Abstract

Drone-based remote sensing combined with AI-driven methodologies has shown great potential for accurate mapping and monitoring of coral reef ecosystems. This study presents a novel multi-scale approach to coral reef monitoring, integrating fine-scale underwater imagery with medium-scale aerial imagery. Underwater images are captured using an Autonomous Surface Vehicle (ASV), while aerial images are acquired with an aerial drone. A transformer-based deep-learning model is trained on underwater images to detect the presence of 31 classes covering various coral morphotypes, associated fauna, and habitats. These predictions serve as annotations for training a second model applied to aerial images. The transfer of information across scales is achieved through a weighted footprint method that accounts for partial overlaps between underwater image footprints and aerial image tiles. The results show that the multi-scale methodology successfully extends fine-scale classification to larger reef areas, achieving a high degree of accuracy in predicting coral morphotypes and associated habitats. The method showed a strong alignment between underwater-derived annotations and ground truth data, reflected by an AUC (Area Under the Curve) score of 0.9251. This shows that the integration of underwater and aerial imagery, supported by deep-learning models, can facilitate scalable and accurate reef assessments. This study demonstrates the potential of combining multi-scale imaging and AI to facilitate the monitoring and conservation of coral reefs. Our approach leverages the strengths of underwater and aerial imagery, ensuring the precision of fine-scale analysis while extending it to cover a broader reef area.

Keywords: coral reef monitoring, computer vision, knowledge distillation, marine biodiversity, multi-scale imaging.

From underwater to aerial: a novel multi-scale knowledge distillation approach for coral reef monitoring

Matteo Contini[1,2,*], Victor Illien[1], Julien Barde[4], Sylvain Poulain[4], Serge Bernard[3], Alexis Joly[2], and Sylvain Bonhommeau[1]

[1]IFREMER Délégation Océan Indien (DOI), Le Port, 97420, La Réunion, France, Rue Jean Bertho
[2]INRIA, LIRMM, Université de Montpellier, CNRS, Montpellier, 34000, France
[3]CNRS, LIRMM, Université de Montpellier, Montpellier, 34000, France
[4]UMR Marbec, IRD, Université de Montpellier, CNRS, Ifremer, Montpellier, 34000, France
[*]corresponding author: Matteo Contini (firstname.lastname at ifremer.fr)

1 Introduction

Coral reefs are among the richest ecosystems on Earth in terms of species diversity. Moreover, they provide a number of key services: they function as natural barriers safeguarding coastlines from erosion and extreme weather events and serve as habitats and breeding grounds for innumerable marine species [1]. Additionally, they support local economies by offering resources for fishing, tourism and potential medicinal compounds [2], [3]. However, these ecosystems are under serious threat from human activities. Destructive and illegal fishing practices [4], anthropogenically derived chemical pollutants [5] and coastal development [6] are some of the main causes of coral reef degradation. Climate change poses an even greater risk through ocean acidification and warming, leading to widespread coral bleaching and habitat loss [7].

In December 2022 the Global Biodiversity Framework was adopted at the 15th Conference of the Parties (COP15) with the objective of protecting 30% of Earth's lands, oceans, coastal areas, and inland water by 2030 [8]. This ambitious goal requires the development of innovative monitoring techniques to assess the status of marine ecosystems and guide conservation efforts.

New techniques based on deep learning are emerging, offering the potential to revolutionize marine monitoring. In [9, 10] the authors developed a deep-learning based method for monitoring marine biodiversity using environmental DNA (eDNA). Besides, [11] proposed the use of convolutional neural networks (CNNs) to predict species distributions in the open ocean by leveraging environmental data and species occurrences. Artificial Intelligence (AI) models can also be used, coupled with Autonomous Surface Vehicles (ASVs), to classify spatialised underwater images generating high resolution species distribution maps [12] This approach enables fine-grained annotations and predictions, allowing the distinction of coral morphotypes and the identification of specific classes, such as algae, that are often difficult to discern. The main drawback of survey methods based solely on ASVs is that, since underwater images are taken at a very fine scale, it is difficult to cover large areas of coral reefs. This implies that, when areas on the order of tens of hectares are to be monitored, the method poses challenges in terms of cost, processing time and ease of deployment.

Recent advances in imaging technologies have opened up new possibilities for large-scale coral reef monitoring. Drone-based imaging has emerged as a valuable tool for coastal habitat mapping and monitoring, providing a cost-effective method for high-resolution habitat classification when combined with machine learning techniques [13]. The main limitation of using only aerial images is that they do not directly provide detailed information on individual benthic organisms. Thus, the annotation of aerial images is usually made on broader classes (e.g., hard bottom, mixed substrate, soft bottom and seagrass). Here we argue that they contain sufficient information to infer the benthic community as a whole if they are combined with fine-grained annotations inferred from underwater images.

To give an illustration of this problem, Figure 1 shows an orthophoto obtained from aerial images on the left and an underwater image on the right, collected by an ASV, corresponding to the zone delimited by the red rectangle. The medium-scale image on the left shows a complex assemblage of corals. However, due to the resolution of the image, although one can guess the presence of several morphotypes of corals, identifying them individually remains a difficult task. This challenge becomes significantly easier when the corresponding underwater images are available. The underwater image on the right provides clear distinctions between coral assemblages and other classes, such as *Algae*, which are often difficult to observe in aerial imagery. Propagating those fine-grained annotations to the aerial image yields a finer classification of the benthic habitat.

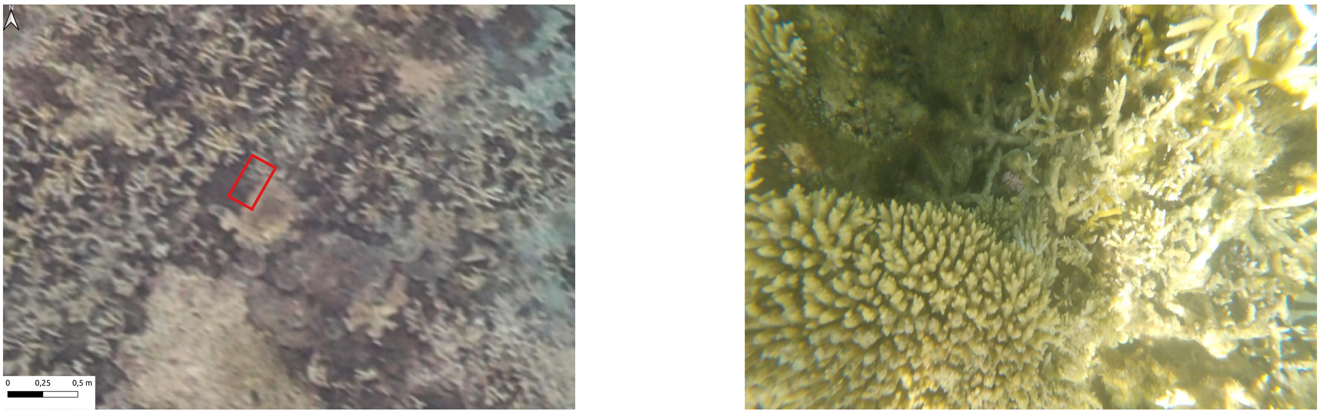


Figure 1: On the left, an aerial orthophoto captures a complex coral assemblage. However, distinguishing between specific morphotypes is challenging because of the limited resolution. The red rectangle highlights the location of the fine-scale underwater image shown on the right. The underwater image provides a higher level of detail, allowing the identification of distinct coral morphotypes and specific classes, such as *Algae*, that are difficult to identify in aerial imagery.

Integrating aerial imagery with human-collected ground truth data can be a first solution to map coastal habitats with high accuracy, as demonstrated in [14]. The authors provide a detailed protocol, from drone imagery collection to orthophoto annotation through GIS softwares, allowing the training of segmentation CNN models on aerial images. In [15], the authors use both aerial and underwater imagery, processed with Support Vector Machines (SVM) and Object-Based Image Analysis (OBIA) for benthic habitat classification. They used UAVs (Unmanned Aerial Vehicles) to capture high-resolution aerial imagery of coastal areas and USVs to collect ultra-high-resolution underwater images. The data from both platforms were processed separately using Structure from Motion (SfM) to create orthophoto mosaics and Digital Surface Models (DSMs). These products were then classified using OBIA and SVM algorithms. However, their approach does not fully leverage the complementarity of the two data sources. Indeed, there is an opportunity to exploit areas where both types of data (fine-scale underwater and medium-scale aerial) are available, in order to improve the model. In particular, they do not employ data from the same zones to train models that can infer fine-scale details from medium-scale data in regions where only aerial imagery is available. This limitation precludes the potential for synergistic use of overlapping datasets to enhance benthic habitat classification on a broader scale.

The aim of this study is to introduce a novel multi-scale deep-learning approach that integrates underwater and aerial imagery for fine-grained assessment of coral reefs at broad spatial scales.

Specifically, we used a transformer-based deep learning model to predict the presence/absence of 31 different classes of corals, associated fauna and habitats in the underwater images. To extend this classification to aerial images, we employed the concept of knowledge distillation [16], where the underwater model acts as the teacher and the aerial model as the student. The objective is for the student model to learn from teacher's outputs, allowing it to achieve comparable performance by mimicking the teacher's knowledge. Concretely, as illustrated in Figure 2, a first fine-scale model (the teacher) is trained on the underwater images associated with fine-grained manual annotations. A second model is then trained on the aerial images, using the underwater predictions and image metadata to generate annotations.

The transfer of information across the two scales is achieved through a weighted footprint method that accounts for partial overlaps between underwater image footprints and aerial image tiles.

By reducing the time-consuming annotation process to a single step on underwater images, this approach allows the aerial model to classify images at a larger scale while maintaining as much as possible the fine-scale

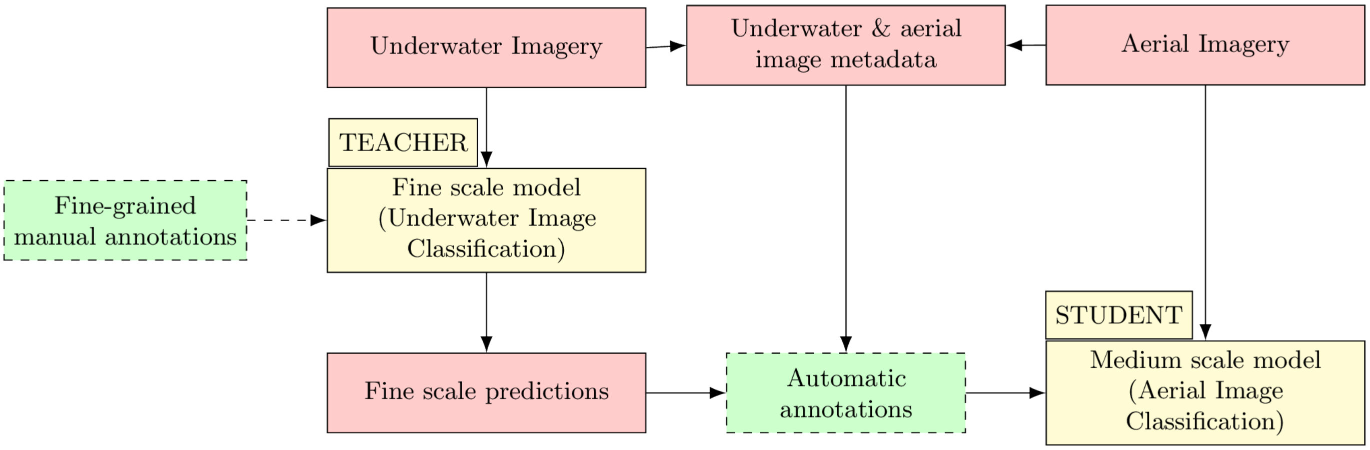

Underwater Imagery
Underwater & aerial image metadata
Aerial Imagery
TEACHER
Fine-grained manual annotations
Fine scale model (Underwater Image Classification)
STUDENT
Fine scale predictions
Automatic annotations
Medium scale model (Aerial Image Classification)

Figure 2: Workflow of the multi-scale approach for coral reef monitoring.

information provided by the underwater model.

This research offers a powerful tool for coral reef monitoring through an accurate classification of coral morphotypes and associated marine organisms. Compared to traditional ASV surveying techniques, this method provides significant advantages in terms of cost efficiency, as it reduces the human time required for ASV deployments to cover the same surface area. UAVs are also easy to deploy; from the beach, we can access sites that are kilometers away without losing control of the drone. Additionally, this approach eliminates the risk of sending ASVs into dangerous zones where they could become stuck or damaged, ensuring safer operations and minimal disturbance to the reef ecosystem. By demonstrating the combination of advanced technologies such as ASVs, drones, and deep learning models, the study contributes to the development of more effective and efficient conservation strategies, showcasing the potential of multi-scale monitoring for environmental protection. To our knowledge, this is the first time that such a cascade of deep learning models has been used to classify aerial images. This opens up new possibilities for upscaling a computer vision model trained on fine-scale images to a larger scale. Although being developed for marine applications, this method could be used for terrestrial ecosystems.

2 Materials and methods

2.1 Underwater image acquisition

Underwater images were collected using an Autonomous Surface Vehicle (ASV) equipped with a *GoPro Hero 8* camera and a differential GPS *Emlid* Reach M2 mounted on a waterproof case. The version of the ASV builds on a previous version developed in [17]. In order to end up with georeferenced images embedded with attitude metadata (roll, pitch and yaw angles), the following steps were taken:

1. Time synchronization between the camera and the GPS clocks.

2. GPS position correction.

3. Bathymetry data correction using local geoid parameters and attitude data of the ASV.

4. Image georeferencing using the corrected GPS position and attitude data.

50 missions were carried out in the lagoon of Reunion Island: 30 in the Saint-Leu lagoon and 20 in the Trou d'eau lagoon. For additional details regarding the processes of time synchronization, metadata correction, and other technical aspects, please refer to Appendix A where the corresponding subsections are discussed in depth. For further details on time synchronization, GPS position correction and image georeferencing, please refer to Appendix A, where these aspects are explained in detail.

2.2 Aerial image acquisition

Aerial drone images were taken with a `DJI Mavic 2 Pro` drone. Images were collected following good practices in use in aerial imagery [18].

Since this drone is not equipped with a differential GPS, once images were taken and the SfM model was built, the orthophoto was georeferenced by collecting ground control points (GCPs) using a differential GPS.

To obtain a high-resolution orthophoto with high positioning precision, the following steps were taken:

1. Mission planning: check the equipment, request authorizations from French authorities, weather conditions and plan the flight mission.
2. Mission execution: fly the drone at an altitude of 60m over the area of interest adapting camera settings to the specific conditions of the day.
3. Image processing: build the Structure from Motion (SfM) model using images taken during the flight mission. This was done using `OpenDroneMap`.
4. GCP collection: collect GCPs using a GPS with centimetric accuracy.
5. Orthophoto georeferencing: georeference the orthophoto using the GCPs.

Two missions were carried out in the lagoon of Reunion Island: one in the *Saint-Leu* lagoon and the other in the *Trou d'eau* lagoon. For further details on mission planning, execution, image processing, and orthophoto georeferencing, please refer to Appendix B, where these aspects are explained in detail.

2.3 Multi-scale positioning

Since the objective is to pass information from a fine scale (underwater images) to a larger scale (drone images), the precision of the relative position between underwater and drone images is crucial.

In order to validate the georeferencing of underwater images with respect to aerial images, we used data from IGN (Institut national de l'information géographique et forestière[1]). Each 3 to 4 years, IGN produces *BD ORTHO®*: a collection of orthophotos with a default resolution of 20 cm. The last orthophoto produced by IGN on Reunion island was in 2022, so we decided to use this data as a reference to validate the georeferencing of our data.

Two visual criteria were then used to confirm the georeferencing of the underwater and aerial images with respect to the *BD ORTHO®* orthophoto:

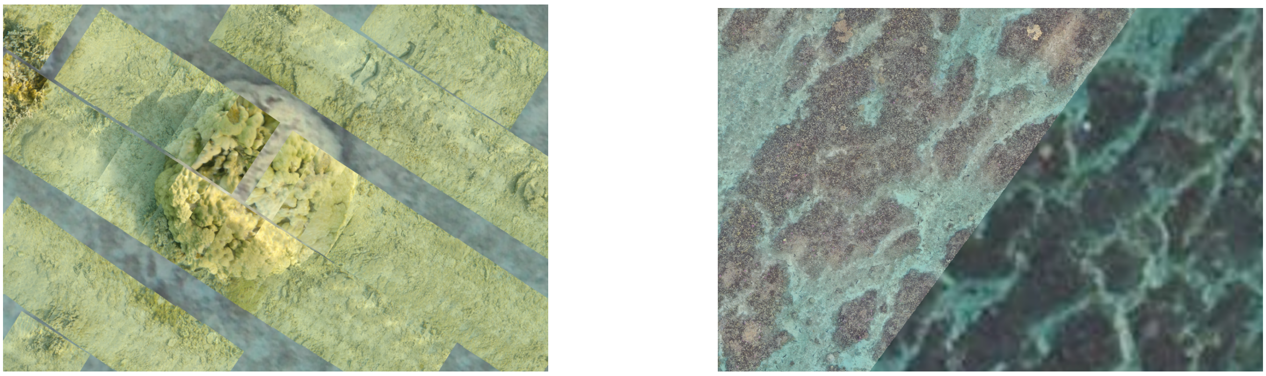
- Relative georeferencing: find the presence of easily recognizable objects in both underwater and aerial orthophoto and compare them on a GIS software. Often the presence of *Porites* corals can be used to compare the two scales since their contours are easily recognizable in both types of images. See Figure 3a.
- Aerial absolute georeferencing: find the presence of easily recognizable coral colonies in both aerial and *BD ORTHO®* orthophoto and compare them on a GIS software. See Figure 3b.

The combination of the relative georeferencing between underwater and aerial images and the aerial absolute georeferencing with respect to the *BD ORTHO®* orthophoto allowed us to crosscheck the georeferencing of the underwater images with respect to national baseline data.

2.4 Underwater image classification

The underwater deep learning model builds on the *DinoV2* architecture, which is a vision transformer model that has been shown to outperform convolutional neural networks on image classification tasks [19]. The model has been trained on the open source dataset Seatizen Atlas image dataset composed of 51 distinct

[1] IGN is the French public state administrative establishment that has the main objective of producing and maintaining geographical information for France and its overseas departments and territories



(a)

(b)

Figure 3: Visual georeferencing criteria to validate the georeferencing of underwater and aerial images with respect to the *BD ORTHO®* orthophoto. On the left (a) underwater images georeferenced with respect to the aerial orthophoto. On the right (b) aerial orthophoto georeferenced with respect to the *BD ORTHO®* orthophoto produced by the French National Geographic institute (IGN). The lighter part on the left corresponds to the drone-based orthophoto and the darker part on the right corresponds to the *BD ORTHO®*.

classes of corals, associated fauna, and habitats. The model architecture and hyperparameters settings are described in Appendix C.

Once the model was trained, we ran inference on 56,653 georeferenced images in *Trou d'eau* lagoon and 58,076 images in *Saint-Leu* lagoon in Reunion Island which are included in the area covered by an aerial drone.

For more information on the data splitting technique used to train the model, please refer to Appendix D.

2.5 Upscaling predictions

Once the orthophoto is georeferenced and underwater inference is done, the key step is to correctly pass the information from the underwater model to the aerial model. The objective is to train an aerial model based on underwater predictions, without spending time on manual annotations of aerial images.

This is achieved by following the steps below:

1. Split aerial orthophoto into tiles, ensuring consistency in ground surface representation of each tile across different sessions. Each tile represents an area of 1.5m x 1.5m. See Section 2.5.1.

2. Filter useless aerial tiles (e.g., black tiles issued from SfM processing errors or tiles without corresponding underwater images). See Section 2.5.2.

3. Associate each aerial tile with underwater images whose camera GPS position is within the tile boundaries. See Section 2.5.3.

4. Compute the footprint of each underwater image and filter aerial tiles with not enough underwater coverage. See Section 2.5.3.

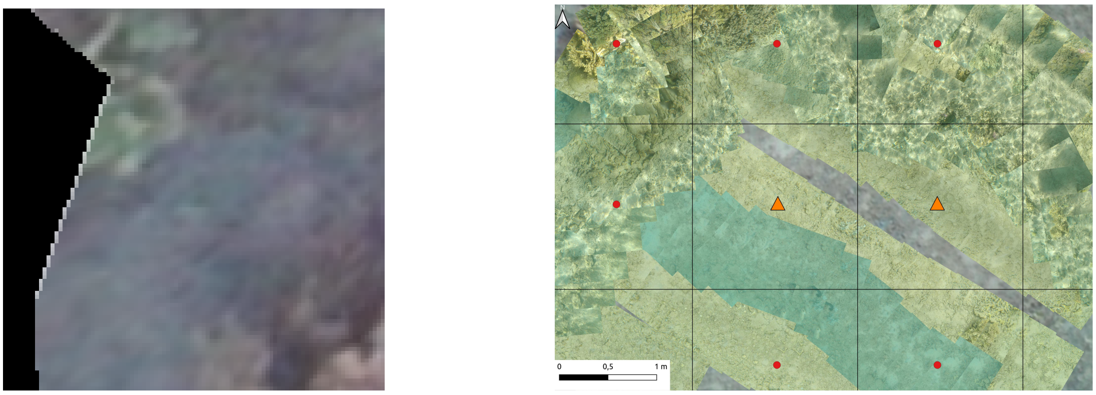
5. Transform underwater predictions into aerial annotations. See Section 2.5.4.

2.5.1 Orthophoto tiling

The first step is to split the aerial orthophoto into tiles. This is done by taking into account the ground sample distance (GSD) of each orthophoto. This approach guarantees that while tile dimensions in pixels may vary due to different GSDs, each tile consistently represents a fixed area on the ground. This method allows for standardized comparison and analysis of images across different datasets and sessions, maintaining a consistent spatial resolution. Splitting the orthophoto into too small tiles results in images without enough context to be correctly classified and/or with an insufficient resolution. On the contrary, splitting the orthophoto into too large tiles results in good classification performances but does not allow for a fine-grained analysis of the data. Searching for the best compromise, we fix this area to be 1.5 m x 1.5 m.

2.5.2 Useless tiles filtering

The second step is to filter out useless tiles. This is done by removing tiles with a high percentage of black pixels (due to errors in SfM processing) and tiles with no corresponding underwater images. Examples of such tiles are shown in Figure 4. In Figure 4a we can see an example of a tile extracted from the aerial orthophoto of the *Saint Leu* lagoon in Reunion Island with a high percentage of black pixels.



(a)

(b)

Figure 4: Examples of useless tiles extracted from the aerial orthophoto of the *Saint Leu* lagoon in Reunion Island: (a) Example of a tile extracted from the aerial orthophoto of the *Saint Leu* lagoon in Reunion Island, with a high percentage of black pixels (b) Example of a group of tiles extracted from the aerial orthophoto of the *Saint Leu* lagoon in Reunion Island, with corresponding underwater images. The tiles in the middle do not have enough coverage of underwater images

2.5.3 Footprint calculation and tile coverage assessment

The third step involves associating underwater predictions with aerial tiles. This assignment is achieved by identifying underwater images whose camera position centre falls within the boundaries of the aerial tiles. After associating underwater images to aerial tiles, the next step is to compute the footprint of each underwater image to filter out aerial tiles with not enough underwater coverage. In Figure 5, we outline the process to calculate the footprint of underwater images based on data from ASV sensors. Using bathymetric data from the echosounder, we measure the distance between the camera and the seabed, which determines the scale of the area captured in each image. The camera orientation in the XYZ axis plane is defined by the roll, pitch, and yaw angles, which determine how the field of view (FOV) is directed relative to the seafloor. Finally the FOV, divided into horizontal (FOV_h) and vertical (FOV_v) angles, defines the area visible to the camera. By projecting these angles down to the seafloor, we calculate the intersection points, forming a polygonal footprint that represents the region covered by the image. This footprint is necessary to associate each underwater image with a specific area of the seafloor. Merging the footprints of all underwater images associated with a tile, we obtain the union of the footprints, which represents the area covered by the underwater images associated with the tile.

This allows us to filter out tiles with not enough underwater coverage. An example of such a tile is shown in Figure 4b, where a group of tiles extracted on the same orthophoto is shown. Tiles whose center is represented by a red point are classified as useful, since they are completely covered by underwater images. On the contrary, tiles represented by an orange triangle are classified as useless, since they do not have enough coverage of underwater images.

2.5.4 Transforming underwater predictions into aerial annotations

The fifth step is to transform underwater predictions into aerial annotations. The trivial approach would be to associate the presence of a class c in a tile t if at least one underwater image associated with the tile is

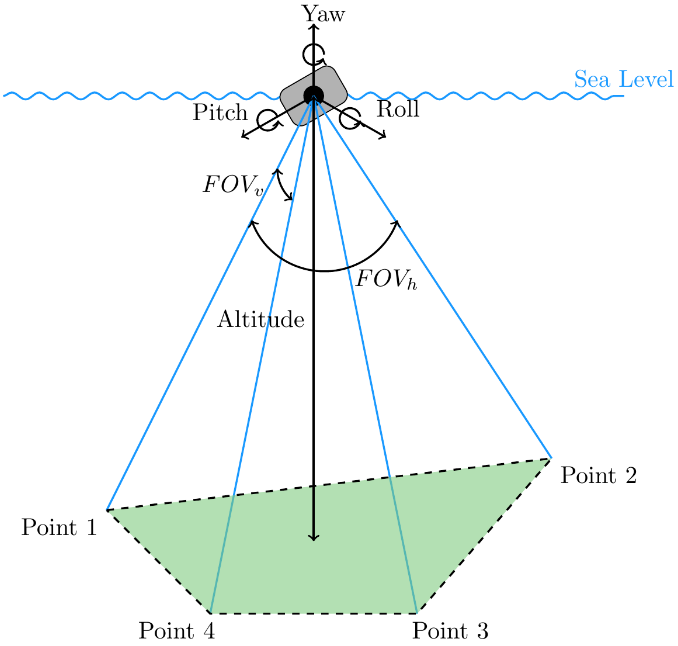


Figure 5: Footprint calculation of underwater images based on echosounder data, camera field of view and ASV angles.

predicted as belonging to the class c by the teacher model. Or, in other words, if class c is not predicted as being absent on all underwater images associated with tile t, which can be formulated as:

$$\forall c \in \mathcal{C}, \forall t \in \mathcal{T}, \quad I(y_c = 1 \mid t) = 1 - \prod_{x \in \mathcal{X}(t)} \left[1 - h_c^{\text{teacher}}(x)\right] \tag{1}$$

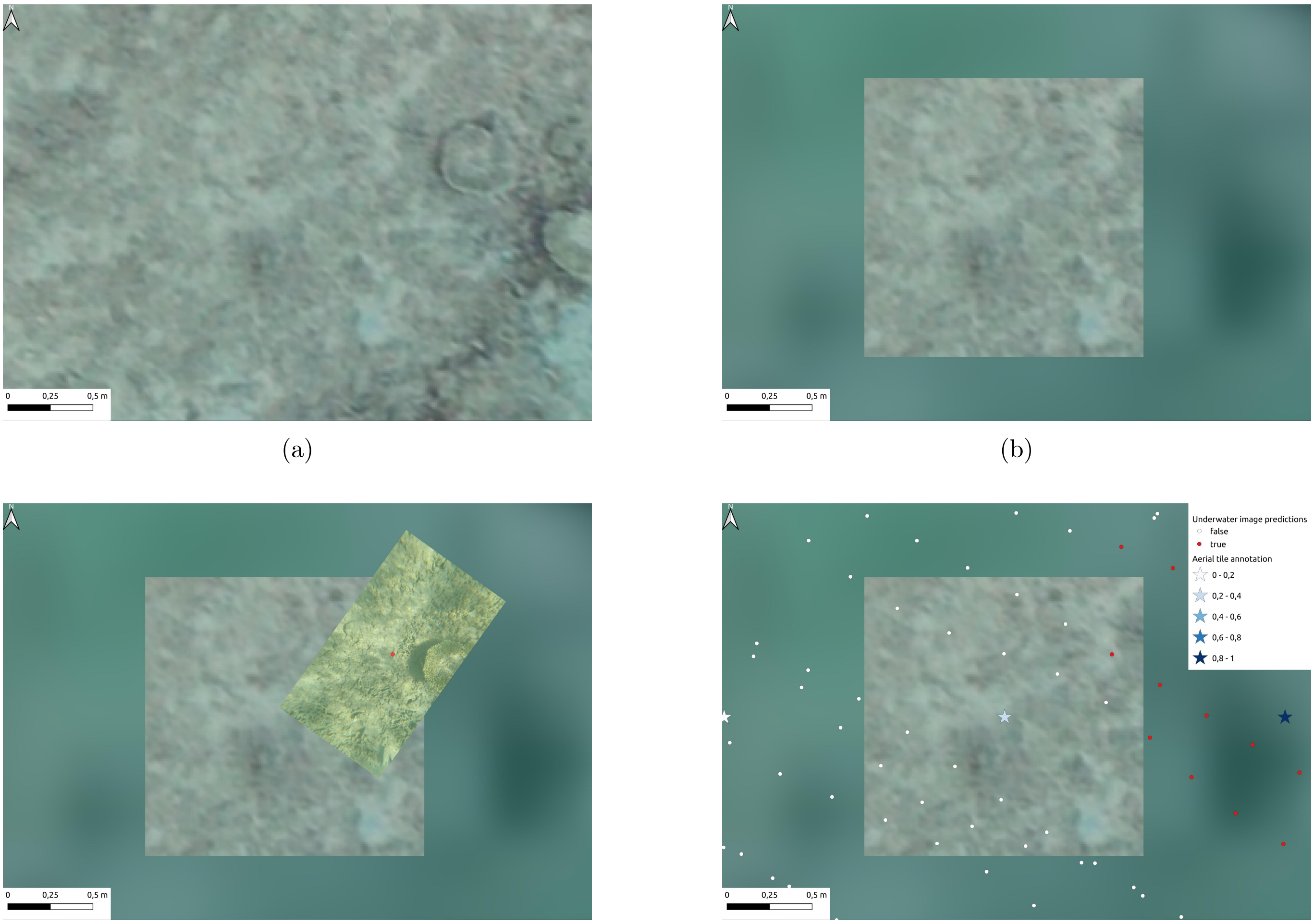
where :

- $\mathcal{C}$ is the set of classes described in Section 2.5.5
- $\mathcal{T}$ is the set of tiles
- $y_c \in \{0; 1\}$ is the binary label associated with the presence/absence of class c
- $I(y_c = 1 \mid t) \in \{0; 1\}$ is a binary function indicating the presence or absence of class c in tile t.
- $\mathcal{X}(t)$ is the set of underwater images associated with tile t
- $h_c^{\text{teacher}}(x) \in \{0; 1\}$ is the binary prediction associated with the presence/absence of class c in underwater image x

The drawback of this approach is that it does not consider the footprint of underwater images, tending to overestimate the presence probability of a class in a tile.

A more realistic approach, since not all underwater images footprint fall entirely within the boundaries of a specific tile, needs to compute the intersection between the underwater image footprint and the corresponding tile. This allows us to give more weight to underwater images that are completely within a tile and less weight to underwater images that are only partially within a tile.

The orthophoto in Figure 6 gives an example with the corresponding predictions on underwater images. In the right part of Figure 6a, a colony of *Acropora Tabular* corals is visible. Proceeding with tile extraction from the orthophoto, we obtain the tile in Figure 6b. Since the *Acropora Tabular* corals do not fall within the tile, we would like that, after computing the tile annotation starting from underwater predictions, the probability for the class *Acropora Tabular* associated with this tile will be weak. Unfortunately, it may happen that underwater images that have the center within the tile (but not all the footprint) include classes that are outside the tile bounds: as shown in Figure 6c, where a part of the *Acropora Tabular* coral colony is visible in the right part of the underwater image. In these cases, weighting underwater predictions based on the intersection between the underwater image footprint and the tile allows reducing the impact of these images on the final aerial annotations. This is shown in Figure 6d where predictions on underwater images are represented with circles on a red ramp and aerial annotations are represented with stars on a blue ramp. Even if in the underwater image in Figure 6c on the right of the tile the presence of the *Acropora Tabular* class is predicted, the overlap between the underwater image and the aerial tile is weak. Consequently, the probability of presence of the *Acropora Tabular* class on the tile is mitigated: ending up with an annotation of 0.4 (while the blue star on the tile just on the right indicates a probability of presence of 0.98).



(a)

(b)

(c)

(d)

Figure 6: Example of the upscaling process from underwater predictions to aerial annotations, depending on the intersection between underwater images footprint and drone tiles, the probability of presence of a class in a tile can be mitigated: (a) Aerial orthophoto of the *Trou d'eau* lagoon in Reunion Island (in the right part of the image a colony of *Acropora Tabular* corals is visible), (b) Drone tile extracted from the aerial orthophoto in Figure 6a (the *Acropora Tabular* corals are not visible in the tile), (c) Superposition of the underwater image in the top right corner of the tile and the drone tile (the *Acropora Tabular* coral colony is outside the drone tile), (d) Underwater predictions associated with the tile in Figure 6b (the *Acropora Tabular* corals is predicted on the top right underwater image of the tile, but weakly predicted on the tile)

To take account of the intersection between the underwater image footprint and the tile bounds, we can consider that the probability of presence is proportional to the intersection area relative to the underwater image area. Therefore, we can modify Equation 1 as follows:

$$\forall c \in \mathcal{C}, \forall t \in \mathcal{T}, \quad P(y_c = 1 \mid t) = 1 - \prod_{x \in \mathcal{X}(t)} \left[1 - \frac{s(x \cap t)}{s(x)} \, h_c^{\text{teacher}}(x) \right] \tag{2}$$

where:

- $P(y_c = 1 \mid t) \in [0, 1]$ is the probability of presence of class c in tile t

- $s(x)$ is the area of the underwater image x
- $s(x \cap t)$ is the area of intersection between the underwater image x and the tile t

The product over all underwater images gives the probability that class c is absent in all underwater images associated with the tile t.

To get a better estimation of the presence probability, we can also take into account the confidence of the teacher model. Therefore, we can replace the binary output of the classifier $h_c^{\text{teacher}}(x)$ by the probabilistic output $p_{\text{teacher}}(y_c = 1 \mid x)$ leading to:

$$\forall c \in \mathcal{C}, \forall t \in \mathcal{T}, \quad P(y_c = 1 \mid t) = 1 - \prod_{x \in \mathcal{X}(t)} \left[1 - \frac{s(x \cap t)}{s(x)}\, p_{\text{teacher}}(y_c = 1 \mid x)\right] \tag{3}$$

where:

- $p_{\text{teacher}}(y_c = 1 \mid x) \in [0, 1]$ is the probabilistic output of the teacher model for class c in the underwater image x, obtained through a sigmoid function on top of the final layer of the model. If the model is trained with the binary cross-entropy loss function (as in our experiments), the output is asymptotically converging to the true conditional probability that class c is present in image x [20].

It is worth noting that when the probability $p_{\text{teacher}}(y_c = 1 \mid x)$ is equal to 1, then Equation 3 is equivalent to Equation 2. But in the general case, it is comprised in the interval $]0, 1[$. In the literature related to knowledge distillation [16], such probabilistic labels passed to the student model are often called *soft labels* in opposition to hard labels such as the one in Equation 1. Training the student model on soft labels rather than hard labels enables a better transfer of information from the teacher to the student. The probability of presence actually captures valuable information on the uncertainty of the teacher model, and allows us to place less weight on ambiguous cases in the loss function of the student.

2.5.5 Aerial dataset

Following the upscaling process detailed in Section 2.5, starting from two aerial orthophotos of the *Trou d'eau* and *Saint-Leu* lagoons in Reunion Island measuring 189,682 m^2 and 204,748 m^2 respectively, we ended up with 4,911 and 6,832 annotated tiles respectively for a total of 11,743 annotated tiles.

Since the upscaling process implies a loss in the image resolution, we made some changes about the classes to be predicted:

1. The first change was to merge *Algae* classes into a single class called *Algae*, indeed distinguishing between the different types of algae (*Algal Assemblage*, *Algae Halimeda*, *Algae Coralline* and *Algae Turf*) is a task that requires a higher resolution than the one we have[2].

2. The second change was to remove underwater classes that do not have a corresponding aerial class: *Blurred* images (an underwater blurred image does not imply a blurred aerial image) and *Homo Sapiens* (since human body parts in underwater images do not imply human body parts in aerial images).

3. The third change was to remove underwater classes that are not relevant for the aerial images, i.e. *Fish*, *Sea cucumber* and *Sea urchin*. The first two classes, even if visible in some aerial images, were removed because underwater and aerial images are not taken at the same time, so that the presence of a sea cucumber or a fish in an underwater image does not imply the presence of those organisms in the corresponding aerial image. The last one was removed since those organisms are not visible at all in aerial images.

Finally, we retained only classes for which there was a sufficient number of annotations. Thus, removing classes that have less than 200 annotations in the aerial dataset, we ended up with 12 classes:

- Coral

[2] In the case of Equation 2 this was done by assigning $h_{Algae}^{\text{teacher}}(x) = 1$ if at least one type of algae (*Algal Assemblage*, *Algae Halimeda*, *Algae Coralline* and *Algae Turf*) was predicted as present on the fine scale image x, otherwise $h_{Algae}^{\text{teacher}}(x) = 0$. In the case of Equation 3 this was done by assigning to $p_{\text{teacher}}(y_{Algae} = 1 \mid x)$ the maximum between all the probabilities predicted by the underwater model for algae classes (*Algal Assemblage*, *Algae Halimeda*, *Algae Coralline* and *Algae Turf*).

1. *Acropora Branching*
2. *Acropora Digitate*
3. *Acropora Tabular*
4. *Dead coral*
5. *Non-acropora Encrusting*
6. *Non-acropora Massive*
7. *Non-acropora Millepora*
8. *Non-acropora Submassive*

- Habitat

  1. *Rock*
  2. *Rubble*
  3. *Sand*

- Other Organisms

  1. *Algae*

2.5.6 Aerial deep learning model (student model)

To train the student model with soft labels, we use the Binary Cross-Entropy (BCE) with logits loss function. This loss measures the divergence between the predicted logits of the student model and the soft labels $P(y_c = 1 \mid t) \in [0, 1]$ generated by the teacher model. Specifically, the loss for class c in tile t is given by:

$$\mathcal{L}_{\text{BCE}}(t, c) = -\Big[P(y_c = 1 \mid t) \cdot \log(p_{\text{student}}(y_c = 1 \mid t)) + (1 - P(y_c = 1 \mid t)) \cdot \log(1 - p_{\text{student}}(y_c = 1 \mid t))\Big] \tag{4}$$

where:

- $P(y_c = 1 \mid t) \in [0, 1]$ is the soft label provided by the teacher model for class c in tile t, as described in equations 2 and 3
- $p_{\text{student}}(y_c = 1 \mid t) \in [0, 1]$ is the probabilistic output of the student model for class c in tile t, obtained through a sigmoid function on top of the final layer of the model.

To maintain consistency with underwater predictions, we used the same architecture for the student model as the one used for the teacher model (i.e. the DinoV2 model [19]).
The only difference is that, since the underwater model was trained with binary values and the aerial model has to be trained on probabilities, when computing evaluation metrics during the training process we cannot use the accuracy, precision, recall and F1-score metrics. Instead, we will compute the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE) and the Kullback-Leibler (KL) divergence metrics.

2.6 Test zone and model evaluation

To evaluate the performance of the aerial deep learning model, we selected a test zone within the *Trou d'eau* lagoon, see Figure 7. This area was chosen due to its diverse composition of coral morphotypes, habitats, and other marine organisms, representing a challenging environment for model validation. The test zone comprises 194 underwater images, corresponding to 28 aerial tiles, for a total of $63m^2$.

The annotation process for the aerial tiles is carried out as follows:

1. For each aerial tile, underwater images with centroids located within the tile boundaries are identified
2. These underwater images are then projected in QGIS to assess the portions of each image that intersect the aerial tile boundaries
3. Each cropped underwater image is manually annotated with fine-grained precision
4. As a result, each aerial tile is annotated with a level of detail comparable to that of underwater imagery and is therefore considered as ground truth data

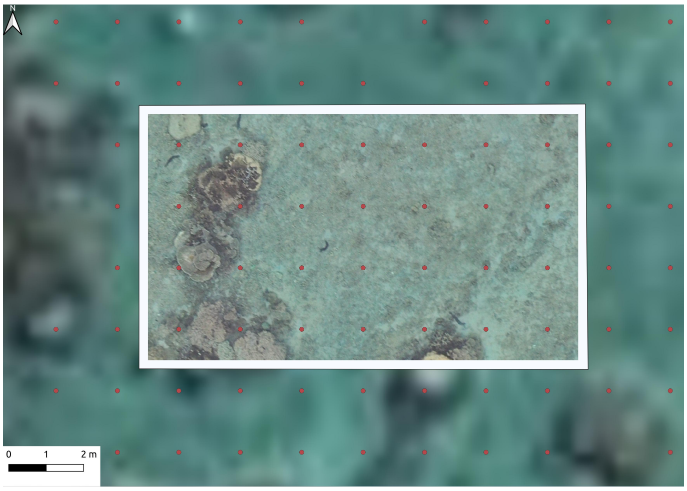


Figure 7: Test zone within the *Trou d'eau* lagoon, selected for model evaluation. The area measures $63m^2$ and comprises 194 underwater images corresponding to 28 aerial tiles.

These aerial tiles were not used during model training, ensuring an unbiased evaluation of the aerial model.

Both equations 2 and 3 were used to generate aerial annotations starting from underwater predictions in the test zone. The generated annotations were then compared with ground truth data to evaluate the goodness of the upscaling process using the AUC (Area Under the Curve) metric, which is commonly used metric in the evaluation of Species Distribution Models (SDMs) [21].

Finally, in order to evaluate the aerial model, we compared the predictions on the test zone with the ground truth data using the AUC metric.

3 Results

3.1 Upscaling process evaluation

In order to evaluate the upscaling process, we compared the generated annotations with ground truth annotations in the geospatial test zone in the *Trou d'eau* lagoon. We first evaluated the quality of the soft labels generated with our methods described in Equations 2 and 3. As they provide a presence probability for each class, we can actually measure their AUC on the ground truth annotations. With a value of 0.9211 for annotations generated through Equation 2 and 0.9251 for annotations generated through Equation 3, both methods show a high level of accuracy in transferring information across scales.

To further evaluate both methods, we then measured the performance of the student model trained with either method. In the following, we will call `Model_spatial_only` the model trained from annotations generated through Equation 2 and `Model_distilled` the model trained from the annotations generated through Equation 3. Only the second model integrates information about the teacher's model confidence (= knowledge distillation). The first model integrates the hard labels predicted by the teacher and the spatial coverage. We first looked at the evaluation metrics measured on the soft labels themselves (using the random test set). The results are shown in Table 1.

Model	RMSE	MAE	KL Divergence
`Model_spatial_only`	0.2019	0.1446	0.9802
`Model_distilled`	**0.1546**	**0.1143**	**0.3931**

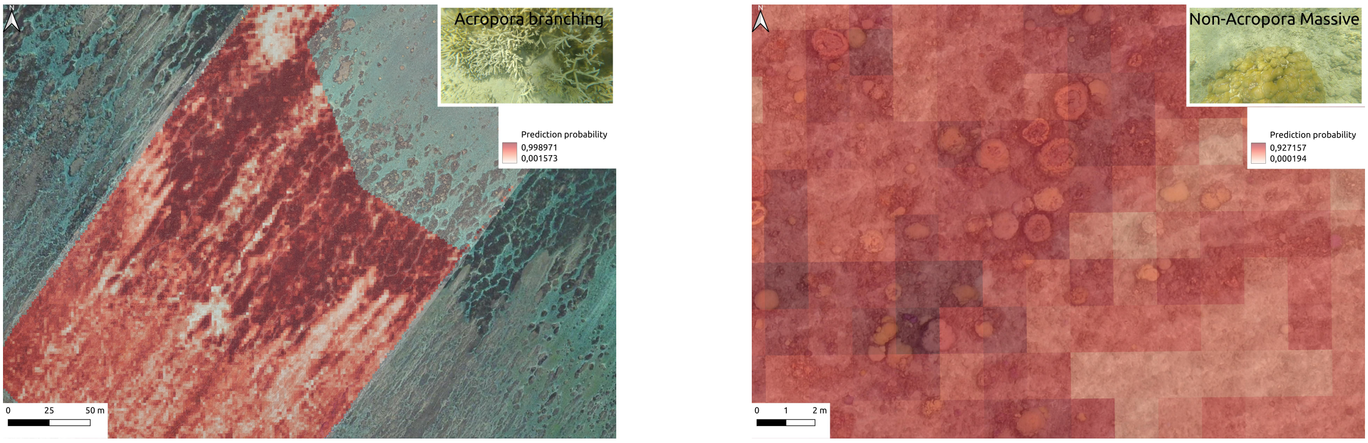
Table 1: Comparison of `Model_spatial_only` and `Model_distilled` on various performance metrics

The results show that `Model_distilled` trained using knowledge distillation (i.e. with Equation 3) allows a better prediction of the soft labels than `Model_spatial_only` trained without knowledge distillation (i.e. Equation 2) on all metrics. This means that the information they contain is more predictable from the aerial image contents.

Finally, comparing the predictions generated with both `Model_spatial_only` and `Model_distilled` on the ground truth data of the geospatial test zone in the *Trou d'eau* lagoon, we obtain an AUC of 0.7753 and 0.7952 respectively. This confirms that the best upscaling method is the one using knowledge distillation (Equation 3) and that high AUC values can be achieved by the aerial model using this method.

3.2 Prediction maps

Once the aerial (student) model is trained, we can use it in inference mode to generate high resolution maps of large areas. In particular, we ran it on 20,027 tiles in the *Trou d'eau* lagoon and 61,059 tiles in the *Saint-Leu* lagoons. For each tile, we used the output of the student model as the probability of presence of each class and then we generated prediction maps for each class.



(a)

(b)

Figure 8: Prediction maps generated by the aerial model superposed on the aerial orthophoto: (a) Prediction map for the *Acropora Branching* class in the *Trou d'eau* lagoon. Red patches indicate the presence of corals. (b) Zoomed-in prediction map for the *Non-acropora Massive* class in the *Saint-Leu* lagoon. Blue patches indicate the presence of corals.

In Figure 8 we show two examples of prediction maps generated by the aerial model for the *Acropora Branching* class in the *Trou d'eau* lagoon and the *Non-acropora Massive* class in the *Saint-Leu* lagoon. The granularity of the prediction raster is fixed at 1.5 m x 1.5 m, which is the same as the aerial tiles. Different spatial scales are shown in Figure 8a and Figure 8b, in order to highlight the model's ability to cover large areas while still being able to capture fine-scale details.

4 Discussion

This study demonstrates the potential of combining underwater and aerial imagery to improve monitoring and management of coral reef ecosystems. Although previous research has highlighted the advantages of using both imaging techniques, this work, to our knowledge, is the first to synergize AI models across different scales. By using high-resolution underwater AI predictions to train a larger-scale aerial model, we ensure the precision of underwater analysis while extending it to cover a broader reef area. This multi-scale approach has the potential to advance marine monitoring, and it can also be applied to other fields such as agriculture, forestry, urban planning, and so forth.

4.1 Transferring information across scales: model independence and flexibility

4.1.1 Upscaling process evaluation

While having a well-performing fine-scale model is needed for reliable medium-scale annotations, working on the model architecture in order to gain a few percentage points of accuracy is not the core of our methodology. The primary objective of our workflow is to train a medium-scale student model to mimic the behaviour of a fine-scale teacher model, without the need to reannotate medium-scale images.

Two different techniques were used to generate aerial annotations starting from underwater predictions: the first one was based on presence/absence values obtained by thresholding underwater predictions (Equation 2), while the second one was based on knownledge distillation, i.e., on integrating the probability values predicted by the underwater model in the soft labels passed to the student (Equation 3).

To compare and validate those methodologies, we evaluated them based on underwater predictions with ground truth data in the test zone in the *Trou d'eau* lagoon. The distillation-based method appeared to be the best with a high value of the AUC metric (0.9251). This indicates that it is a reliable method for transferring information from fine-scale to medium-scale images, allowing for a more nuanced estimate of class presence compared to using binary predictions 2.

Indeed, the soft labels generated by the teacher model provide a more accurate representation of class presence compared to the hard targets, since they capture the probability distribution across all possible classes, offering continuous values between 0 and 1. This approach allows the student model to better learn from the teacher model's uncertainty, leading to improved generalization capabilities and performances [22]. This result is consistent with the performance of the aerial model trained with annotations generated by Equation 3, which outperformed the model trained with annotations generated by Equation 2 on all metrics.

A key advantage of this methodology is its model independence. If a more accurate or advanced model becomes available in the future (thanks to new annotated images or improved algorithms), our framework allows for easy adaptation. By applying the new model to our existing fine-scale data, we can regenerate high-resolution predictions, which can then be seamlessly transferred to the medium scale. This adaptability ensures that the model continues to benefit from the most accurate fine-scale insights.

4.1.2 Aerial model evaluation

The model's performance on the neural network test set underscores its robustness. As shown in Table 1, the RMSE, MAE, and KL Divergence metrics are all low, indicating that the model's predictions closely match annotations on the test set. The relatively low MAE compared to the RMSE (0.1143 vs. 0.1546) suggests that predictions on average deviate from the true values by less than 12%, indicating a high level of accuracy. Occasional larger discrepancies make the RMSE slightly higher, but the overall results are still very promising. The high AUC value (0.7952) computed on the test zone further confirms the model's strong performance, indicating that the model can reliably generate probability estimates that closely align with the true distributions of the labels.

With regard to possible improvements concerning the deep learning model, as previously mentioned, in this study we used *DinoV2* as a backbone: one of the SOTA (State Of The Art) computer vision models that currently performs the best on benchmark datasets. Applying transfer learning, we take advantage of the model's strong generalization capabilities while fine-tuning it to address our specific problem. However, we recognize that with the rapid advancements in artificial intelligence, our model may already be on the path to obsolescence [23]. Continuously updating the model predictions to train the medium-scale model would enable us to incorporate the latest breakthroughs in AI, ensuring increasingly refined annotation quality over time.

4.2 Georeferencing challenges in multi-scale monitoring

A framework that enables the transfer of information from fine-scale to a broader-scale imagery relies essentially on achieving precise alignment between the two layers of data. As shown in Figure 6 the benthic substrate can vary significantly across small distances, so that in order to upscale underwater predictions to aerial annotations, data need to be accurately georeferenced.

Using differential GPS technology is the first step to achieve precise positioning, but does not guarantee a centimetric accuracy. To improve the ASV positioning, we used PPK techniques thanks to the CentipedeRTK network [3], ending up with a centimetric accuracy in the ASV positioning. Unfortunately, the position of the ASV is not the same as the position of the image, since waves can change the attitude of the ASV by tilting the direction of the camera from the vertical axis (see Figure 5). To correct this, we used the camera angles on the three axes (Roll, Pitch and Yaw) and the echosounder data to compute the footprint of underwater images. Ending up with the latitude and longitude of the four corners defining the footprint

[3] https://docs.centipede.fr/

on the seabed of each underwater image. A check on the quality of the georeferencing of underwater images is then necessary in order to validate the image positioning accuracy with an unbiased approach. In our case we chose to compare our data with data produced by the French National Institute of Geographic and Forest Information (IGN), which is a reference in the field of georeferencing. Thanks to a visual comparison between the two datasets, we were able to validate the accuracy of the fine scale georeferencing process.

Although using a differential GPS on the platform acquiring images is a significant advantage, it is neither a necessary nor a sufficient condition to achieve centimetric accuracy. For instance, our methodology demonstrated that combining PPK techniques, GCPs, and validation against an external reference such as IGN data can deliver the required accuracy for coral reef monitoring. However, since our current setup involves UAVs without embedded differential GPS, precise aerial image alignment still depends on the manual collection of GCPs, making the process time-consuming and labor-intensive. Transitioning to UAVs equipped with embedded differential GPS could significantly streamline the data collection process. By applying PPK techniques directly to aerial images, the need for GCP collection could be minimized to a few validation points, reducing mission planning time and increasing efficiency. This advancement would enable faster and more scalable reef monitoring over large areas while maintaining accurate positioning.

4.3 Expanding spatial coverage and species identification

4.3.1 Satellite imagery upscaling

This study highlights several areas for future improvement. Although the classification accuracy was high across most classes, certain coral types remain challenging to differentiate at the aerial scale due to image resolution constraints. Addressing this limitation may involve refining aerial image resolution by reducing the flight altitude (respecting the regulations in force in the country where the data is collected), employing more advanced image processing techniques, using higher-quality drones / cameras with better sensors [24] or even using hyperspectral cameras to capture more detailed information about the reef [25].

Finally, mimicking the idea presented in this study, we could extend the methodology to the satellite scale. In [26] the authors discuss the complementary nature of UAV and satellite data, pointing out that integrating these technologies can improve spatial and temporal resolution in remote sensing applications. Successful integration would allow rapid monitoring of large areas, significantly reducing the need for field data collection, as satellite imagery is often freely available and collected at regular intervals. Furthermore, access to a historical archive of satellite imagery provides a unique opportunity to study ecosystem evolution over past decades, while supporting long-term monitoring efforts in the future. This extended temporal and spatial coverage could greatly improve our understanding of ecological change on a global scale.

4.3.2 Expanding to slow-moving species identification through synchronized imaging

The collection of both fine-scale and aerial images simultaneously has the potential to enable the identification of certain benthic species, which would otherwise be challenging to recognise. To illustrate, slow-moving organisms such as sea cucumbers are frequently visible in aerial images (e.g., Figure 7). Given that these species move at a slow speed, synchronised collection of both image types would provide the temporal and spatial alignment necessary for passing the information from the fine-scale model to the medium-scale model. Although this data collection approach imposes additional constraints compared to the method used in this study, where fine-scale and medium-scale data can be collected days or even months apart, it offers the advantage of providing additional ecological insights that would otherwise be inaccessible [27].

5 Conclusion

In this study, we presented a novel methodology for transferring information across scales in coral reef monitoring. By combining high-resolution underwater imagery with medium-scale aerial data, we were able to train a deep learning model to predict benthic substrate composition over a large reef area. Our approach leverages the strengths of both imaging techniques, ensuring the precision of fine-scale analysis while extending it to cover a broader reef area. This multi-scale framework has the potential to revolutionize marine monitoring, providing a more comprehensive and efficient way to assess coral reef health.

Our methodology is not only innovative but also highly adaptable. By training a medium-scale model to mimic the behaviour of a fine-scale model, we ensure that the system can easily incorporate new advances in AI without the need for data re-annotation. This flexibility allows us to continuously improve the model's predictions, ensuring that it remains at the cutting edge of coral reef monitoring. Using standardized annotation protocols and adhering to FAIR (Findable, Accessible, Interoperable, and Reusable) data principles can broaden the range of ecosystems monitored with our methodology.

Looking ahead, we see great potential for our methodology to be extended to the satellite scale. By integrating UAV and satellite data, we can enhance spatial and temporal resolution in remote sensing applications, providing a more comprehensive view of coral reef ecosystems. This extended coverage could greatly improve our understanding of ecological change on a global scale, supporting long-term monitoring efforts and contributing to the conservation of these vital marine ecosystems.

Moreover, the proposed upscaling methodology shows promise for applications in a number of different fields beyond the monitoring of coral reefs. In terrestrial environments, it could be used to support forestry management by extending detailed ground-based observations to regional scales. In a completely different domain, like in urban planning, ground-level observations of pedestrians, traffic or vegetation could inform city-wide analyses using aerial or satellite imagery. These examples demonstrate the versatility of this approach, which could be applied across a range of different disciplines and environments.

6 CRediT authorship contribution statement

Matteo Contini: Conceptualization, Data curation, Methodology, Formal analysis, Writing – original draft. **Victor Illien**: Conceptualization, Data curation, Methodology, Software, Writing – review and editing. **Julien Barde**: Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Writing – review and editing. **Sylvain Poulain**: Data curation, Methodology, Software. **Serge Bernard**: Conceptualization, Data curation, Funding acquisition, Methodology, Supervision, Writing – review and editing. **Alexis Joly**: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Writing – review and editing. **Sylvain Bonhommeau**: Conceptualization, Data curation, Funding acquisition, Methodology, Project administration, Writing – review and editing.

7 Acknowledgement

The authors acknowledge the Pôle de Calcul et de Données Marines (PCDM) for providing DATARMOR storage, support services and computational resources. We extend our heartfelt thanks to Laurence Maurel, Cam Ly Rintz, Leanne Carpentier, Magali Duval, Laura Babet, Belen De Ana, Anne-Elise Nieblas, Arthur Lazennec, Victor Russias, Mervyn Ravitchandirane, Mohan Julien, Pierre Gogendeau, Thomas Chevrier and Justine Talpaert Daudon, involved in the annotation and collection of data for this study. Their contributions were indispensable to our research efforts. We also thank Emilien Alvarez for his valuable advice on drone imaging.

8 Funding

This work was supported by several projects: Seatizen (Ifremer internal grant), Plancha (supported by the Contrat de convergence et de transformation 2019-2022, mesure 3.3.1.1 de la Préfecture de la Réunion, France), IOT project (funded by FEDER INTERREG V and Prefet de La Réunion: grant #20181306-0018039 and the Contrat de Convergence et de Transformation de la Préfecture de La Réunion), Ocean and Climate Priority Research Programme, FISH-PREDICT project (funded by the IA-Biodiv ANR project: ANR-21-AAFI-0001-01), B1-4SEA funded by Explorations de Monaco and G2OI FEDER INTERREG V (grant #20201454-0018095).

9 Conflict of Interest statement

The authors declare no conflict of interest.

10 Data availability

The data that support the findings of this study are openly available in Zenodo at Seatizen Atlas. All code for data processing associated with the current submission is available on drone-upscaling Github.

The code for downloading data associated with the current submission is available on zenodo-tools Github.

The code used to train the neural network model used in the current submission is available on DinoVdeau Github.

11 Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT-4o and GitHub Copilot in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article. This tools were not involved in the design, implementation, data analysis, or manuscript preparation of the study.

References

[1] Ove Hoegh-Guldberg, Peter Mumby, A.J. Hooten, R.S. Steneck, Paul Greenfield, Erick Gomez, Catherine Harvell, Peter Sale, Alasdair Edwards, Ken Caldeira, Nancy Knowlton, C. Mark Eakin, Roberto Iglesias-Prieto, Nyawira Muthiga, Roger Bradbury, Alfonse Dubi, and M Hatziolos. Coral reefs under rapid climate change and ocean acidification. *Science (New York, N.Y.)*, 318:1737–42, 01 2008. doi: 10.1126/science.1152509.

[2] Alice Rogers, Julia Blanchard, and Peter Mumby. Fisheries productivity under progressive coral reef degradation. *Journal of Applied Ecology*, 55, 12 2017. doi: 10.1111/1365-2664.13051.

[3] Andrew W Bruckner. Life-saving products from coral reefs. *Issues in Science and Technology*, 18(3): 39–44, 2002.

[4] Deny Hidayati et al. The importance of the sustainable use of fishery resources to improve the livelihoods of fishermen on the islands of sumatra and sulawesi, indonesia. In *Proceedings of the 5th Conference on Agribusiness, Green Energy, Environment, and Sustainable Development (CAGEES-V5)*, pages 120–141, Jun 2022.

[5] Joost W van Dam, Andrew P Negri, Sven Uthicke, and Jochen F Mueller. Chemical pollution on coral reefs: exposure and ecological effects. In Francisco Sanchez-Bayo, Paul J. van den Brink, and Reinier M. Mann, editors, *Ecological Impacts of Toxic Chemicals*, pages 187–211. Bentham Science Publishers, 2011.

[6] Terry P Hughes, Andrew H Baird, David R Bellwood, Margaret Card, Sean R Connolly, Carl Folke, Richard Grosberg, Ove Hoegh-Guldberg, Jeremy BC Jackson, Janice Kleypas, et al. Climate change, human impacts, and the resilience of coral reefs. *science*, 301(5635):929–933, 2003.

[7] Terry P Hughes, James T Kerry, Mariana Álvarez-Noriega, Jorge G Álvarez-Romero, Kristen D Anderson, Andrew H Baird, Russell C Babcock, Maria Beger, David R Bellwood, Ray Berkelmans, et al. Global warming and recurrent mass bleaching of corals. *Nature*, 543(7645):373–377, 2017.

[8] COP15: Nations Adopt Four Goals, 23 Targets for 2030 In Landmark UN Biodiversity Agreement, 2022. URL https://www.cbd.int/article/cop15-cbd-press-release-final-19dec2022.

[9] Letizia Lamperti, Théophile Sanchez, Sara Si Moussi, David Mouillot, Camille Albouy, Benjamin Flück, Morgane Bruno, Alice Valentini, Loïc Pellissier, and Stéphanie Manel. New deep learning-based methods for visualizing ecosystem properties using environmental dna metabarcoding data. *Molecular Ecology Resources*, 23(8):1946–1958, 2023.

[10] Letizia Lamperti, Olivier François, David Mouillot, Laëtitia Mathon, Théophile Sanchez, Camille Albouy, Loïc Pellissier, and Stéphanie Manel. A spatial matrix factorization method to characterize ecological assemblages as a mixture of unobserved sources: An application to fish edna surveys. *Methods in Ecology and Evolution*, n/a(n/a), 2024. doi: https://doi.org/10.1111/2041-210X.14430. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14430.

[11] Gaétan Morand, Alexis Joly, Tristan Rouyer, Titouan Lorieul, and Julien Barde. Predicting species distributions in the open ocean with convolutional neural networks. *bioRxiv*, 2024. doi: 10.1101/2023.08.11.551418. URL https://www.biorxiv.org/content/early/2024/07/08/2023.08.11.551418.

[12] Benjamin Misiuk and Craig J. Brown. Benthic habitat mapping: A review of three decades of mapping biological patterns on the seafloor. *Estuarine, Coastal and Shelf Science*, 296:108599, 2024. ISSN 0272-7714. doi: https://doi.org/10.1016/j.ecss.2023.108599. URL https://www.sciencedirect.com/science/article/pii/S027277142300389X.

[13] Michaela Doukari and Konstantinos Topouzelis. Overcoming the uas limitations in the coastal environment for accurate habitat mapping. *Remote Sensing Applications: Society and Environment*, 26:100726, 2022. ISSN 2352-9385. doi: https://doi.org/10.1016/j.rsase.2022.100726. URL https://www.sciencedirect.com/science/article/pii/S2352938522000349.

[14] Kristina Øie Kvile, Hege Gundersen, Robert Nøddebo Poulsen, James Edward Sample, Arnt-Børre Salberg, Medyan Esam Ghareeb, Toms Buls, Trine Bekkby, and Kasper Hancke. Drone and ground-truth data collection, image annotation and machine learning: A protocol for coastal habitat mapping and classification. *MethodsX*, 13:102935, 2024. ISSN 2215-0161. doi: https://doi.org/10.1016/j.mex.2024.102935. URL https://www.sciencedirect.com/science/article/pii/S2215016124003868.

[15] Daniele Ventura, Luca Grosso, Davide Pensa, Edoardo Casoli, Gianluca Mancini, Tommaso Valente, Michele Scardi, and Arnold Rakaj. Coastal benthic habitat mapping and monitoring by integrating aerial and water surface low-cost drones. *Frontiers in Marine Science*, 9:1096594, 2023.

[16] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[17] Pierre Gogendeau, Sylvain Bonhommeau, Hassen Fourati, Mohan Julien, Matteo Contini, Thomas Chevrier, Anne Elise Nieblas, and Serge Bernard. An open-source autonomous surface vehicle for acoustic tracking, bathymetric and photogrammetric surveys, 2024. URL https://arxiv.org/abs/2406.18760.

[18] Richard K Slocum, W Wright, C Parrish, B Costa, M Sharr, and TA Battista. Guidelines for bathymetric mapping and orthoimage generation using suas and sfm, an approach for conducting nearshore coastal mapping, 2019. Available online: https://coastalscience.noaa.gov/data_reports/guidelines-for-bathymetric-mapping-and-orthoimage-generation-using-suas-and-sfm-an-approach-for-conducting (accessed on 7 September 2024).

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, April 2023. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

[20] Titouan Lorieul. *Uncertainty in predictions of deep learning models for fine-grained classification*. PhD thesis, Université Montpellier, 2020.

[21] J Elith, CH Graham, RP Anderson, M Dudik, S Ferrier, A Guisan, RJ Hijmans, F Huettmann, JR Leathwick, A Lehmann, J Li, LG Lohmann, BA Loiselle, G Manion, C Moritz, M Nakamura, Y Nakazawa, J. M. Overton, A.T. Peterson, SJ Phillips, K Richardson, R Scachetti-Pereira, RE Schapire, J Soberon, S Williams, Mary Wisz, and NE Zimmermann. Novel methods improve prediction of

species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006. ISSN 0906-7590. doi: 10.1111/j.2006.0906-7590.04596.x.

[22] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1503.html#HintonVD15.

[23] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2024. URL https://arxiv.org/abs/2405.14093.

[24] Elisa Giusti, Amerigo Capria, Anna Lisa Saverino, Samuele Gelli, Jorge Muñoz-Castañer, Raquel Dosil, Jorge Naya, and Javier Menéndez. A drone-based multisensory payload for maritime pollutants detections. *IEEE Aerospace and Electronic Systems Magazine*, 38(3):4–18, 2023. doi: 10.1109/MAES.2022.3232071.

[25] Thomas Rossiter, Thomas Furey, Tim McCarthy, and Dagmar B. Stengel. Uav-mounted hyperspectral mapping of intertidal macroalgae. *Estuarine, Coastal and Shelf Science*, 242:106789, 2020. ISSN 0272-7714. doi: https://doi.org/10.1016/j.ecss.2020.106789. URL https://www.sciencedirect.com/science/article/pii/S0272771419308431.

[26] Emilien Alvarez-Vanhard, Thomas Corpetti, and Thomas Houet. Uav & satellite synergies for optical remote sensing applications: A literature review. *Science of remote sensing*, 3:100019, 2021.

[27] Chantal Conand, Sonia Ribes-Beaudemoulin, Florence Trentin, Thierry Mulochau, and Emilie Boissin. Marine Biodiversity of La Reunion Island: Echinoderms. *Western Indian Ocean Journal of Marine Science*, 17(1):111–124, 2018. URL https://hal.univ-reunion.fr/hal-01906874.

[28] Julien Ancelin, Sylvie Ladet, and Wilfried Heintz. Le Real Time Kinematic collaboratif, lowcost et open source. Positionnement GNSS temps réel, cinématique, collaboratif et en accès libre et à faible coût. In Thierry Badard, Jacynthe Pouliot, Ma8hieu Noucher, and Marlène Villanova-Oliver (Eds), editors, *Spatial Analysis and GEOmatics 2023*, Actes de la conférence Spa1al Analysis and GEOma1cs (SAGEO) 2023, pages 184–197, Québec, Canada, June 2023. GDR MAGIS Méthodes et Applications pour la Géomatique et l'Information Spatiale and Centre de Recherche en Données et Intelligence Géospatiales de l'Université Laval (Québec). URL https://hal.inrae.fr/hal-04144737.

[29] M.J. Westoby, J. Brasington, N.F. Glasser, M.J. Hambrey, and J.M. Reynolds. 'structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300–314, 2012. ISSN 0169-555X. doi: https://doi.org/10.1016/j.geomorph.2012.08.021. URL https://www.sciencedirect.com/science/article/pii/S0169555X12004217.

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

[31] Camilo L M Morais, Marfran C D Santos, Kássio M G Lima, and Francis L Martin. Improving data splitting for classification applications in spectrochemical analyses employing a random-mutation Kennard-Stone algorithm approach. *Bioinformatics*, 35(24):5257–5263, 05 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz421. URL https://doi.org/10.1093/bioinformatics/btz421.

[32] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23808-6.

A ASV supplementary informations

A.1 Time synchronization

Videos were cut into frames with a rate of 2.997 fps, so that the cutting frame rate $f_c = 2.997$ fps is a divisor of the video frame rate $f_v = 23.976$, ensuring that the ratio $\frac{f_v}{f_c}$ is an integer.

Since we use time in order to synchronize metadata and images, we need a method to assign a precise timestamp to each frame. Before each data acquisition, as differences of several seconds/minutes can be observed between the clocks of the different devices (the GPS receiver clock is not the same as the camera), the user films the time given by a GPS application on his mobile phone with the camera in order to associate the exact satellite time (UTC+0) to a specific frame or image. In the case where the time filmed with the camera follows UTC standards, leap seconds caused by the difference between UTC time and GPS time must be taken into account when synchronizing the GPS position with the images. This specific frame can then be used as a starting point to correct the timestamp of all images by using the frame rate f_c and the number of frames between the starting frame and the frame of interest.

Cutting frames with a rate that is a divisor of the video frame rate is particularly important when working with precise position accuracy.

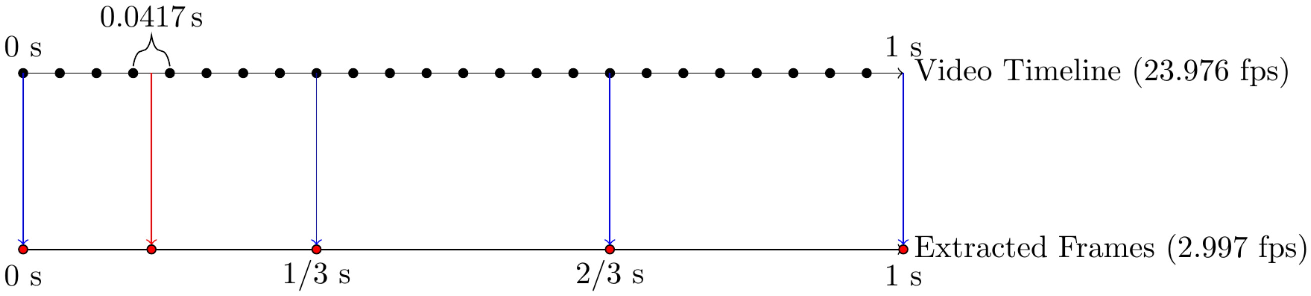


Figure 9: Frame rate extraction from GoPro video. If the cutting frame rate is not a divisor of the video frame rate, a misalignment between the timestamp and the corresponding frame is introduced. Later causing an error in assigning the GPS position to the frame.

Indeed if the cutting frame rate is not a divisor of the video frame rate, it may happen that during the cutting process of the video a frame is skipped or duplicated, causing a misalignment between the timestamp and the corresponding frame. This is represented in Figure 9, where blue arrows represent frames extracted from the video timeline with a rate of $f_c = 2.997$ fps and red arrow represent a frame extracted with a random frame rate. In the first case, there is a perfect alignment between video frames and frames extracted from it. In the second case, we can see that for the required frame rate there is no corresponding frame in the original video timeline. So that, depending on the chosen option, either the frame is skipped or the closest frame is duplicated. In both cases, an error is introduced in the extracted frame timestamp. A difference in the timestamp will result in a misalignment between the real GPS position and the calculated one, which is proportional to the speed at which the ASV acquired the data. For more information about camera video setting, please refer to Appendix A.3.

A.2 Metadata correction

Since the ASV is equipped with a differential GPS, PPK (Post-Processed Kinematic) corrections can be applied to the GPS position of the rover in order to get a centimetric accuracy.

Indeed, for each data collection event a mobile base station has been strategically deployed near the field mission. The mobile base station, connected to the *CentipedeRTK* network which provides real-time corrections to the GPS base station, ensures a high precision of the GPS base position [28]. This allows to refine the GPS position of the rover in a post processing step using corrections from the base station.

We can then attach to each frame the corresponding position, using the timestamp as a reference.

Underwater images positioning was checked in two different ways:

1. Firstly we computed the standard deviation on the east and north axis of the GPS position of the rover. If the value was below the centimeter for both axis then the session was considered as a good one.

2. Secondly a visual check was done by visually comparing images that had a very close GPS position but a different timestamp. If the two images represent the same zone then the session was considered as a good one. An example is given in Figure 10. The two images taken in the *Trou d'eau* lagoon in Reunion Island are at a distance of 6.062 cm from each other and are taken 24 minutes and 49 seconds apart. It is clear how, except from the sea cucumber that has moved a little bit between the two images, the two frames represent the same zone, validating in this way the data collection event.

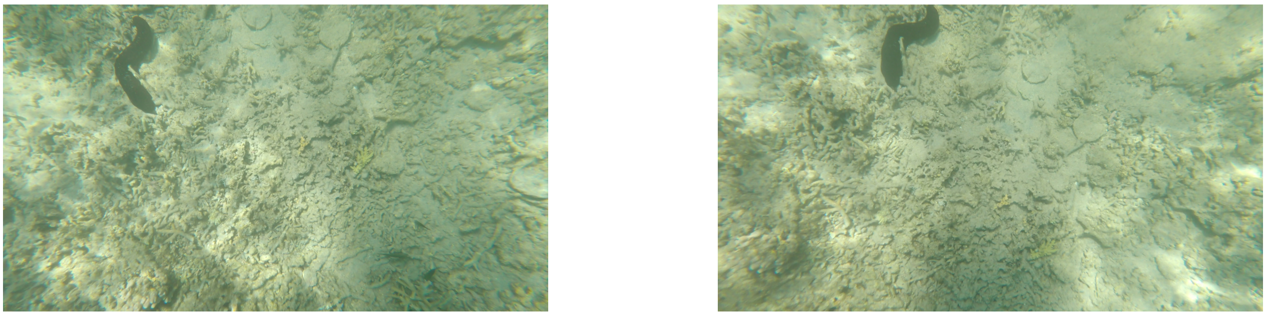


Figure 10: Example of two images taken in the *Trou d'eau* lagoon in Reunion Island at a distance of 6.062 cm from each other and 24 minutes and 49 seconds apart. The similarity between the two images is an important visual criterion to validate a data collection event.

Moreover, since the ASV is equipped with an IMU (Inertial Measurement Unit) that provides the roll, pitch and yaw angles of the rover, it could be possible to correct the bathymetry data using local geoid parameters and the attitude data of the ASV. This data were then attached to each frame using, again, the timestamp as a reference.

A.3 Camera settings

GoPro camera setting can be found in Table 2.

Parameter	Value
File Type Extension	MP4
MIME Type	video/mp4
Time Scale	60000
Preferred Rate	1
Preferred Volume	1
Firmware Version	HD8.01.02.51.00
Camera Model Name	HERO8 Black
Auto Rotation	U
Digital Zoom	Y
Pro Tune	Y
White Balance	6500K
Sharpness	HIGH
Color Mode	FLAT
Auto ISO Max	400
Auto ISO Min	100
Rate	2_1SEC
Field Of View	L
Sensor Readout Time	7.9200005531311
Electronic Image Stabilization	N/A
Image Width	1920
Image Height	1080
Graphics Mode	0
X Resolution	72
Y Resolution	72
Compressor Name	GoPro AVC encoder
Bit Depth	24
Video Frame Rate	59.9400599400599
Avg Bitrate	45266194

Table 2: Video File Parameters

B UAV supplementary informations

B.1 Mission planning

Imaging the seabed, even in tropical environments with very clear waters, is often complicated by reflections of sunlight from the water surface. Direct sun rays reflections can be extremely bright and cause oversaturated areas on aerial images. These reflections overexpose the image, making it difficult to see the seafloor structure, as shown in Figure 11.

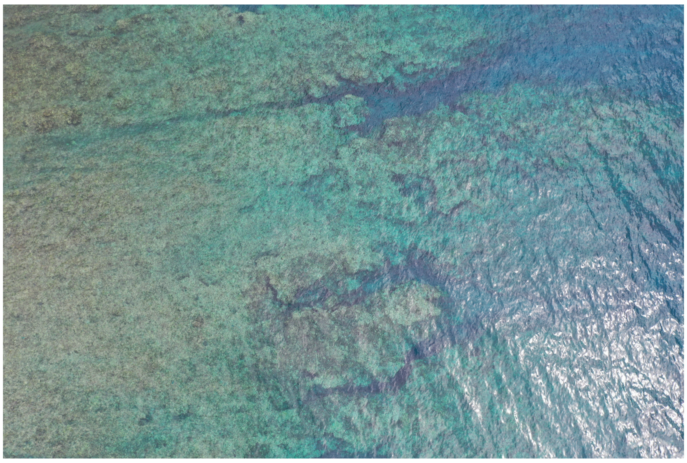


Figure 11: Example of an overexposed image, taken by a Mavic 2 Pro drone in the *Saint-Leu* lagoon in Reunion Island. The sun reflection on the water surface makes it impossible to see the seafloor structure.

If the seafloor is not visible in the image, SfM algorithms cannot map these areas effectively. It is thus necessary to avoid these reflections in the images. A solution is to take images when the sun is at a low angle, which reduces the reflection of sunlight on the water surface [18]. This can be achieved by surveying the desired area early in the morning or late in the afternoon, or on a cloudy day. In our case, images were taken between daybreak and sunrise, which is the best time to avoid reflections on the water surface on the west part of Reunion island.

B.2 Mission execution

In SfM surveys, overlap and sidelap refer to the percentage to which each consecutive image overlaps with the preceding image and the images on adjacent flight lines. In order to obtain a good 3D model, it is important to have a high overlap and sidelap between images. The overlap and sidelap should be at least 70% for a good 3D model [29]. In our case, the overlap and sidelap were both set to 80%.

For what concerns the flight altitude, a lower one implies a higher resolution point cloud in the SfM process since more details of the seabed will be visible but on the other hand the wave induced refraction is more visible. On the contrary, setting a too high flying altitude will reduce the resolution of the point cloud, making it difficult to distinguish characteristics of study objects (e.g., coral colonies, habitats, etc.). In our case, the flight altitude was set to 60 meters.

B.3 Image processing

Many photogrammetry softwares are available to build a 3D model from images. In our case, since the objective was to create an open-source pipeline, the software package `OpenDroneMap` was used. `OpenDroneMap` is a commercial-grade open-source software package for SfM photogrammetric processing (initially developed for aerial images) that can be used to generate georeferenced orthophotos, point clouds, elevation models and textured 3D models from aerial images.

Settings used for the processing images from both missions are shown in Table 3.

Setting	Value
Auto-boundary	True
DEM Resolution	2.0
DSM	True
Orthophoto Resolution	1.0
Point Cloud Quality	Ultra
Rolling Shutter	True

Table 3: OpenDroneMap Processing Settings

B.4 Orthophoto georeferencing

Since the drone was not equipped with a differential GPS, once the orthophoto was built, Ground Control Points (GCPs) were chosen on fixed and easily distinguishable objects on land (e.g. manhole covers, corners of basketball courts, etc.) and easy-to-distinguish corals (e.g., large *Porites* or *Acropora Tabular* corals). GCPs position was then collected using a GPS with centimetric accuracy and then the orthophoto was reconstructed by forcing pixels representing GCPs to be at the ground truth collected position. GCPs examples are shown in Figure 12 with a pink point on the image.

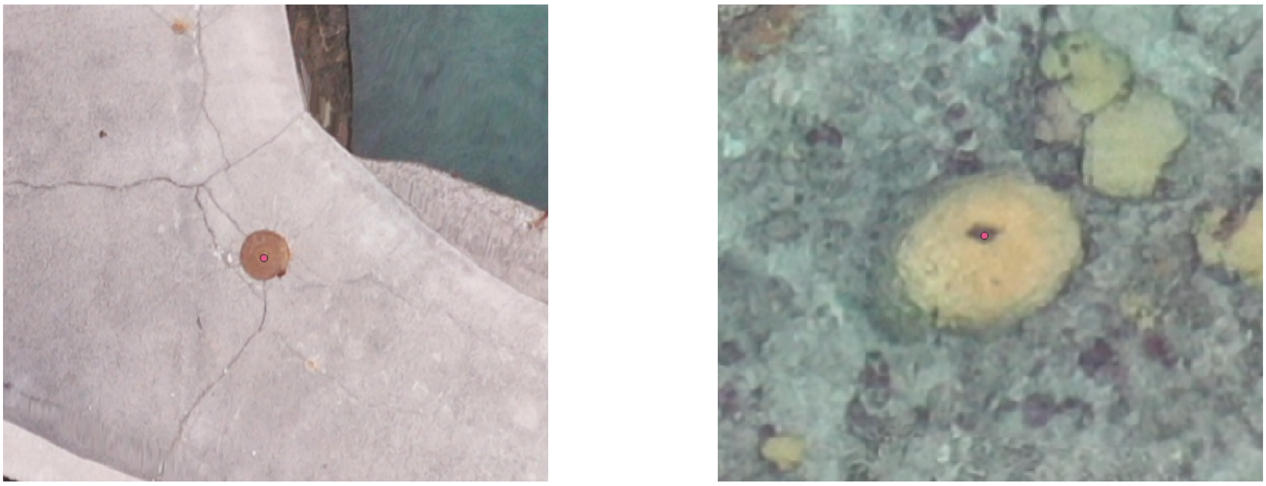


Figure 12: Examples of Ground Control Points (GCPs) collected in Reunion island in order to correct the aerial orthophoto position. Both human made objects (on the left) and easily distinguishable corals (a *Porites* coral on the right) can be used to set GCPs.

The *Trou d'eau* lagoon mission measured a surface of 189,682 m^2 and 10 GCP were collected. The *Saint-Leu* lagoon mission measured a surface 204,748 m^2 and 9 GCP were collected.

Final orthophoto positioning was checked by computing the Root Mean Square Error (RMSE) between the GCPs and the orthophoto.

Lagoon	Error Type	Mean	Standard Deviation	RMS Error
Trou d'eau	X Error (meters)	0.001	0.011	0.011
	Y Error (meters)	-0.001	0.010	0.010
	Z Error (meters)	-0.011	0.028	0.030
	Total			**0.022**
Saint-Leu	X Error (meters)	-0.001	0.003	0.003
	Y Error (meters)	-0.000	0.001	0.001
	Z Error (meters)	0.000	0.004	0.004
	Total			**0.003**

Table 4: GCP errors statistics for *Trou d'eau* Lagoon and *Saint-Leu* Lagoon

In Table 4 we show the error statistics for GCPs collected in the *Trou d'eau* and the *Saint-Leu* lagoons. We can observe that the RMSE is below 2.5 cm for both lagoons, which is a satisfactory level of precision for the continuation of the study.

C Fine scale deep-learning model

DinoV2, a family of state-of-the-art transformer models in computer vision produces general-purpose visual features (i.e., features that work across image distributions and tasks without fine-tuning), and is compatible with classifiers as simple as linear layers. Meaning that the model can be readily applied to various tasks like image classification or segmentation without necessitating encoder fine-tuning [19]. This implies that instead of training the entire model on a specific dataset, we can simply fine-tune a classification head atop the frozen encoder.[4] *DinoV2* comprises four distinct models based on different sizes: small, base, large, and giant.

In our study, following a performance evaluation against training time experiment, we opted for the large model. Transitioning from the small to base version and from base to large version, demonstrates a performance increase for all the metrics (F1 Micro, F1 Macro, Roc Auc, and Accuracy) with a reasonable uptick in training time, see Table 5. The giant version of the model, despite having slightly lower loss than the large model, has lower performances for all the metrics and the training steps per second are almost multiplied by three, inducing a considerable increase in total training time.

Model	Loss	F1 Micro	F1 Macro	Roc Auc	Accuracy	Steps per second
DinoVdeau-small	0.1320	0.8009	0.6614	0.8649	0.2903	**0.164**
DinoVdeau-base	0.1260	0.8131	0.6976	0.8760	0.3014	0.207
DinoVdeau-large	0.1209	**0.8228**	**0.7175**	**0.8813**	**0.3111**	0.21
DinoVdeau-giant	**0.1208**	0.8209	0.7101	0.8812	0.3080	0.66

Table 5: Model size comparison

For the classification head, we used a more complex model than a simple linear classifier in order to improve the expressiveness of the model and capture more complex interactions between the variables extracted by the backbone model. Specifically, we added a bottleneck layer block consisting of a linear layer, a ReLU, a batch normalization, and a dropout layer. The resulting model is called *DinoVdeau* and the architecture is shown in Figure 13.

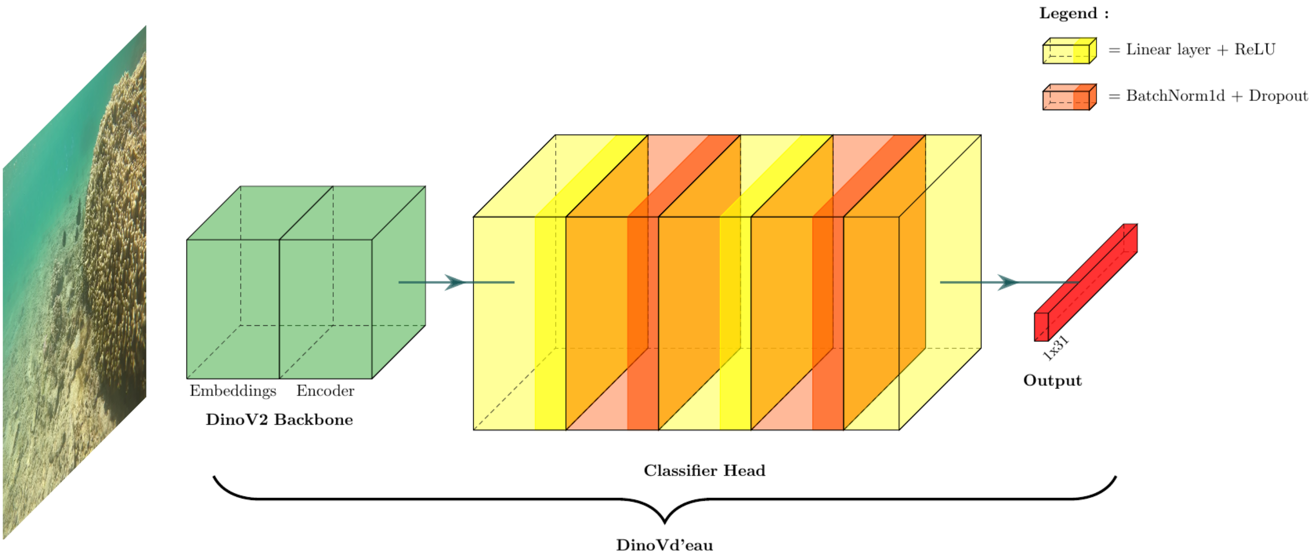


Figure 13: Architecture of *DinoVdeau* model

Introducing the bottleneck layer block instead of a single linear classifier results in a performance gain of approximately 0.006% and 0.015% for the F1 Micro and the Accuracy metrics respectively.

Using *DinoV2* as the backbone compared to a *Resnet50* baseline model results in a performance gain of approximately 0.098% and 0.140% for the F1 Micro and the Accuracy metrics respectively.

The model was trained for 93 epochs, incorporating an early stopping mechanism with a patience of 10 epochs to mitigate overfitting. Maintaining a fixed batch size of 32, the initial learning rate was established

[4] To give an idea, the large version of the model (`Vit-L`) has 0.3B of parameters and the giant version (`Vit-g`) 1.1B parameters.

at 10^{-3}, and was decreased by a factor of 0.1 whenever the model's performance plateaued for more than 5 epochs. Network weight updates were executed using the Adam optimization algorithm with a weight decay of 10^{-4}.

The training lasted 75 hours, on the *Datarmor* supercomputer equipped with an `NVIDIA Tesla V100 PCIe 32 GB` GPU and `32 Intel Xeon-Gold 4216 (2.1GHz/16-core/100W)` CPUs. All the trainings were done using the Huggingface `Transformers` library [30].

D Data splitting

As explained in [31], the process of data splitting is a crucial step in the construction of image classification models. This is essential to assess their effectiveness with an unbiased approach. To achieve this, we employ preprocessing techniques and leverage prior knowledge of the dataset. The approach involves dividing the samples into three distinct parts: the training set D_{tr}, the validation set D_{val}, and the test set D_{test}. It's important to note that our task involves multilabel classification, where an input can have multiple labels. This complicates the stratification process compared to monolabel classification. Traditional single-label approaches to stratifying data fall short in providing balanced dataset divisions in the multilabel case because it's not feasible to create tuples (image, label), given that each image corresponds to a variable number of labels. Consequently, once an image is assigned to a dataset, all corresponding labels are assigned to it.

To build a model with high generalization capabilities, we opt for a temporal criterion to independently divide the dataset into three subdatasets. The goal is to maximize the diversity of the training dataset by including images from different periods and islands in the Indian Ocean. This temporal splitting criterion corresponds to a spatial one as well, considering that images are typically collected during data collection campaigns in specific locations. This approach ensures that the training dataset is not only temporally diverse but also representative of various spatial contexts. Subsequently, we employ the `scikit multilabel` data stratification technique to split each subdataset into three subsets [32]. For the purpose of achieving optimal predictive performance for our neural network, we implement the algorithm represented by the following pseudo-code:

Algorithm 1 Dataset splitting algorithm

```
for $i = 1, \ldots, Nb\_years$ do
    1.Split the $D_i$ into $train_i$ and $val - test_i$ sets using the Multi-label data stratification technique
    2.Split the $val - test_i$ set into $val_i$ and $test_i$ sets using the Multi-label data stratification technique.
    3.Concatenate the current $train_i$, $val_i$, and $test_i$ datasets to the overall $D_{train}$, $D_{val}$ and $D_{test}$datasets.
end for
```

In Table 6, we present the total number of images for each class, along with their corresponding distribution in the training, validation, and test sets. The results indicate a well-balanced class distribution.

Class	**Train Frequency**	**Validation Frequency**	**Test Frequency**	**Total**
Acropore_branched	0.666666	0.167702	0.165632	1206
Acropore_digitised	0.690607	0.160059	0.149552	674
Acropore_tabular	0.640284	0.183073	0.177643	1507
Algae_assembly	0.609174	0.194251	0.195574	3562
Algae_limestone	0.601715	0.198876	0.199408	2207
Algae_sodding	0.606493	0.197214	0.196293	3426
Dead_coral	0.613217	0.194557	0.193226	1839
Fish	0.643513	0.169074	0.169413	1359
Human_object	0.599117	0.198823	0.200059	678
Living_coral	0.604631	0.198982	0.195387	2916
Millepore	0.612976	0.208060	0.178964	571
No_acropore_encrusting	0.602345	0.207921	0.188734	682
No_acropore_foliaceous	0.742105	0.119298	0.136842	285
No_acropore_massive	0.595306	0.204966	0.200388	1548
No_acropore_sub_massive	0.623316	0.187305	0.188379	1930
Rock	0.606011	0.197692	0.196297	6171
Sand	0.599165	0.200334	0.199501	5990
Scrap	0.590874	0.201673	0.207453	3586
Sea_cucumber	0.600769	0.195385	0.204615	1300
Sea_urchins	0.589441	0.186291	0.224267	321
Sponge	0.581487	0.192027	0.226486	389
Syringodium_isoetifolium	0.600307	0.198256	0.201437	1949
Thalassodendron_ciliatum	0.600921	0.199232	0.199846	1304
Useless	0.601836	0.199179	0.199985	977

Table 6: Class frequency distribution across training, validation, and test sets.