARACATI 2017 dataset. 2c displays a satellite image highlighting the places where the images 2a and 2b were captured. Satellite images from Google©, Digital Globe© 08-06-2017, 32°01'30.1"S 52°06'24.1"W.

Both datasets were recorded by a mini Remotely Operated Vehicle (ROV) Seabotix LBV-300 with a Teledyne BlueView P900-130 Multibeam Forward Looking Sonar (MFLS), a magnetic compass and a SOUTH S82T Differential Global Position System (DGPS). A floating board is attached on the vehicle so that it follows the vehicle and remains on the surface of the water during the trajectory. The DGPS is installed on top of the floating board and records the 2D vehicle position with high precision.

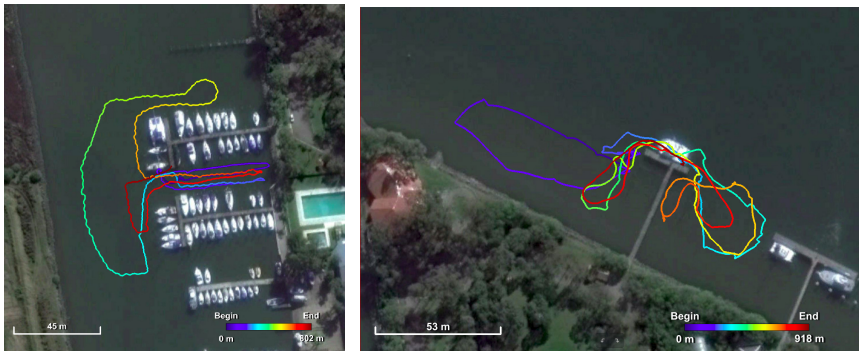
3.1 ARACATI 2014

This dataset was first published in the work of Silveira et al. (2015). In 75 minutes, the vehicle travels a total of 802 meters acquiring 10232 images. Figure 3a shows the path travelled by the robot. The MFLS was configured to cover a range of 30 meters. The dense presence of structures such as pier and boats are the main features of this dataset because of the place and the path travelled by the vehicle.

3.2 ARACATI 2017

As previously mentioned, this dataset was collected at one of the harbors of the Yacht Club of Rio Grande, Brazil by an underwater robot. The MFLS was con-

figured to cover a range of 50 meters. Figure 3b shows the path travelled by the vehicle, alongside further information regarding the length of the voyage. In total 24676 images were captured in 77 minutes. Unlike ARACATI 2014, the main characteristic of this dataset is the sparse presence of structures that involves a smaller area on the images because of the increased coverage range of the sonar.



(a) (b)

Fig. 3: Robot path of the adopted datasets. 3a shows the path of ARACATI 2014. 3b shows the path of ARACATI 2017. Satellite images from Google©, Digital Globe© 08-06-2017, (a) 32°01'33.7"S 52°06'30.7"W (b) 32°01'30.1"S 52°06'24.1"W.

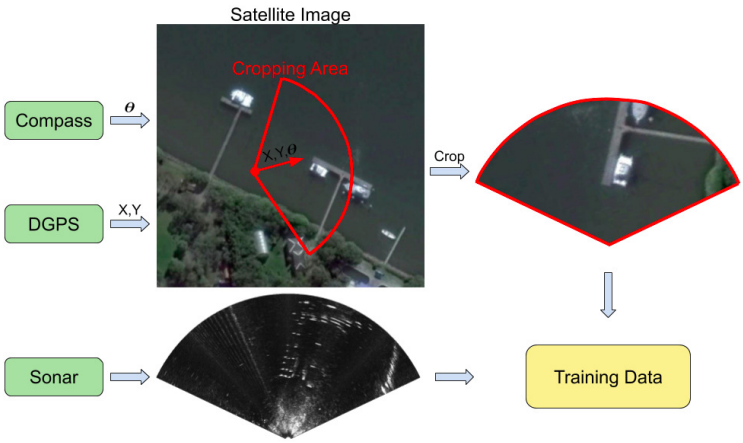
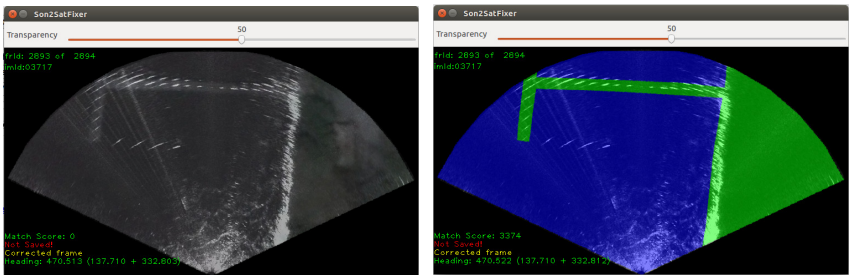


Fig. 4: Diagram for data preprocessing workflow which generates the ground truth data. Satellite images from Google©, Digital Globe© 08-06-2017, 32°01'30.1"S 52°06'24.1"W.

3.3 Data Preprocessing

To create the training dataset, a satellite image of the Yacht Club provided by Google Earth was used. The satellite image is automatically cropped considering the position from the DGPS, heading from the magnetic compass and the coverage field of each acoustic image[1] as shown in Figure 4.



(a) (b)

Fig. 5: Manual compass correction tool. Each satellite image is manually rotated by using mouse commands until it correctly matches the correspondent sonar image.

After initial processing, problems were discovered with the compass data. Some cropped satellite images did not correctly match with the sonar images, worsening the learning process of the neural network. In order to fix this problem, a tool was developed for manual correction of the compass data. Figure 5 displays the interface of the tool.

A fixed offset is not enough to solve the misalignment problem of all images because the compass is affected by magnetic interferences by the ship hulls or even by the vehicle motor. Therefore, each image had to be manually corrected.

An image selection criteria was adopted where images with a time-stamp difference lower than 0.13 seconds in the DGPS and compass data were selected. This procedure resulted in 2894 images. Since DGPS, compass and sonar images have different acquisition rate, the criteria ensures a selection of the most synchronized data. Figure 6 shows the selected images partially cover the entire dataset.

After preprocessing of the datasets, the ARACATI 2017 dataset contained 2894 pairs of acoustic and ground truth satellite images that were used for training purposes. On the other hand, the ARACATI 2014 dataset contained 839 pairs of acoustic and ground truth satellite images used exclusively for testing purposes.

[1] A video showcasing the cropping of the dataset is available at https://youtu.be/92yNiGjQLLo.

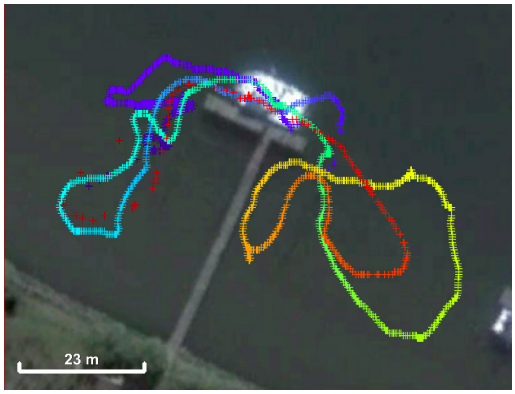


Fig. 6: The position of the 2894 selected images that were manually corrected. Image provided by Google©, Digital Globe© 08-06-2017, 32°01'30.1"S 52°06'24.1"W.

4 Methodology

The formal definition of sonar-to-satellite translation is a function $\mathcal{G}\colon \mathbb{R}^{H\times W} \rightarrow \mathbb{R}^{C\times H\times W}$, where $\mathcal{G}$ is, in this case, a generative CNN and C, H, W are the depth, height and width of the satellite image, respectively.

To attack the sonar-to-satellite problem, this section proposes a trainable end-to-end CNN, using state-of-the-art techniques from the Deep Learning literature. A generator and a discriminator network operate jointly to build up the architecture. The generator is a custom U-Net architecture, making use of encoding and decoding layers, as well as skip connections. Additionally, the generator uses trainable end-to-end guided filters to transfer structure from acoustic images to aerial ones. On the other hand, the discriminator is a Deep Convolutional Generative Adversarial Network (DCGAN) and exists only for training purposes, *i.e.*, it does not exist during evaluation. Both of these networks and their details will be outlined in this section.

4.1 Generator

4.1.1 U-Net based network augmented with guided filter

One of the most critical pieces of the proposed architecture is the guided filter from He et al. (2013). Since the guided filter is a general linear translation-variant filter, the following equation describes its output at a pixel i:

$$q_i = \sum_j W_{ij}(I)p_j. \tag{1}$$

In this equation, p_j is the input pixel, W_{ij} is the filter kernel, I is the guide and q_i is the output pixel.

As defined in He et al. (2013), the following function defines the guided filter for color images:

$$q_i = a_k^T I_i + b_k, \forall i \in \omega_k, \tag{2}$$

where I_i is a 3×1 color vector, ω_k is a window centered in pixel k, a_k is a 3×1 vector of coefficients, q_i and b_k are scalars.

By minimizing a linear ridge regression model, as in Draper and Smith (2014), the coefficients for the local linear model can be defined as follows:

$$a_k = (\Sigma_k + \epsilon U)^{-1} \left(\frac{1}{|\omega|} \sum_{i \in \omega_k} I_i p_i - \mu_k \overline{p}_k \right), \tag{3}$$

$$b_k = \overline{p}_k - a_k^T \mu_k, \tag{4}$$

where Σ_k is the 3×3 covariance matrix of I in ω_k and U is a 3×3 identity matrix.

By manipulating Equations 1, 2, 3 and 4, it can be proven that the kernel weights are given by:

$$W_{ij}(I) = \frac{1}{|\omega|^2} \sum_{k:(i,j) \in \omega_k} \left(1 + \frac{(I_i - \mu_k)(I_j - \mu_k)}{\sigma_k^2 + \epsilon} \right), \tag{5}$$

where μ_k and σ_k^2 are the mean and variance of I in the filter and $|\omega|$ is the number of pixels in ω_k.

For this architecture, an implementation of the guided filter provided by Wu et al. was used to integrate the filter with CNNs to form deep guided filtering networks.

Another vital piece of the architecture is the Generative Adversarial Network (GAN). It works by using a discriminator network, which is described in the next section, to calculate the probability of images in a training batch being real. After that, the sigmoid cross-entropy of these probabilities is computed and used as part of the objective function.

The network model is inspired by GuidedNet from Gonçalves et al. (2018). Similarly to GuidedNet, our generator uses guided filters as a way to transfer structure from the original acoustic image to the output satellite image. Also, dilated convolutions, proposed by Yu and Koltun, were used to increase the receptive field dramatically. Therefore, the proposed network is capable of capturing more context and structure than the original CNN used in Giacomo et al. (2018).

Our general architecture is a U-Net based CNN, as in Ronneberger et al. (2015). The network expects a 256×128 acoustic image as input. A schematic of our generator architecture is presented in Figure 7.

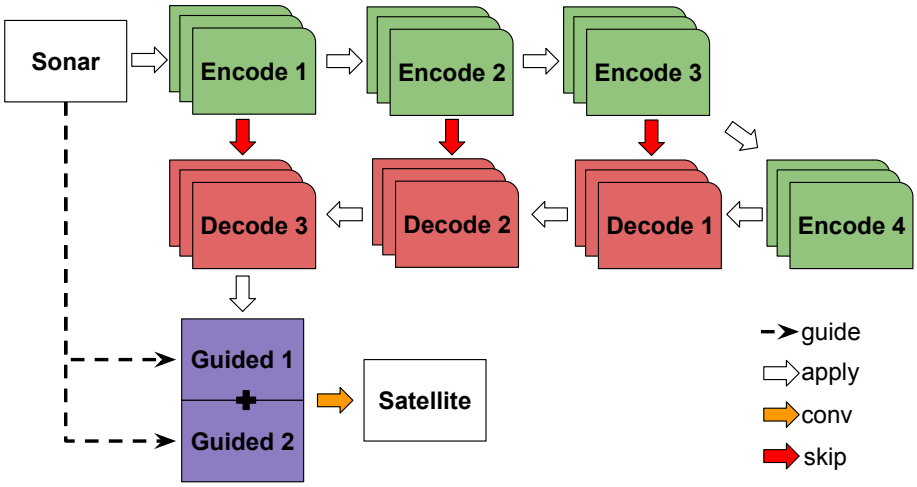


Fig. 7: A diagram that describes the model of our generator network.

As displayed in Figure 7, the network possesses three high-level layers. Namely, the encode, decode and guided layers. These layers are formulated in the following manner:

- Encode is a layer starting with four dilated convolutions of kernel 3×3 with dilation rates: 1, 2, 4 and 8. Then, these convolutions are concatenated, and a max pooling and ReLU are applied. Finally, the layer finishes with a batch normalization step.
- Decode is the layer that starts with an up convolution, *i.e.*, up-sampling followed by a convolution of kernel size 4×4. Then, a dropout of rate 0.2 and a ReLU are applied. Afterwards, a batch normalization step is employed. Finally, the layer performs a skip connection with the feature map of the equivalent encoding layer.
- Guided is a layer starting with a convolution of kernel 3×3 followed by a ReLU activation. Finally, a batch normalization step and a guided filter are performed using the input acoustic image as a guide.

After going through all the layers presented in Figure 7, the network applies the Rectified Linear Unit (ReLU) activation function. Previously, Giacomo et al. (2018) had used the hyperbolic tangent and concluded that it whitened the output images.

In the end, the network will produce a 256×128 satellite image from a given acoustic image. Therefore, the model described constitutes a trainable end-to-end solution to the sonar-to-satellite problem.

4.1.2 Loss function based on discriminative network

For the cost function of our architecture, a linear combination of three different losses was used.

If you consider the acoustic image to be x, the ground truth aerial image y, the generator $\mathcal{G}$, the discriminator D and the VGG-16 neural network to be ϕ. Then, the first loss, the L1 distance is given by:

$$\mathcal{L}_{L1}(\mathcal{G}) = \mathbb{E}_{x,y}\left[|y - \mathcal{G}(x)|\right] \tag{6}$$

The second loss function, the style reconstruction loss, proposed by Johnson et al. (2016), requires the computation of the *Gram matrix* of the feature maps, given by the following mathematical function:

$$G_j(x) = \frac{\psi\psi^T}{C_j H_j W_j}, \tag{7}$$

where ψ is $\phi_j(x)$ reshaped into a matrix of dimensions $C_j \times H_j W_j$.

Then, the style reconstruction loss can be calculated by the squared Frobenius norm of the difference between the Gram matrices of the output and target images, as shown below:

$$\mathcal{L}_{style}(\mathcal{G}) = \mathbb{E}_{x,y}\left[\|G_j(\mathcal{G}(x)) - G_j(y)\|_F^2\right]. \tag{8}$$

The third loss function is the cGAN that can be expressed in the following way:

$$\mathcal{L}_{GAN}(\mathcal{G}, D) = \mathbb{E}_{x,y}\left[\log D(x,y)\right] + \mathbb{E}_x\left[\log\left[1 - D(x, \mathcal{G}(x))\right]\right]. \tag{9}$$

The final objective function is then derived by linearly combining Equations 6, 8 and 9. Each of these individual losses is weighted by a hyperparameter, as follows:

$$\mathcal{L}(\mathcal{G}, D) = \arg\min_{\mathcal{G}} \max_{D} \lambda_1 \mathcal{L}_{GAN}(\mathcal{G}, D) + \lambda_2 \mathcal{L}_{style}(\mathcal{G}) + \lambda_3 \mathcal{L}_{L1}(\mathcal{G}). \tag{10}$$

This equation is a minimax two-player game, where the generator attempts to minimize the function and the discriminator to maximize it.

4.2 Discriminator

The input of the discriminator network is a concatenation of the input acoustic image with either the target satellite image or the generated satellite image outputted by the generator network. On the other hand, the discriminator outputs a probability vector that estimates the chance of a given image in the batch belonging to the training set, *i.e.*, being real as understood by the discriminator.

The architecture of the discriminator network can be visualized in Figure 8. This architecture uses some ideas from Radford et al. (2015), such as batch normalization and strided convolutions in the discriminator.

Our discriminator consists of three convolutional steps followed by a flattening and then a dense layer. Each convolution is applied with a 5×5 kernel size, stride 2 and followed by a batch normalization. Afterwards, the contents of the feature maps are flattened and thrown into a Multilayer Perceptron (MLP). Finally, the network applies the sigmoid activation function to acquire the probability of each image in the batch having come from the training data.

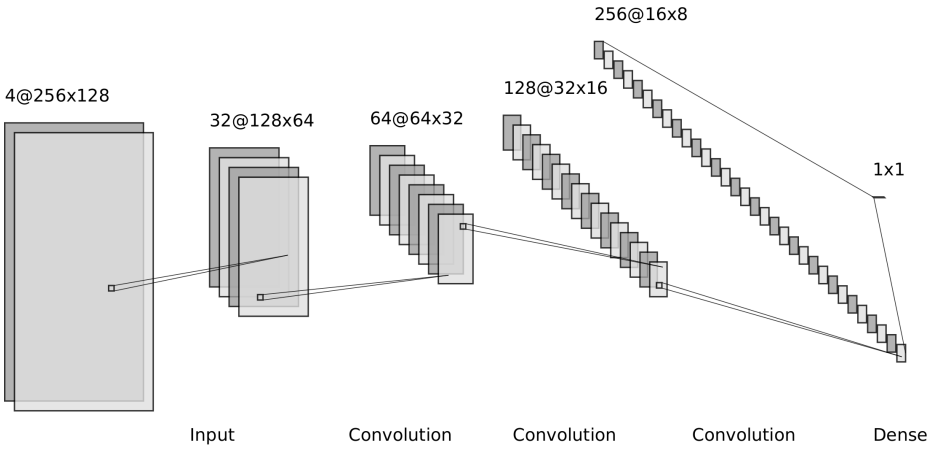


Fig. 8: The schematic that represents the model of our discriminator network.

4.3 Optimization and Training

To train the network, one gradient descent step on the generator G and then one step on the discriminator D were alternated. For updating the weights, the Adam optimizer, introduced in Kingma and Ba (2014), was used.

Implementation of the networks[2] was made using the TensorFlow (TF) library. An NVIDIA Titan X was used for the majority of the conducted experiments.

Training ran for exactly 100 epochs. Hyperparameters used were as suggested in Isola et al. (2017): a learning rate of 0.0002 and Adam momentum parameters of $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Each epoch elapsed approximately 5 minutes of training on an NVIDIA Titan X or NVIDIA GTX 1080. After training, new acoustic images can be evaluated at a frequency of about 20 Hz.

5 Experimental Results

In this section, some results for the two datasets, ARACATI 2014 and 2017 will be presented. Also, qualitative and quantitative analyses of the results will be performed to identify the strengths and weaknesses of the proposed method.

5.1 ARACATI 2017

ARACATI 2017 was divided into two parts: 90% to be used for training purposes and 10% for validation. Figure 9 showcases some samples from the testing set that were used to evaluate the method.

[2] TensorFlow implementation of the model and the ARACATI 2017 dataset are available for download at: https://github.com/giovgiac/son2sat.

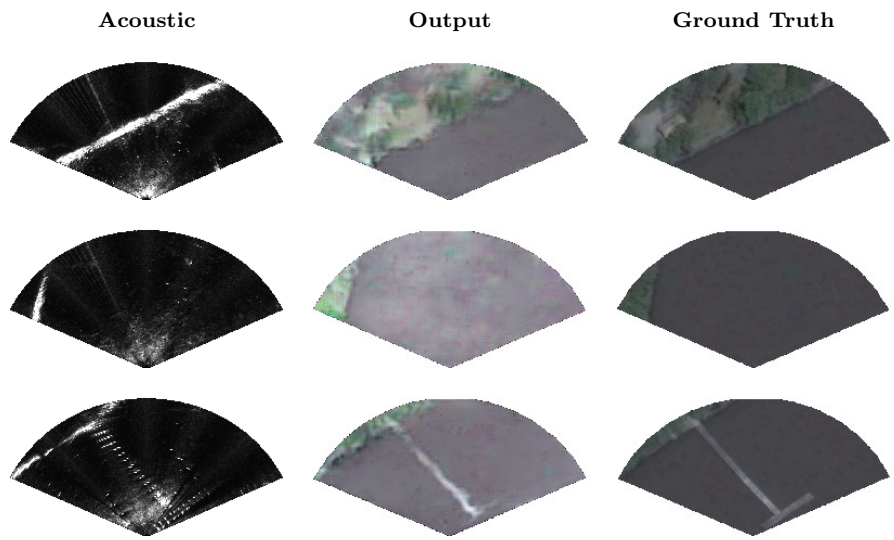


Fig. 9: Results extracted when running our method on the testing set of the ARACATI 2017 dataset.

The testing set consisted of about 289 sonar images propagated through the trained generator. Afterwards, a few output images that highlighted the strengths and weaknesses of the method were picked.

As is visible in Figure 9, the network manages to properly transfer structures from the acoustic image in the corresponding satellite image.

From Figure 9, it is perceivable that the CNN encounters some issues when dealing with the pier. It can be inferred that incorrect GPS and compass data cause these issues that remain in the dataset, even after manual correction. However, the network still manages to transfer the pier, leading to impressive results.

Methods	MSE	PSNR	SSIM
Giacomo et al. (2018)	0.0176	18.9372	**0.8310**
Ours	**0.0122**	**21.8455**	0.8213

Table 1: Quantitative results for the ARACATI 2017 dataset in three image quality metrics.

Table 1 lays out the results of three image quality metrics: Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) from both the presented method and the one proposed in Giacomo et al. (2018). In general the new method performs better, however it loses by a small percentage margin in the SSIM metric.

5.2 ARACATI 2014

ARACATI 2014 was used as a testing dataset where reliable data was available for producing ground truth images. Therefore, satellite images were extracted from a total of 839 acoustic images in a location that the network had never seen before. This dataset was introduced to test if the method would generalize when encountering different scenarios. In particular, the acoustic images from these two datasets are quite different, as ARACATI 2017 was captured with a forward distance of 50m and ARACATI 2014 with 30m.

Figure 10 presents a few samples that were chosen to highlight the performance of the network in this dataset.

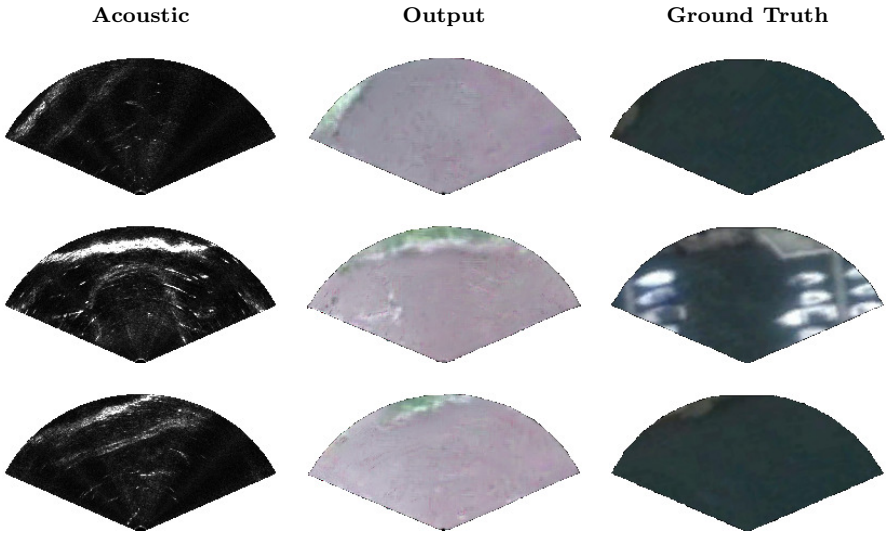


Fig. 10: Results extracted when running our method on the ARACATI 2014 dataset.

As observable in Figure 10, the network does not perform as well in the ARACATI 2014 dataset as it did in the ARACATI 2017. However, it is also noticeable that the main strengths of the method were maintained. Thus, the generator successfully transferred essential features from the acoustic image to the satellite image. Also, it is perceivable that the CNN can capture contextual information from the images and take appropriate advantage of that.

It is important to note that the network missed several of the piers from the ARACATI 2014 dataset. However, that outcome is expected, due to the acoustic images having different forward distances. Also, the ARACATI 2014 dataset has a much larger density of objects when compared to the ARACATI 2017 dataset. Therefore, the 2014 images are significantly more polluted.

Table 2 once again lays out the results of three image quality metrics, for the presented method and Giacomo et al. (2018). As observed in the ARACATI 2017

dataset, the proposed method still loses by a small percentage difference in the SSIM, but performs better in the other two metrics.

Methods	MSE	PSNR	SSIM
Giacomo et al. (2018)	0.0404	14.1208	**0.6699**
Ours	**0.0316**	**15.0925**	0.6035

Table 2: Quantitative results for the ARACATI 2014 dataset in three image quality metrics.

Key features are translated from the source image into the target image successfully. To exemplify, it is possible to adequately visualize the borders between water and land in both the ARACATI 2014 and 2017 datasets. Also, the generator manages to avoid pitfalls that could occur due to the noisy nature of sonar images. However, one may notice that some details, such as boats and piers, are often missed. With all that considered, these results still succeed in reaching our goal, *i.e.*, adequately transferring structure from a sonar image to a generated satellite image to allow for easier image processing down the line.

6 Conclusions

In this paper, we introduced a novel method was introduced, which improves the one proposed in Giacomo et al. (2018), for acquiring satellite images from given acoustic images that were captured in the same region. Our proposal consists of using a U-Net based CNN augmented with guided filters and dilated convolutions to train a generator neural network attached to a DCGAN discriminator. Also, we train and validate our proposed network with two real datasets, which were captured by underwater vehicles in the coast of Brazil. We qualitatively and quantitatively analyze the generated results with samples from the testing sets of the datasets.

We believe our method can help facilitate traditionally difficult robotic tasks like underwater localization and navigation. Using our proposed methodology, AUVs can acquire acoustic images, convert them to satellite images and then use that data to locate themselves or map the environment around them. Since satellite images are of easier interpretation, robots should be able to achieve superior results with less time.

For future work, we intend to consider whether it is possible to use drones to capture aerial images. In an affirmative scenario, it would be beneficial to cooperate drones and underwater robot for effective localization techniques. Finally, we want to follow up on different applications that open up when successfully translating acoustic images into aerial ones. These applications might include, for example, underwater localization and navigation, among others.

Acknowledgements This research is partly supported by CNPq, CAPES and FAPERGS. We also would like to thank the colleagues from NAUTEC-FURG for helping with the experimental data and for productive discussions and meetings. Finally, we would like to thank

NVIDIA for donating high-performance graphics cards. All authors are with NAUTEC, Intelligent Robotics and Automation Group, Universidade Federal do Rio Grande - FURG, Rio Grande - Brazil.

7 Declarations

7.1 Ethical Approval

Not applicable.

7.2 Consent to Participate

Not applicable.

7.3 Consent to Publish

Not applicable.

7.4 Authors Contributions

- Giovanni G. De Giacomo: implementation and execution of the Deep Learning experiments; writing of the manuscript.
- Matheus M. dos Santos: development of the dataset and associated tools; helped writing the manuscript.
- Paulo L. J. Drews-Jr: theoretical support on the idea; revising the manuscript.
- Silvia S. C. Botelho: theoretical support on the idea; revising the manuscript.

7.5 Funding

This study was partly supported by the National Council for Scientific and Technological Development (CNPq) and Coordenacao de Aperfeioamento de Pessoal de Nivel Superior - Brasil (CAPES) - Finance Code 001. This paper is also a contribution of the INCT-Mar COI funded by CNPq Grant Number 610012/2011-8.

7.6 Conflicts of Interest

The authors declare that they have no conflict of interest.

7.7 Availability of data and materials

TensorFlow implementation of the model and the ARACATI 2017 dataset are available for download at: https://github.com/giovgiac/son2sat.

References

Deng X, Zhu Y, Newsam S (2018) What is it like down there? generating dense ground-level views and image features from overhead imagery using conditional generative adversarial networks. arXiv preprint arXiv:180605129

Dos Santos MM, De Giacomo GG, Drews P, Botelho SS (2019a) Satellite and underwater sonar image matching using deep learning. In: 2019 Latin American Robotics Symposium (LARS), 2019 Brazilian Symposium on Robotics (SBR) and 2019 Workshop on Robotics in Education (WRE), IEEE, pp 109–114

Dos Santos MM, De Giacomo GG, Drews PL, Botelho SS (2019b) Underwater sonar and aerial images data fusion for robot localization. In: 2019 19th International Conference on Advanced Robotics (ICAR), IEEE, pp 578–583

Draper NR, Smith H (2014) Applied regression analysis, vol 326. John Wiley & Sons

Giacomo G, Machado M, Drews P, Botelho S (2018) Sonar-to-satellite translation using deep learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, pp 454–459

Gonçalves LT, de Oliveira Gaya JF, Junior PJLD, da Costa Botelho SS (2018) Guidednet: Single image dehazing using an end-to-end convolutional neural network. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, pp 79–86

He K, Sun J, Tang X (2013) Guided image filtering. IEEE transactions on pattern analysis and machine intelligence 35(6):1397–1409

Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision, Springer, pp 694–711

Kim D, Walter MR (2017) Satellite image-based localization via learned embeddings. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp 2073–2080, DOI 10.1109/ICRA.2017.7989239

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434

Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer, pp 234–241

Silveira L, Guth F, Drews-Jr P, Ballester P, Machado M, Codevilla F, Duarte-Filho N, Botelho S (2015) An open-source bio-inspired solution to underwater slam. IFAC-PapersOnLine 48(2):212–217

Steffens C, Messias L, Drews-Jr P, Botelho S (2020) Cnn based image restoration: Adjusting ill-exposed srgb images in post-processing. Journal of Intelligent & Robotic Systems DOI 10.1007/s10846-019-01124-9

Steffens CR, Messias LRV, Drews-Jr P, Botelho SSdC (2019) Contrast enhancement and image completion: A cnn based model to restore ill exposed images. In: 2019 IEEE 17th International Conference on Industrial Informatics (INDIN),

IEEE, vol 1, pp 226–232
Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Iccv, vol 98, p 2
Viswanathan A, Pires BR, Huber D (2014) Vision based robot localization by ground to satellite matching in gps-denied situations. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 192–198, DOI 10.1109/IROS.2014.6942560
Wu H, Zheng S, Zhang J, Huang K (2018) Fast end-to-end trainable guided filter. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1838–1847
Yu F, Koltun V (2015) Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:151107122