



OPEN

DATA DESCRIPTOR

# Seatizen Atlas: a collaborative dataset of underwater and aerial marine imagery

Matteo Contini<sup>1,2✉</sup>, Victor Illien<sup>1</sup>, Mohan Julien<sup>1</sup>, Mervyn Ravitchandirane<sup>1</sup>, Victor Russias<sup>1</sup>, Arthur Lazennec<sup>1</sup>, Thomas Chevrier<sup>1,3</sup>, Cam Ly Rintz<sup>1</sup>, Léanne Carpentier<sup>1</sup>, Pierre Gogendeau<sup>1</sup>, César Leblanc<sup>2</sup>, Serge Bernard<sup>4</sup>, Alexandre Boyer<sup>5</sup>, Justine Talpaert Daudon<sup>5</sup>, Sylvain Poulain<sup>5</sup>, Julien Barde<sup>5</sup>, Alexis Joly<sup>2</sup> & Sylvain Bonhommeau<sup>1</sup>

Citizen Science initiatives have a worldwide impact on environmental research by providing data at a global scale and high resolution. Mapping marine biodiversity remains a key challenge to which citizen initiatives can contribute. Here we describe a dataset made of both underwater and aerial imagery collected in shallow tropical coastal areas by using various low cost platforms operated either by citizens or researchers. This dataset is regularly updated and contains >1.6 M images from the Southwest Indian Ocean. Most of images are geolocated, and some are annotated with 51 distinct classes (e.g. fauna, and habitats) to train AI models. The quality of these photos taken by action cameras along the trajectories of different platforms, is highly heterogeneous (due to varying speed, depth, turbidity, and perspectives) and well reflects the challenges of underwater image recognition. Data discovery and access rely on DOI assignment while data interoperability and reuse is ensured by complying with widely used community standards. The open-source data workflow is provided to ease contributions from anyone collecting pictures.

## Background & Summary

It is widely known that oceans are being affected by human activities through over-fishing<sup>1</sup>, pollution<sup>2–5</sup> and global warming<sup>6,7</sup>. Biodiversity and habitats in coastal ecosystems have suffered a significant decline, ranging from 30% to 60%<sup>8,9</sup>. In particular, coral reefs are home to a great level of biodiversity (e.g., approximately 7,000 fish species<sup>10</sup>) and provide vital ecosystem services, protecting coastlines from storms and erosion, provide jobs for local communities and offer opportunities for recreation<sup>11</sup>. According to the National Oceanic and Atmospheric Administration (NOAA), over half a billion people depend on coral reefs for food, income, and protection<sup>12</sup>.

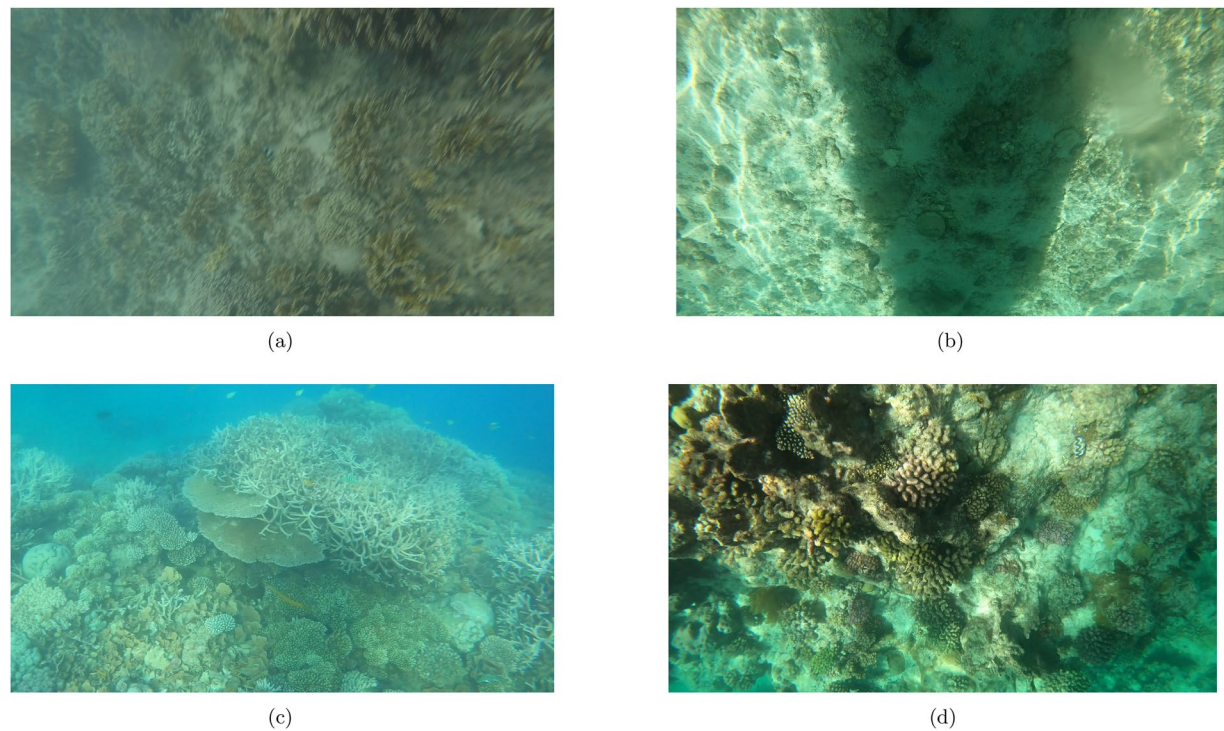
Several nations have implemented monitoring programs through underwater visual censuses (UVC)<sup>13–15</sup>. This method requires highly trained scientific divers who note visual observations under water as well as a substantial amount of time to collect the data<sup>16,17</sup>.

Most recent methods use autonomous underwater vehicles (AUV)<sup>18</sup> or remotely operated vehicles (ROV)<sup>19</sup> in order to take underwater images and videos. These new techniques make it possible to considerably increase the amount of available data. In addition, places that were previously inaccessible to humans, such as regions requiring long decompression stops or areas where diving is dangerous due to environmental conditions, are now reachable, as human limits of divers are exceeded by underwater robots. This technology requires highly specialized scientists to annotate the considerable amount of collected images<sup>20</sup>. Therefore, there is a need to develop solutions to reduce the involved time and costs.

Fortunately, the great advances that have been made in artificial intelligence over the last few decades have encouraged more and more people to focus their efforts on automatic image classification using neural networks<sup>21</sup>. Protocols combining the use of submarine robots for image acquisition with artificial intelligence

<sup>1</sup>IFREMER Délégation Océan Indien (DOI), Le Port, 97420, La Réunion, Rue Jean Bertho, France. <sup>2</sup>INRIA, LIRMM, Université de Montpellier, CNRS, Montpellier, 34000, France. <sup>3</sup>Company for Open Ocean Observations and Logging (COOOL), Saint-Leu, 97436, La Réunion, France. <sup>4</sup>CNRS, LIRMM, Université de Montpellier, Montpellier, 34000, France. <sup>5</sup>UMR Marbec, IRD, Université de Montpellier, CNRS, Ifremer, Montpellier, 34000, France.

✉e-mail: [matteo.contini@ifremer.fr](mailto:matteo.contini@ifremer.fr)



**Fig. 1** Examples of underwater images taken with four different types of platforms: **(a)** Kitesurf **(b)** Paddleboard **(c)** Scuba diving **(d)** Autonomous Surface Vehicle.

| CountryCode          | Number of Images |
|----------------------|------------------|
| Reunion island (REU) | 1,468,481        |
| The Seychelles (SYC) | 133,979          |
| Mauritius (MUS)      | 24,370           |

**Table 1.** Number of images per country.

techniques are certainly the best solution (at the moment) for achieving large-scale monitoring<sup>22</sup>. In upcoming years, the number of places monitored with these techniques will increase considerably, allowing a broader and more precise analysis of ecosystem evolution. This method will provide one of the most rigorous, efficient and cost-effective key to monitor coastal ecosystems in the future.

Nevertheless, compared to classic image recognition, there are a number of difficulties that come into play when one wish to classify underwater images. First of all, the non-uniformity of the light leads to numerous shadows and plays of light that can easily be confused by the neural network with marine species. The turbidity of water and, in the case of images taken near a coral reef, background confusion add noise that makes the task more difficult. In addition, depending on the depth at which the photo was taken (with the same light conditions), the colors of the image may vary considerably because water absorbs different wavelengths of light to different degrees (the first to be absorbed is red, followed by orange and yellow). An additional part of image preprocessing is therefore strongly recommended in order to reduce color distortion<sup>23,24</sup>. The similarity in shape, color and texture of different species of corals (inter-class similarity) adds another challenge in accurate species classification. To address these challenges we introduce a dataset that combines data collected through a participatory science project with a large volume of data collected through scientific platforms. This dataset is unique both in the types of data-collecting devices and the geographical locations covered. It integrates images captured using a variety of platforms, including kitesurfs, paddleboards, snorkeling masks, autonomous surface vehicles and aerial drones. Furthermore, the dataset covers a broad geographical area, featuring data collected from various regions in the Southwest Indian Ocean, such as Réunion Island, Seychelles, and Mauritius. We provide a full data workflow available at <https://github.com/SeatizenDOI> which enable users to process the data, make predictions on images with an AI algorithm, and publish the dataset on Zenodo. The dataset currently contains 1,626,830 images, with 2.6% of these being citizen images. All images are accurately described with rich meta-data (exif, GPS) and 14,492 have been manually annotated with the information about the habitat and species observed following the Global Coral Reef Monitoring Network (GCRMN) and Global Biodiversity Information Facility (GBIF) standards. Figure 1 shows sample images collected with different platforms and Table 1 and Table 2 show the distribution of images along different countries and years.

| Year | Number of Images |
|------|------------------|
| 2015 | 3487             |
| 2021 | 38,926           |
| 2022 | 158,349          |
| 2023 | 1,104,995        |
| 2024 | 321,073          |

**Table 2.** Number of images per year.

Whereas in previous works<sup>4,25</sup> cameras were stationary or rotating around a fixed point<sup>26</sup>, here, cameras are in motion. This increases the variability of the type of images collected which, depending on the speed of the user, may be more or less blurred. Moreover, the diversity of platforms and observer points of view (see Fig. 1) introduces diversity among images of the same marine species (intra-class dissimilarity). Indeed, pictures taken from boards are always taken from above, whereas pictures taken while snorkeling may be taken from the side.

All images provide geolocation data, which significantly enhances the dataset. However, the precision of the geolocation can greatly vary depending on the platform used to collect data (please refer to the Image Georeferencing section). For annotated images, this precision makes it possible to directly associate the occurrence of a taxon with latitude and longitude coordinates. These data can typically be shared as part of open international platforms such as the Global Biodiversity Information Facility (GBIF), which we intend to do in the near future. In the case of unannotated images, the visible taxa can be inferred through the use of artificial intelligence techniques. This data being very dense and precisely geolocated, it can then be used to construct species distribution maps at very high resolution and high survey frequencies. To this end, each underwater image in the dataset has a prediction of presence or absence for classes described in the Data annotation section, made using the computer vision model described in the Session deposit section.

Differing from many existing public datasets in the marine biology domain such as CoralNet<sup>27</sup>, DeepFish<sup>28</sup> and UIEB<sup>29</sup>, the proposed dataset and benchmark do not only aim to advance and evaluate the performance of a specific computer vision task. Instead, we think that this new dataset opens up many possibilities of use, such as:

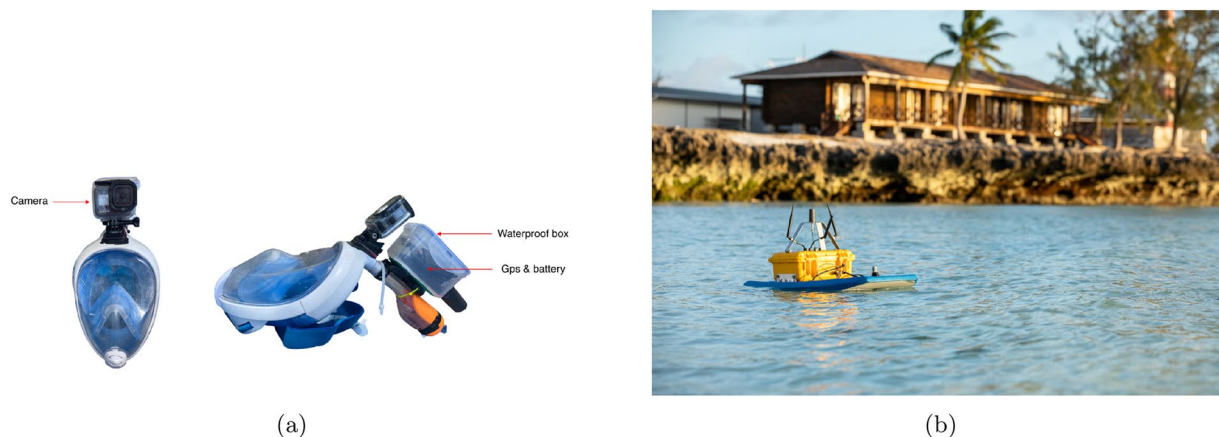
- providing historical pictures of marine ecosystems with high precision location which enables to come back at the same location and assess the evolution of these ecosystems.
- evaluating existing computer vision models on spatialized images, achieving important ecological results (e.g., species distribution maps) on areas of great interest (marine protected areas, strict natural reserves or simply areas that are difficult to access).
- training/testing AI models on a subset of annotated images (e.g., train from scratch or fine-tuning computer vision models on a curate annotated dataset).
- combining individual photos to generate aerial or underwater photogrammetric models using imagery acquired by scientific platforms with sufficient horizontal and vertical overlap, as well as accurate geolocation and attitude (Roll, Pitch and Yaw) data.
- validate remote sensing models with geolocated field data, enabling rapid photo interpretation.
- provide a data structure that fosters the good practices for open science and FAIR data management by implementing various community standards (e.g., Darwin Core for biodiversity, GCRMN for scientific diving, EXIF & XMP for cameras and COCO for computer vision) to disseminate the same dataset.

Images have numerous metadata that may vary depending on how they were collected. This metadata are a valuable resource that can be used to improve the performance of neural networks, reconstruct photogrammetry models or be used as stand-alone data to produce analysis (e.g., bathymetric maps using depth data, considering that more than 85% of images are associated with a depth measure from an echosounder). Protocols and codes are also described and openly shared for scientists and citizens for the sake of reproduction and contribute building a bigger collaborative resource.

## Methods

**Instrumentation.** Data were collected from kitesurf, paddleboard, snorkel, Autonomous Surface Vehicle (ASV), Unmanned Aerial Vehicle (UAV) DJI Mavic 2 Pro. Cameras used to collect underwater images are GoPro HERO5, GoPro HERO7, GoPro HERO8. The GPS module to georeference images is a centimeter-resolution *Emlid* Reach M2.

**Data acquisition system.** Underwater data acquisition was carried out using instrumented marine platforms. These platforms can be divided into two groups: citizen and scientific platforms. The first type of platforms (paddleboards, kitesurfs and snorkeling masks) are designed for being used by citizens practicing marine sports. These platforms are equipped with a camera and, most of the time, with a GPS module (differential GPS module when possible), allowing the acquisition of georeferenced images with centimeter accuracy. Citizen images have been collected by the “Seatizen” team with the objective of creating platforms, standards, and a methodology to measure the impact that citizen science can have on scientific projects. The second type of platform is an Autonomous Surface Vehicle (ASV) equipped with a GPS module, an echosounder and a GoPro<sup>30</sup>. Aerial images have been collected by commercial drones (e.g., Mavic 2 pro).



**Fig. 2** Examples of instrumented platforms: **(a)** citizen platform mask equipped with a camera and a differential GPS **(b)** scientific platform Autonomous Surface Vehicle.

Figure 2 shows an example of both citizen and scientific platforms. Aerial data acquisition was carried out using a DJI Mavic 2 Pro aerial drone.

**Video / camera system.** Underwater images in the dataset come from photos (classic or timelapse mode) or videos taken with action cameras, like GoPro cameras. Videos have been first split into frames at a fixed time interval. The frame extraction frequency can be adjusted during the post-processing phase. This flexibility allows to extract more frames when identifying interesting moments, such as when a rare species appears briefly, or fewer frames during unimportant parts of the session. The resulting frame resolution is either  $1920 \times 1080$  or  $3840 \times 2160$  pixels, LRV files from GoPro are also usable instead of MP4 files. It results in less resolution ( $854 \times 480$ ) than MP4 but photogrammetry can still be done with them.

**Data synchronization.** For most of the platforms, cameras are often immersed in the water. As GPS waves cannot propagate in water, the GPS system (or antenna) has to be placed on the surface of the platform. Some images of the dataset were collected along with a differential GPS module to get centimeter-level position. We use time in order to synchronize positions and images. 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 to infer their accurate location (by using the frequency at which the frames of the video were extracted). The inferred GPS position is saved in a CSV file to perform geospatial analysis on images. For frames extracted from videos, the GPS position is recorded in the image exif metadata thanks to the *Exiftool* Perl library.

**Image georeferencing.** The positional accuracy of images varies depending on the platform used to collect the data, such as paddleboards, scuba diving, ASVs or UAVs. To ensure transparency and allow validation of the position data for each data collection event, when differential GPS was used, the `metadata.csv` file located in the `METADATA` folder (see the Session deposit section for more information about a generic folder diagram) contains two columns, `GPSsde` and `GPSsdr`, which provide the standard deviations of the GPS position on the East and North axes respectively. In the `GPS/DEVICE` folder of each data collection event, graphs describing the position accuracy can be founded. These values give an estimation of the accuracy of the GPS receiver position.

**Citizen platforms.** For sessions gathered by citizen scientists using platforms like paddleboards, kitesurfs or snorkel, attitude (Roll, Pitch and Yaw) and depth data are not provided. Therefore, depending on the angle at which the image was taken and the depth at which the camera was, the difference between the position of the GPS receiver and the actual position of the image can vary greatly. In future versions of the dataset, an additional uncertainty can be introduced by the offset between the GPS receiver position and the camera position, that varies in space and time within the same session (e.g., when the GPS receiver is a watch worn by the user and the camera is on the board, during a kitesurf or paddleboard session).

**ASV.** For sessions conducted with ASVs, the GPS data is corrected using Post-Processed Kinematic (PPK) techniques, ensuring a high positional precision on the order of centimeters. However, this centimeter-level precision applies to the GPS data itself and does not necessarily reflect the exact position of the image footprint on the seabed. Factors such as the platform's attitude can affect the precise location of the image footprint. To account for angles of rotation on the three axis for ASV sessions, each image metadata includes attitude data collected from the ArduPilot system<sup>30</sup>, along with corresponding echosounder data (depth). These additional data can be used to compute the image footprint on the seabed and further refine the image position.



**UAV.** The UAV data included in the dataset were acquired without differential GPS, resulting in positional accuracy on the order of meters. If Ground Control Points (GCPs) were used to refine the positioning of the SfM model, their coordinates are saved in the `METADATA` folder, along with a description of the chosen points.

To assist users in evaluating the positioning accuracy of each SfM model, a detailed photogrammetry report `000_photogrammetry_report.pdf` automatically generated by OpenDroneMap is provided at the root of each data collection event. Specifically, in the `Processing Summary` section of this file, the georeferencing accuracy is estimated by the `GPS errors` field. Further details on this error, including GPS, GCP (when available), and 3D errors, are developed in the `GPS/GCP/3D Errors Details` section of the report.

The influence of water refraction was not considered during SfM processing.

**Data annotation.** Some images of the dataset have also been annotated to describe their content and enable image recognition. Three different types of annotation have been used in the current dataset:

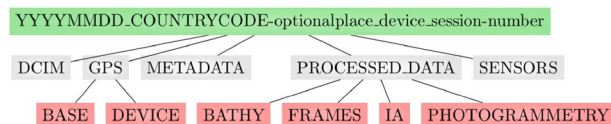
- the binary convention: images are classified as “useful” or “useless” based on their ecological interest (those annotated useful are considered of interest from an ecological perspective, while those taken out of the water, too blurry or unusable are considered useless). Images where people can be identified are also removed automatically. This type of annotation is not shared because of privacy rights, as they often contain identifiable people.
- the multilabel convention: the annotator indicates all the classes that he/she observes on the image. The classes are derived from the GCRMN standards (see below).
- the instance segmentation convention: the annotation is the same as the multilabel one, but in addition to indicating all the classes present on the image, the annotator must also highlight the contours of each instance for each class.

Annotation classes are based on GCRMN scenario 2 standards, a widely accepted framework for coral reef monitoring<sup>31</sup>. To apply GCRMN classes for deep learning and computer vision tasks, we modify the description of two classes: *Dead coral* and *Living coral*:

- The *Living coral* category was refined to only include healthy corals that could not be assigned to a specific morphotype due to image limitations (such as distance from the camera or water turbidity). This adjustment prevents all morphotype corals (e.g., *Acropora Branching*, *Acropora Digitate*) from being broadly categorized under *Living coral*, thus ensuring that no overlap occurs between different classes.
- The *Dead coral* category was modified to include all corals that have died, regardless of their original morphotype. By placing all dead corals into a single category, we avoid potential ambiguity and overlap that could arise from attempting to maintain morphotype distinctions for non-living corals.

In order to further improve the interoperability and applicability of our dataset for both AI experts and marine biologists, we have integrated the corresponding taxon for each class in the GBIF format. This ensures that the classification system is standardised, thus facilitating the work of both communities with the data. Further information can be found in the `seatizen_atlas_db.gpkg` file and in the `gbif_class_mapping.csv` file. Classes were then completed with classes added by the Seatizen project team (underlined classes). In total, 38 classes were retained for annotation. Classes can be divided into 6 groups:

- Algae
  - (1) *Algal Assemblage*
  - (2) *Algae Halimeda*
  - (3) *Algae Coralline*
  - (4) *Algae Turf*
- Coral
  - (1) *Acropora Branching*
  - (2) *Acropora Digitate*
  - (3) *Acropora Submassive*
  - (4) *Acropora Tabular*
  - (5) *Bleached coral*
  - (6) *Dead coral*
  - (7) *Gorgonian*
  - (8) *Living coral*
  - (9) *Non-acropora Millepora*
  - (10) *Non-acropora Branching*
  - (11) *Non-acropora Encrusting*
  - (12) *Non-acropora Foliose*
  - (13) *Non-acropora Massive*
  - (14) *Non-acropora Coral free*
  - (15) *Non-acropora Submassive*



**Fig. 3** Folder diagram of a data collection session.

- Seagrass

- (1) *Syringodium isoetifolium*
- (2) *Thalassodendron ciliatum*

- Habitat

- (1) *Rock*
- (2) *Rubble*
- (3) *Sand*

- Other Organisms

- (1) *Thorny starfish*
- (2) *Sea anemone*
- (3) *Ascidians*
- (4) *Giant clam*
- (5) *Fish*
- (6) *Other starfish*
- (7) *Sea cucumber*
- (8) *Sea urchin*
- (9) *Sponges*
- (10) *Turtle*

- Custom classes

- (1) *Blurred*
- (2) *Homo Sapiens*
- (3) *Human object*
- (4) *Trample*
- (5) *Useless*
- (6) *Waste*

More details about each class can be found in the Supplementary Information file.

## Data Records

Data are stored and shared using the Zenodo multi-disciplinary open repository maintained by CERN (European Organization for Nuclear Research). The dataset<sup>32</sup> is currently made of 291 deposits which are data collection events (surveys) that have been operated either by an ASV (234) or an UAV (33), by scuba diving (22) or paddleboard (2). Each survey has been assigned its own DOI (Digital Object Identifier) and complies with standardized data structure and naming conventions. The main dataset is finally built on top of all surveys. The next three subsections describe these three types of data deposits and the structure of the related DOI.

**Session deposit.** The first type of data deposit refers to a specific survey which is a data collection event. Two distinct versions may be available for each repository:

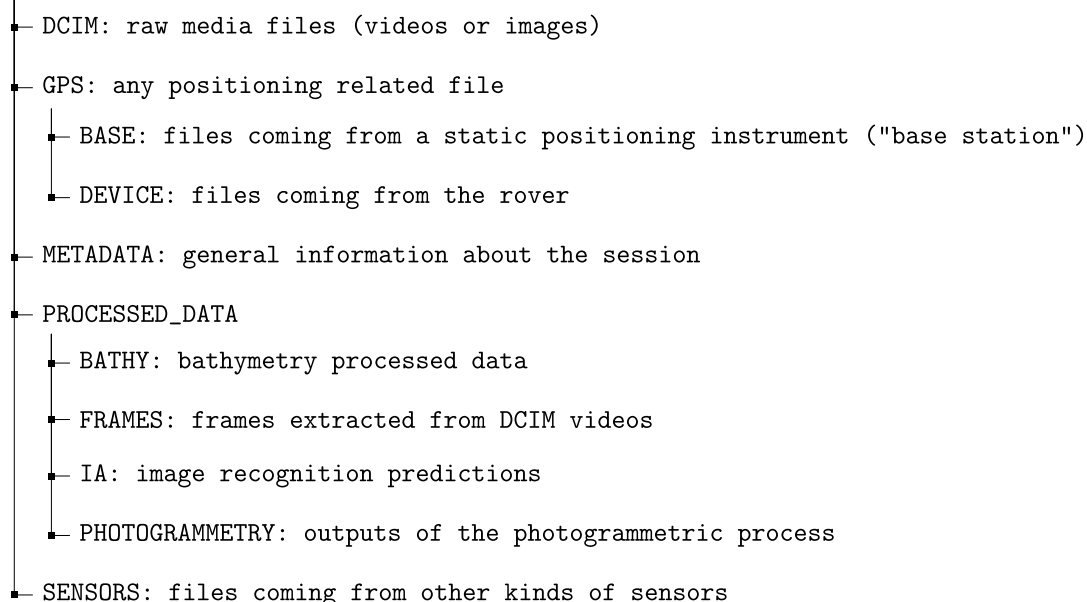
- RAW\_DATA: a version containing raw data that can be either open or restricted (open under request) if containing images that are not shareable with the community (e.g., human faces or blurry images). Raw data are kept safe to ensure the work can be reproduced or improved if needed.
  - PROCESSED\_DATA: a version containing the processed data.
- Each data survey complies with a file structure and a naming convention which is illustrated in Fig. 3:
- each data survey has a unique identifier which comes from the concatenation of the following four items (separated by the underscore symbol, e.g., YYYYMMDD\_COUNTRYCODE-optionalplace\_device\_session-number):
- (a) the date of the session which complies with the standard ISO 8601 date format YYYYMMDD.
  - (b) the Alpha-3 code of the country where data were collected and, if the user deems it appropriate, an

additional tag describing the location where the data collection took place separated by a dash symbol (e.g., if the session took place in Le Morne area in Mauritius then the user may write MUS-LeMorne or MUS-LEMORNE or MUS-Le-Morne).

- (c) the type of platform used in order to collect data (e.g., scuba diving, kitesurf, paddleboard, surfboard, ASV, UAV).
- (d) the sequential number of the survey (needed to distinguish different surveys within the same day in the same location with the same platform).

Each data deposit is divided into five folders: DCIM, GPS, METADATA, PROCESSED\_DATA and SENSORS. Below is a description of each of them:

YYYYMMDD\_COUNTRYCODE-optionalplace\_device\_session-number



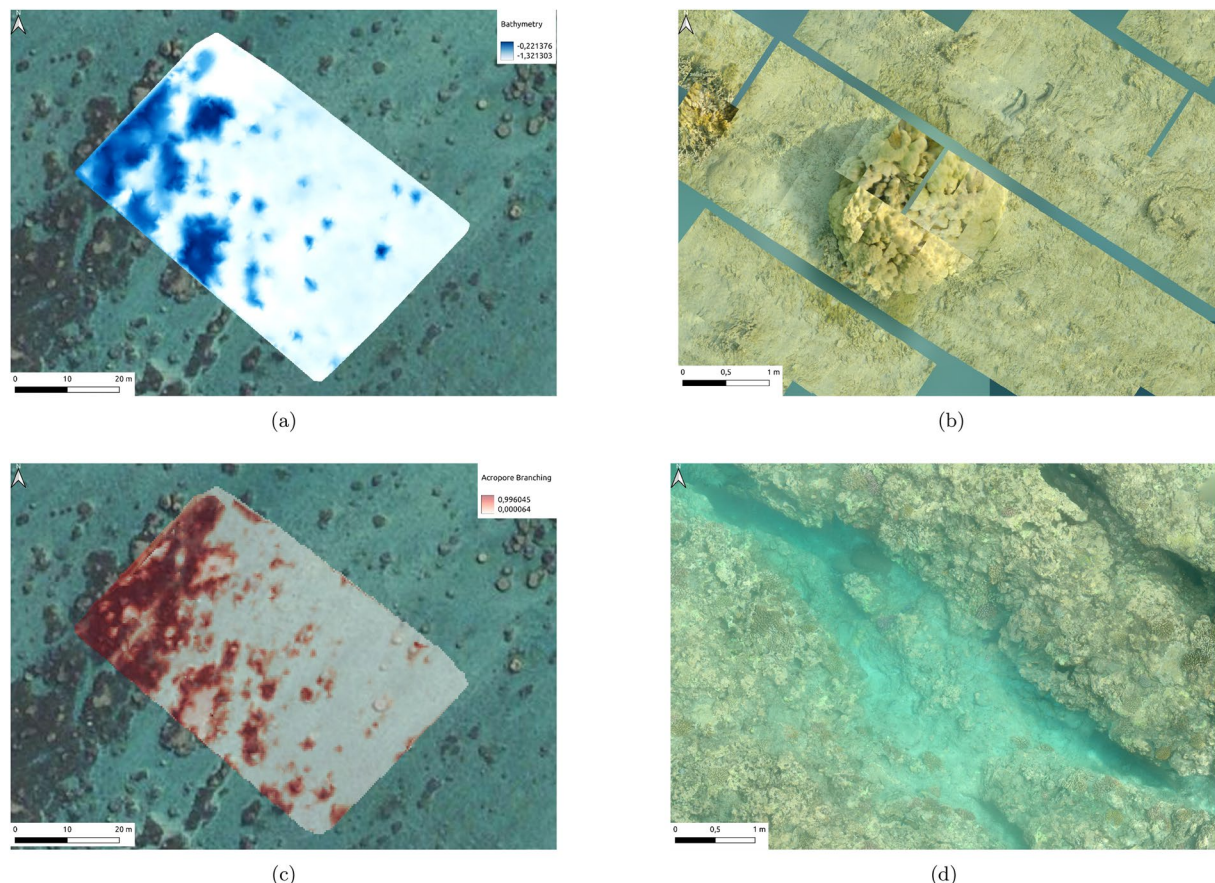
- DCIM: folder to store videos and photos depending on the media collected.
- GPS: folder to store any positioning related file. If any kind of correction is possible on files (e.g., Post-Processed Kinematic thanks to rinex data) then the distinction between device data and base data is made: data are then sorted in the BASE folder for files coming from rtk station or any static positioning instrument and DEVICE when coming from the device. On the other hand, if only device position data are present and the files cannot be corrected by post-processing techniques (e.g., gpx files), then the distinction between base and device is not necessarily made and the files may be placed directly at the root of the GPS folder.
- METADATA: folder with general information files about the session, exif metadata
- PROCESSED\_DATA: contains all the folders needed to store the results of the data processing of the current session.

- BATHY: destination folder for bathymetry processed data. See for instance Fig. 4.
- FRAMES: destination folder for frames extracted from DCIM videos, enriched with metadata (when present). See for instance Fig. 4.
- IA: destination folder for image recognition predictions. Inference has been made with *DinoVdeau* model. More informations are available in the Code availability Section. See for instance Fig. 4.
- PHOTOGRAMMETRY: destination folder for reconstructed models in photogrammetry. See for instance Fig. 4.

- SENSORS: folder to store files coming from other sources (bathymetry data from the echosounder, log file from the autopilot, mission plan, etc.)

Figure 4 illustrates four examples of use of the dataset based on advanced technologies. These include the generation of detailed bathymetric maps, the integration of underwater images with aerial orthophotos thanks to precise georeferencing, the use of deep learning for underwater image classification and the creation of orthophotos using Structure from Motion (SfM) techniques.

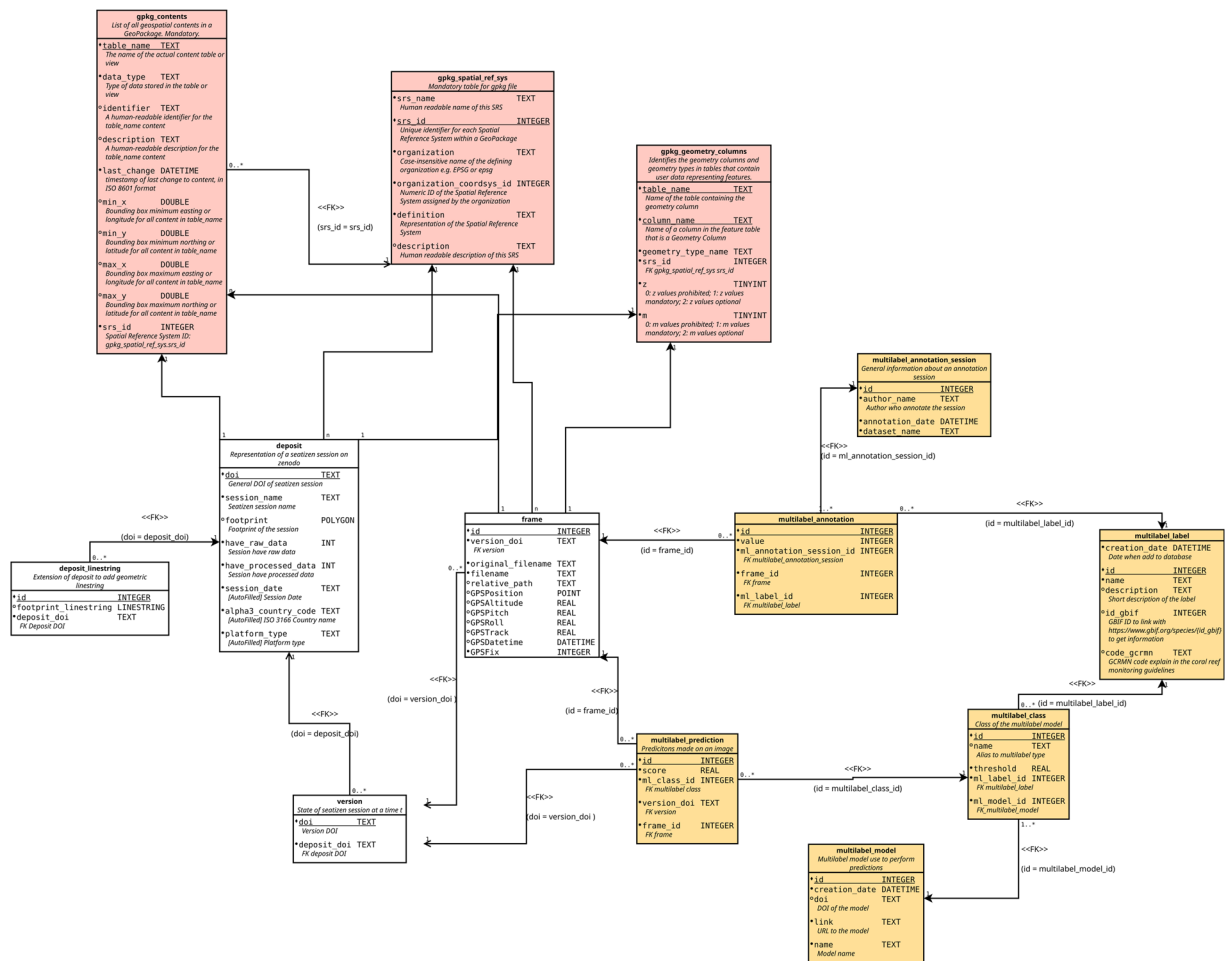
**Global deposit.** The second type of data deposit is the main one and is called [Seatizen Atlas](#)<sup>32</sup>, it provides a description of the dataset as a whole. This repository contains the following files:



**Fig. 4** Examples of processed data obtained from ASV and UAV raw data in the west part of Reunion island: **(a)** Example of a bathymetric map obtained from data collected by the ASV echosounder **(b)** Example of underwater ASV images georeferenced using differential GPS **(c)** Example of a distribution map of the presence of *branching acropora* computed with a *DinoV2* vision transformer trained on the annotated data of Seatizen Atlas and used to automatically make predictions on unannotated images **(d)** Example of an underwater orthophoto obtained from georeferenced ASV images, using SfM techniques and the opensource software OpenDroneMap.

- `seatizen_atlas_db.gpkg`: geopackage file that stores extensive geospatial data, allowing for efficient management and analysis of spatial information. More information about the file structure can be found in the Unified Modeling Language diagram in Fig. 5.
- `session_doi.csv`: a CSV file listing all sessions published on Zenodo. This file contains the following columns:
  - `session_name`: identifies the session.
  - `session_doi`: indicates the URL of the session.
  - `place`: indicates the location of the session.
  - `date`: indicates the date of the session.
  - `raw_data`: indicates whether the session contains raw data or not.
  - `processed_data`: indicates whether the session contains processed data.
- `metadata_images.csv`: a CSV file describing all metadata for each image published in open access. This file contains the following columns:
  - `OriginalFileName`: indicates the original name of the photo.
  - `FileName`: indicates the name of the photo adapted to the naming convention adopted by the Seatizen team (i.e., `YYYYMMDD_COUNTRYCODE-optionalplace_device_session-number_original-imagename`).
  - `relative_file_path`: indicates the path of the image in the deposit.
  - `frames_doi`: indicates the DOI of the version where the image is located.
  - `GPSLatitude`: indicates the latitude of the image (if available).
  - `GPSLongitude`: indicates the longitude of the image (if available).
  - `GPSAltitude`: indicates the depth of the image (if available).
  - `GPSRoll`: indicates the roll of the image (if available).





**Fig. 5** UML of geopackage file.

- GPSPitch: indicates the pitch of the image (if available).
- GPSTrack: indicates the track of the image (if available).
- GPSDatetime: indicates when images were taken (if available).
- GPSFix: indicates GNSS quality levels (if available).
- metadata\_multilabel\_predictions.csv: a CSV file describing all predictions from last multilabel model with georeferenced data.
  - FileName: indicates the name of the photo adapted to the naming convention adopted by the Seatizen team (i.e., YYYYMMDD\_COUNTRYCODE-optionalplace\_device\_session-number\_original-imagename).
  - frames\_doi: indicates the DOI of the version where the image is located.
  - GPSLatitude: indicates the latitude of the image (if available).
  - GPSLongitude: indicates the longitude of the image (if available).
  - GPSAltitude: indicates the depth of the frame (if available).
  - GPSRoll: indicates the roll of the image (if available).
  - GPSPitch: indicates the pitch of the image (if available).
  - GPSTrack: indicates the track of the image (if available).
  - GPSFix: indicates GNSS quality levels (if available).
  - prediction\_doi: refers to a specific AI model prediction on the current image (if available).
  - A column for each class predicted by the AI model.
- metadata\_multilabel\_annotation.csv: a CSV file listing the subset of all the images that are annotated, along with their annotations. This file contains the following columns:
  - FileName: indicates the name of the photo.
  - frame\_doi: indicates the DOI of the version where the image is located.
  - relative\_file\_path: indicates the path of the image in the deposit.
  - annotation\_date: indicates the date when the image was annotated.

- A column for each class with values:
  - \* 1: if the class is present.
  - \* 0: if the class is absent.
  - \* -1: if the class was not annotated.
- seatizen\_atlas.qgz: a qgis project which formats and highlights the geopackage file to facilitate data visualization.
- darwincore\_multilabel\_annotations.zip: a Darwin Core Archive (DwC-A) file listing the subset of all the images that are annotated, along with their annotations.

These files bring together all key information in an accessible and easy-to-read format, ensuring that key data are easily available and well-organized.

**Annotation deposit.** In order to facilitate the use of annotated images, a third deposit containing all annotated images and their annotation files was created: [Seatizen Atlas image dataset](#)<sup>33</sup>. This repository contains the following files:

- 20241016\_132819\_multilabel\_annotations.csv: a CSV file describing all multilabel annotations. This file contains the following columns:
  - FileName: indicates the name of the photo.
  - A column for each class with values:
    - \* 1: if the class is present.
    - \* 0: if the class is absent.
    - \* -1: if the class was not annotated.
- segmentation\_annotations.json: a JSON file describing all instance segmentation annotations, formatted according to the COCO standard.
- gbif\_class\_mapping.csv: a CSV file mapping the dataset classes to their corresponding GBIF taxa. This file contains the following columns:
  - label: the name of the class in the current dataset.
  - old\_label: the original name used in previous versions of the dataset.
  - code\_gcrmn: the GCRMN code<sup>31</sup> (if available).
  - id\_gbif: a link to the corresponding taxon in GBIF (if available).
- images.zip: a compressed file containing all annotated images.

This deposit currently includes 14,492 multilabel and 1,200 instance segmentation annotated images.

## Technical Validation

**Annotation methodology.** Multilabel annotations were made by trained marine scientists: Thomas Chevrier, Léanne Carpentier, Cam Ly Rintz, Julien Barde, Belen de Ana, Laurence Maurel, and Pascal Mouquet. In order to standardise the annotations and eliminate annotator-related bias, Victor Russias (biologist working in the field of marine biology of tropical ecosystems) later took over all the annotations to make them consistent. The annotator asked expert benthologists (Laurence Maurel, Lionel Bigot, Thierry Mulochau) to confirm his/her annotations.

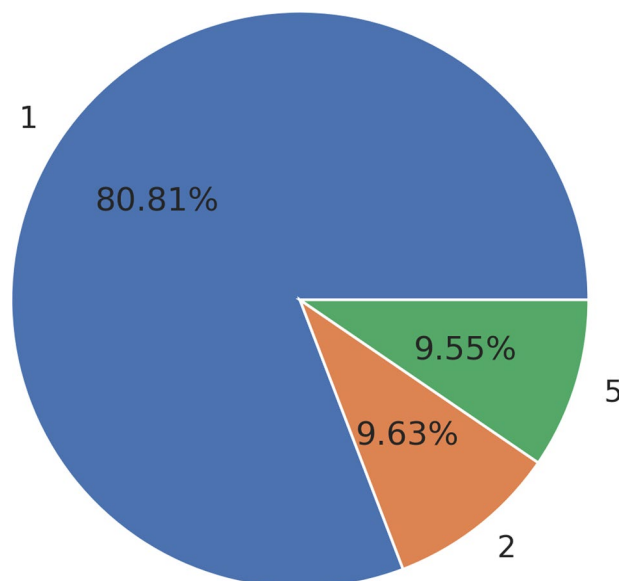
Instance segmentation annotations were made by Laurence Maurel, expert biologist with 20-year experience in the field.

**Spatial data accuracy.** In the current version of the *Seatizen Atlas* dataset, 97.91% of underwater images are embedded with a position gathered with a differential GPS receiver Emlid Reach M2. GPS data are saved in RINEX (Receiver Independent Exchange Format) files, which allow a spatial correction using Post-Processed Kinematic (PPK) techniques. In order to do so, a base station is needed. In Réunion island data were corrected using data from the centipede collaborative open network<sup>34,35</sup>.

For each data collection event involving a differential GPS receiver, the GPSfix, GPSsde and GPSsdn columns recorded in the metadata.csv file for each image give an estimate of the accuracy of the positioning of the GPS:

- GPSfix: indicates GNSS quality levels (1, 2 or 5). A value of 1 indicates a high-precision fixed solution of the order of centimeters. A value of 2 indicates a float solution of the order of decimeters. A value of 5 indicates a single-point positioning solution of the order of meters.
- GPSsde: indicates the standard deviation of the GPS position on the East axis.
- GPSsdn: indicates the standard deviation of the GPS position on the North axis.

In Fig. 6, the percentages of each GNSS quality level are shown for underwater images of the dataset. As can be seen, 80.81% of the images have a GNSS quality level of 1, indicating a high-precision fixed solution of the order of centimeters.



**Fig. 6** GNSS quality levels for underwater images of the dataset.

To better understand the accuracy of the GPS position, we compute the mean of the standard deviation of the GPS position on the East and North axes for each image. The mean of the GPSsde and GPSsdn values are 0.065 meters and 0.087 meters, respectively. However, computing them only on images with a GNSS quality level of 1, the mean values of both GPSsde and GPSsdn are of 5 millimeters.

### Usage Notes

The authors are available for consultations about and collaborations involving the data.

To facilitate the reuse and sharing of data, a GitHub repository has been created at [zenodo-tools Github](#). After selecting the sessions you are interested in by filtering on the global repository [Seatizen Atlas](#), you can download them directly by following the instructions in the README file. This repository also aims to allow those who wish to contribute to the growth of the dataset to participate in a simple and intuitive manner. Instructions for uploading new sessions to Zenodo can be found in the repository.

### Code availability

All codes for data processing associated with the current submission are available on [plancha-workflow Github](#) under the **v1.0.3** release.

The code for downloading data associated with the current submission is available on [zenodo-tools Github](#).

The code to train the neural network model used in the current submission is available on [DinoVdeau Github](#).

Received: 27 August 2024; Accepted: 9 December 2024;

Published: 14 January 2025

### References

- Nieto, A., Ralph, G., Comeros-Raynal, M., Heessen, H. & Rijnsdorp, A. *European Red List of marine fishes*. Publications Office of the European Union, 2015.
- Cinner, J. E. *et al.* Gravity of human impacts mediates coral reef conservation gains. *Proceedings of the National Academy of Sciences* **115**, E6116–E6125, <https://doi.org/10.1073/pnas.1708001115> (2018).
- Graham, N. *et al.* Extinction vulnerability of coral reef fishes. *Ecology Letters* **14**, 341–348, <https://doi.org/10.1111/j.1461-0248.2011.01592.x> (2011).
- Villon, S. *et al.* A deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecological Informatics*, **48**, <https://doi.org/10.1016/j.ecoinf.2018.09.007> (2018).
- Robinson, J. P. W. *et al.* Fishing degrades size structure of coral reef fish communities. *Global Change Biology* **23**, 1009–1022 (2017).
- Hoegh-Guldberg, O. & Bruno, J. The impact of climate change on the world's marine ecosystems. *Science* **328**, 1523–1528, <https://doi.org/10.1126/science.1189930> (2010).
- Halpern, B. *et al.* A global map of human impact on marine ecosystems. *Science* **319**, 948–952, <https://doi.org/10.1126/science.1149345> (2008).
- Harley, C. *et al.* The impacts of climate change in coastal marine systems. *Ecology Letters* **9**, 228–241, <https://doi.org/10.1111/j.1461-0248.2005.00871.x> (2006).
- Hoegh-Guldberg, O. *et al.* Coral reefs under rapid climate change and ocean acidification. *Science* **318**, 1737–1742, <https://doi.org/10.1126/science.1152509> (2008).
- Kulbicki, M. *et al.* Global biogeography of reef fishes: A hierarchical quantitative delineation of regions. *PloS One* **8**, e81847, <https://doi.org/10.1371/journal.pone.0081847> (2013).
- Rogers, A., Blanchard, J. & Mumby, P. Fisheries productivity under progressive coral reef degradation. *Journal of Applied Ecology*, **55**, <https://doi.org/10.1111/1365-2664.13051> (2017).
- Watch, R. Coral reefs dashboard. <https://resourcewatch.org/dashboards/coral-reefs>, Accessed: 2024-08-05 (2024).

13. Cinner, J. *et al.* Gravity of human impacts mediates coral reef conservation gains. *Proceedings of the National Academy of Sciences* **115**, 201708001, <https://doi.org/10.1073/pnas.1708001115> (2018).
14. Thresher, R. & Gunn, J. Comparative analysis of visual census techniques for highly mobile, reef-associated piscivores (carangidae). *Environmental Biology of Fishes* **17**, 93–116 (2004).
15. Brock, V. E. A preliminary report on a method of estimating reef fish populations. *The Journal of Wildlife Management* **18**, 297–308 (1954).
16. Mallet, D. & Dominique, P. Underwater video techniques for observing coastal marine biodiversity: A review of sixty years of publications (1952–2012). *Fisheries Research* **154**, 44–62, <https://doi.org/10.1016/j.fishres.2014.01.019> (2014).
17. Villon, S. *et al.* Coral reef fish detection and recognition in underwater videos by supervised machine learning: Comparison between deep learning and HOG + SVM methods. In Blanc-Talon, J., Distant, C., Philips, W., Popescu, D. & Scheunders, P. (eds.), *Advanced Concepts for Intelligent Vision Systems*, pages 160–171. Springer International Publishing, Cham, 2016.
18. Di Ciaccio, F. & Troisi, S. Monitoring marine environments with autonomous underwater vehicles: A bibliometric analysis. *Results in Engineering* **9**, 100205, <https://doi.org/10.1016/j.rineng.2021.100205> (2021).
19. Aird, P. Chapter 5 - Deepwater: Essentials and differences. In Aird, P. (ed.), *Deepwater Drilling*, pages 165–224. Gulf Professional Publishing, <https://doi.org/10.1016/B978-0-08-102282-5.00005-3> (2019).
20. Francour, P., Liré, C. & Harvey, E. Comparison of fish abundance estimates made by remote underwater video and visual census. *Naturalista Siciliano* **23**, 155–168 (1999).
21. Al-Saffar, A. A. M., Tao, H. & Talab, M. A. Review of deep convolution neural network in image classification. In *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, pages 26–31, <https://doi.org/10.1109/ICRAMET.2017.8253139> (2017).
22. Siddiqui, S. *et al.* Automatic fish species classification in underwater videos: Exploiting pretrained deep neural network models to compensate for limited labelled data. *ICES Journal of Marine Science*, **75**, <https://doi.org/10.1093/icesjms/fsx109> (2017).
23. Galdran, A., Pardo, D., Picon, A. & Alvarez-Gila, A. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, **26**, <https://doi.org/10.1016/j.jvcir.2014.11.006> (2014).
24. Wang, K. *et al.* Underwater image restoration based on a parallel convolutional neural network. *Remote Sensing* **11**, 1591, <https://doi.org/10.3390/rs11131591> (2019).
25. Jalal, A., Mian, A., Shortis, M. & Shafait, F. Fish detection and species classification in underwater environments using deep learning with temporal information. *Ecological Informatics* **57**, 101088, <https://doi.org/10.1016/j.ecoinf.2020.101088> (2020).
26. Pelletier, D., Carpentier, L., Roman, W. & Bockel, T. La vidéo rotative autonome pour l'observation des habitats et de la macrofaune côtiers. Guide méthodologique des systèmes stavro et micado. 2016.
27. Chen, Q., Beijbom, O., Chan, S., Bouwmeester, J. & Kriegman, D. A new deep learning engine for coralnet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3693–3702, 2021.
28. Qin, H., Li, X., Liang, J., Peng, Y. & Zhang, C. Deepfish: Accurate underwater live fish recognition with a deep architecture. *Neurocomputing* **187**, 49–58, <https://doi.org/10.1016/j.neucom.2015.10.122> (2016).
29. Li, C. *et al.* An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing* **29**, 4376–4389, <https://doi.org/10.1109/TIP.2019.2955241> (2020).
30. Gogendeau, P. *et al.* An open-source autonomous surface vehicle for acoustic tracking, bathymetric, and photogrammetric surveys, 2024. arXiv: <https://arxiv.org/abs/2406.18760>.
31. Conand, C., Chabanet, P., Quod, J.-P., Bigot, L. & de Grissac, A. J. *Guidelines for coral reef monitoring in the south-west region of the Indian Ocean*. Indian Ocean Commission, (1999).
32. Contini, M., Barde, J., Bonhommeau, S., Illien, V. & Joly, A. Seatizen atlas. *Zenodo* <https://doi.org/10.5281/zenodo.13951435> (2024).
33. Contini, M., Barde, J., Bonhommeau, S., Illien, V. & Joly, A. Seatizen atlas image dataset. *Zenodo* <https://doi.org/10.5281/zenodo.13951614> (2024).
34. Ancelin, J., Ladet, S. & Heintz, W. CentipedeRTK, un réseau pour la géolocalisation haute précision au service de l'environnement, 2024.
35. Ancelin, J. *et al.* jancelin/docs-centipedeRTK: v2.4. *Zenodo* <https://doi.org/10.5281/zenodo.6760153> (2022).

## Acknowledgements

We would like to express our sincerest gratitude to the following individuals and organizations for their invaluable contributions to this project: SIF (Seychelles Islands Foundation), RNMR (Réserve Naturelle Marine de la Réunion), Government of Mauritius, Government of Seychelles, Monaco Explorations, De Oliveira Denis, Morand Gaétan, De Ana Belen, Babet Laura, Fari Chloé, Roumagnac Marie, Nieblas Anne-Elise, Chanut Jeremie, Kerzerho Vincent, Bigot Lionel, Mulochau Thierry, Broudic Léo, Pinault Mathieu, and Maurel Laurence. For research conducted in Saya de Malha, authorisation was given by the Memorandum of Understanding between Société des Explorations de Monaco and the designated authority of the joint management area, signed on 13/06/2022. For research conducted in the Seychelles ZEE, authorisation was given by the Memorandum of Understanding between Société des Explorations de Monaco and the Seychelles Government, signed on 14/10/2022. The project's success is in no small part due to the invaluable support and dedication of those involved. Data collection was part of 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).

## Author contributions

Sylvain B., J.B., A.J and M.C. conceived the work, M.C., V.I., Sylvain B., J.B. and S.P. standardized and published the data, M.C., V.I., M.J., M.R., V.R., A.L., T.C., C.L.R., L.C., P.G., Serge B., S.P., J.B. and Sylvain B. collected the data, M.C., V.I., M.J., T.C., C.L.R., L.C., P.G., C.L., Serge B., A.B., J.D., S.P., J.B., A.J. and Sylvain B. curated the data. V.I. standardized and published the code. All authors reviewed and approved the final version of the manuscript.

## Competing interests

The authors declare no competing interests.



## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-024-04267-z>.

**Correspondence** and requests for materials should be addressed to M.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025, corrected publication 2025