Deep-Sea Fauna Segmentation: A Comparative Analysis of Convolutional and Vision Transformer Architectures at Lucky Strike Vent Field

Pedro Soto Vega, Gustavo Andrade-Miranda, Gilson da Costa, Panagiotis Papadakis, Marjolaine Matabos, Thibault Napoleon, Ayoub Karine, Henrique Fagundes Gasparoto

Deep-Sea Fauna Segmentation: A Comparative Analysis of Convolutional and Vision Transformer Architectures at Lucky Strike Vent Field

Pedro J. Soto Vega[1, 3], Gustavo X. Andrade-Miranda[4], Gilson A. O. P. da Costa[5], Panagiotis Papadakis[6], Marjolaine Matabos[7], Thibault Napoleon[1], Ayoub Karine[2], Henrique Fagundes Gasparoto[3]

[1] L@bISEN, Vision-AD, ISEN Yncréa Ouest, 20 rue Cuirassé Bretagne, 29200 Brest, France - (pedro-juan.soto-vega, thibault.napoleon)@isen-ouest.yncrea.fr
[2] L@bISEN, Vision-AD, ISEN Yncréa Ouest, 33 Quater Chemin du Champ de Manœuvre, 44470 Carquefou, France - ayoub.karine@isen-ouest.yncrea.fr
[3] L@bISEN, AutoROB, ISEN Yncréa Ouest, 20 rue Cuirassé Bretagne, 29200 Brest, France - henrique.gasparoto@isen-ouest.ycrea.fr
[4] University Brest, LaTIM, INSERM, UMR 1101, Brest, France - andradema@univ-brest.fr
[5] Institute of Mathematics and Statistics, State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil – gilson.costa@ime.uerj.br
[6] IMT Atlantique, Lab-STICC, UMR 6285, Team RAMBO, F-29238 Brest, France – panagiotis.papadakis@imt-atlantique.fr
[7] University Brest, CNRS, Ifremer, UMR6197 Biologie et Ecologie des Ecosystèmes marins Profonds, 29280 Plouzané, France - marjolaine.matabos@ifremer.fr

Keywords: Fauna Segmentation, Deep Learning, Hydrothermal Vents, Vision Transformers, Convolutional Networks

Abstract

Due to recent technological developments, the acquisition and availability of deep-sea imagery has increased exponentially in the last years, leading to an increasing backlog in image annotation and processing, attributable to limited specialized human resources. In this work, we investigate the performance of well-established convolutional neural networks and Vision Transformer (ViT) based architectures, namely, DeepLabv3+ and UNETR, for the segmentation of fauna in deep-sea images. The dataset consists of images captured at the Lucky Strike Vent field, located on the mid-Atlantic ridge, of three edifices named Montsegur, White Castle, and Eiffel Tower. Our experimental investigation reveals that the Vision Transformer consistently outperforms the fully convolutional deep learning architecture, by approximately 14% in terms of F1-Score, demonstrating the effectiveness of ViTs in capturing intricate patterns and long-range dependencies present in deep-sea imagery. Our findings highlight the potential of ViTs as a promising approach for accurate semantic segmentation in challenging environmental contexts, paving the way for improved understanding and analysis of deep-sea ecosystems.

1. Introduction

Since the discovery of deep-sea hydrothermal vents, there has been a growing wave of economic and ecological interest in those environments. The related anthropogenic activities have, however, raised concerns among scholars and specialists about the proper analysis and preservation of such ecosystems. Fortunately, alongside the increasing availability of remote sensing data captured from drones, planes and satellites, there has been a rapid increase in the quantity and sizes of deep underwater image datasets. But such an increase in volume and quality, particularly in terms of image resolution, has generated a strong demand for annotating deep-sea images, which is a costly and time-consuming process that requires highly trained professionals (Schoening et al., 2016). In order to cope with such a demand, several solutions have been proposed to automate the analysis of the ever-expanding datasets.

To date, a number of methods following the traditional processing chains of computer vision (CV) and machine learning (ML) have been proposed. We can mention a few exemplary cases. Schoening et al. (2012) proposed a semi-automatic image analysis system for assessing megafaunal densities at the Artic Deep Sea Observatory. The system comprises an ensemble of Support Vector Machine (SVM) classifiers, each associated with a particular species. A Maximum Likelihood Classifier (MLC) combined with two decision tree methods – Quick Unbiased Efficient Statistical Tree (QUEST) (Loh and Shih, 1997) and Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE) (Kim and Loh, 2001) – were employed in Ierodiaconou et al. (2011) for detecting benthic biological communities, using video imagery among other capturing systems. Also using different imaging systems, Schmid et al. (2016); Faillettaz et al. (2016), employed Random Forest (RF) for zooplankton analysis. The spatio-temporal distribution of shrimps was the objective of Osterloff et al. (2016). In that work, images were automatically pre-processed using a super-pixel segmentation algorithm named Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2012), and RF was used to classify the super-pixels. Sharma et al. (2010) used shallow Artificial Neural Networks (ANN) to estimate the occurrence of deep-sea minerals using seafloor images. However, regardless of their specific objectives, the aforementioned efforts were mainly based on traditional image analysis approaches that rely on hand-crafted features and shallow-learning techniques, which are deficient in producing expressive image representations for proper pattern recognition.

Deep Learning (DL) techniques emerged in the last decade as the state-of-the-art in computer vision and image-based pattern

recognition tasks, including deep-sea applications. Villon et al. (2018) used Convolutional Neural Networks (CNNs) to identify coral reef fish species. In Durden et al. (2021) CNNs were trained to classify fauna in seabed images. Xue et al. (2021) studied the performance of several state-of-the-art DL architectures for identifying deep-sea debris. A method for recognition and tracking of deep-sea organisms was proposed in Lu et al. (2020), using the YOLO model (Redmon et al., 2016) as an object detector. For another deep-sea application, i.e., visual monitoring, Gradient Generation Adversarial Networks (GGAN) were proposed in Ma et al. (2021) to restore noisy images from the bottom of the sea. Juliani and Juliani (2021) employed a model based on the U-Net architecture (Ronneberger et al., 2015) for segmenting seafloor mounts directly over raw bathymetry data. More recently, Katija and co-authors introduced the FathomNet (Katija et al., 2022), which provides annotated and localized imagery for developing ML algorithms. They also provide a set of ML models trained to detect the fauna present in the underwater image data.

Recently, Vision Transformers (ViTs) (Dosovitskiy et al., 2021) have demonstrated remarkable capabilities in various computer vision applications. Unlike conventional CNNs, which primarily focus on local image features, ViTs excel at capturing the relationships among distant parts of an image. This holistic perspective has proved to be very useful for tasks like object detection, image classification and semantic segmentation. Consequently, ViTs have delivered state-of-the-art performance and have even surpassed CNNs in those domains. In deep-sea applications, architectures composed of transformer blocks have been employed for underwater image enhancement (Peng et al., 2023; Yang et al., 2023). A convolutional vision transformer was proposed in Rajani et al. (2023) for segmenting substrata in sonar images. Target categories in underwater images have been identified in Sun et al. (2022) through object detection supported by the Mask-RCNN, utilizing a backbone comprised of Swin-Transformer (Liu et al., 2021) blocks.

Regarding the specific application of this work – semantic segmentation of benthic fauna communities – a few DL-based methods have been proposed. For instance, Pavoni et al. (2021) employed the DeepLabv3+ model (Chen et al., 2018) for automatic segmentation of corals. Shashidhara et al. (2020) leveraged the U-Net (Ronneberger et al., 2015)for scale worms segmentation. Addressing the scarcity of labeled data for deep-sea applications, Lütjens and Sternberg (2021) proposed innovative data augmentation techniques involving modifications to the entire image composition and additional alterations for synthetically generating them within a CNN-based instance segmentation approach aimed at counting a limited set of morphotypes.

Despite ongoing efforts, the use of deep learning-based automatic recognition techniques remains limited by the wide variety of species and complex underwater environments found in benthic fauna communities. Such limited capacity hinders the ability to monitor and understand the related ecosystems efficiently.

This work seeks to investigate the use of ViTs-based DL architectures for semantic segmentation of fauna in deep-benthic environments. For that purpose we compare a CNN and a ViT-based architecture for the task. To the best of our knowledge, no prior research has employed Vision Transformers (ViTs) for the semantic segmentation of fauna in deep-benthic environments, nor has comprehensively compared the performance of convolutional and vision transformer-based architectures in that application.

The deep learning models were evaluated using three image datasets (Ramière et al., 2023), which contain RGB images acquired in different locations at the 1700m deep Lucky Strike vent field region (Langmuir C et al., 1993).

The remainder of this paper is organized as follows. Section 2 presents the DL-based architectures evaluated in this work. Section 3 describes the datasets, the experimental setup, the network implementations and the adopted performance metrics. Section 4 presents the obtained results, and finally, Section 5 presents conclusions and directions for future research.

2. Methods

Semantic segmentation is most effectively accomplished using fully convolutional components. These networks are usually made up of an encoder stage that reduces spatial resolution through convolution and pooling operations across layers, followed by a decoder stage that recovers the original spatial resolution. Architectures with encoders composed of consecutive ViT blocks have recently been proposed.

In the next sections, we will briefly describe the neural network models evaluated in this work, namely DeepLabv3+ (Chen et al., 2018) and UNETR (Hatamizadeh et al., 2022).

Among several alternatives that can be explored in semantic segmentation, DeepLabv3+ is renowned for its strong performance in this task, especially in scenarios where high-resolution segmentation is required. On the other hand, UNETR has demonstrated state-of-the-art performance in various segmentation benchmarks, showcasing its effectiveness in capturing complex spatial relationships within images.

2.1 Fully Convolutional Network: DeepLabv3+

The DeepLab series of fully convolutional CNN architectures progressively refined semantic image segmentation through a number of innovations: atrous convolution in version 1, which enlarged the filter receptive fields for capturing broader context; Atrous Spatial Pyramidal Pooling (ASPP) in version 2, which captured multi-scale information; and image pooling in version 3, which incorporated global context.

DeepLabv3+ (Chen et al., 2018) builds upon DeepLabv3 (Gao, 2023) by adding a decoder to improve segmentation quality, especially at object boundaries. It upsamples the encoder output, i.e., high-level features, and combines it with low-level features from the backbone network to preserve spatial details. The model allows using different architectures as encoders, such as Xception (Chollet, 2017). In this work, we used the ResNet-101 (He et al., 2016) as the model's backbone. Figure 1 depicts the components of DeepLabv3+ architecture.

2.2 Hybrid Convolutional and Transformer Network: UNETR

Originally proposed for medical image applications, the UNETR (Hatamizadeh et al., 2022) is a hybrid network that adopts a U-Net (Ronneberger et al., 2015) architecture for segmenting 3D images. It uses ViT blocks as encoder and a fully convolutional network as a decoder. The encoder learns features

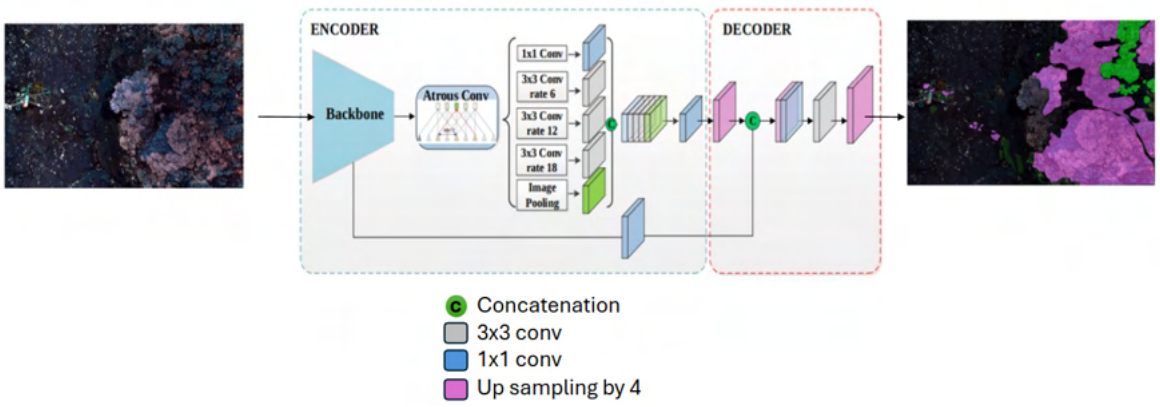

Figure 1. DeepLabv3+ architecture (Torres et al., 2021).

from input image patches in such a way that allows it to capture global information and long-range spatial dependencies. The encoder directly communicates with the decoder through skip connections to extract multi-scale information and integrate it to make pixel-wise predictions.

The UNETR encoder operates over 1D sequences, whereby the input image is partitioned into patches of equal size and represented in an embedding space. Then, a 1D learnable positional embedding is added to the embeddings to preserve spatial information. Subsequently, the patches undergo multiple multi-head self-attention blocks before reaching the decoding stage. The transformation to the segmentation mask space is achieved through CNN up-sampling combined with multi-level feature aggregation.

The details of UNETR architecture are depicted in Figure 2 and described as follows. The encoder is made up of 12 ViT blocks where each block consists of 12 Multi-Head Self-Attention (MHSA) and MultiLayer Perceptron (MLP) components. Additionally, Norm stages carry out layer normalization operations, while patching and flattening procedures are performed in Linear projection and Embedded Patches stages. The decoder is assembled through the integration of convolution-based operations, specifically convolution (conv) and deconvolution (deconv) techniques, along with batch normalization (BatchNorm) layers and concatenation blocks.

3. Experimental analysis

The experiments conducted in this work aim to evaluate convolutional and ViT-based architectures, i.e., DeepLabv3+ and UNETR, in a particular semantic segmentation problem, namely ridge and hydrothermal vent fauna classification. The datasets used in this work comprise images taken from three different locations on a particular vent field (Vega et al., 2024), as shown in Figure 3.

3.1 Study area

The study areas are located at the Lucky Strike (LS) vent field along the Mid-Atlantic Ridge (MAR, $37°17N, 32°16W$). LS is a basalt-hosted hydrothermal vent field located near the Azores Triple Junction on the slow-spreading MAR at a depth of approximately 1700 meters (Langmuir C et al., 1993). This large hydrothermal field extends over more than 1 km^2 and lies at a seamount’s summit, harboring a central fossilized lava lake surrounded by three volcanic cones and faults (Ondréas et al., 2009). Figure 3 shows the LS localization as well as the positions of the Eiffel Tower (ET), Montsegur (MS), and White Castle (WC) vent edifices.

3.2 Dataset

The dataset[1] (Ramière et al., 2023) publicly available, comprises RGB images collected during the MoMARSAT 2018 cruise (Cannat and Sarradin, 2018) using the Remotely Operated Vehicle (ROV) *Victor6000* over and around the following edifices (areas hereafter called sites): Montsegur (MS, see Marticorena et al. (2021)), White Castle (WC) and Eiffel Tower (ET, see Girard et al. (2020)). Each image has a dimension of 4000×6000 pixels with a spatial resolution of 0.001 m/pixel. Images of the seabed have been acquired at one image every three seconds with a downward-looking HD camera OTUS2 with navigation tracks. Constant ROV altitude (5 ± 2 m) planned in parallel transects spaced 1.8 m apart to ensure overlap between each captured image at a constant speed of 0.5 m.s^{-1}.

The acquired images were pre-processed in the following order. First, blurred and obscured samples were removed. Second, a non-overlapped set of pictures was selected using the MATISSE 3D software (Arnaubec et al., 2015) (Ifremer). The MATISSE 3D computes image overlaps through geo-referencing, using the ROV’s navigation parameters and camera positions. Third, the set of non-overlapped images of each site was corrected by attenuating the blue color and homogenizing the light conditions, contrast, and saturation in MATISSE 3D.

Figure 4 shows the categories considered in this study. Human experts manually labeled each image at pixel scale, considering an overall set of categories distributed differently among the three sites – MS: 35, WC: 32, ET: 44. We have focused on a subset of all possible categories, primarily those that are common to the three sites and with a number of images that allow the splitting of the training, validation, and test sets.

As shown in Figure 4, the datasets associated with each site are quite unbalanced regarding the number of samples in the different classes. For example, the classes of ”Blue glass sponge” and ”Cataetyx laticeps” are represented less frequently in MS and WC, while the ET site has more images than both MS and WC combined. It is important to observe that when training a classifier, if the problem of imbalanced classes is not explicitly considered, undesirable biases may be introduced, and strongly

[1] https://www.seanoe.org/data/00838/95015

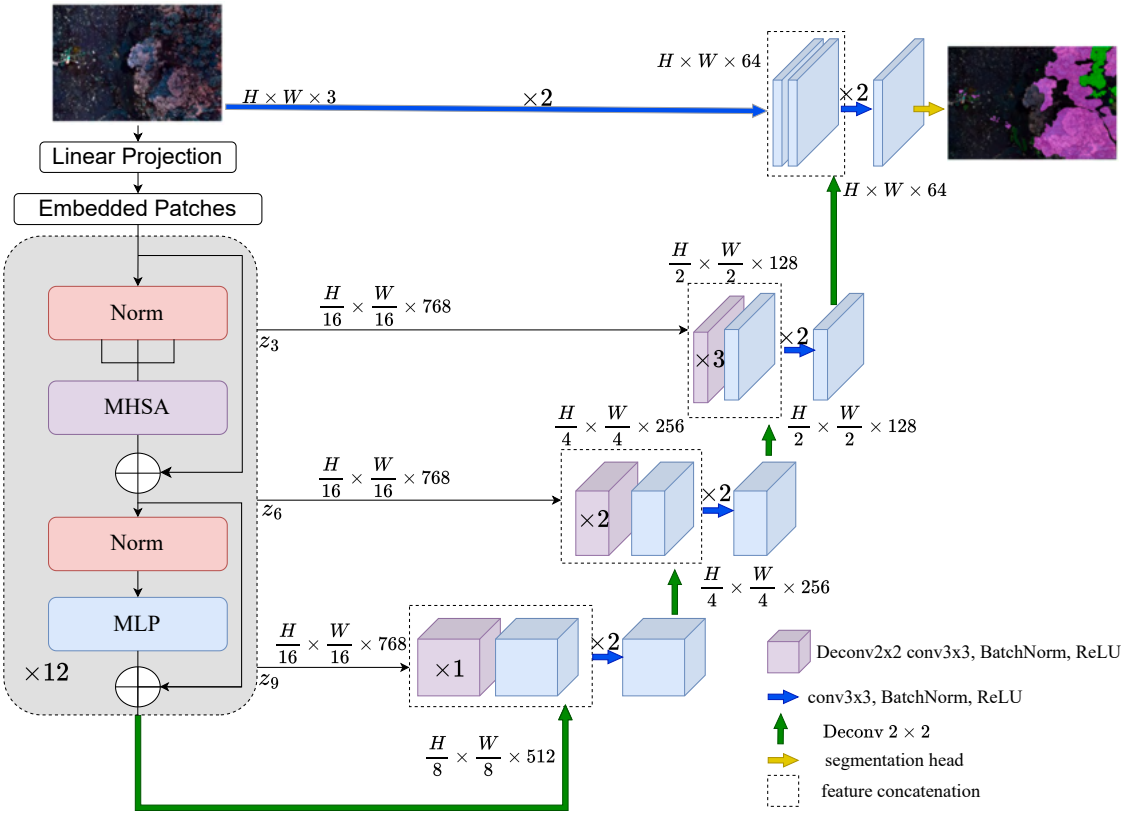

Figure 2. UNETR architecture (Hatamizadeh et al., 2022).

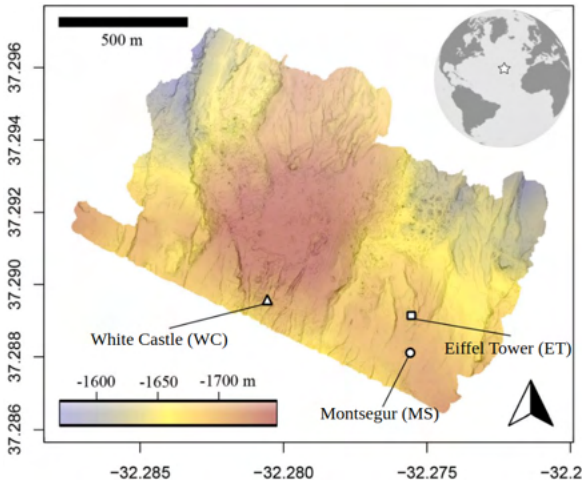

Figure 3. Map of the Lucky Strike vent field (northern Mid-Atlantic Ridge) (Vega et al., 2024).

affect the classifier’s performance. In such cases, the classifier will tend to predict the over-represented classes more frequently due to the smaller impact of the errors associated with their samples in computing overall accuracy. In the following section, we will explain how we addressed class imbalance in this work.

3.3 Classifiers training setup

For the accuracy assessment of fauna characterization, we split every set of images into training, validation, and testing subsets. More specifically, 60 % of the images from each site were randomly selected for training, 20% for validation, and the remaining 20% for testing.

To compensate for class imbalance (see Figure 4), we adopted a *weighted cross-entropy* cost function to train the networks. The intention was to force the DNNs not to be biased towards the over-represented classes by assigning larger weights to the underrepresented ones. Equation 1 shows the loss function ($\mathcal{L}$) employed in the training of the networks.

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C}\sum_{i<I,j<J} w_c(y_n(i,j)\log(h(x_n(i,j)))) \quad (1)$$

In the equation, N stands for the number of training images, x_n represents the n^{th} training image, while y_n represents the respective true label (or labels) codified into a one-hot vector. Furthermore, $h(x_n)$ corresponds to a vector comprising the predicted likelihood values for each class of x_n, computed with the learned function $h(\cdot)$. Additionally, $w_c = \frac{N}{N_c}$ is the weight of each class $c \in C$, which comprises N_c images. Furthermore, i, j represent the pixel coordinates, and I, J are the number of pixels in rows and columns of the input images.

During training, the network backbones – ResNet-101 and ViT Base – of DeepLabv3+ and UNETR, respectively, were initialized with parameters trained on ImageNet datasets. The inputs to the networks were patches with dimensions 512×512×3, extracted from the original full-resolution images. The patches were extracted using a sliding window procedure, with an overlap of 5% in each direction. As in applications such as those addressed in (Vega et al., 2024), splitting the images into patches functions as a data augmentation strategy and also eases GPU memory handling. Additionally, the loss function was minimized using the Adam optimizer (Kingma, 2017), with an ini-

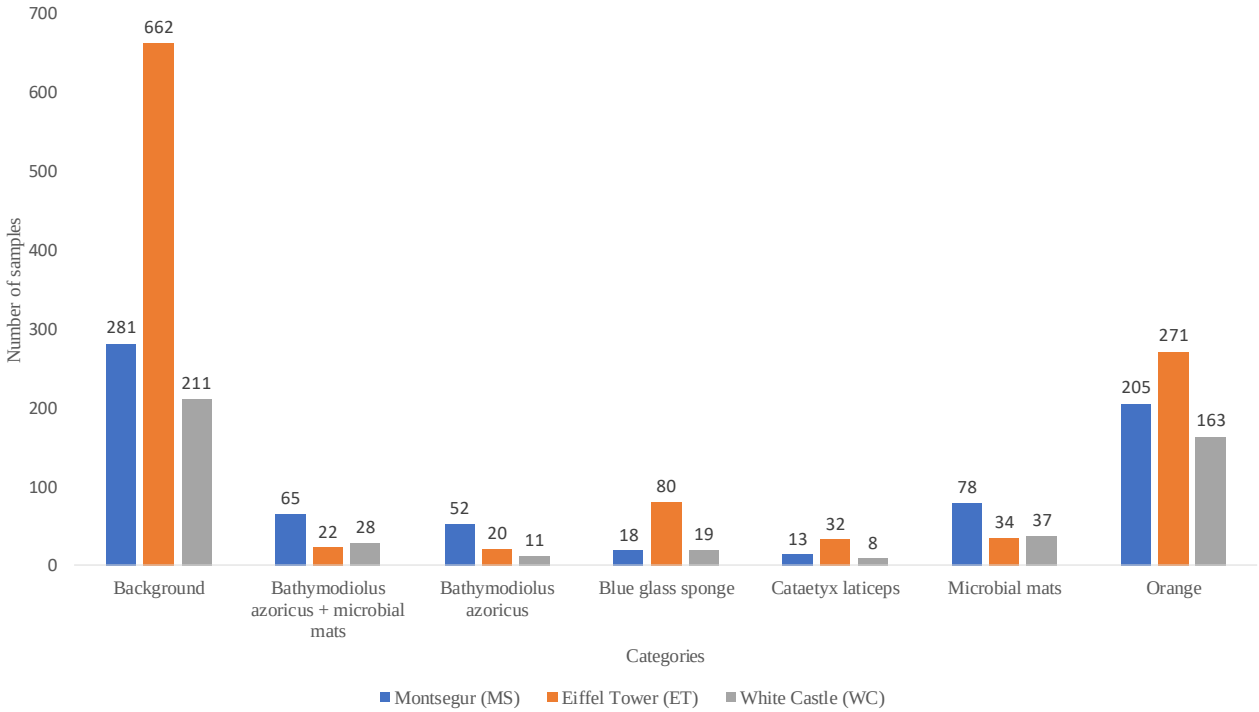

Figure 4. Number of images in each site and in fauna category.

Sites	**MS**		**WC**		**ET**	
Categories/DNNs	DeepLabv3+	UNETR(Base)	DeepLabv3+	UNETR(Base)	DeepLabv3+	UNETR(Base)
Background	98.3	98.3	96.5	*97.8*	*98*	97.8
Bathymodiolus azoricus + microbial mats	69.3	*64.7*	20	*40.1*	52.1	*58*
Bathymodiolus azoricus	46.9	*48.3*	3.1	*9.5*	*16.2*	13.9
Blue glass sponge	62	*66.4*	0	0	9	*59.5*
Cataetyx laticeps	10.2	*70*	0.3	*70.6*	1.9	*59.1*
Microbial mat	66.1	*69.7*	2.3	*9.9*	14.3	*36.7*
Orange	46.1	*47.3*	46	*47.6*	*30.1*	26.4
Average F1 Score	56.9	**66.3**	24	**39.3**	31.6	**50.2**
mIoU	44.3	**52**	19.6	**31**	24.5	**38**

Table 1. Accuracy scores (%) of different semantic segmentation methods.

tial learning rate μ_0 and momentum β_1 equal to 0.0001 and 0.9, respectively. Aiming at better convergence during training, we adopted a learning rate decay procedure following a cosine function. During test time, the networks predict input patches of the same dimension as in training, extracted in the same way but with an overlap of 25 %.

The networks evaluated in this work were implemented using the Pytorch deep learning framework on a hardware platform with the following configuration: Intel(R) Xeon(R) processor, 64 GB of RAM, and NVIDIA GeForce RTX 2060Ti GPU. The batch size was 4, and the early stopping procedure was used to avoid overfitting. The patience parameter, which controls the number of epochs without improvements in the validation loss, was set to 50 for a total of 200 epochs. Some data augmentation operations were applied to all extracted patches, they were color transformations for changing brightness, contrast, hue and saturation.

3.4 Metrics

The performances of the classifiers in all experiments were expressed in terms of the average F1-scores and Intersection over Union (IoU) computed for each individual class. Specifically, the F1-score is expressed as the harmonic mean of Precision (P_c) and Recall (R_c), for each class, as follows:

$$F1 - score_c = \frac{2 \times P_c \times R_c}{P_c + R_c}, \tag{2}$$

where:

$$P_c = \frac{t_p}{t_p + f_p}. \tag{3}$$

$$R_c = \frac{t_p}{t_p + f_n}. \tag{4}$$

The IoU on the other hand is expressed as:

$$IoU_c = \frac{t_p}{t_p + f_n + f_p}. \tag{5}$$

In equations 3, 4, and 5, t_p denotes the number of pixels correctly assigned to class c (true positives); f_p represents the number of pixels erroneously classified as class c (false positives);

and f_n corresponds to the number of pixels of class c incorrectly classified as another class (false negatives).

4. Results and discussion

Table 1 presents the accuracy results achieved by the DNN models in terms of F1 scores and mIoU. The models were trained and tested using data from the same site, namely, MS, WC, and ET. The results obtained for each fauna class are shown in Table 1 in terms of F1 scores. The second to last row of the table shows the average F1 class (considering all classes). Also considering all classes, the last row of the table shows the mean IoU. The highest performances achieved for the average F1 score and mean IoU for each site are shown in bold digits. Also, the best results for each class in each site are highlighted in italics.

Figure 5 shows some exemplary images of the different sites, together with the respective ground truths and predicted maps, computed by both networks. The first column (figures 5(a)(d)(g)) shows the ground truth information; the second (figures 5(b)(e)(h), shows the DeepLabv3+ predictions; and the third column (figures 5(c)(f)(i)) shows the UNETR predictions. In all images the class maps (ground truth and predictions) are overlaid on the original images, leaving the background class transparent.

The results show that the UNETR consistently outperformed the DeepLabv3+ models. Specifically, UNTER surpassed DeepLabv3+ performance by +9.4 %, +15.3 %, and + 18.6% for MS, WC, and ET, respectively, in terms of F1 scores. These results are consistent with previous deep semantic segmentation works in which ViTs-based architectures, like UNETR, excel in capturing global information and long-range dependencies as compared to architectures based on convolutional kernels, such as DeepLabv3+. Moreover, when comparing the results obtained for each class, UNETR outperformed DeepLabv3+ in the vast majority of cases. DeepLabv3+ achieved slightly better scores in specific categories, namely ”bathimodiolus azoricus” and ”orange,” but only within the ”ET edifice” site.

When analyzing the per class results, it is interesting to observe that for the WC dataset, both models failed to recognize the ”blue glass sponge” category. This is likely due to the low number of training samples representing that species. Additionally, the species occupy very small areas in the images, which may make their recognition particularly difficult. Similarly, poor results were also obtained for the ”bathimodiolus azoricus” and ”microbial mats” classes, for the WC dataset, in which the occurrence of those classes are also low (refer to Figure 4). It is worth highlighting that the classes ”bathimodiolus azoricus” and ”microbial mats” use to appear together or in close proximity (see figures 5(g)(h)(i)), so that the specialists created a category denominated ”bathimodiolus azoricus + microbial mats”. This represents a challenge in terms of the models’ generalization capacity. On the other hand, both architectures performed more accurately when recognizing ”b. azoricus + microbial mats” than separately (refer to Table 1). An interesting case is class ”cataetyx laticeps” for which DeepLabv3+ performed poorly, while UNETR consistently achieved fair recognition scores in all sites. The scores were as follows: 10.2% vs. 70%, 0.3% vs. 70.6%, and 1.9% vs. 59.1%. Again, these results underscore the efficiency of transformers in capturing global information, which seems to be particularly beneficial in scenarios where object characteristics align with morphotypes present within this category. Figures 5(b)(c) in the bottom left corner of both images shows examples of segmentation maps produced by the networks for this category (which is a fish species).

In general, considering results along the sites, models trained and evaluated in MS achieved the highest accuracies, and the lowest in WC. The results for the ET site were in between those obtained for MS and WC. After carefully analyzing such a behavior, it is worth noting that the results for ”bathimodiolus azoricus”, ”bathimodiolus azirocus + microbial mats”, ”blue glass sponge” and ”microbial mats” in WC and ET were lower than in MS, for both architectures. Interestingly, such categories are more frequent in MS than in WC and ET, and are less represented in WC (refer to Figure 4). Thus, such underrepresentation adversely affected the overall performance of the models trained with images from the WC and ET sites. Additionally, both architectures struggle to classify ”orange” accurately and have obtained similar results among the three sites. The latter should be due to the abundance of shells and white fragments on the sea floor (see figures 5(e)(f)(h)(i)).

5. Conclusions

In this work, we compared convolutional and Vision Transformer (ViT)-based deep neural network architectures in the identification ofsix fauna species within the Lucky Strike Vent field. Utilizing RGB images obtained from a Remotely Operated Vehicle (ROV), we evaluated the performance of models with those architectures across three distinct sites within the hydrothermal vent environment: Montesegur (MS), White Castle (WC), and Eiffel Tower (ET).

The study unveiled the potential of the tested networks as an automated alternative for fauna characterization in deep-sea environments, where significant visual interpretation is often required. On average, UNETR consistently exhibited the highest accuracy across the conducted experiments, although DeepLabv3+ occasionally surpassed it in a small number of cases.

Analyzing the results for the individual classes, it became apparent that the limited availability of training samples and the intricacies of each morphotype influenced both UNETR and DeepLabv3+. Notably, ”microbial mats”, ”bathimodiolus azoricus”, and ”blue glass sponge” emerged as the most challenging species for semantic segmentation methods.

Given the success of Vision Transformers (ViTs) in the task addressed in this study, their capacity to capture long-term dependencies and global information proved crucial. However, we believe that exploring other approaches, such as Graph Neural Networks (GNNs), could be a valuable direction for further research in this context.

The study presented here represents a step in the monitoring of deep-sea environments, offering a more agile, less subjective, and more accurate approach. However, definitive and universally applicable conclusions regarding the strengths and limitations of automatic mapping methods necessitate further investigation, encompassing data that captures the full spectrum of environmental diversity. Our ongoing research endeavors will be dedicated to advancing towards this overarching goal.

6. Acknowledgments

We thank the captain and crew of the RV L’Atalante and the pilots of the ROV *Victor6000*. We also thank chief scientists

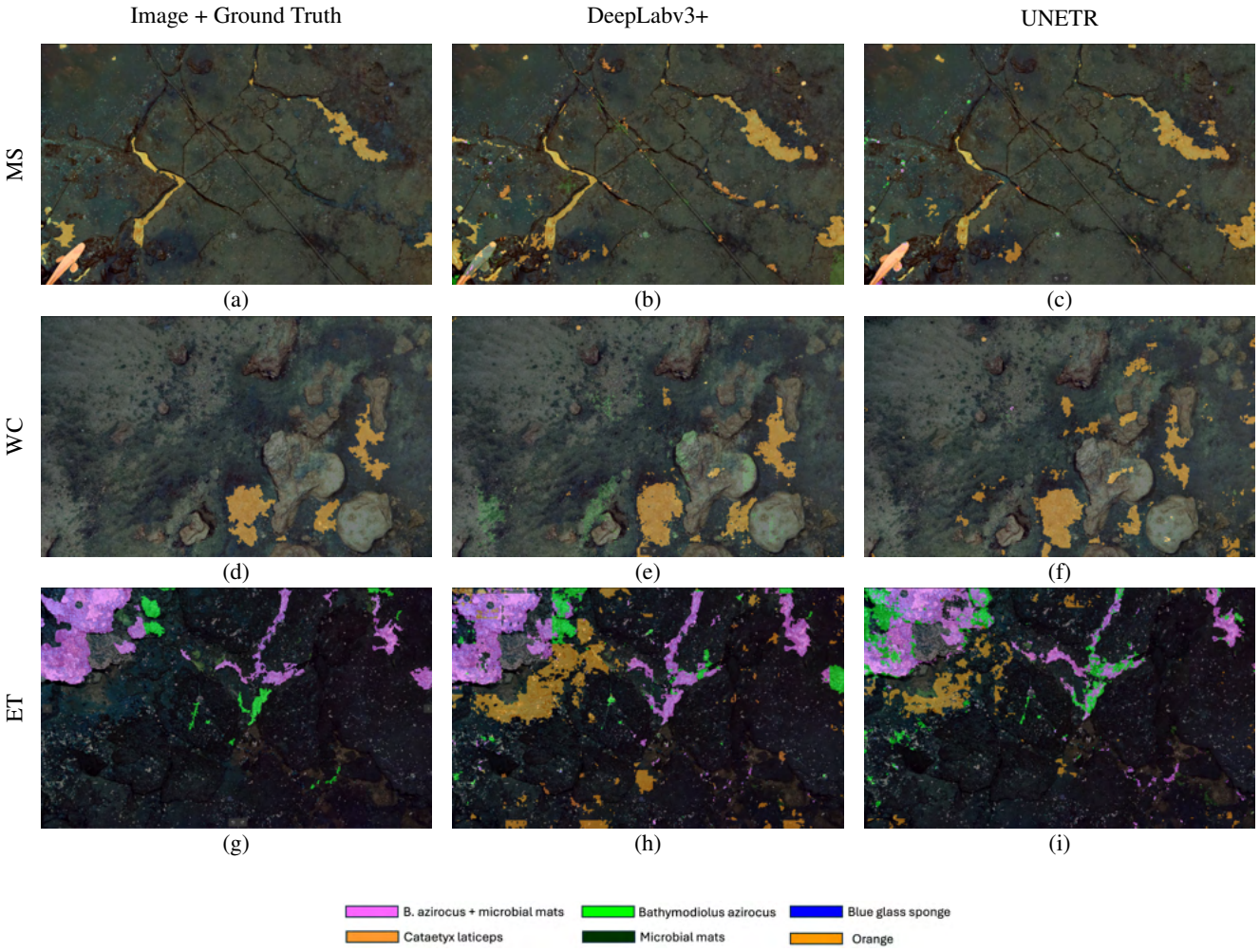

Figure 5. Semantic segmentation visual results obtained by DeepLabv3+ and UNETR.

Mathilde Cannat and Julien Legrand for co-leading the Momarsat 2018 cruise.

This work was supported by the European Union’s Horizon 2020 research and innovation project iAtlantic under Grant Agreement No. 818123. This output reflects only the authors’ view, and the European Union cannot be held responsible for any use that may be made of the information contained therein. The work is also supported by the region of Bretagne as part of the ABYSSES project. We also acknowledge financial support from the EU project EMSO (http://www.emso-eu.org/), EMSO-France, and the French observatory EMSO-Azores, funded by IFREMER and CNRS.

Finally, we acknowledge the support of the Brazilian Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S., 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2274–2281.

Arnaubec, A., Opderbecke, J., Allais, A.-G., Brignone, L., 2015. Optical mapping with the ariane hrov at ifremer: The matisse processing tool. *OCEANS 2015 - Genova*, 1–6.

Cannat, M., Sarradin, P., 2018. MOMARSAT2018 cruise, RV L’Atalante. *French Oceanographic Cruises doi*, 10, 18000514.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chollet, F., 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1251–1258.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Durden, J. M., Hosking, B., Bett, B. J., Cline, D., Ruhl, H. A., 2021. Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance. *Progress in Oceanography*, 196.

Faillettaz, R., Picheral, M., Luo, J. Y., Guigand, C., Cowen, R. K., Irisson, J. O., 2016. Imperfect automatic image clas-

sification successfully describes plankton distribution patterns. *Methods in Oceanography*, 15-16, 60–77.

Gao, R., 2023. Rethinking dilated convolution for real-time semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4675–4684.

Girard, F., Sarrazin, J., Arnaubec, A., Cannat, M., Sarradin, P.-M., Wheeler, B., Matabos, M., 2020. Currents and topography drive assemblage distribution on an active hydrothermal edifice. *Progress in Oceanography*, 187, 102397.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., Xu, D., 2022. Unetr: Transformers for 3d medical image segmentation. *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 574–584.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, 770–778.

Ierodiaconou, D., Monk, J., Rattray, A., Laurenson, L., Versace, V. L., 2011. Comparison of automated classification techniques for predicting benthic biological communities using hydroacoustics and video observations. *Continental Shelf Research*, 31(2 SUPPL.).

Juliani, C., Juliani, E., 2021. Deep learning of terrain morphology and pattern discovery via network-based representational similarity analysis for deep-sea mineral exploration. *Ore Geology Reviews*, 129, 103936.

Katija, K., Orenstein, E., Schlining, B., Lundsten, L., Barnard, K., Sainz, G., Boulais, O., Cromwell, M., Butler, E., Woodward, B. et al., 2022. FathomNet: A global image database for enabling artificial intelligence in the ocean. *Scientific Reports*, 12(1), 1–14.

Kim, H., Loh, W.-Y., 2001. Classification Trees with Unbiased Multiway Splits. *Journal of the American Statistical Association*, 96(454), 589–604.

Kingma, D. P., 2017. Variational Inference & Deep Learning: A New Synthesis. *PhD Thesis*, 1–162.

Langmuir C, Charlou Jean-Luc, Colodner D, C. I., Desbruyeres Daniel, Desonie D, Emerson T, Fornari D, Fouquet Yves, Humphris S, Fiala-Medioni A, Saldanha L, Sours-Page R, Thatcher M, Tivey M.K, Van Dover C, von Damm K, Wiese K, Wilson C, 1993. Lucky Strike - A newly discovered hydrothermal site on the Azores platform. *Ridge Events*, 4(2), 3–5. https://archimer.ifremer.fr/doc/00070/18096/.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loh, W.-Y., Shih, Y.-S., 1997. Split selection methods for classification trees. *Statistica sinica*, 7(4), 815–840.

Lu, H., Uemura, T., Wang, D., Zhu, J., Huang, Z., Kim, H., 2020. Deep-sea organisms tracking using dehazing and deep learning. *Mobile Networks and Applications*, 25(3), 1008–1015.

Lütjens, M., Sternberg, H., 2021. Deep Learning based Detection, Segmentation and Counting of Benthic Megafauna in Unconstrained Underwater Environments. *IFAC-PapersOnLine*, 54(16), 76–82.

Ma, C., Li, X., Li, Y., Tian, X., Wang, Y., Kim, H., Serikawa, S., 2021. Visual information processing for deep-sea visual monitoring system. *Cognitive Robotics*, 1, 3–11.

Marticorena, J., Matabos, M., Ramirez-Llodra, E., Cathalot, C., Laes-Huon, A., Leroux, R., Hourdez, S., Donval, J.-P., Sarrazin, J., 2021. Recovery of hydrothermal vent communities in response to an induced disturbance at the Lucky Strike vent field (Mid-Atlantic Ridge). *Marine Environmental Research*, 168, 105316.

Ondréas, H., Cannat, M., Fouquet, Y., Normand, A., Sarradin, P.-M., Sarrazin, J., 2009. Recent volcanic events and the distribution of hydrothermal venting at the Lucky Strike hydrothermal field, Mid-Atlantic Ridge. *Geochemistry, Geophysics, Geosystems*, 10(2).

Osterloff, J., Nilssen, I., Nattkemper, T. W., 2016. A computer vision approach for monitoring the spatial and temporal shrimp distribution at the LoVe observatory. *Methods in Oceanography*, 15-16, 114–128.

Pavoni, G., Corsini, M., Pedersen, N., Petrovic, V., Cignoni, P., 2021. Challenges in the deep learning-based semantic segmentation of benthic communities from Ortho-images. *Applied Geomatics*, 13, 131–146.

Peng, L., Zhu, C., Bian, L., 2023. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*.

Rajani, H., Gracias, N., Garcia, R., 2023. A convolutional vision transformer for semantic segmentation of side-scan sonar data. *Ocean Engineering*, 286, 115647.

Ramière, A., Matabos, M., Sarrazin, J., Borremans, C., Vega Pedro Juan, S., Marcillat, M., Cannat, M., Wheeler, B., Van Audenhaege, L., 2023. Seabed images and substrata of the southern lucky strike hydrothermal vent field. Technical report, SEANOE.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

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

Schmid, M. S., Aubry, C., Grigor, J., Fortier, L., 2016. The LOKI underwater imaging system and an automatic identification model for the detection of zooplankton taxa in the Arctic Ocean. *Methods in Oceanography*, 15-16, 129–160.

Schoening, T., Bergmann, M., Ontrup, J., Taylor, J., Dannheim, J., Gutt, J., Purser, A., Nattkemper, T. W., 2012. Semi-automated image analysis for the assessment of megafaunal densities at the Artic deep-sea observatory HAUSGARTEN. *PLoS ONE*, 7(6).

Schoening, T., Osterloff, J., Nattkemper, T. W., 2016. RecoMIA—Recommendations for marine image annotation: Lessons learned and future directions. *Frontiers in Marine Science*, 3, 59.

Sharma, R., Sankar, S. J., Samanta, S., Sardar, A. A., Gracious, D., 2010. Image analysis of seafloor photographs for estimation of deep-sea minerals. *Geo-Marine Letters*, 30(6), 617–626.

Shashidhara, B. M., Scott, M., Marburg, A., 2020. Instance segmentation of benthic scale worms at a hydrothermal site. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1314–1323.

Sun, Y., Wang, X., Zheng, Y., Yao, L., Qi, S., Tang, L., Yi, H., Dong, K., 2022. Underwater object detection with swin transformer. *2022 4th International Conference on Data Intelligence and Security (ICDIS)*, IEEE, 422–427.

Torres, D. L., Turnes, J. N., Juan, P., Vega, S., Feitosa, R. Q., Silva, D. E., Junior, J. M., Almeida, C., 2021. Deforestation Detection with Fully Convolutional Networks in the Amazon Forest from Landsat-8 and Sentinel-2 Images. 1–20.

Vega, P. J. S., Papadakis, P., Matabos, M., Van Audenhaege, L., Ramiere, A., Sarrazin, J., da Costa, G. A. O. P., 2024. Convolutional neural networks for hydrothermal vents substratum classification: An introspective study. *Ecological Informatics*, 102535.

Villon, S., Mouillot, D., Chaumont, M., Darling, E. S., Subsol, G., Claverie, T., Villéger, S., 2018. A Deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics*, 48, 238–244.

Xue, B., Huang, B., Chen, G., Li, H., Wei, W., 2021. Deep-sea debris identification using deep convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 8909–8921.

Yang, G., Kang, G., Lee, J., Cho, Y., 2023. Joint-ID: Transformer-based Joint Image Enhancement and Depth Estimation for Underwater Environments. *IEEE Sensors Journal*.