Vision-based Obstacle Avoidance Using Deep Learning

Joel O. Gaya, Lucas T. Gonçalves, Amanda C. Duarte, Breno Zanchetta, Paulo Drews-Jr, Silvia S. C. Botelho
Intelligent Robotics and Automation Group - NAUTEC
Center of Computational Sciences - C3
Universidade Federal do Rio Grande - FURG
Rio Grande, Brazil
Emails: {joelfelipe, lucas.teixeira, a.duarte, bfzanchetta, paulodrews, silviacb}@furg.br

***Abstract*—This paper describes a vision-based obstacle avoidance strategy using Deep Learning for Autonomous Underwater Vehicles (AUVs) equipped with a simple colored monocular camera. For each input image, our method uses a deep neural network to compute a transmission map that can be understood as a relative depth map. The transmission map is estimated for each patch of the image to determine the obstacles nearby. With this map we are able to identify the most appropriate Region of Interest (RoI) and to find a direction of escape. This direction allows the robot to avoid obstacles by performing a control action. We evaluate our approach in two underwater video sequences. The results show the approach is able to successful find a RoI that avoids coral reefs, fish, the seafloor and other object present in the scene.**

***Keywords*—Deep learning; Obstacle Avoidance; AUV; Transmission Map.**

I. Introduction

Robots and machines are becoming increasingly autonomous with the advances of technology. In the field of underwater robotics, the number of Autonomous Underwater Vehicles (AUVs) available commercially has increased considerably in the last years. These vehicles tend to be small and cheap, but with limited sensing capabilities. Some of them have only a color camera on board in its standard configuration, but even these AUVs can be used for research purposes.

Computer vision has been utilized broadly to achieve various underwater robotics tasks such as: habitat and animal classification [1], mapping [2], 3D scene reconstruction [3], visualization [4], docking [5], tracking [6], inspection [7] and localization [8]. Nevertheless, sonar-based sensors are usually adopted in large vehicles [9], and the usual approach to obstacle avoidance in these vehiclies [10]. However, this kind of sensor is expensive and heavy that limits its applicability for small vehicles.

There are few studies addressing vision-based obstacle avoidance for this field and even less using a monocular camera. However, it is important to emphasize that methods like [11], based on binocular vision, require a calibration step and have a higher cost than monocular. Recently, [12] presented a method based on a monocular camera that uses superpixel segmentation [13].

Nowadays, the application of deep learning largely improves the performance of many tasks such as: object classification [14] [15] [16] [17], segmentation [18], spatial transformation [19], among others.

In this paper, we propose a real-time obstacle avoidance approach applicable for small AUVs equipped with a single monocular camera. We estimated a transmission map using a deep neural network to obtain a direction of escape.

Underwater images carry information about the relative depth of objects due to the relation between medium effects and deepness. We can take advantage of this property to estimate a relative depth map. Based on this information, we control the robot to the direction of escape. This direction is determined by the highest distance mean in an area previously determined based on the robot's dimensions and camera characteristics.

This work contributes not only providing an underwater obstacle avoidance method but also proposing a new convolutional neural network (CNN) topology to estimate a transmission map of a input image. In addition, we were able to correctly choose a direction of escape in experiments with a real image sequence.

Considering [20], the estimated transmission map can be used to address the obstacle avoidance problem. [21] proposed a method to estimate the transmission of a small *hazy image* patch. According to the authors, the usage of a neural network to estimate the transmission map provides better results than classic methods found in the literature. Even though, underwater and hazy images present similarities [22], the work presented in [21] is not adopted for underwater scenes. Taking this into account, we used a neural network to estimate a transmission map in underwater scenes. Thus, the estimated transmission map obtained by the network can be used to choose a possible escape direction.

The remainder of this paper is organized in the following way: Section II explains our methodology to process the image and to avoid obstacles; Section III evaluates our approach on real underwater sequences. Finally, in Section IV, we summarize the paper's contributions and present the future research directions.

II. Methodology

Our method estimates a transmission map of underwater images provided by monocular color camera using a deep neural network. This transmission map could be understood as

978-1-5090-3656-1/16 $31.00 © 2016 IEEE
DOI 10.1109/LARS-SBR.2016.9

the relative depth of objects in the scene. Then, we can provide a valid direction of escape to achieve obstacle avoidance based on those depths. This process is depicted in Fig 1.

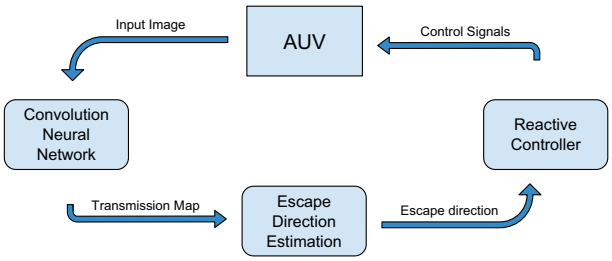


Fig. 1. Method overview. Input image from a video sequence which is adopted to estimate a transmission map using Deep Learning. This transmission map allow us to compute a direction of escape and control the signals of the robot.

A. Image Formation Background

In underwater environments, the light can be affected by *scattering* and *absorption* during its propagation in the medium. These two phenomena culminate into an effect called *attenuation*.

To make matters worse, *scattering* also degrades the image formation by adding noisy information. *Forward scattering* happens when light rays coming from scene are scattered in small angles and reach the image plane, creating a blurry effect on the image. This effect, however, has a small contribution on the final result and is frequently neglected [23]. Another effect, *backscattering*, happens when the light coming from outside the captured scene is scattered over the camera plane. The backscattering creates a characteristic veil on the image that reduces the contrast. We can describe an image captured in a underwater medium, in each color channel $\lambda \in \{r, g, b\}$, as [24] [25]:

$$I_\lambda(x) = E_{d_\lambda}(x) + E_{f_\lambda}(x) + E_{bs_\lambda}(x), \tag{1}$$

where $E_{d_\lambda}(x)$ is the direct component (signal), $E_{f_\lambda}(x)$ is the forward scattering component and $E_{bs_\lambda}(x)$ the backscattering component. As mentioned earlier, we can ignore the influence of the second component in the final image leading us to:

$$I_\lambda(x) = E_{d_\lambda}(x) + E_{bs_\lambda}(x). \tag{2}$$

a) Direct Component: The direct component $E_{d_\lambda}(x)$ represents the amount of light that reaches the camera and is defined as:

$$E_{d_\lambda}(x) = J_\lambda(x)\, \mathrm{e}^{-cd(x)}, \tag{3}$$

where $J_\lambda(x)$ is the signal with no degradation, which is attenuated by the term $\mathrm{e}^{-cd(x)}$, named transmission $t(x)$.

b) Backscattering Component: Following [24] [25], and their respective simplifications [23], the backscattering component $E_{bs_\lambda}(x)$ can be defined as:

$$E_{bs_\lambda}(x) = A_\lambda\,(1 - t(x)), \tag{4}$$

where A_λ is the veiling light that represents the color and radiance characteristics of the media. This constant is related to the volume of water on the Line Of Sight (LOS). Also, this constant is altered by the depth and influenced by the light source [26][11]. The $(1 - t(x))$ portion weights the effect of backscattering as a function of the depth[1] $d(x)$ between the scene object and the camera. As much as the distance is higher, the influence of A_λ over the final image will be higher.

c) Final Model: We obtain the final model by applying Eq. 3 and Eq. 4 on Eq. 2 resulting into:

$$I_\lambda(x) = J_\lambda(x)\, t(x) + A_\lambda\,(1 - t(x)). \tag{5}$$

This model is commonly used in image restoration methods to isolate $J_\lambda(x)$, which contains the image information with no degradation. However, we are interested in $t(x)$ due to its relation with the depth. The convolutional neural network provide us an estimation and is described in the next section.

B. Transmission Estimation

Since it is not possible to estimate the depth d without previous knowledge about the scene, we estimate a transmission for each patch of the image. The values are in the interval $[0; 1]$, which is adopted to estimate the relative depth of objects within the image.

In underwater environment, the transmission estimation is an important step to compute the relative depth of an object. As [21] proposed, the transmission can be estimated using a convolutional neural network. Our model follows the same principle, but with a different topology. Also, [21] is only adopted to hazy images and our data is composed exclusively by underwater turbid images. Thus, our model needs to learn the relation between the underwater turbid image patches and their respective transmission maps.

1) Architecture and Layer Design: Our network's architecture is summarized in Fig. 2. The network is composed by six layers as shown in Table I. The first four layers are divided in two pairs of asymmetric convolutions, followed by one pooling and one convolutional layer.

TABLE I. LAYER SPECIFICATIONS OF THE MODEL PROPOSED.

Layer Type	Input Size	Times Applied	Kernel Size
Asymmetric	$16 \times 16 \times 3$	16	3×1
Asymmetric	$14 \times 16 \times 16$	16	1×3
Asymmetric	$14 \times 14 \times 16$	32	5×1
Asymmetric	$10 \times 14 \times 32$	32	1×5
MaxPool	$10 \times 10 \times 32$	32	5×5
Conv	$2 \times 2 \times 32$	1	2×2

[1]The term depth is associated to the distance between the scene and the camera.

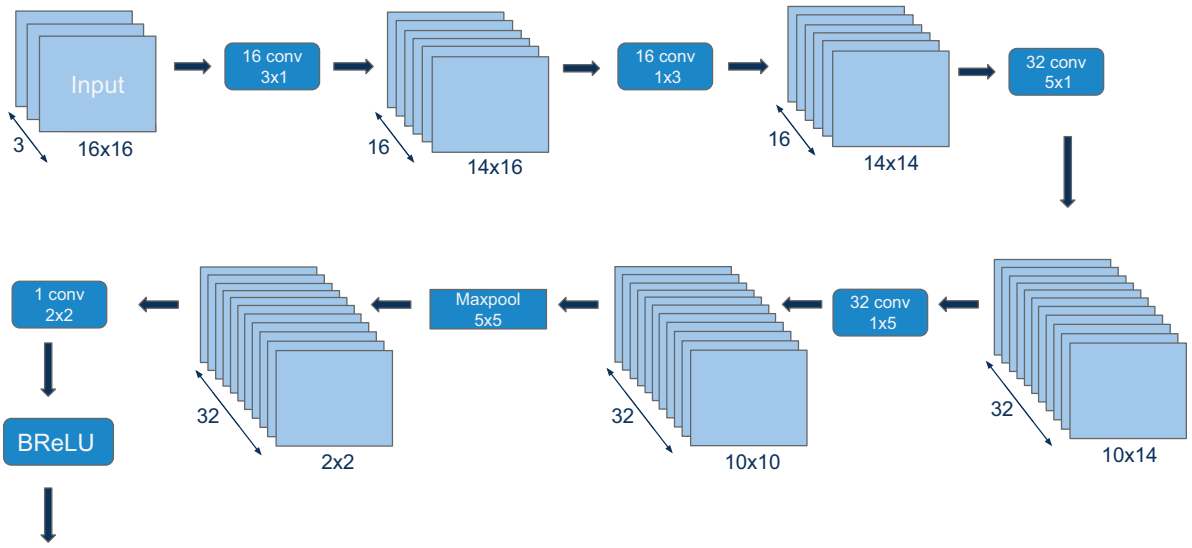


Fig. 2. Convolutional Neural Network architecture displaying the layers and the generated feature maps. Between each group of feature maps is shown a layer with the layer type and kernel size. Convolution layers also inform the number of convolution.

For the sake of computation and memory efficiency, we replaced normal squared kernels by pairs of asymmetric kernels. These kernels produce the same results with a small computational burden and memory usage. As detailed in [15], this spatial factorization results in a 33% cheaper solution with the same receptive field.

As we convolve across the image, we end up losing spatial dimensions by reducing the feature maps sizes. The size reduction can be avoided using padding. However, our convolutions are applied without any padding due to our goal is to convert the original 16×16 patch to only one value. Also, we use a pooling layer that summarizes the information stored in groups of nearby neurons in the same feature map.

Usually, the neuron's output is modeled with the hyperbolic tangent (tanh) function or the Rectifier Linear Unit (ReLU) [14]. Since the ReLU activation function outputs values are larger than zero and the transmission is limited to one, we limited our activation output to be a number in the interval $[0; 1]$. Therefore, we use BReLU [21]. This activation function defines an upper and a lower limit to the output. The function is defined by:

$$A(x) = \min(t_{max}, \max(t_{min}, t(x)), \tag{6}$$

where $A(x)$ is the activation of the output neuron for a input patch x, $t(x)$ is the last convolutional layer's output. t_{min} and t_{max} are, respectively, the lower and the upper limit values which, in our case, are set to 0 and 1.

2) Gathering Data and Training the Model: The feasibility of gathering and labeling data to train deep neural networks is generally low. The requirement of pairs of turbid images and their associated accurate transmission maps hinders even more the ability to train the model with genuine data. For the training of our method, we assume that the depth along a patch is constant, so it is possible to estimate only one transmission per patch without the requirement of knowing its depth map. This assumption is usually adopted in almost all transmission estimation methods in the literature [26] [27] [21].

Following that principle and knowing the adversities of collecting data, we decided to generate synthetic data. First, we gathered a set of 680 clear underwater images and partitioned them in many 16×16 patches. In each one of those, we simulated a scattering media with a randomly generated transmission, limited between 0 and 1. The result was a data set consisting of a 1,000,000 turbid underwater patches and their respective ground truths, *i.e.* the simulated transmission, in which 800,000 are used for training, 100,000 for validation and 100,000 for testing. This division has been chosen to avoid overfitting.

We adopted the *back propagation* algorithm to compute the gradient with respect to the model's parameters to optimize them. The model is trained with batches of 256 square patches. After processing each batch, we compare the outputs with the ground truth values and compute the loss function L by using the L_1 distance function as follows:

$$L = \sum_{i=1}^{256} |y_i - f(x_i)|, \tag{7}$$

where y_i is the ground truth transmission value of the patch and $f(x_i)$ its output value. Lastly, we use the Adam [28] optimizer to readjust the weights, repeating this process in order to minimize the loss L.

We apply the model to an underwater turbid image after the network is trained. We analyze every patch in the image with a stride of one. That process ends up reducing our spatial dimensions by 16 pixels both to the width and height of the image. Since each square patch becomes one pixel, we are analyzing each pixel's surroundings and estimating that pixel's transmission, for every pixel in the image.

C. Direction of Escape Estimation and Control Scheme

1) Direction of Escape: In order to find the most appropriate direction of escape, we must have a previous knowledge about the adopted vehicle and the camera. Based on this, we empirically define a shape to describe the robot's silhouette in the image. Then, the transmission map is analyzed that allow us to find the best place in the image to fit the shape. In our method, we defined the preferred position to fit the shape as the lowest transmission mean value on pixels enclosed by the shape. The method finds a position using a rectangular shape due to its small computational burden. We can compute the direction of escape by:

$$d_i = argmin(S_p * t(x)), \tag{8}$$

where d_i is the direction of escape, S_p is the function that defines our empirical shape and $t(x)$ is the transmission. As proposed by [12], the direction of escape could be not valid. Thus, the solution is to set the pitch angle to an upward direction. Finally, we compute the average between the current and previous valid values to avoid sudden changes and to smooth the vehicle motion.

2) Reactive Controller: Let a valid and stable direction of escape be defined as $d_i = (d_x, d_y)$. The thruster angles are computed using the direction d_i based on the position error $e = (e_x, e_y, e_z)$, based on the center of the image $P_c = (c_x, c_y)$ and the following equations:

$$\begin{aligned} e_x &= D_{RoI}, \\ e_y &= \frac{d_x - c_x}{c_x}, \\ e_z &= -\frac{d_y - c_y}{c_y}, \end{aligned} \tag{9}$$

where D_{RoI} is the selected RoI's average depth. We implemented a P controller for each one of the AUV's degrees of freedom based on these references. The controllers estimate *heave* and *surge* motions and *yaw* rotation:

$$\begin{aligned} u_s &= Kp_s \cdot e_x, \\ u_y &= Kp_y \cdot e_y, \\ u_h &= Kp_h \cdot e_z, \end{aligned} \tag{10}$$

where Kp_s, Kp_y and Kp_h are their proportional gains [29].

III. Experimental Results

We tested our approach on a AUV Surveying Mission video sequence from Australian Centre for Field Robotics [30] and a diving video at Guanacaste Coast [31], both acquired in a real oceanic environment. These videos present some typical characteristics common in underwater sequences such as marine fauna approaching the camera, changes in lightning and some floating particles.

Fig. 3(a) and Fig. 3(d) show one input example from each video sequence. These images are the inputs of the convolutional neural network. The generated transmission map are shown in Fig. 3(b) and Fig. 3(e). The direction of escape estimated by our approach are shown in Fig. 3(c) and Fig. 3(f), where the red squares highlight the RoI and the cross defined the optimal direction of escape.

According to Fig. 3(c) and Fig. 3(f), the AUV would successfully avoid obstacles and navigate in a safety way. However, our experimental results were achieved in an offline estimation, *i.e.* analyzing a previously acquired video.

Applied in an underwater environment, an AUV captures images of the scene and processes them to establish the areas of its field of view appropriate to navigate avoiding collisions. Furthermore, the vehicle is able to decide to go towards areas that the objects are farthest.

IV. Conclusions and Future Work

In this paper we proposed a novel automatic obstacle avoidance method for images captured with a single monocular camera in underwater environments and an innovative method to compute the transmission *on underwater images*. For each incoming video frame and no prior knowledge about the environment, our approach uses a previously trained convolutional neural network to compute a transmission map. The transmission provides an estimation of the depths in relation to the camera. Our proposed approach is able to find free areas and to establish a direction of escape which later ended up becoming a reactive obstacle avoiding method.

As a future work, we intend to test our model in real-time with the underwater vehicle autonomously navigating. Furthermore, we are developing an end-to-end convolutional neural network to learn the whole process. The network will process an image and provide direct pulse-width modulation (PWM) controls to the vehicle without the requirement of the intermediate procedures.

Acknowledgments

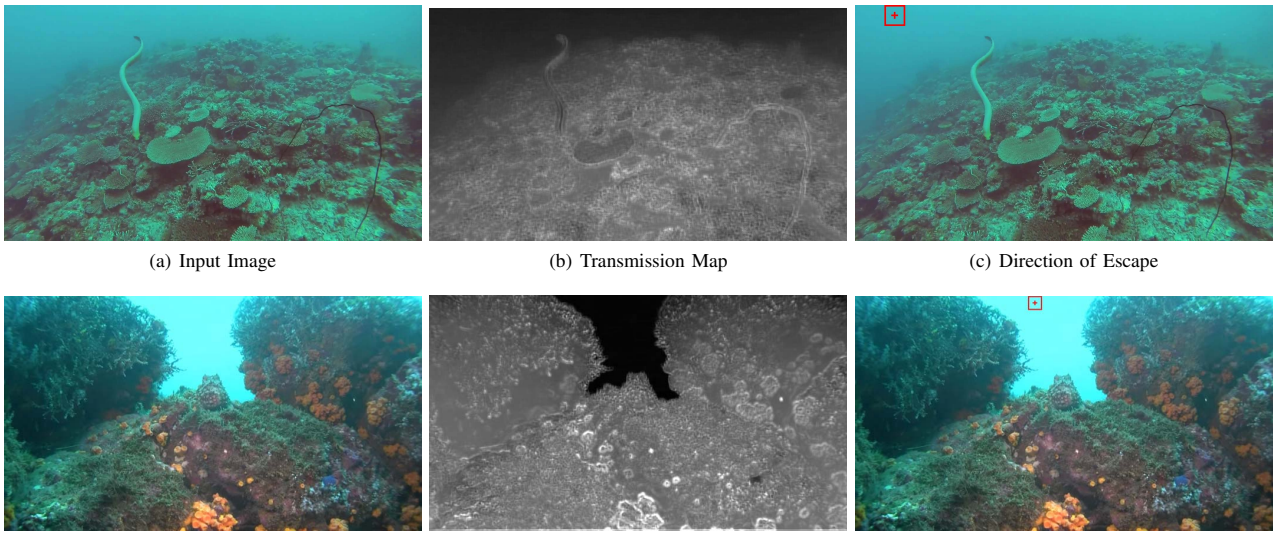
The authors would like to thank the Brazilian Petroleum Corporation - Petrobras and the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) to Funding Authority for Studies and Projects (FINEP) and to Ministry of Science and Technology (MCT) for their financial support through the Human Resources Program of ANP to the Petroleum and Gas Sector - PRH-ANP/MCT. This paper is also a contribution of the Brazilian National Institute of Science and Technology - INCT-Mar funded by CNPq Grant Number 610012/2011-8. This work is partly funded by CNPQ, CAPES and FAPERGS.

References

[1] F. Codevilla, S. S. Botelho, N. Duarte, S. Purkis, A. Shihavuddin, R. Garcia, and N. Gracias, "Geostatistics for context-aware image classification," in *Computer Vision Systems*. Springer, 2015, pp. 228–239.

[2] R. Campos, R. Garcia, P. Alliez, and M. Yvinec, "A surface reconstruction method for in-detail underwater 3d optical mapping," *The International Journal of Robotics Research*, vol. 34, no. 1, pp. 64–89, 2015.

[3] A. Concha, P. Drews-Jr, M. Campos, and J. Civera, "Real-time localization and dense mapping in underwater environments from a monocular sequence," in *IEEE OCEANS*, 2015, pp. 1–5.



(a) Input Image (b) Transmission Map (c) Direction of Escape

(d) Input Image (e) Transmission Map (f) Direction of Escape

Fig. 3. Direction of Escape Estimation. The first row shows the results obtained by using a video extracted in a AUV Surveying Mission from the Australian Centre for Field Robotics. The second row shows the results obtained in a video collected in a diving activity at Guanacaste Coast, Costa Rica. Each column shows: sample images (a,d), the respective transmission map (b,e) and the estimated direction of escape (c,f).

[4] P. Drews-Jr, E. R. Nascimento, M. F. Campos, and A. Elfes, "Automatic restoration of underwater monocular sequences of images," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 1058–1064.

[5] F. D. Maire, D. Prasser, M. Dunbabin, and M. Dawson, "A vision based target detection system for docking of an autonomous underwater vehicle," in *Australasian Conference on Robotics and Automation*, 2009.

[6] P. Drews-Jr, E. R. Nascimento, A. Xavier, and M. Campos, "Generalized optical flow model for scattering media," in *22nd International Conference on Pattern Recognition (ICPR)*, 2014, pp. 3999–4004.

[7] F. S. Hover, R. M. Eustice, A. Kim, B. Englot, H. Johannsson, M. Kaess, and J. J. Leonard, "Advanced perception, navigation and planning for autonomous in-water ship hull inspection," *The International Journal of Robotics Research*, vol. 31, no. 12, pp. 1445–1464, 2012.

[8] S. S. C. Botelho, P. Drews-Jr, G. L. Oliveira, and M. D. S. Figueiredo, "Visual odometry and mapping for underwater autonomous vehicles," in *6th Latin American Robotics Symposium (LARS)*, 2009, pp. 1–6.

[9] M. M. Santos, P. Ballester, G. B. Zaffari, P. Drews-Jr, and S. Botelho, "A topological descriptor of acoustic images for navigation and mapping," in *12th Latin American Robotics Symposium (LARS)*, 2015, pp. 289–294.

[10] Y. Petillot, I. T. Ruiz, and D. M. Lane, "Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar," *Oceanic Engineering, IEEE Journal of*, vol. 26, no. 2, pp. 240–251, 2001.

[11] M. Roser, M. Dunbabin, and A. Geiger, "Simultaneous underwater visibility assessment, enhancement and improved stereo," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3840–3847.

[12] F. G. Rodríguez-Telles, R. Pérez-Alcocer, A. Maldonado-Ramirez, L. Abril Torres-Mendez, B. B. Dey, and E. A. Martinez-Garcia, "Vision-based reactive autonomous navigation with obstacle avoidance: Towards a non-invasive and cautious exploration of marine habitat," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3813–3818.

[13] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

[16] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[19] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2008–2016.

[20] P. Drews-Jr, E. Hernandez, A. Elfes, E. R. Nascimento, and M. Campos, "Real-time monocular obstacle avoidance using underwater dark channel prior," *International Conference on Intelligent Robots and Systems*, 2016.

[21] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *arXiv preprint arXiv:1601.07661*, 2016.

[22] P. Drews-Jr, E. R. Nascimento, S. S. C. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, Mar 2016.

[23] Y. Schechner and N. Karpel, "Recovery of underwater visibility and structure by polarization analysis," *Oceanic Engineering, IEEE Journal of*, vol. 30, no. 3, pp. 570–587, July 2005.

[24] J. S. Jaffe, "Computer modeling and the design of optimal underwater imaging systems," *Oceanic Engineering, IEEE Journal of*, vol. 15, no. 2, pp. 101–111, 1990.

[25] B. McGlamery, "A computer model for underwater camera systems," in *Ocean Optics VI*. International Society for Optics and Photonics, 1980, pp. 221–231.

[26] P. Drews-Jr, E. Nascimento, F. Moraes, S. Botelho, and M. Campos, "Transmission estimation in underwater single images," in *IEEE Inter-*

national Conference on Computer Vision Workshops (ICCVW), 2013, pp. 825–830.

[27] F. Codevilla, J. D. O. Gaya, A. C. Duarte, and S. S. da Costa Botelho, "General participative media single image restoration," *CoRR*, vol. abs/1603.01864, 2016. [Online]. Available: http://arxiv.org/abs/1603.01864

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] V. N. Kuhn, P. L. J. Drews-Jr, S. C. P. Gomes, M. A. B. Cunha, and S. S. da Costa Botelho, "Automatic control of a ROV for inspection of underwater structures using a low-cost sensing," *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, vol. 37, no. 1, pp. 361–374, 2015.

[30] A. C. for Field Robotics, "Auv surveying mission gbr hd," Youtube. [Online]. Available: https://youtu.be/Da7_iDdBxKQ

[31] P. de la Montagne, "Costa rica diving at guanacaste coast pacific," Youtube. [Online]. Available: https://youtu.be/vKduIggPd38