scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Robotic monitoring of European habitats: a labeled dataset for plant detection in Annex I habitats of Italy

Giovanni Di Lorenzo[1] ✉, Franco Angelini[1], Michele Pierallini[1], Simone Tolomei[1], Davide De Benedittis[1], Agnese Denaro[2], Giovanni Rivieccio[2], Maria Carmela Caria[2,3], Federica Bonini[4], Anna Grassi[4], Leopoldo de Simone[5], Emanuele Fanfarillo[3,5], Tiberio Fiaschi[5], Simona Maccherini[3,5], Barbara Valle[3,5,6], Marina Serena Borgatti[6], Simonetta Bagella[2,3], Daniela Gigante[4], Claudia Angiolini[3,5], Marco Caccianiga[6] & Manolo Garabini[1]

**The present data descriptor presents a dataset designed for the detection of plant species in various habitats of the European Union. This dataset is based on images captured using multiple different hardware including quadrupedal robot ANYmal C, referring to ecologically important species to assess the presence and conservation status in Annex I habitats 2110, 2120, 6210*, 8110, 8120, and 9210*. Plant scientists and robotic engineers gathered the data in key Italian protected areas and labeled it using YOLOtxt format. Researchers in vegetation science, habitat monitoring, robotics, machine learning, and biodiversity conservation can access the dataset through Zenodo. The ultimate goal of this collaborative effort was to create a dataset that can be used to train artificial intelligence models to assess parameters that enable robotic habitat monitoring. The availability of this dataset may enhance future studies and conservation initiatives for Annex I habitats inside and outside the Natura 2000 network. The dataset and the methods used to obtain it are fully described, highlighting the significance of interdisciplinary cooperation in habitat monitoring.**

## Background & Summary

In the European Green Deal[1], the European Union (EU) highlights the central role of habitat and species conservation, as per its directive on habitats[2]. The goal is to expand the areas covered and protected by the Natura 2000 network[3] (N2000N) and to increase the capacity to collect sufficient data to assess their conservation status. Data obtained from habitat monitoring activities provides valuable insights into ongoing changes and enables the development of more effective mitigation and adaptation strategies. The fact that more and more ecosystems and species are in jeopardy[4,5] makes these activities even more crucial.

Currently, the monitoring task in key areas of the N2000N is carried out mostly by human operators, particularly in terrestrial environments. In fact, only highly skilled workers have the knowledge and expertise to evaluate the presence and conservation status of specific habitats, operating simultaneously in unstructured natural environments not known a priori, like natural habitats.

The increasing pressures exerted by direct and indirect anthropogenic factors negatively influence several EU habitats. Land use changes and fragmentation of landscapes are significant issues, that lead to the degradation and loss of biodiversity[6]. Human activities, such as intensive land management[7] practices and mechanical

[1]Centro di Ricerca 'Enrico Piaggio', and Dipartimento di Ingegneria dell'Informazione, University of Pisa, Largo Lucio Lazzarino 1, 56122, Pisa, Italy. [2]Department of Chemical, Physical, Mathematical and Natural Sciences, Via Piandanna 4, 07100, Sassari, Italy. [3]NBFC, National Biodiversity Future Center, Palermo, Italy. [4]University of Perugia, Department of Agricultural, Food and Environmental Sciences, Borgo XX giugno 74, I-06121, Perugia, Italy. [5]University of Siena, Department of Life Sciences, Via Mattioli, 4, 53100, Siena, Italy. [6]Università degli Studi di Milano, Department of Biosciences, Via Celoria 26, 20133, Milano, Italy. ✉e-mail: gdilorenzo.research@gmail.com

| MACRO CATEGORY | INTERPRETATION MANUAL OF EUROPEAN UNION HABITATS NAME | ANNEX I CODE | EXTENDED HABITAT NAME ACCORDING TO ANNEX I |
|---|---|---|---|
| Mediterranean Coastal Dunes | Sea dunes of the Atlantic, North Sea and Baltic coasts | 2110 | Embryonic shifting dunes |
| Mediterranean Coastal Dunes | Sea dunes of the Atlantic, North Sea and Baltic coasts | 2120 | Shifting dunes along the shoreline with *Ammophila arenaria* (white dunes) |
| Grasslands | Semi-natural dry grasslands and scrubland facies | 6210* | Semi-natural dry grasslands and scrubland facies on calcareous substrates (Festuco-Brometalia) (*important orchid sites) |
| Alpine Screes | Scree | 8110 | Siliceous scree of the montane to snow levels (*Androsacetaliaalpinae* and *Galeopsetalia ladani*) |
| Alpine Screes | Scree | 8120 | Calcareous and calcshist screes of the montane to alpine levels (*Thlaspietearotundifolii*) |
| Mediterranean Deciduous Forests | Mediterranean deciduous forests | 9210* | Apennine beech forests with *Taxus* and *Ilex* |

**Table 1.** For each of the six habitats are noted the macro category used in the work, the name referring to "Interpretation manual of European Union Habitats[81] (version EUR 28)", the Annex I of the Habitat Directive[2] code, and the extended habitat name according to Annex I.

disturbances, further increase the vulnerability of ecosystems[8]. Moreover, invasive alien species[9], coupled with the impacts of climate change[10,11], including rising temperatures, droughts, and extreme weather events, pose significant challenges to the preservation of these environments[12–15].

Given the circumstances, it is imperative to closely observe the habitats of N2000N to the maximum extent possible, despite the often insufficient[16,17] economic resources that would be necessary to monitor using traditional means. The Natural Intelligence[18] (NI) (https://www.nih2020.eu/) project aims to enhance human monitoring capabilities by integrating robotic technologies[19]. This is achieved by leveraging legged robots and Artificial Intelligence (AI) algorithms to gather data in natural habitats, effectively replicating the survey methods typically carried out by human scientists. Employing such robots represents a significant advancement in optimizing habitat monitoring efforts while providing uniform and standardized methods[20]. In general, the monitoring task necessitates the ability to identify objects of interest in the environment, making it natural to utilize AI[21] for object detection[22] or segmentation[23] based on the specific scenario. The subject of this data descriptor is the dataset developed within the NI framework to enable training of AI algorithms capable of processing data acquired by autonomous legged robots and providing insights into the characterizing species of critical habitats. This data descriptor introduces a novel labeled dataset created for the identification of plant species in six Annex I habitat types found within the N2000N, included in the macro categories Mediterranean Coastal Dunes, Grasslands, Alpine Screes, and Mediterranean Deciduous Forests, as per Table 1. To the Author's best knowledge, there are no other labeled datasets suitable for object detection for those habitats or acquired with a ground robot. Nevertheless, several related datasets have been developed for vegetation monitoring tasks. The vast majority of these datasets pertain to precision agriculture[24], focusing on the detection of plant diseases and weed classification[25–27], with data acquired mostly through unmanned aerial vehicles[28].

Each habitat comprises characteristic species that embody that habitat type's features and functions called typical species (TS). Other species, including native species (NS), invasive alien species (IAS), and early warning species (EWS), may also suggest a poor conservation status of the habitat and that it is the object of pressure and deviated from its optimal condition. Consequently, it is crucial to identify all these groups of plant species and their associated coverage. The primary objective of this dataset is to make data available for public use in developing and training new object detection algorithms.

In general, the monitoring task necessitates the ability to identify objects of interest in the environment, making it natural to utilize AI[21] for object detection[22] or segmentation[23] based on the specific scenario.

Our team of plant scientists and robotic engineers successfully collected images and videos of the selected plant species across four different macro categories in Italy, as depicted in Fig. 1, utilizing the robot's cameras while in teleoperated or autonomous missions and by human operators with multiple hardware. These diverse habitats included the Mediterranean coastal dunes of Sardinia[29], the grasslands of central Apennine[30], the rocky slopes of Stelvio National Park[31], and the beech forests of Tuscany[32]. While previous data descriptors[29–32] were each associated with a single habitat and included a variety of data obtained via legged robots within that habitat, this dataset encompasses all four habitats and provides only images acquired by both robots and human operators, along with the corresponding labels. Consequently, although there may be an overlap on a few raw images, the novelty of this dataset resides in its extensive expert-validated YOLO annotations, which facilitate direct implementation for object detection tasks.

The necessary data was retrieved, and labeling was performed by a group of plant scientists who were experts in the respective habitats. Habitat monitoring requires differentiating and identifying individual plants, which can be achieved through the application of image segmentation or object detection algorithms, rather than image classification methods. While image segmentation necessitates a time-consuming and extensive labeling phase, object detection offers the optimal balance between the simplicity of the labeling phase and the comprehensiveness of the results. After data collection, domain experts annotated all of the human- and robot-collected images using bounding box annotations in YOLOtxt format.

This data descriptor presents the inaugural dataset for computer vision focusing on the designated pilot habitats for the N2000N. The dataset emphasizes the significance of interdisciplinary collaboration in habitat

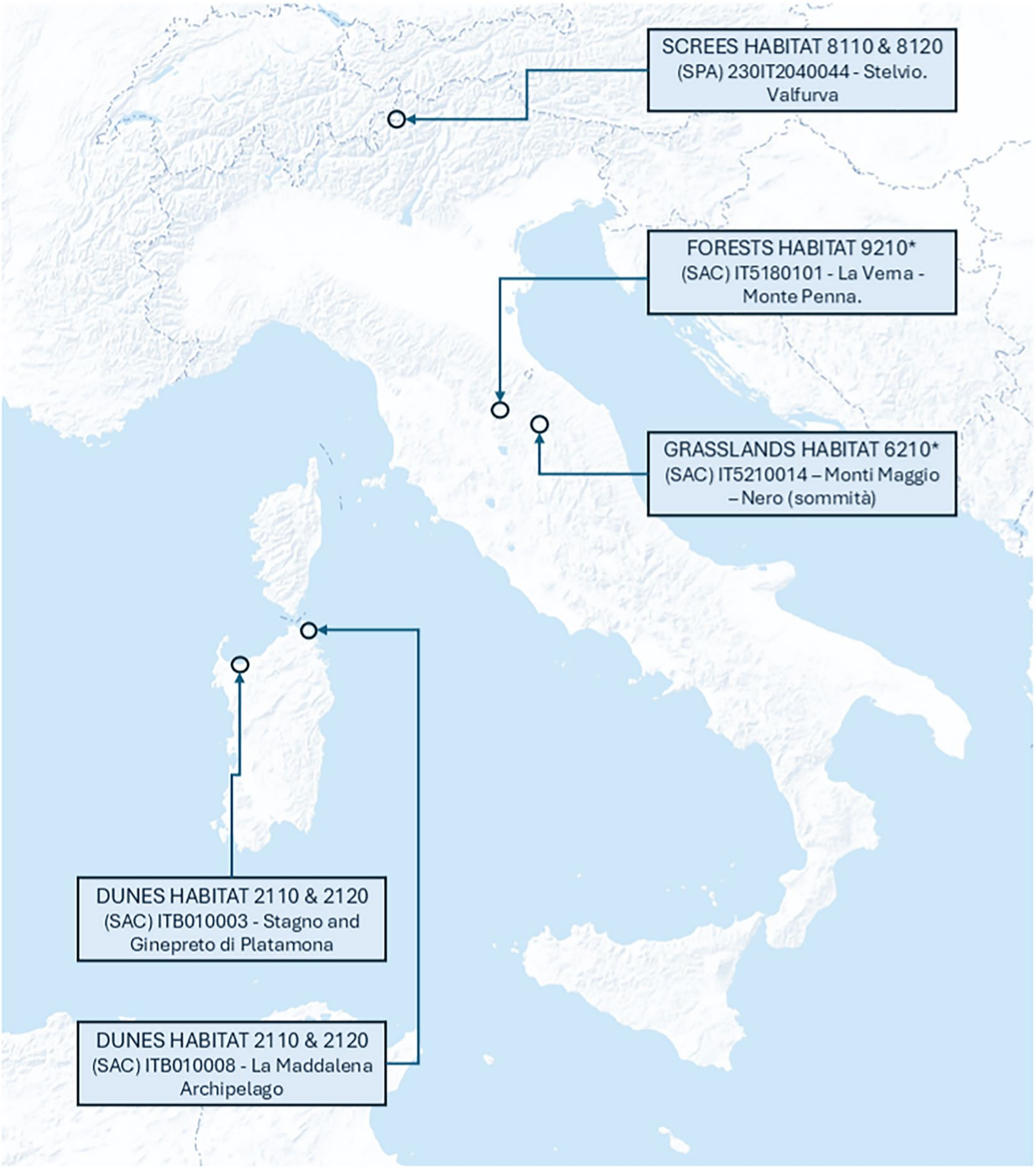

SCREES HABITAT 8110 & 8120
(SPA) 230IT2040044 - Stelvio.
Valfurva
FORESTS HABITAT 9210*
(SAC) IT5180101 - La Vema -
Monte Penna.
GRASSLANDS HABITAT 6210*
(SAC) IT5210014 – Monti Maggio
– Nero (sommità)
DUNES HABITAT 2110 & 2120
(SAC) ITB010003 - Stagno and
Ginepreto di Platamona
DUNES HABITAT 2110 & 2120
(SAC) ITB010008 - La Maddalena
Archipelago

**Fig. 1** A map of Italy with displayed the habitats and the sites covered during the field missions where the vast majority of data presented in this data descriptor was acquired.

monitoring and conservation, illustrating how the combination of robotics and plant science can result in more comprehensive and efficient environmental data collection. Each dataset adds to an expanding catalog of resources intended to enhance the tools and methodologies available for ecological research and habitat conservation.

## Methods

The four target macro categories (Mediterranean coastal dunes, grasslands, Alpine screes, and Mediterranean deciduous forests) considered in this work consist of six different habitat types listed in Annex I of the Habitats Directive. To adhere to Annex I of the Habitat Directive[2], habitats are identified by a four-digit code that includes the first digit, which represents the type of the environment, along with the subsequent digits that provide more specific information. In this dataset, we consider habitats 2110, 2120, 6210*, 8110, 8120, and 9210*. In the six habitats considered, 2110 refers to Embryonic shifting dunes, 2120 refers to Shifting dunes along the shoreline with *Ammophila arenaria* (white dune), 6210* refers to Semi-natural dry grasslands and scrubland facies on calcareous substrates (*Festuco-Brometalia*) (*important orchid sites), 8110 refers to Siliceous scree of the montane to snow levels (*Androsacetalia alpinae* and *Galeopsetalia ladani*), 8120 refers to Calcareous and calcshist screes of the montane to alpine levels (*Thlaspietea rotundifolii*), and 9210* refers to Apennine beech forests with *Taxus* and *Ilex*. Table 1 synthesizes the information relative to each habitat.

| MEDITERRANEAN COASTAL DUNES HABITAT (2110–2120) | | | |
|---|---|---|---|
| **Class** | **Species** | **Ecological role** | **Annotations** |
| all | all | — | 24378 |
| Achillea maritima | *Achillea maritima* | Typical species | 2548 |
| Calamagrotis arenaria | *Ammophila arenaria* subsp. *arundinacea* | Typical species | 473 |
| Carpobrotus acinaciformis | *Carpobrotus acinaciformis* | Invasive alien species | 5641 |
| Eryngium maritimum | *Eryngium maritimum* | Native species | 4407 |
| Pancratium maritimum | *Pancratium maritimum* | Native species | 6626 |
| Thinopyrum junceum | *Thinopyrum junceum* | Typical species | 4683 |

| GRASSLANDS HABITAT (6210*) | | | |
|---|---|---|---|
| **Class** | **Species** | **Ecological role** | **Annotations** |
| all | all | — | 11397 |
| Asphodelus macrocarpus | *Asphodelus macrocarpus* | Early Warning species | 5945 |
| Dactylorhiza sambucina | *Dactylorhiza sambucina* | Typical species | 2761 |
| Anacamptis morio | *Anacamptis morio* | Typical species | 2691 |

| ALPINE SCREES HABITAT (8110-8120) | | | |
|---|---|---|---|
| **Class** | **Species** | **Ecological role** | **Annotations** |
| all | all | — | 11988 |
| Cerastium spp. | *Cerastium uniflorum* and *Cerastium pedunculatum* | Typical species | 2116 |
| Luzula alpino-pilosa | *Luzula alpinopilosa* | Early Warning species | 1210 |
| Saxifraga | *Saxifraga bryoides* | Typical species | 2258 |
| Ranunculus glacialis | *Ranunculus glacialis* | Typical species | 2432 |
| Geum reptans | *Geum reptans* | Typical species | 2576 |
| Papaver alpinum | *Papaver alpinum* | Typical species | 1396 |

| MEDITERRANEAN DECIDUOUS FORESTS HABITAT (9210*) | | | |
|---|---|---|---|
| **Class** | **Species** | **Ecological role** | **Annotations** |
| all | all | — | 13475 |
| Anemonoides nemorosa | *Anemonoides nemorosa* | Typical species | 2882 |
| Corydalis cava | *Corydalis cava* | Typical species | 1182 |
| Doronicum columnae | *Doronicum columnae* | Early Warning species | 2867 |
| Anemonoides ranunculoides | *Anemonoides ranunculoides* | Typical species | 6544 |

**Table 2.** For each of the four macro categories are recapped the classes (species) used in the annotations along with their ecological role and the relative number of annotations.

Plant species with a different ecological role were selected for each macro category. Typical species (TS) serve as an indicator for the "structure and functions" metrics which determine how successfully the habitats, where NS live, are being conserved. TS must be selected according to the monitoring manuals in a manner that reflects the favorable structure and functions of the habitat type. This means that TS should simultaneously be an effective indicator of the favorable habitat quality, unique to the habitat or present across a significant portion of the habitat range, and sensitive to changes in the habitat condition. EWS are plant species that indicate the ongoing processes that alter the structure and function of their habitat. This concept is particularly applicable to non-typical taxa, whose presence can serve as an informative proxy for detecting the initial community functional shifts[33]. This also took into account IAS, as they pose a significant risk to biodiversity and worsen the conservation status of habitats that are important to the European Community.[9]. Table 2 recaps the macro categories, the species selected, and their ecological role, while details are given in the subsequent habitat-specific methods sections.

While standardizing conditions (e.g., controlled lighting and fixed camera settings) might yield a more homogeneous dataset with potentially higher performance on in-laboratory tests, such uniformity would fail to capture the variability inherent in real-world monitoring scenarios. Our dataset intentionally encompasses a wide range of environmental conditions, including varying lighting, exposure, angles, and focus. To ensure this we used autonomous robotic missions, teleoperated operations, and human-operated cameras with different hardware, ensuring that artificial intelligence algorithms trained on this dataset can generalize robustly across unpredictable field conditions.

**Employed Hardware and Procedures.** The vast majority of data was acquired in the spring and summer of 2022 and 2023 (starting April 27th, 2022, and ending July 15th, 2023) when fieldwork was carried out to

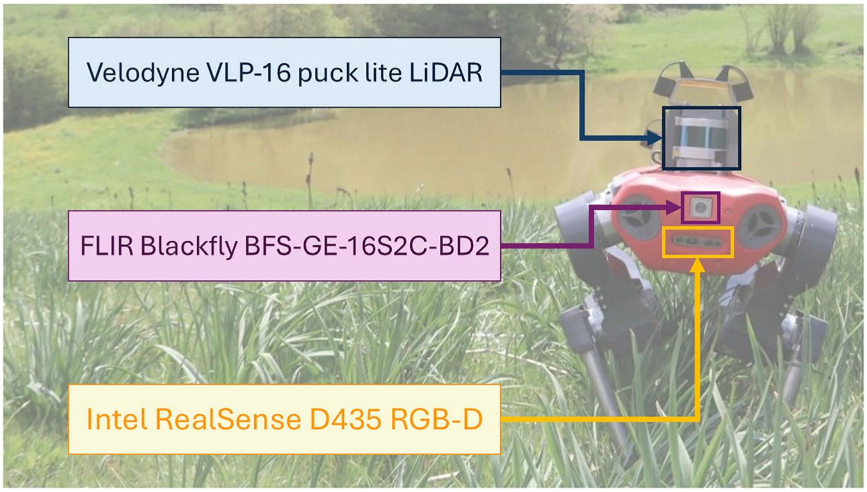

Velodyne VLP-16 puck lite LiDAR
FLIR Blackfly BFS-GE-16S2C-BD2
Intel RealSense D435 RGB-D

**Fig. 2** The robot ANYmal C operating in Grasslands environment, with a clear display of the sensor providing it with environmental perception capabilities.

verify the occurrence and distribution of the selected target species in the study areas and to carry out explorative vegetation plots. Moreover, we collected a large quantity of pictures and videos of the plants in different phenological stages to later be labeled. The guidelines followed were in general the same for each habitat as per the Data Acquisition Pipeline subsection in the Technical Validation section. The procedure was generally different for each macro category and is detailed in the relative Method sections. The species identified as indicators differ for each habitat, as differ the procedures of data acquisition that need to mimic the ones effectuated by the human operators, and the fieldwork periods also differ as they depend on the directive and the lifecycle of the chosen species. Because of this, the method utilized differs between each habitat and has been described in a distinct method subsection for Mediterranean Dunes, Grasslands, Alpine Screes, and Mediterranean Deciduous Forests. The raw data used in the work presented in this data descriptor were acquired following three different procedures: robot in autonomous mission for around 7.6% of the images, teleoperated robot for around 27.9% of the images, and by human operators with different cameras for the remaining 64,5% of the images. This difference in amounts is related to the technical difficulty of the data acquisition: the more difficult the procedure, the less time we were able to operate it. While autonomous missions yield systematic, grid-based images with relatively uniform angles, teleoperated and human-operated images inherently introduce greater diversity in viewpoints and exposure conditions. This variability, rather than being detrimental, enhances the dataset's robustness by more accurately simulating real-world conditions and enabling artificial intelligence models to generalize more effectively. Furthermore, the expert annotation process and subsequent quality assurance protocols helped normalize any differences in image quality or species representation. Needs to be noted that the acquisition of the raw data in an autonomous fashion is more time-effective, but we chose to not label most of the data acquired in this stage of the project willing to use it as a benchmark for the performance of the framework and of the models.

*Data acquired employing robot in autonomous mission.* Robot's autonomous missions are conducted to mimic the procedures used for habitat monitoring carried out by human operators. First, for data acquisition, we selected sites accessible by both human operators and robots. The legged robot used in this work is ANYmal C[34], which was developed by ANYbotics AG based in Zurich, Switzerland. This robot possesses dimensions of 1.05 m by 0.52 m and weighs 50 kilograms. It is capable of performing various movements, including hip abduction and adduction, hip flexion and extension, and knee flexion and extension, due to its three actuated joints per leg. ANYmal C can be operated wirelessly through a dedicated remote controller or a PC, or it can operate autonomously in unstructured environments, such as dunes. It is powered by a 932.4 Wh lithium-ion battery, which provides an operational window of 2–4 hours on a single charge. The robot is also equipped with an array of exteroceptive sensors, including a Velodyne VLP-16 puck lite LiDAR, four depth cameras Intel RealSense D435 RGB-D, and two wide-angle cameras FLIR Blackfly BFS-GE-16S2C-BD2, which provide the robot with environmental perception capabilities. The LiDAR sensor uses laser pulses to create a detailed three-dimensional representation of the surroundings, while the depth cameras use stereo vision to capture both color images and range information as shown in Fig. 2.

During the mapping phase, the robot is teleoperated in the area following a non-predefined trajectory, reconstructing a three-dimensional map of the surrounding environment. Next, the robot performs an autonomous mission, it locates itself in the previously created operative space scenario map and it moves following the waypoints, that vary for each habitat according to the different procedures. During the entire autonomous mission, videos are acquired with the onboard cameras, and data relating to the robot's status is recorded. At each waypoint, the robot oriented itself consistently before taking four photographs utilizing its four RGB-D cameras. In the presence of obstacles, the robot was teleoperated for security concerns: otherwise, it moved autonomously.

*Data acquired employing robot in teleoperated mission.* The robot's remote-controlled missions were conducted for the sole purpose of acquiring data on the target species to be labeled. Consequently, spaces within the study area corresponding to each habitat were selected where it was not feasible to replicate the plot carried out by botanists as traditional monitoring via an autonomous mission but where at least one target species was present. Indeed, the data acquired via an autonomous mission were utilized for benchmarking the quality of the framework and the associated artificial intelligence models. Therefore, it was essential to ensure that the data acquired via a teleoperated robot were obtained in a location distinct from the one where the autonomous missions were conducted, in order to avoid positive bias. In this procedure, the robot is teleoperated by an operator in a manner that enables the capture of photographs of the species identified by plant scientists by varying position, focal distance, number of individuals, and angle within the image plane.

*Data acquired employing human operators with multiple hardware.* Regarding the data acquired by a human operator, different hardware was employed, including photographic cameras (OLYMPUS E-M5, SONY ILCE-6000, NIKON D7200, NIKON D300, NIKON COOLPIX S3300, CANON EOS 80D) and smartphone cameras (HUAWEI FIG-LX1, SAMSUNG SM-J250Y, SAMSUNG SM-A536B, XIAOMI 21061119DG, XIAOMI 22101316G, APPLE iPhone SE). This procedure focused on maximizing the generalization capacity of the dataset, beginning with the utilization of diverse hardware for the acquisition of raw data, while also considering the diversification of various environmental conditions of the photographs (e.g., weather, space, time, exposure, brightness, number of individuals).

**Mediterranean Coastal Dunes: methods for Annex I habitat 2110 and 2120.** The data were primarily collected in two sites located in North Sardinia (Italy). The first one is the dunal system located inside the Special Areas of Conservation (SAC) ITB010003 - Stagno and Ginepreto di Platamona. It is a 17-kilometer-long sandy beach with longitudinal and parabolic dunes aligned north-west/southeast. The second one, spiaggia del Relitto, is located in the La Maddalena Archipelago which is a SAC (ITB010008) and a National Park. A series of communities that extend along parallel transects from the shoreline to the most stable regions behind the dune characterize the vegetation of the emergent portion of the beach in both locations. The morphology of the dunes varies depending on the exposure to wind; the dunes in Platamona have varying heights and forms, whereas the dunes in Spiaggia del Relitto are flat.

Fieldwork was conducted over the course of four days during the first campaign, from May 16th to May 19th, 2022, and over the course of seven days during the second campaign, from May 16th to May 20th, 2023. All the fieldwork was conducted in Platamona excluding one day of the latter campaign spent at the Spiaggia del Relitto. Those periods have been chosen as they represent the ideal conditions for sampling dune vegetation[35].

The zonation of coastal dunes is shaped by a combination of ecological factors such as salt spray, burial, stability, and nutrient availability. Almost all types of dune vegetation are recognized as habitats of community interest in the European Union under Annex I of the Habitat Directive. We considered the habitats 2110 and 2120.

*Habitat 2110 - Embryonic shifting dunes.* This habitat represents the initial stages of dune formation, found as a seaward fringe at the base of taller dunes or as ripples and elevated sand surfaces on the upper beach. We considered the two TS *Thinopyrum junceum* (L.) Á.Löve and *Achillea maritima* (L.) Ehrend. & Y.P.Guo,. Moreover, we considered other two NS, *Eryngium maritimum* L. and *Pancratium maritimum* L.

*Habitat 2120 - Shifting dunes along the shoreline with Ammophila arenaria (white dunes).* These are mobile dunes forming the seaward cordon or the cordons within coastal dune systems. In this habitat, we considered the TS, the geophyte *Ammophila arenaria* (L.) Link subsp. *arundinacea* (Husn.) H. Lindb., a perennial plant with both horizontal and vertical rhizomes that can reach up to 120 cm which plays a crucial role in the formation and maintenance of dunes. Other species commonly found in this habitat include *Eryngium maritimum* and *Pancratium maritimum*.

We also considered the presence of the IAS *Carpobrotus acinaciformis* L. L.Bolus[36,37], which poses a significant threat to biodiversity and negatively impacts the conservation status of these habitats. For reference, see Table 2.

TS and NS living in Mediterranean coastal dunes are psammophilous species that are adapted to live on sand in extreme conditions determined by high salinity, wind exposure, and scarcity of water and nutrients. They have rigid, hairy leaves with thickened cuticles and extensive, highly branched root systems resistant to mechanical pull[38]. The IAS *C. acinaciformis* has a facultative C3-CAM photosynthetic strategy, high phenotypic plasticity, resistance to high salinity concentrations, and intense vegetative clonality. It shows strong adaptability and dispersal ability in sandy dunes[39].

Vegetation surveys are conducted within adjacent 1m × 1m plots along a transect perpendicular to the coastline, documenting the changes in vegetation across environmental gradients for each of the two habitats. During its autonomous mission, the robot followed the linear transect, halting every meter to capture images and assess each contiguous 1 $m^2$ plot.

**Grasslands: methods for Annex I habitat 6210*.** The data were primarily collected in the Valsorda area, in central Italy, on the Apenninic mountain ridge at around 1000 m a.s.l. within the municipality of Gualdo Tadino in the province of Perugia, Italy. This location is part of the N2000N and is specifically designated as the Special Area of Conservation (SAC) IT5210014 - Monti Maggio - Nero (sommità). The area comprises a mosaic of grasslands and sparse beech forests. The vegetation in Valsorda is composed of diverse communities that are adapted to the varying altitudes and slopes of the terrain. Grassland habitats dominate the mountain tops, particularly Annex I habitat 6210*, which includes semi-natural dry grasslands and scrubland on calcareous substrates. These grasslands are interspersed with patches of beech forests (habitat 9210*), which are more prevalent on cooler, north-facing slopes. The area's geomorphology is shaped by its rugged terrain, with sporadic steep slopes and rocky outcrops.

The orchid species play a particularly important role in this habitat type, and the timing of their flowering is crucial for monitoring their occurrence. Because of this, data collection was carried out in May, which is the ideal period for detecting both the target TS and EWS species in the field in habitat 6210*, as they typically flower between April and June[40]. Also, during the early stages of their life cycle, the orchid species are difficult to distinguish, while at the time of flowering, they become easier to recognize. Fieldwork was conducted from May 10th to May 13th, 2022, during the first campaign, and from May 8th to May 10th, 2023, during the second campaign.

**Habitat 6210* - Semi-natural dry grasslands and scrubland facies on calcareous substrates (Festuco-Brometalia) (* important orchid sites).** This habitat is typically found on calcareous soils in open, sunny locations, often on slopes and hillsides, and is characterized by species-rich semi-natural grasslands. The vegetation includes a variety of grasses and herbaceous plants that are well-adapted to dry conditions. These grasslands are known for their high biodiversity and often support a range of orchids, including *Anacamptis morio* (L.) R.M.Bateman, Pridgeon & M.W.Chase and *Dactyloriza sambucina* (L.) Soó. The habitat also provides important ecological functions, such as supporting pollinators and serving as a refuge for many species of invertebrates as well as feeding areas for herbivorous mammals. The diversity and structure of the vegetation can vary depending on factors such as grazing pressure, which influences the presence and abundance of species within the habitat.

There are no generally valid lists of Typical Species (TS) at the European or national level because of the variability in grassland flora. Official lists of species are included in reference physiognomic combinations at national level[41], but they are suitable for habitat identification, not for habitat monitoring. It is advised to identify TS at a regional or even local scale[42]. The presence of orchid species is typically regarded as an indicator of favorable conservation status, and their abundance within a surveyed area is particularly significant as an indicator of the priority status of habitat 6210*[2]. In accordance with the guidelines and the recommendation to select TS at the local level for habitat 6210*, we chose two typical orchid species, *Anacamptis morio* (L.) R.M.Bateman, Pridgeon & M.W.Chase and *Dactylorhiza sambucina* (L.) Soó (in both its yellow and pink forms), which are among the most commonly found orchids in habitat 6210* at the regional level. *A. morio* grows in different environments, predominantly in unimproved dry grasslands, often in large populations, in full sunlight or partial shade. *D. sambucina* mainly occurs in meadows and pastures on the mountain range, though it can occasionally be encountered in clearings or bright woodlands[43]. Both species are rather indifferent to the substrate type; *A. morio* generally prefers nutrient-poor soils, while *D. sambucina* the moister, nutrient-rich ones[44]. As Early Warning Species (EWS) we selected *Asphodelus macrocarpus* Parl. This tall, rhizomatous geophyte is known for its vigorous vegetative growth in the spring, typically spreading from forest edges into semi-natural grassland habitats, in the Apenninic areas. The presence and impacts of this species on habitat 6210* in the central Apennines are well-documented, particularly in areas where traditional agropastoral activities have been reduced or abandoned. *A. macrocarpus* colonizes grasslands through direct invasion and rapid expansion, facilitated by its heliophilic nature and efficient vegetative propagation[44–47]. The invasive characteristics of this species are further intensified by its unpalatability to grazing animals[47–49]. In its native range, *A. macrocarpus*grows in grasslands, woodlands, and open shrub areas, on well structured soils[46].

In grassland monitoring, the surveyed area typically consists of a 4 m × 4 m plot or a transect of varying length. The objective during this phase is to gather habitat data using the four onboard cameras. Navigation is achieved through a series of waypoints that the robot autonomously determines in real-time: upon reaching a waypoint, the robot sets the next one at a distance of 1 m. This approach is chosen for two main reasons. Firstly, botanists generally divide the area into multiple 1 × 1 $m^2$ blocks for detailed surveys. Secondly, not pre-selecting the waypoints allows the robot to adjust to any variations in terrain slope.

**Alpine Screes: methods for Annex I habitat 8110 and 8120.** The data were primarily collected in the Valfurva area, situated within the Stelvio National Park, in the province of Sondrio, Italy. This location is part of the N2000N and is specifically designated as the Special Protection Area (SPA) IT2040044 - Stelvio. Valfurva is characterized by its alpine landscape, which is located within the high altitudes of the Italian Alps. The area comprises a combination of rocky screes and sparse vegetation, with elevations ranging from montane to alpine zones. The scree habitats in this region include habitats 8110 and 8120. These habitats are found on steep, rocky slopes, where the substrate is composed of loose rock debris, either siliceous or calcareous. The vegetation is sparse due to the unstable and nutrient-poor conditions, but it includes specialized plant communities that are adapted to the harsh environmental conditions of these high-altitude areas. These scree habitats, classified as unfavorable-inadequate (U1) in the fourth EU report, are critical for the conservation of specialized alpine flora. The microclimate of Valfurva, influenced by its elevation and exposure, fosters a unique combination of species that vary significantly depending on altitude, rock type, and exposure to sunlight and wind.

Fieldwork was conducted from July 19th to July 21st, 2022, during the first campaign, and from July 10th to July 15th, 2023, during the second campaign. The days selected were determined because in high-altitude habitats, this period is ideal for observing the phenological stages of vascular plants, as it coincides with their peak blooming and development.

*Habitat 8110 - Siliceous scree of the montane to snow levels (Androsacetalia alpinae and Galeopsetalia ladani).* This habitat consists of loose rock debris found on steep slopes at elevations ranging from montane to alpine levels. The scree is primarily composed of siliceous rock, and the vegetation is sparse due to the unstable substrate. Typical species in this habitat include pioneer plants that are adapted to the shifting conditions. These species often have deep root systems to colonize the loose, rocky environment. Notable species include *Geum reptans* and *Ranunculus glacialis*, both of which are well-adapted to the high altitudes and poor soil conditions. The habitat prevents soil erosion and acts as a refuge for specialized alpine flora and fauna.

*Habitat 8120 - Calcareous and calcshist screes of the montane to alpine levels (Thlaspietea rotundifolii).* This habitat is found on slopes with calcareous or calcshist scree, typically at montane to alpine elevations. The substrate consists of loose, unstable rock fragments, which support specialized vegetation. The flora is characterized by its ability to stabilize the scree, reducing the movement of rocks and contributing to the gradual development of soil. These habitat is important for the conservation of alpine plant species that are adapted to extreme environmental conditions and for maintaining biodiversity.

The considered TS for the Screes habitats are Cerastium spp. (which includes *Cerastium uniflorum* Clairv. and *Cerastium pedunculatum* Gaudin), *Geum reptans* L., *Papaver alpinum* L., *Ranunculus glacialis* L., and *Saxifraga bryoides* L. Conversely, the only considered EWS is *Luzula alpinopilosa* (Chaix) Breistr.

All these species share the need for cold, high-altitude environment, with intense light and low nutrient content. *Papaver alpinum*, being linked to calcareous screes (habitat 8120) grows on soils with high pH, while all the other species need acidic soils typical of habitat 8110[50]. All the TS are strictly linked to the coarse soil characteristic of active scree slopes, while *Luzula alpinopilosa*, the only EWS, shows a preference for substrates rich in fine fraction[50], which testify an incoming stabilization of the scree slope.

In screes settings, the surveyed area typically consists of a 5 m × 5 m plot[51]. The objective during this phase is to gather habitat data using the four onboard cameras. Navigation is achieved through a series of waypoints that the robot autonomously determines in real-time: upon reaching a waypoint, the robot sets the next one at a distance of 1 m. This approach is chosen for two main reasons. Firstly, botanists generally divide the area into multiple 1 × 1 $m^2$ blocks for detailed surveys. Secondly, not pre-selecting the waypoints allows the robot to adjust to any variations in debrits size.

**Mediterranean Deciduous Forests: methods for Annex I habitat 9210*.** The data were primarily collected in the La Verna forest, located within the "Foreste Casentinesi" National Park, and part of the N2000N as the Special Area of Conservation (SAC) IT5180101 - "La Verna - Monte Penna." This area is situated in the Apennine region of Italy and is characterized by its montane landscape, which includes dense beech forests. The forested area is primarily composed of habitat 9210*, known as Apennine beech forests with *Taxus* and *Ilex*, which are particularly significant due to their high species diversity and the presence of taxa with important conservation value. The vegetation in La Verna forest is dominated by beech (*Fagus sylvatica*) trees, with a well-developed understory that includes sporadic shrubby individuals of *Abies alba* and an almost continuous herbaceous layer rich in species. The fieldwork for this study was timed to coincide with the flowering season of target nemoral understory species, which serve as indicators of the conservation status of habitat 9210*, to ensure that the most relevant phenological stages of the indicator species were captured.

Fieldwork was conducted from April 27th to April 28th, 2022, during the first campaign, and from May 2nd to May 6th, 2023, during the second campaign. The days selected were determined by taking into account the examined indicator species' blooming season, which runs from March to May.

*Habitat 9210* - Apennine beech forests with Taxus and Ilex.* This habitat is of prioritary interest according to the habitat Directive and includes termophilous beech forest, which occur in the submontane belt and show regression into the montane belt whit an oceanic bioclimate (from meso to supratemperate) rich in spring flowering geophytes. The vegetation is dominated by beech (*Fagus sylvatica* L. subsp. *sylvatica*) trees, forming dense forests with a well-developed understory. Characteristic species include *Taxus baccata* L. and *Ilex aquifolium* L. These forests are significant for their ecological functions, including soil stabilization, carbon sequestration, and providing habitat for a diverse range of flora and fauna. The structure and composition of the forest can vary depending on factors such as altitude, soil type, and forest management practices, which influence the presence and abundance of typical and associated species within the habitat[52].

In the forest habitat 9210*, we concentrated on four species: *Anemonoides nemorosa* (L.) Holub, *Corydalis cava* (L.) Schweigg. & Körte, and *Anemonoides ranunculoides* (L.) Holub as Typical Species (TS), and *Doronicum columnae* Ten. as an Early Warning Species (EWS).

The TS belonging to the genera Anemonoides and Corydalis are vernal sciaphilous geophytes that share a high tolerance for dense summer shade and a preference for alkaline to slightly acidic substrates, typical of habitat 9210*[53]. Their presence is closely associated with unmanaged and ancient forests, where stable shade conditions support their growth and persistence[54]. In contrast, *Doronicum columnae*, the only EWS, tolerates light in the understory and also contributes to the formation of megaforb-rich forest edges[55]. Its presence may then indicate canopy openings, which can result from anthropogenic disturbances in the habitat.

For the study, two different kind of plots were selected. Smaller plots were chosen in areas free from obstacles, with flat, even terrain, devoid of trees, to conduct preliminary floristic surveys that served as a testbed for evaluating the feasibility of robot deployment. Subsequently, larger plots were designated for comprehensive structural surveys and also to gather data for the floristic survey, as of interest of this data descriptor. The monitoring procedure was designed to replicate typical botanist fieldwork, with circular plots of approximately 200 $m^2$ (radius = 8 m) being selected, a commonly used size and shape for monitoring forest habitats[56].

For the data acquisition during the autonomous mission phase, the robot followed a predefined grid pattern to capture images and videos of the indicator species. This grid was composed of waypoints arranged in a bottom-right to top-left configuration, with each waypoint spaced 1 m apart. This methodology was chosen for two primary reasons: first, botanists typically divide the survey area into multiple 1 × 1 $m^2$ blocks for detailed analysis, and second, allowing the robot to autonomously place waypoints ensures adaptability to any variations in terrain slope.

**Data Annotation.** To the best of the Author's knowledge, labeled datasets for object detection relating to habitats 2110, 2120, 6210*, 8110, 8120, and 9210* are not available. Once the necessary data had been retrieved, a group of plant scientists who are experts in the target habitat carried out labeling. Images have been acquired through the robot's teleoperated cameras and by human operators with multiple hardware. Part of the images have also been recovered from botanical archives. The robot acquisitions have been concentrated between mid-April 2022 and late July 2023. Acquisitions carried out by operators have been distributed over a longer period and also cover different areas. Following data collection, expert botanists annotated all of the human and robot-collected images using bounding box annotations, with the aid of online tools such as Labelbox and Roboflow, and offline tools such as ModifiedOpenLabelling.

The dataset described in this data descriptor comprises four datasets for dunes, grasslands, forests, and screes. A `class` in the context of object detection refers to a distinct category or type of object that the model is trained to recognize and differentiate. In this dataset, each class represents a specific plant species relevant to the labeled environments. The complete dataset encompasses the 19 classes corresponding to TS, NS, IAS, or EWS for each habitat. It encompasses a total of 10264 images with a resolution of 640 × 640, containing 61238 annotations, yielding an average of 5.97 annotations per class. There is a significant class imbalance, which stems from the fact that the datasets were developed separately for each habitat. Some images without bounding boxes were added for each habitat acting as true negatives, constituting around 7% of the total dataset. The number of images and annotations for each habitat and the relative classes and species are the following, as per Table 2:

- **Mediterranean Coastal Dunes**: 2891 labeled images, with 24378 annotations as shown in Fig. 3 and Fig. 4, across six classes representing *Ammophila arenaria* subsp. *arundinacea*, *Thinopyrum junceum*, *Achillea maritima*, *Pancratium maritimum*, *Eryngium maritimum*, and *Carpobrotus acinaciformis*.
- **Grasslands**: 2091 labeled images, with 11397 annotations as shown in Fig. 5 and Fig. 6, across three classes representing *Asphodelus macrocapus*, *Anacamptis morio*, and *Dactylorhiza sambucina*.
- **Alpine Screes**: 2725 labeled images, with 11988 annotations as shown in Fig. 7 and Fig. 8, across six classes representing *Geum reptans*, *Ranunculus glacialis*, *Saxifraga bryoides*, *Cerastium spp*, *Papaver alpinum*, and *Luzula alpino-pilosa*.
- **Mediterranean Deciduous Forests**: 2557 labeled images, with 13475 annotations as shown in Fig. 9 and Fig. 10, across four classes representing *Anemonoides ranuncoloides*, *Anemonoides nemorosa*, *Doronicum columnae*, and *Corydalis cava*.

Due to the procedure followed, each image could include one of the relative target species, but also some images act as a true negative showing only the background. If there is a label and so at least one instance of a TS, NS, IAS, or EWS is included, also other target species instances may be present

## Data Records

The dataset is publicly available on Zenodo[57] and can be accessed at https://doi.org/10.5281/zenodo.11504938. This dataset includes labeled images of TS, NS, IS, and EWS from four distinct habitats: Dunes, Grasslands, Screes, and Forests. Another dataset, containing the raw images obtained during the 2023 campaigns, is also publicly available on Zenodo[58] and can be accessed at https://doi.org/10.5281/zenodo.15050728.

We chose the YOLOtxt format for the labels, as it is popular for object detection and in general in machine learning. The YOLOtxt format is used for annotating objects within images. It was developed in particular to train and test the object detection algorithms of the You Only Look Once (YOLO) family of models[59–62]. Upon conversion, the data can be used with any other mode. To each image in the dataset corresponds a *.txt* file that contains the annotation data that has the same name as the images, differing only for the file extension (*.txt* and not *.jpg*, in our case). In each annotation file, there is a line for each instance of an object detected within the image with the class and the bounding box coordinates and dimension for the detected object. So each *.txt* file includes one line for each object, with the following structure: `<class_id> <x_center> <y_center> <width> <height>.`

- `<class_id>`: an integer that represents the class (the specie in our case) as defined in the corresponding *.yaml* configuration file.
- `<x_center>`: the x-coordinate of the center of the bounding box, relative to the width of the image (normalized between 0 and 1).
- `<y_center>`: the y-coordinate of the center of the bounding box, relative to the height of the image (normalized between 0 and 1).
- `<width>`: the width of the bounding box, relative to the width of the image (normalized between 0 and 1).
- `<height>`: the height of the bounding box, relative to the height of the image (normalized between 0 and 1).

An example of the plotting of the bounding boxes on an original image according to the relative label file is displayed in Fig. 11.

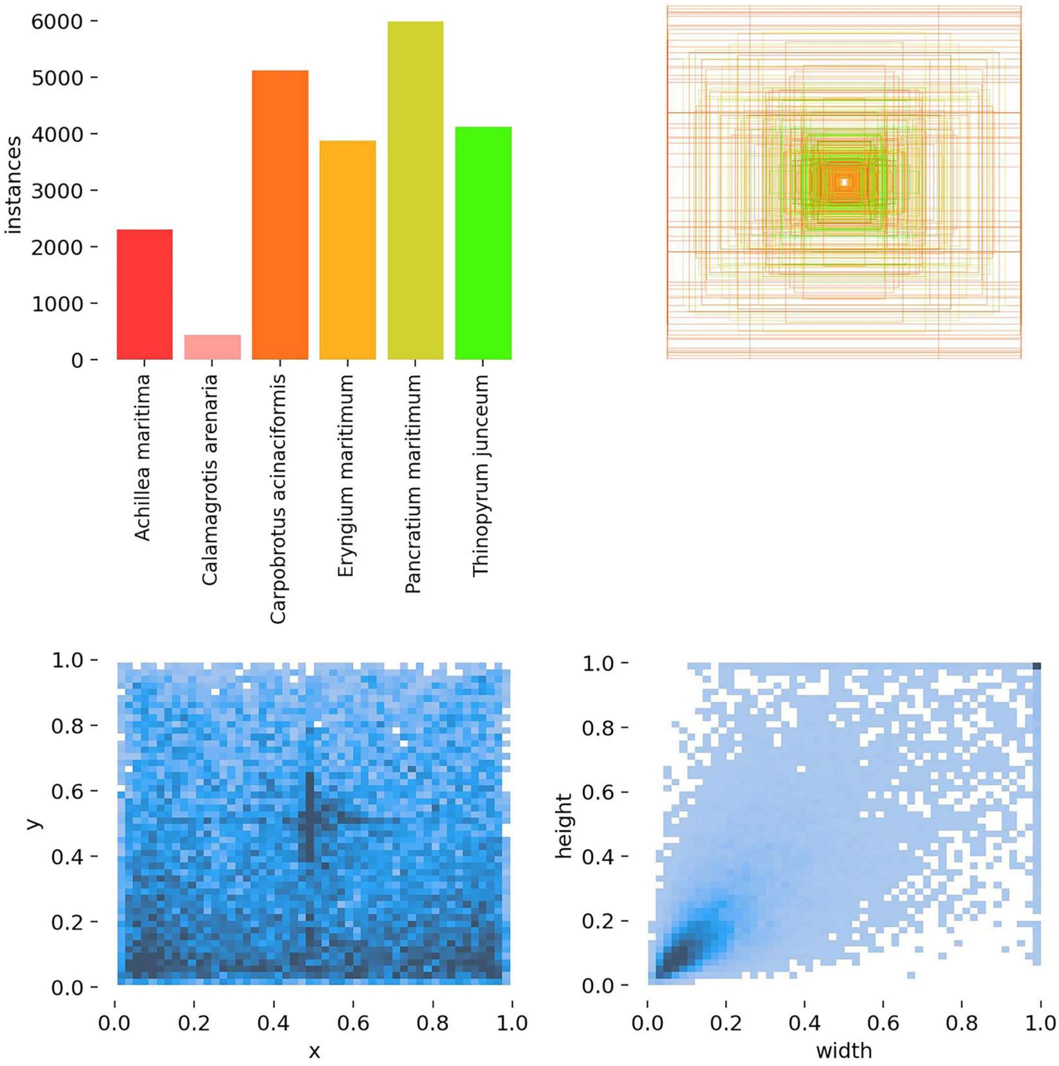


**Fig. 3** Overview of the Mediterranean Coastal Dunes dataset's labels. Top left corner, a bar chart that shows the instances per class is displayed. Top right, an overlapping triangle plot representing bounding box dimensions distribution is displayed. Bottom left, a heatmap showing the distribution of detections across the image space is presented. Bottom right, a heatmap showing the distribution of bounding box dimensions (width and height) is displayed. In the *.yaml* file, the class Calamagrotis arenaria is associated with the species *Ammophila arenaria* subsp. *arundinacea*.

The dataset is organized into a README.txt file, and four main subfolders, each corresponding to a different habitat, as per Fig. 12. Each habitat-specific subfolder contains all the necessary data in YOLOtxt format for training and validating AI models aimed at detecting and identifying species within these ecological niches. The four subfolders are named after the four categories:

- **Dunes**: Images and annotations for species in EU habitats 2110 and 2120, located in Italian dunes. Species include: *Achillea maritima* (TS), *Ammophila arenaria* subsp. *arundinacea* (TS), *Carpobrotus acinaciformis* (IAS), *Eryngium maritimum* (NS), *Pancratium maritimum* (NS), and *Thinopyrum junceum* (TS).
- **Grasslands**: Images and annotations for species in EU habitat 6210* in the Italian Central Apennines. Species include: *Asphodelus macrocarpus* (EWS), *Dactylorhiza sambucina* (TS), and *Anacamptis morio* (TS).
- **Screes**: Images and annotations for species in EU habitats 8110 and 8120 in the Italian Alps. Species include: *Cerastium spp.*(TS), *Luzula alpino-pilosa* (EWS), *Saxifraga bryoides* (TS), *Ranunculus glacialis* (TS), *Geum reptans* (TS), and *Papaver alpinum* (TS).
- **Forests**: Images and annotations for species in EU habitat 9210* in the Italian Apennines. Species include: *Anemonoides nemorosa* (TS), *Corydalis cava* (TS), *Doronicum columnae* (EWS), and *Anemonoides ranunculoides* (TS).

Each of the four habitat-specific subfolders containing the data in YOLOtxt format is organized as follows, as per Fig. 12:

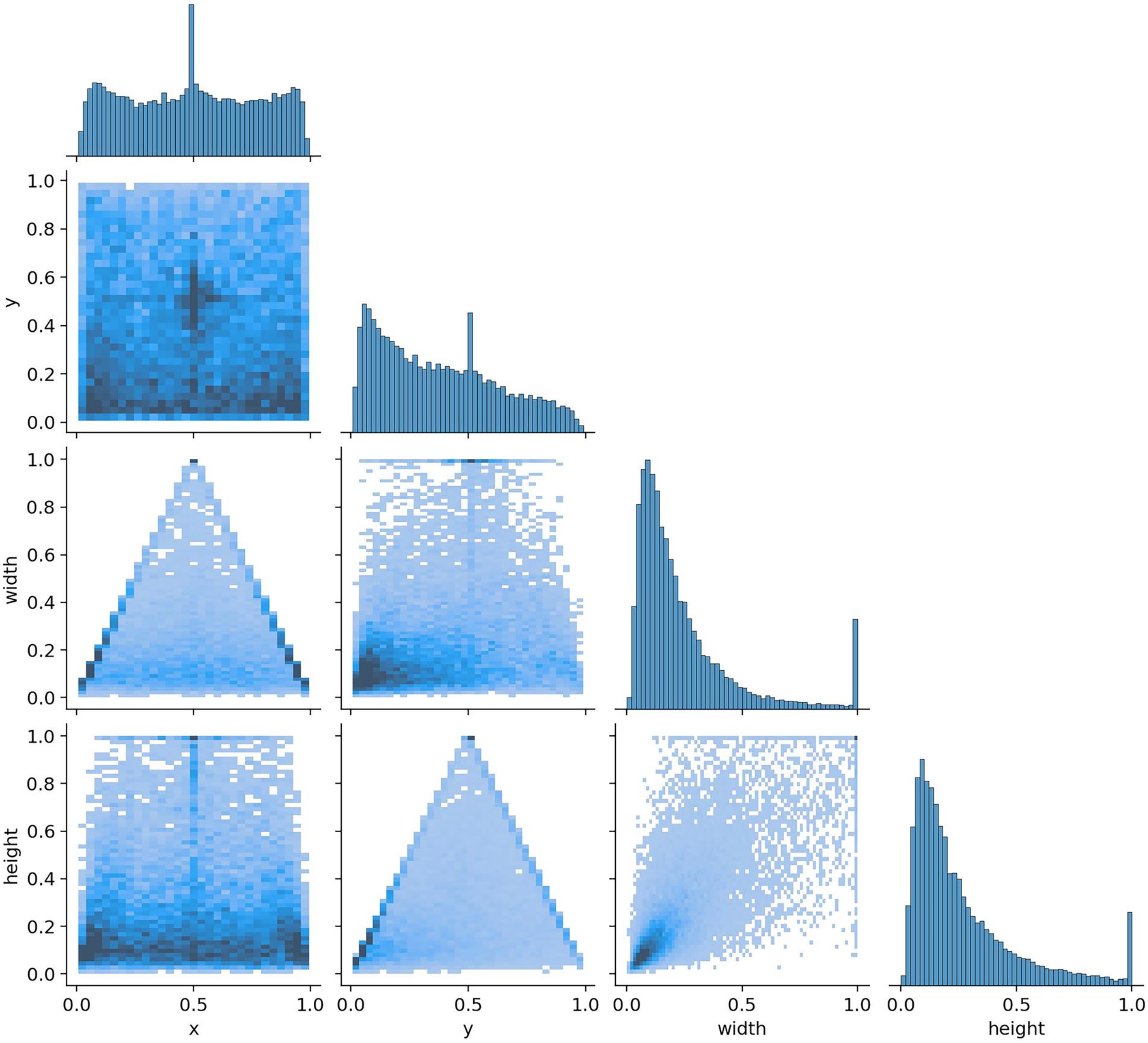


**Fig. 4** Complete correlogram of the Mediterranean Coastal Dunes dataset's labels.

- **Species data images**: This folder contains 640 × 640 images in *.jpg* format relative to the habitat. As specified by the corresponding label (a *.txt* file with the same name as the *.jpg* file) in the YOLOtxt format, it may or not be one or more instances of one or more of the classes (species) in the image.
- **Labels**: For each image, a corresponding *.txt* file is provided. The first entry in each *.txt* file identifies the species (as defined in the accompanying *.yaml* file), and the remaining entries describe the vertices of the bounding box that outlines the species in the image.
- **Configuration file**: Each subfolder contains a single *.yaml* file that lists the parameters used to describe the dataset. This file includes the path to the dataset's root directory, the path to the directory containing training, validation, or test images and labels, and the number of classes (species) included in the dataset and relative class identifier `<class_id>`. In this dataset, there are no train-validation-test splits pre-imposed, and in general the paths relative to folders should be relative to the machine.

### Data formats.

In this section are provided the technical details about the various file formats used across the dataset.

*.jpg.* The *.jpg* format is used in object detection and machine learning due to its efficient compression, which reduces file size while maintaining sufficient image quality for model training and inference. This allows for faster data loading and processing which is necessary with large datasets. Additionally, the *.jpg* format is widely supported across various platforms and tools, making it an accessible and practical choice for diverse machine learning workflows. The *.jpg* files in this dataset include images collected by the robot and by human operators using other hardware. All images have been resized to 640 × 640 pixels to maintain consistency.

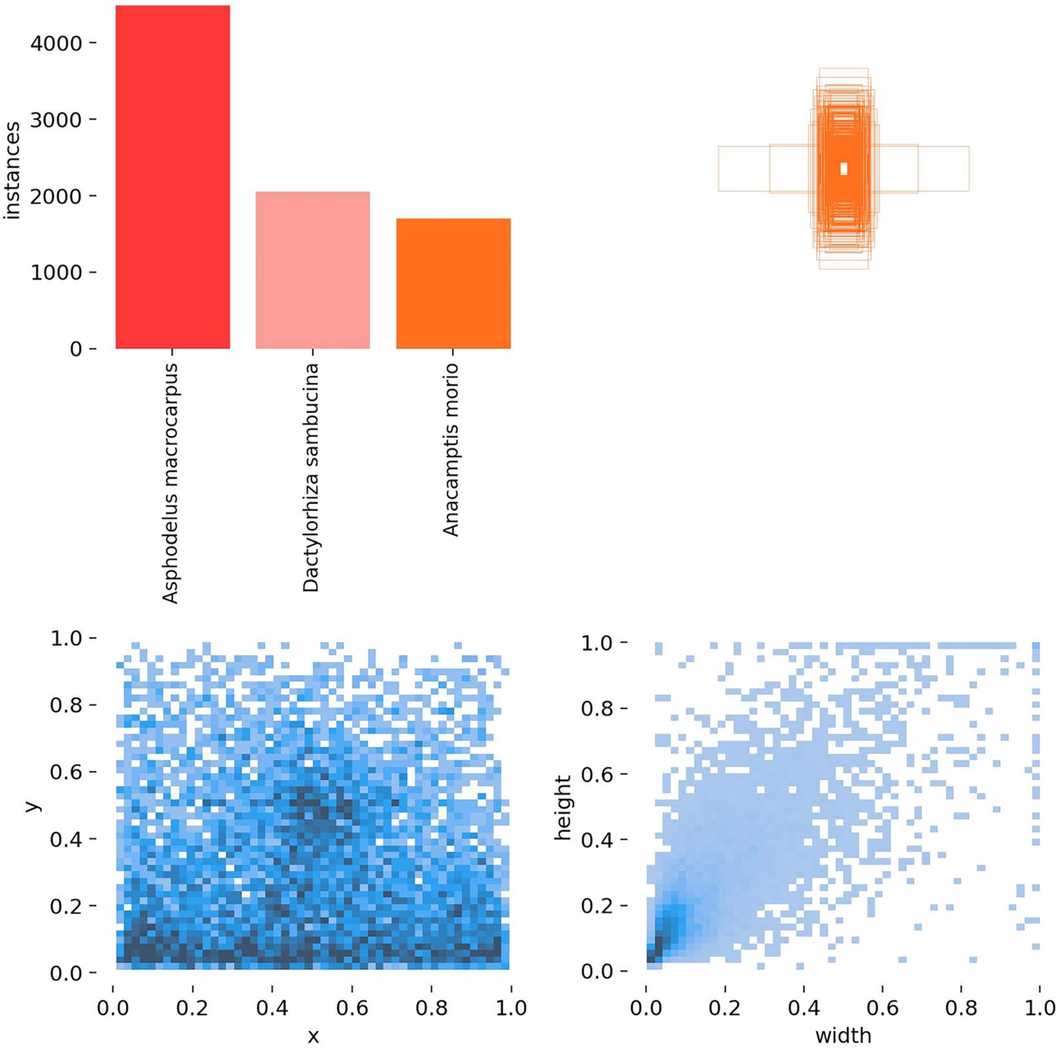


**Fig. 5** Overview of the Grasslands dataset's labels. Top left corner, a bar chart that shows the instances per class is displayed. Top right, an overlapping triangle plot representing bounding box dimensions distribution is displayed. Bottom left, a heatmap showing the distribution of detections across the image space is presented. Bottom right, a heatmap showing the distribution of bounding box dimensions (width and height) is displayed.

*.txt.* The *.txt* format contains text without formatting or style. Those basic text files are widely compatible and can be opened on various hardware and operating systems. This format is commonly used for writing simple text, managing raw data, and ensuring compatibility across platforms. It is used in machine learning and object detection for storing simple data such as labels or configuration parameters. It is popular because it is simple and human-readable.

*.yaml.* The *.yaml* (YAML Ain't Markup Language) format is a strict superset of JSON, but unlike JSON it does not require quotes around most string values as it uses Python-style indentation to indicate nesting. It is commonly used in machine learning for defining the structure of datasets due to its readability and flexibility. It is favored because it is both human-readable and machine-parsable. It can be used also to specify configuration files, such as those that define model parameters or dataset paths.

## Technical Validation

During the various fieldworks, the robotics engineering team composed of Giovanni Di Lorenzo (G.D.L.), Franco Angelini (F.A.), Michele Pierallini (M.P.), Simone Tolomei (S.T.), Davide De Benedittis (D.D.B.), and Manolo Garabini (M.G.) cooperated with four different plant scientists teams, one for each habitat:

- Dunes: Simonetta Bagella (S.B.), Agnese Denaro (A.D.), Giovanni Rivieccio (G.R), and Maria Carmela Caria (M.C.C.).
- Grasslands: Daniela Gigante (D.G.), Federica Bonini (F.B.), and Anna Grassi (A.G.).
- Screes: Marco Caccianiga (M.C.), Barbara Valle (B.V.), and Marina Serena Borgatti (M.S.B.).

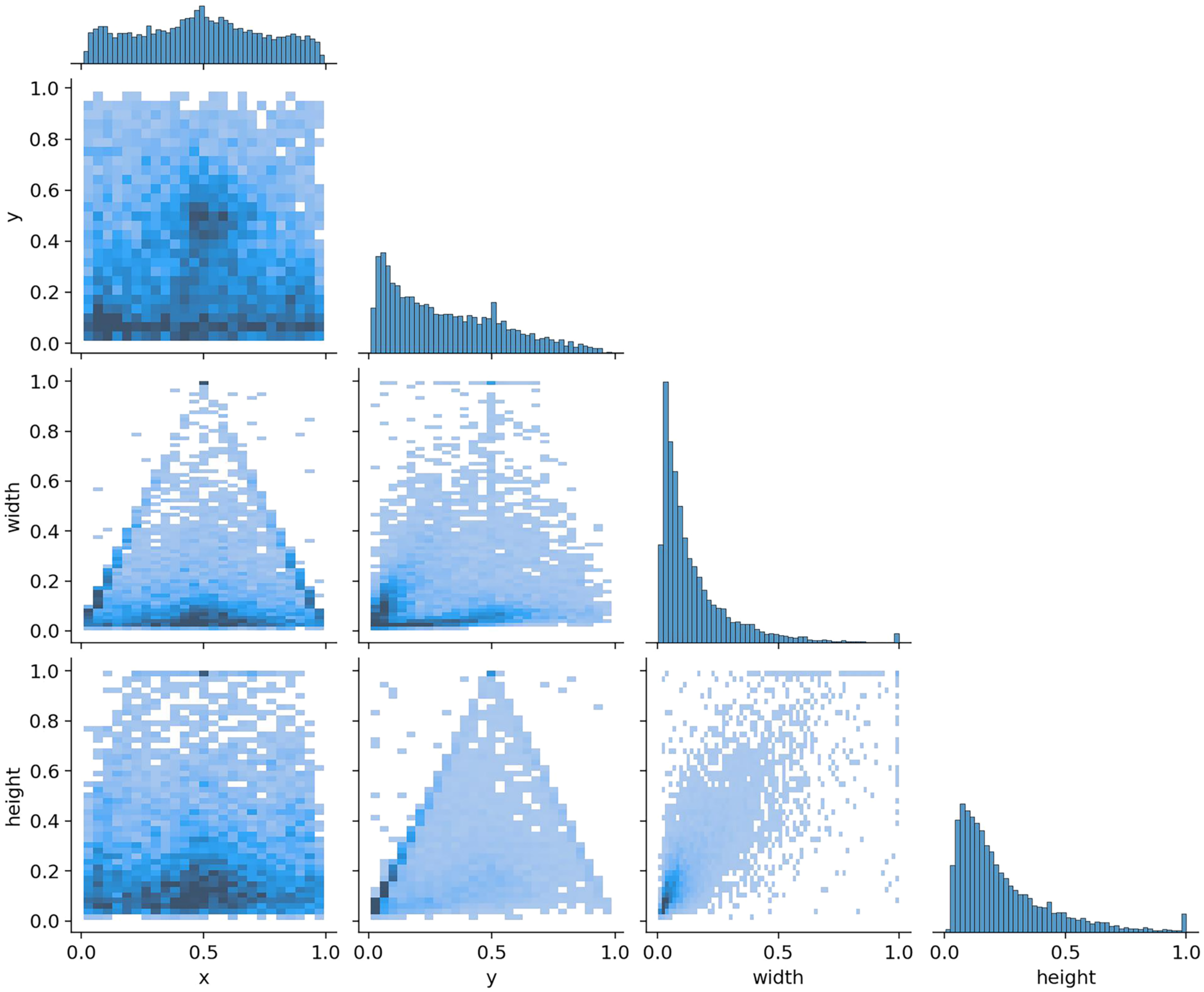


Fig. 6 Complete correlogram of the Grasslands dataset's labels.

- Forests: Claudia Angiolini (C.A.), Leopoldo de Simone (L.D.S.), Emanuele Fanfarillo (E.F.), Tiberio Fiaschi (T.F.), and Simona Macherini (S.M.).

The data collection process was conducted during peak phenological periods to capture the target species in their most relevant stages. Each image was annotated using the YOLOtxt format. The annotations were then subjected to a review process where the team of plant scientist cross-checked the labels to ensure their accuracy. Any discrepancies in species identification were discussed and then the final decision was obtained through a consensus among the experts. This ensures that the presented dataset is reliable and scientifically accurate. Further quality assurance was conducted by manually rechecking random samples of annotations. Then, object detection algorithms were run on the dataset to verify that the annotations led to accurate species identification.

The data collection and labeling procedures were managed by all Authors, who also carefully examined the finished dataset to look for errors and inconsistencies.

**Data Acquisition Pipeline.** To tackle the difficulties in obtaining accurate and dependable environmental data, a comprehensive guide has been developed. This guide covers the numerous elements of the data acquisition process, taking into account the objective of creating a dataset for object detection, and is the result of collaboration between various fields of expertise. It is important to ensure that the quantity, diversity, and quality of data collected are maintained while trying to establish a powerful computer vision dataset. The procedures followed in order to achieve this dataset are given in this section with particular emphasis on extensive data acquisition, labeling, and quality assurance procedures.

First of all we needed to ensure the data quality: to build a dataset to later train AI models for computer vision are necessary high-quality images and videos. We employed different high-resolution cameras positioned and calibrated to minimize distortion of the images. The captured images have resolutions of several megapixels, ensuring that fine details and textures are preserved during data collection, even though the final dataset consists of images resized to 640 × 640 pixels for consistency and computational efficiency in model training. To downsample these images, we employed OpenCV's cv2.resize function with high-quality interpolation, specifically the bicubic interpolation[63] method. This technique is widely recognized in computer vision for its ability to

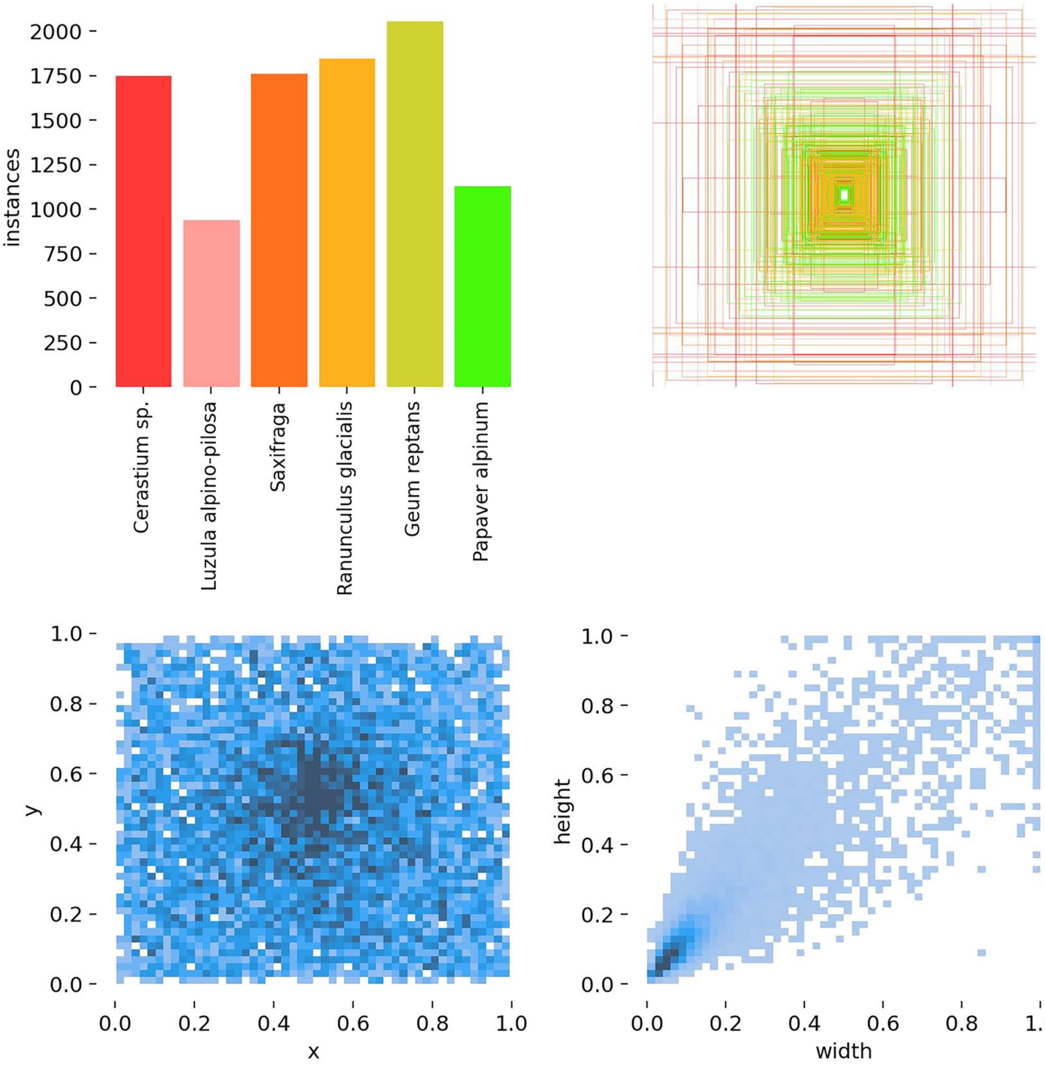


**Fig. 7** Overview of the Alpine Screes dataset's labels. Top left corner, a bar chart that shows the instances per class is displayed. Top right, an overlapping triangle plot representing bounding box dimensions distribution is displayed. Bottom left, a heatmap showing the distribution of detections across the image space is presented. Bottom right, a heatmap showing the distribution of bounding box dimensions (width and height) is displayed. In the *.yaml* file, the class Luzula alpino-pilosa corresponds to the species *Luzula alpinopilosa*, and the class Saxifraga corresponds to *Saxifraga bryoides*.

maintain edge sharpness and texture details, thereby preserving critical visual cues essential for object detection tasks.

Periodically the hardware needed to be checked to ensure the same performance throughout the whole data collection period[5]. Then it was necessary to ensure a sufficient quantity of data. We looked to maximize the data quantity by conducting extensive data gatherings at multiple sites, with multiple hardware, and at multiple times. The automated systems of the legged robot ANYmal C were leveraged to capture data at high frequencies, and data has been acquired by human operators too. We made efforts to extend the data collection period to capture seasonal changes and long-term ecological shifts[64,65] to obtain a diverse dataset to help the generalization. To ensure variety along with the maximizing of the data amount we acquired data at different locations, times, and with different hardware. On the times side, various weather conditions and times of day were taken into account to improve the dataset's robustness.

Once data were collected, we performed a rigorous labeling process. Expert plant scientists, each specialized in one or more of the habitats, annotated the images with bounding boxes identifying the target species (TS, NS, EWS, and IS depending on the habitat). This was facilitated by specialized software tools (online tools such as Labelbox and Roboflow, and offline tools such as ModifiedOpenLabelling that streamlined the annotation process and ensured consistency. The labeled data were then reviewed and validated by multiple experts to ensure accuracy and reliability[66,67]. We then implemented quality assurance protocols to maintain the integrity of the dataset, like cross-referencing labeled data with field observations and periodically reviewing the dataset

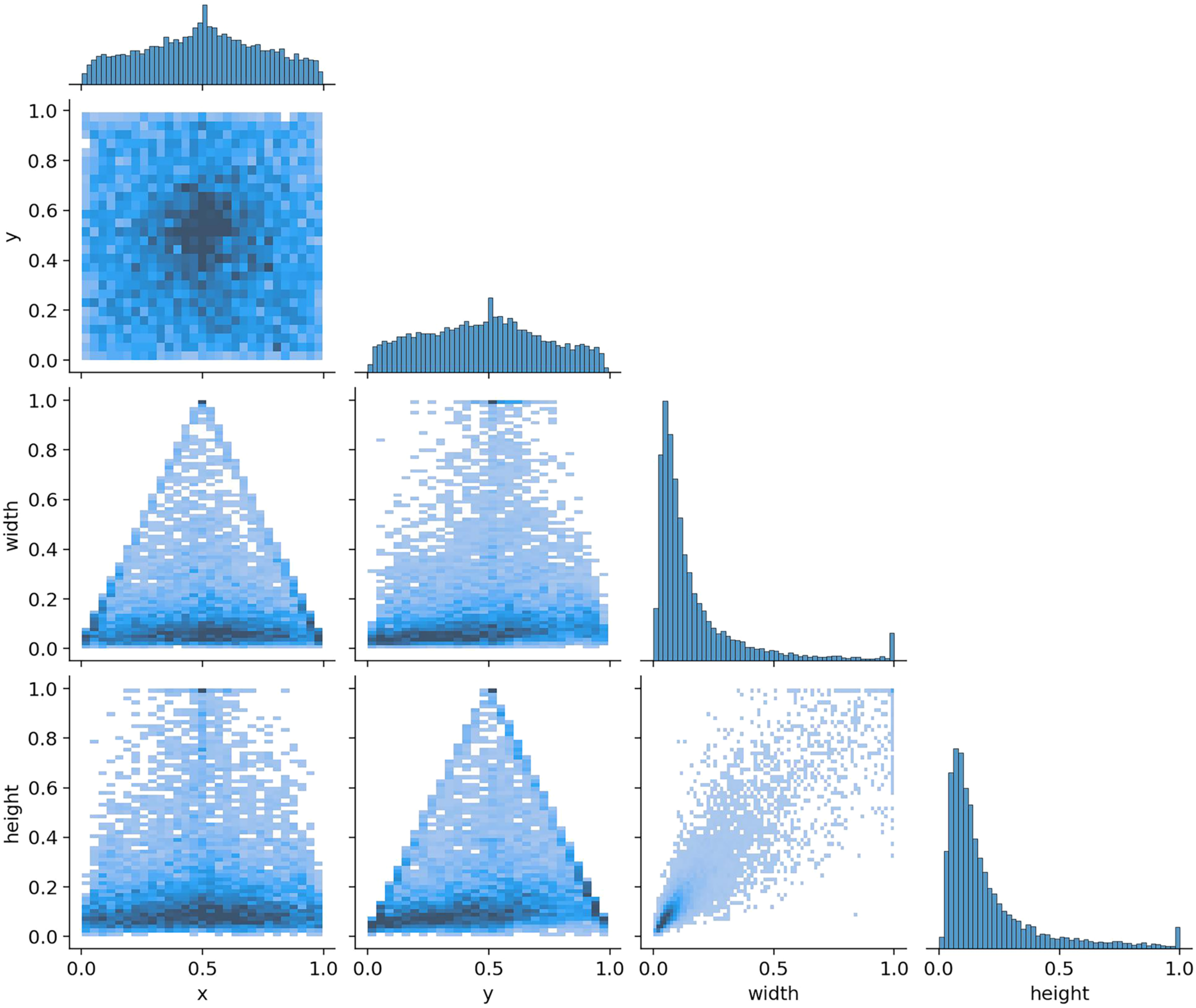


**Fig. 8** Complete correlogram of the Alpine Screes dataset's labels.

manually. This ensures that the dataset remains accurate, comprehensive, and suitable for training advanced computer vision algorithms[68].

A series of decisions, distinct for each of the four habitats, further ensure the validity of the dataset. The technical validation for each one of the habitats is explained in detail below.

### Mediterranean Coastal Dunes: technical validation for habitat 2110 and 2120.

*Mediterranean Coastal Dunes: Area selection.* The study areas for data collection in Mediterranean coastal dunes were selected by the relative team of plant scientists, specialized in the local flora. The primary sites chosen were Platamona, situated in the northwest of Sardinia in the Gulf of Asinara, and Spiaggia del Relitto within the La Maddalena Archipelago, located in the northeast of Sardinia. Platamona features a 17-kilometer-long sandy shoreline with longitudinal and parabolic dunes aligned in a northwest-southeast orientation. The dunes vary in height and form depending on wind exposure, offering a diverse landscape for study. Conversely, Spiaggia del Relitto presents a flatter dune morphology, influenced by the region's distinct environmental factors.

In those sites are both relevant habitats 2110 and 2120. The selection of these areas was based on their significance within the N2000N, being Special Areas of Conservation (SAC) ITB010003 - Stagno and Ginepreto di Platamona and ITB010008 - La Maddalena Archipelago National Park.

*Mediterranean Coastal Dunes: Time selection.* Fieldwork for the Mediterranean coastal dunes was conducted during two years: from May 16th to May 19th, 2022, and from May 16th to May 20th, 2023. These dates were selected by the plant scientists based on the phenological stages of the dune vegetation. This timing ensured that the vegetative structures of the target species were easily distinguishable. We ensured the collection of data across varied time, weather, and environmental conditions.

*Mediterranean Coastal Dunes: Target species selection and validity.* Target species for habitats 2110 and 2120 were selected according to the guidelines outlined by the European Union in Annex I of the Habitats Directive.

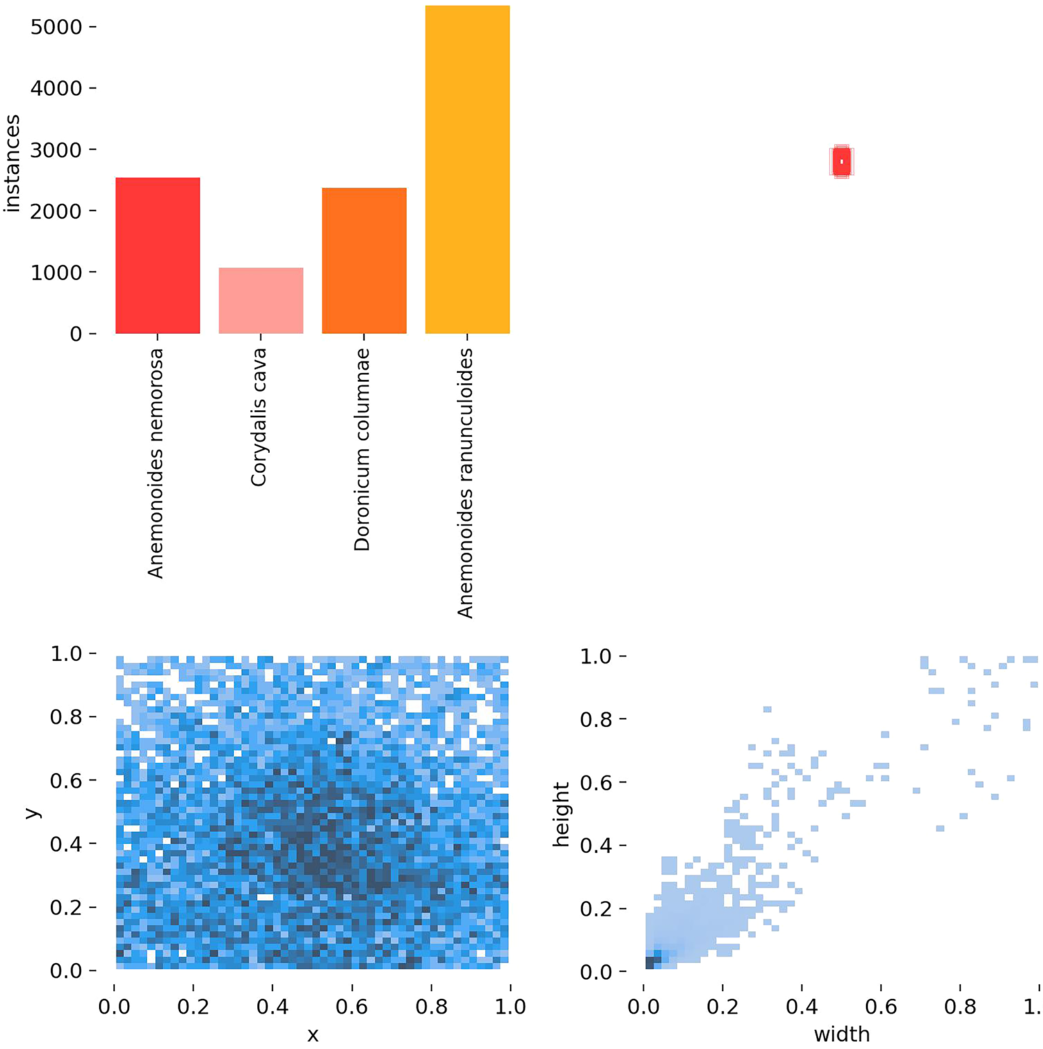


**Fig. 9** Overview of the Mediterranean Deciduous Forests dataset's labels. Top left corner, a bar chart that shows the instances per class is displayed. Top right, an overlapping triangle plot representing bounding box dimensions distribution is displayed. Bottom left, a heatmap showing the distribution of detections across the image space is presented. Bottom right, a heatmap showing the distribution of bounding box dimensions (width and height) is displayed.

The selection process was carried out by the team of plant scientists who identified these species both in the field and later in the labeling phase. The identification process was supported by the use of adequate flora manuals to ensure accurate classification of each species.

For habitat 2110, the selected TS include *Thinopyrum junceum* and *Achillea maritima*, which are characteristic of the embryonic dune systems. Habitat 2120 is represented by the TS *Ammophila arenaria* subsp. *arundinacea* that plays a critical role in the stabilization and formation of white dunes. Additionally, two NS, *Eryngium maritimum* and *Pancratium maritimum*, were considered due to their prevalence in these dune systems. The IS *Carpobrotus acinaciformis* was included as it poses a significant threat to the conservation of these habitats. As guidelines were used a floristic manual[69] and scientific publications.

### Grasslands: technical validation for habitat 6210*.

*Grasslands: Area selection.* The data collection for the grasslands was conducted within the N2000N site Special Area of Conservation (SAC) IT5210014 "Monti Maggio-Nero (sommità)" and Valsorda." This site is located in the Valsorda area within the municipality of Gualdo Tadino in the province of Perugia, Italy, within the central Apennines. The location was selected because the SAC includes extensive areas of the target habitat 6210*. Plant scientist previously identified and assessed according to European and national standards that the selected plots are representative examples of habitat 6210*.

*Grasslands: Date selection.* Fieldwork in the grasslands was carried out from May 10th to May 13th, 2022, and again from May 8th to May 10th, 2023. These dates were chosen based on the phenological stages of the target species, particularly the orchid species, to acquire data in the peak flowering period. This ensures that the target species are distinguishable.

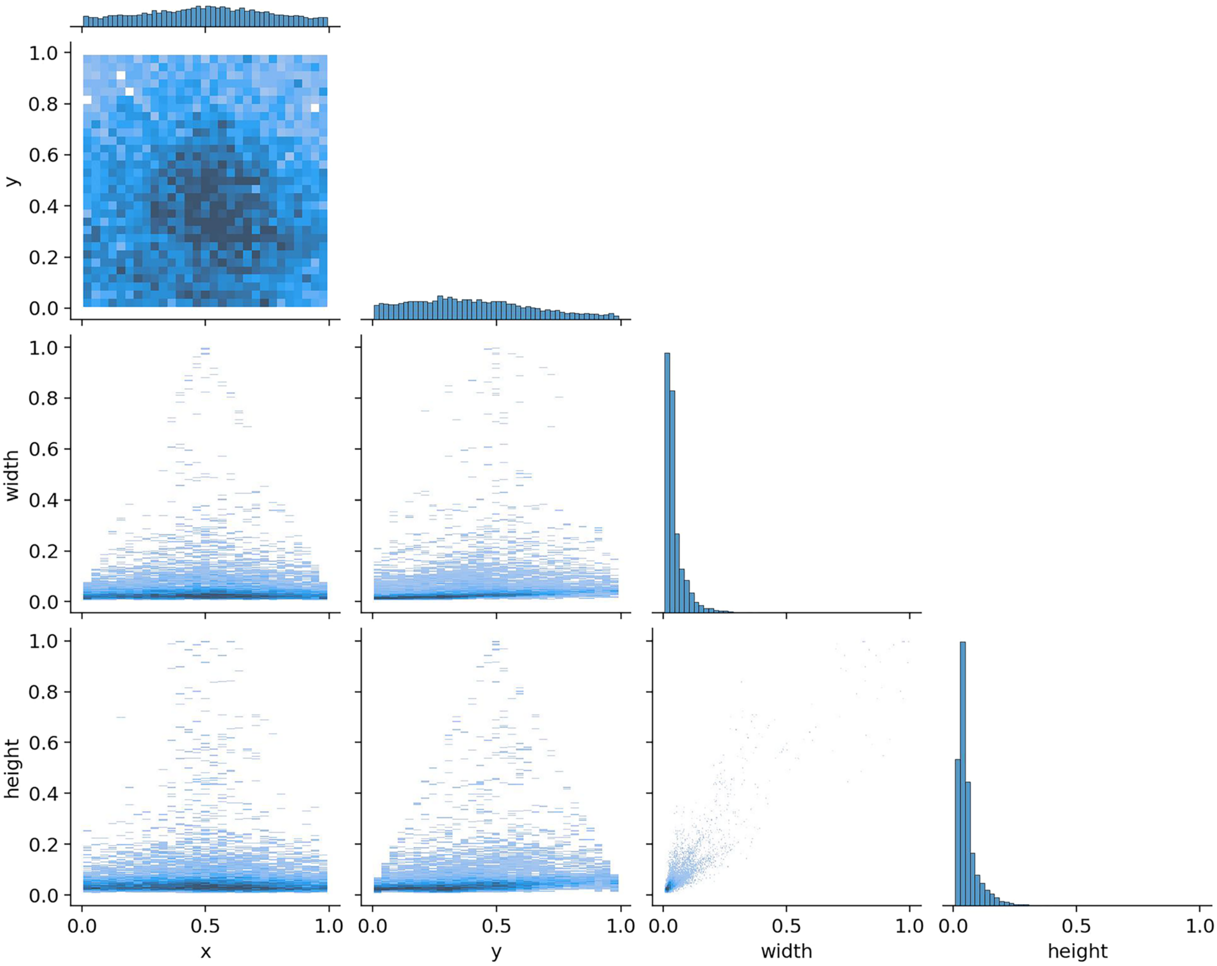
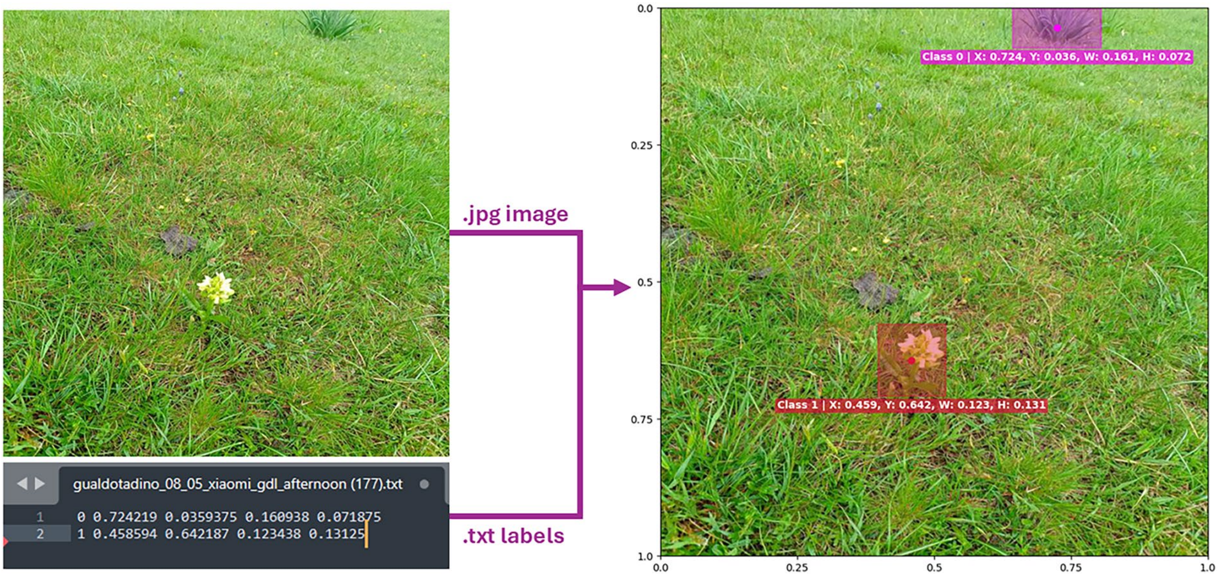


**Fig. 10** Complete correlogram of the Mediterranean Deciduous Forests dataset's labels.


.jpg image
.txt labels
Class 0 | X: 0.724, Y: 0.036, W: 0.161, H: 0.072
Class 1 | X: 0.459, Y: 0.642, W: 0.123, H: 0.131
0.0
0.25
0.5
0.75
1.0

gualdotadino_08_05_xiaomi_gdl_afternoon (177).txt

```
0 0.724219 0.0359375 0.160938 0.071875
1 0.458594 0.642187 0.123438 0.13125
```

**Fig. 11** The figure shows a side-by-side comparison of a raw image with the corresponding txt file in YOLOtxt format (on the left) and the relative visualization of the YOLO object detection annotations on the same image (on the right). E.g. in the shown image class 0 corresponds to *Asphodelus macrocarpus*, while class 1 corresponds to *Dactylorhiza sambucina*, as specified in the relative *.yaml* file.

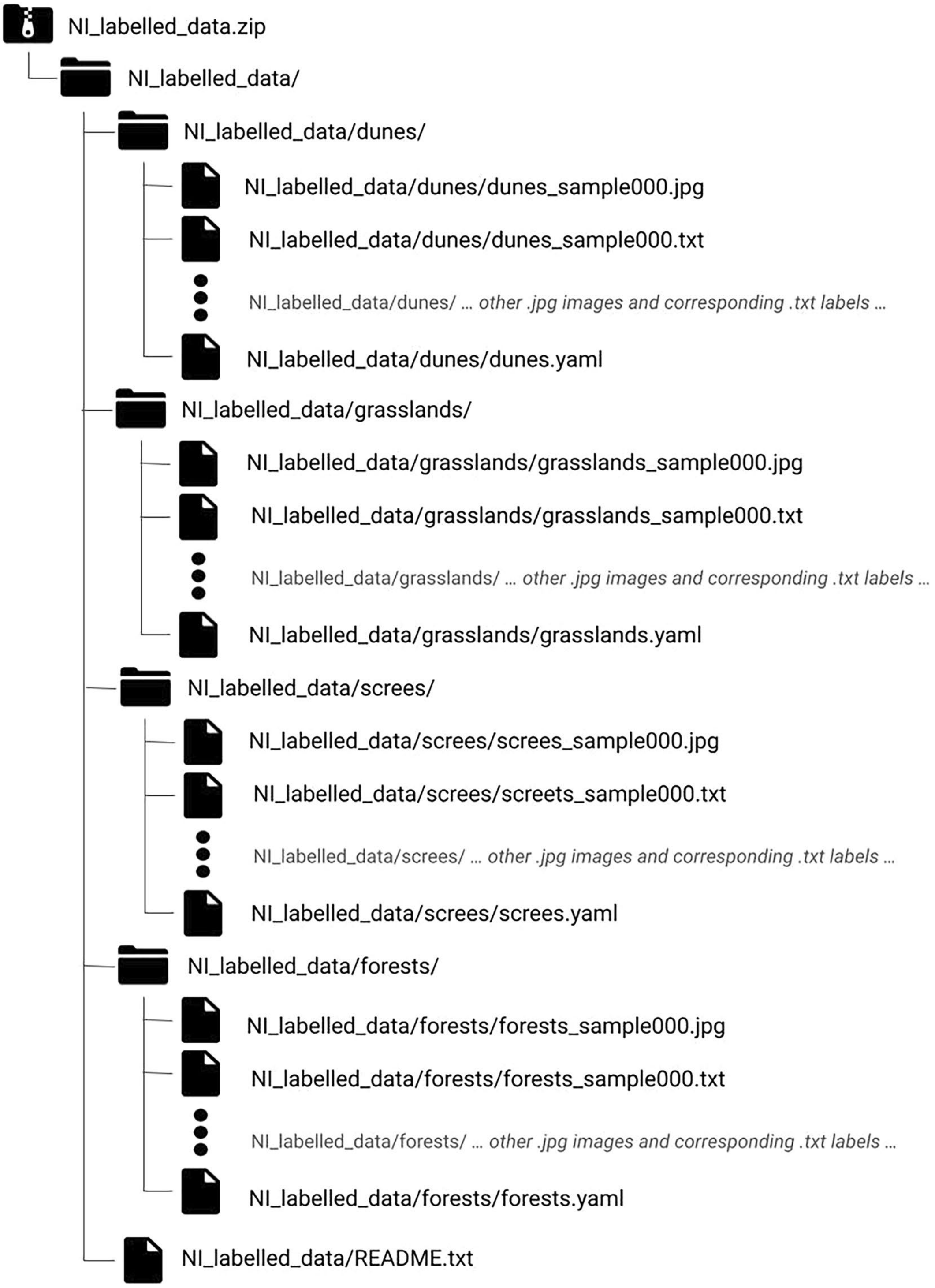

NI_labelled_data.zip
NI_labelled_data/
NI_labelled_data/dunes/
NI_labelled_data/dunes/dunes_sample000.jpg
NI_labelled_data/dunes/dunes_sample000.txt
NI_labelled_data/dunes/ ... other .jpg images and corresponding .txt labels ...
NI_labelled_data/dunes/dunes.yaml
NI_labelled_data/grasslands/
NI_labelled_data/grasslands/grasslands_sample000.jpg
NI_labelled_data/grasslands/grasslands_sample000.txt
NI_labelled_data/grasslands/ ... other .jpg images and corresponding .txt labels ...
NI_labelled_data/grasslands/grasslands.yaml
NI_labelled_data/screes/
NI_labelled_data/screes/screes_sample000.jpg
NI_labelled_data/screes/screets_sample000.txt
NI_labelled_data/screes/ ... other .jpg images and corresponding .txt labels ...
NI_labelled_data/screes/screes.yaml
NI_labelled_data/forests/
NI_labelled_data/forests/forests_sample000.jpg
NI_labelled_data/forests/forests_sample000.txt
NI_labelled_data/forests/ ... other .jpg images and corresponding .txt labels ...
NI_labelled_data/forests/forests.yaml
NI_labelled_data/README.txt

**Fig. 12** Overview of the folder structure of the dataset avalaible through Zenodo[57]. The dataset contains labeled pictures of target species of four different macro categories, divided into four subfolders, with a README.txt file recapping the relevant information. The four subfolders are named after the four categories. Each subfolder contains: species data images in *.jpg* format, corresponding *.txt* files containing the labelling information, and a single *.yaml* file recapping the classes.

*Grasslands: Target species selection and validity.* The selection of TS for habitat 6210* was based on regional and local traits of the habitat[70]. The chosen TS include two orchid species, *Anacamptis morio* and *Dactylorhiza sambucina*, both of which are common in habitat 6210* within the central Apennines and are considered indicators of favorable conservation status. Additionally, *Asphodelus macrocarpus* was selected as an EWS due to its rapid vegetative growth and its known impact on habitat 6210* in areas where traditional agropastoral activities have declined. As guidelines were used a floristic manual[69] and scientific publications[42,47,49,71].

**Alpine Screes: technical validation for habitat 8110 and 8120.** *Alpine Screes: Area selection.* The data collection for habitats 8110 and 8120, associated with alpine screes, was conducted in the Valfurva area within the Stelvio National Park, located in the province of Sondrio, Italy. This area is designated as the Special Protection Area (SPA) IT2040044 under the N2000N. Valfurva is characterized by its high-altitude alpine landscape, which includes steep, rocky slopes composed of both siliceous and calcareous screes. The habitats 8110 and 8120 are crucial for conserving specialized alpine flora, as these habitats provide a refuge for plant species adapted to the extreme conditions found at high elevations. The selection of this site was guided by its representation of these critical habitats, ensuring that the selected areas are typical examples of habitat 8110 (Siliceous scree of the montane to snow levels) and habitat 8120 (Calcareous and calcshist screes of the montane to alpine levels).

*Alpine Screes: Time selection.* Fieldwork in the alpine screes was conducted during two primary periods: from July 19th to July 21st, 2022, and from July 10th to July 15th, 2023. The selected period corresponds to the peak blooming season for many of these alpine species, which is critical for accurate identification and monitoring. Conducting the surveys during this time ensured that the phenological stages of the vascular plants were captured, coinciding with their peak development and making the species more easily identifiable.

*Alpine Screes: Target species selection and validity.* The TS selected for habitats 8110 and 8120 include Cerastium spp. (which includes *Cerastium uniflorum* Clairv. and *Cerastium pedunculatum* Gaudin), *Geum reptans* L., *Papaver alpinum* L., *Ranunculus glacialis* L., and *Saxifraga bryoides* L. These species are indicative of the alpine scree environments, which are characterized by loose rock debris and harsh growing conditions. Additionally, *Luzula alpinopilosa* Chaix Breistr. was chosen as an EWS due to its sensitivity to environmental changes and its role in indicating the ongoing stabilization of the scree ecosystems. As guidelines were used a floristic manual[69] and scientific publications[51,72–75].

**Mediterranean Deciduous Forests: technical validation for habitat 9210*.** *Mediterranean Deciduous Forests: Area selection.* The data collection for habitat 9210* (Apennine beech forests with *Taxus* and *Ilex)* was conducted within the La Verna-Monte Penna forest, located in the "Foreste Casentinesi, Monte Falterona e Campigna" National Park, and is part of the Natura 2000 network. Specifically, this area is designated as the Special Area of Conservation (SAC) IT5180101. The site was selected due to its significance as an example of habitat 9210*.

*Mediterranean Deciduous Forests: Time selection.* Fieldwork in the La Verna forest was conducted during two periods: April 27th to April 28th, 2022, and May 2nd to May 6th, 2023. These dates were chosen based on the flowering season of the target nemoral understory species.

*Mediterranean Deciduous Forests: Target species selection and validity.* In habitat 9210*, four species were selected as indicators of the habitat's conservation status: *Anemonoides nemorosa*, *Corydalis cava*, and *Anemonoides ranunculoides* as TS, and *Doronicum columnae* as EWS. The selection of these species was guided by their prevalence and ecological significance within the Apennine beech forests. The plant scientists ensured accurate identification in the field, supported by taxonomic references, and in the labeling process. Plant species were identified according to the book "Flora d'Italia"[69] and their nomenclature was updated according to "World Flora Online[75]". The Italian Manual for EU Annex I Habitats Monitoring was also consulted[76].

## Usage Notes

This data descriptor presents a dataset designed to be used by researchers across various disciplines, particularly those working in robotics, vegetation science, habitat monitoring, machine learning, and biodiversity conservation. A possible robotic future direction is integrating vision AI-based algorithms developed leveraging this dataset with path planning into a framework for intelligent robotic monitoring. Once the algorithm that guarantees the best performance has been chosen or developed, the vision information can be used not only to respond to some of the monitoring indicators but also to integrate the information within the path planning phase in a visual servoing[77] like approach. The additional information given by the detection or segmentation models can so be integrated to refine visual SLAM[78] algorithms and leveraged to optimize information-driven path planning[79]. Future work could also explore the use of this dataset in proximal sensing or in combination with remote sensing to cover larger areas. In this direction the dataset could contribute to the development of multi-robot collaboration frameworks, allowing ground and aerial robots to work together in real time, optimizing habitat coverage, and improving the accuracy of biodiversity assessments.

Below we present some recommendations to assist with the use of the dataset to train machine learning models.

**Recommended Software and Tools.** The dataset is formatted in YOLOtxt. Researchers can utilize the preferred machine learning framework (TensorFlow, PyTorch, or Darknet) to train models using this dataset. Good tools to visualize and pre-process images are libraries like OpenCV (URL: https://opencv.org/) or PIL (Python Imaging Library) (URL: https://pillow.readthedocs.io/en/stable/).

To visualize or edit the annotations, we used online tools like Labelbox (URL: https://labelbox.com/) or Roboflow (URL: https://roboflow.com), or offline tools like ModifiedOpenLabelling (URL: https://github.com/ivangrov/ModifiedOpenLabelling). To convert labels between different formats, researchers can use platforms

| HUMAN ACQUIRED RAW IMAGES (JPG) | | | |
|---|---|---|---|
| RESOLUTION | TOTAL IMAGES PER RESOLUTION | CAMERA MODEL | TOTAL IMAGES PER CAMERA MODEL |
| 3072 × 4096 | 1399 | 21061119DG | 34 |
| 3120 × 4160 | 278 | 22101316G | 5701 |
| 3264 × 2448 | 305 | COOLPIX S3300 | 28 |
| 4000 × 2250 | 8 | E-M5 | 543 |
| 4000 × 3000 | 292 | FIG-LX1 | 308 |
| 4032 × 3024 | 180 | ILCE-6000 | 787 |
| 4080 × 3072 | 15 | NIKON D300 | 245 |
| 4096 × 3072 | 4302 | NIKON D7200 | 269 |
| 4160 × 3120 | 30 | SM-A536B | 91 |
| 4240 × 2832 | 601 | SM-J250Y | 286 |
| 4288 × 2848 | 245 | SM-M215F | 292 |
| 4608 × 3456 | 571 | iPhone SE | 180 |
| 4624 × 2604 | 83 | — | — |
| 6000 × 4000 | 455 | — | — |

| ROBOT ACQUIRED RAW IMAGES (PNG) | | | |
|---|---|---|---|
| RESOLUTION | TOTAL IMAGES PER RESOLUTION | CAMERA MODEL | TOTAL IMAGES PER CAMERA MODEL |
| 1920 × 1080 | 3469 | Realsense D435 RGB-D | 3469 |

**Table 3.** Metadata recap for the raw images acquired during the 2023 missions by robot/operator, with the number of images per resolution and hardware. The dataset is publicly available through Zenodo[58] and can be accessed at https://doi.org/10.5281/zenodo.15050728.

like Roboflow or Label Studio (URL: https://labelstud.io/), or Python libraries such as pyLabel (URL: https://github.com/pylabel-project/pylabel). These tools support common formats such as YOLO, COCO, and Pascal VOC. We recommend researchers to also do data augmentation leveraging libraries like Albumentations[80] (URL: https://github.com/albumentations-team/albumentations) to artificially expand the dataset.

While these suggested tools are open-source, Roboflow provides more advanced features beyond the free plan, available through paid options.

**Suggested Processing.** Could be needed to optimize the dataset for training. Depending on their needs, researchers may consider the following preprocessing steps:

- Normalization: Normalize the pixel values of the images to a standard range (e.g., 0–1) to ensure consistency across the dataset.
- Data Augmentation: Apply data augmentation techniques such as rotation, flipping, scaling, and color adjustments to increase the robustness of models trained on this dataset.
- Class Balancing: Given the natural imbalance in species representation across the dataset, researchers may need to employ techniques such as oversampling, undersampling, or class weighting to address this issue during model training.

**Integration with Other Datasets.** First of all, note that, to the Author's best knowledge, there are no other labeled datasets for habitats 2110, 2120, 6210*, 8110, 8120, and 9210*, and because of this the integration with other floristic or habitat-specific datasets should be handled carefully. It should be noted that such integration could enhance model training or comparative studies. For example, combining this dataset with other annotated image datasets of European flora could provide a more comprehensive training base for AI models aimed at recognizing flora in general. When merging datasets, researchers should ensure consistency in image resolution and format, annotation format, and criteria used in the drawing of the bounding boxes for the different species.

**Additional Dataset Containing Unlabeled Raw Data.** While the primary objective of this data descriptor is to introduce a specifically labeled dataset designed for object detection, it is posited that the dissemination of raw data (images) can also benefit the research community. Consequently, the complete dataset obtained during the 2023 campaigns mentioned above has been made accessible on Zenodo[58] and can be accessed at https://doi.org/10.5281/zenodo.15050728. These images are presented in their original high resolution and were captured using various hardware, as detailed in Table 3 and displayed in Fig. 13. The alignment with the labeled data set discussed in this study is partial, as not all raw data were labeled and data from additional sources were used, as previously noted. This dataset can serve as a resource for researchers to evaluate computer vision methodologies beyond those proposed herein.

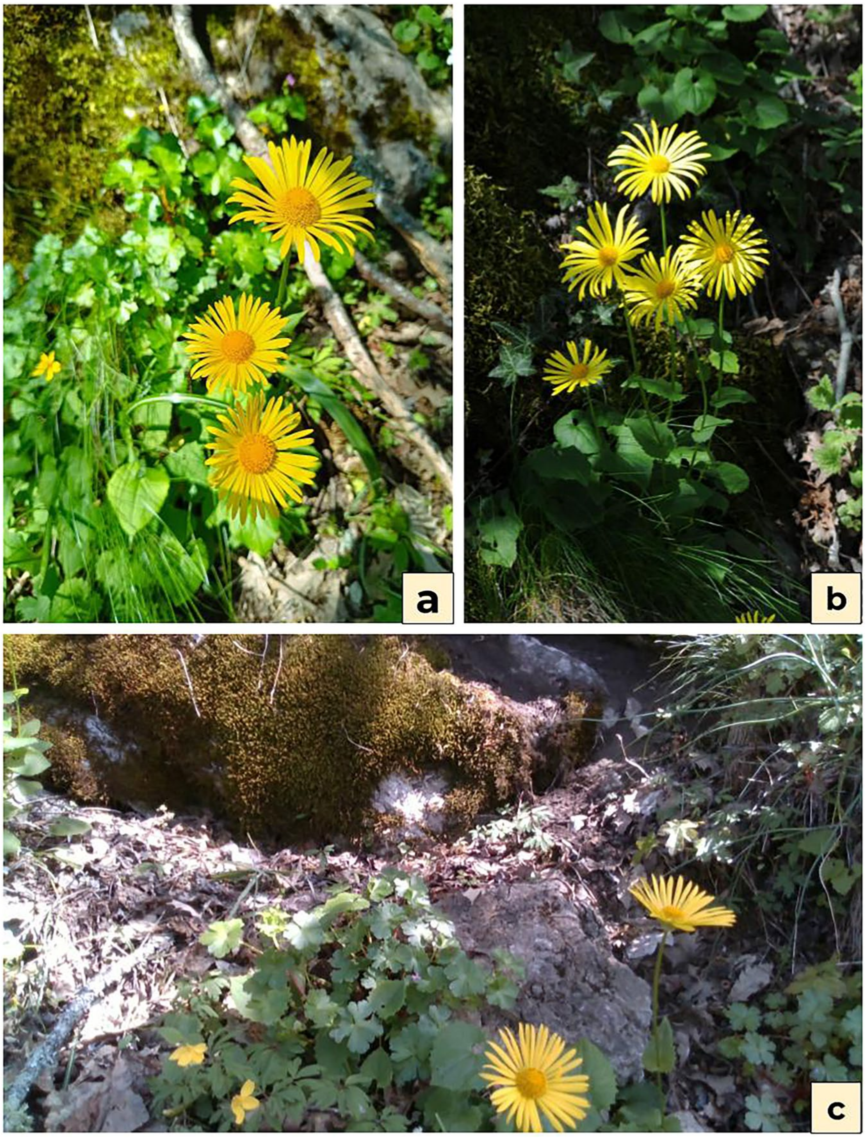


Fig. 13 Examples of the same species (*Doronicum columnae*) acquired in the same location via different modalities and hardware. **(a,b)** are both human-acquired and in *.jpg* format, while **(c)** has been acquired with the robot and is in *PNG* format. **(a)** has 2072 × 4096 resolution and has been acquired with the 22101316G camera relative to a Redmi Note 12 Pro smartphone. **(b)** has 4288 × 2848 resolution and has been acquired with a Nikon D300 camera. **(c)** has 1920 × 1080 resolution and has been acquired with a Realsense D435 RGD-D camera. All the information relative to the metadata of the images are recapped in Table 3.

**Privacy and Safety Considerations.** There are no privacy or safety restrictions associated with this dataset. It is publicly accessible and can be freely downloaded and used by researchers, educators, and conservationists. Users are encouraged to cite this dataset appropriately in any publications or projects that utilize the data.

## Code availability

No custom code was used in generating or processing the dataset presented in this data descriptor.

Received: 6 December 2024; Accepted: 9 May 2025;
Published online: 20 May 2025

## References

1. Fetting, C. The european green deal. *ESDN report* **2**, 53 (2020).
2. Directive, H. *et al*. Council directive 92/43/eec of 21 may 1992 on the conservation of natural habitats and of wild fauna and flora. *Official Journal of the European Union* **206**, 50 (1992).
3. Evans, D. Building the european union's natura 2000 network. *Nature conservation* **1**, 11–26 (2012).
4. European Environment Agency. State of nature in the EU 2020 Accessed: 2024-12-03 (2020).
5. Pimm, S. L. *et al*. The biodiversity of species and their rates of extinction, distribution, and protection. *science* **344**, 1246752 (2014).

6. Hoekstra, J. M., Boucher, T. M., Ricketts, T. H. & Roberts, C. Confronting a biome crisis: global disparities of habitat loss and protection. *Ecology letters* **8**, 23–29 (2005).
7. Gomes, E. *et al*. Future land-use changes and its impacts on terrestrial ecosystem services: A review. *Science of The Total Environment* **781**, 146716 (2021).
8. Sala, O. E. *et al*. Global biodiversity scenarios for the year 2100. *science* **287**, 1770–1774 (2000).
9. Lazzaro, L. *et al*. Impact of invasive alien plants on native plant communities and natura 2000 habitats: State of the art, gap analysis and perspectives in italy. *Journal of Environmental Management* **274**, 111140 (2020).
10. Araújo, M. B., Alagador, D., Cabeza, M., Nogués-Bravo, D. & Thuiller, W. Climate change threatens european conservation areas. *Ecology letters* **14**, 484–492 (2011).
11. Elsen, P. R., Monahan, W. B., Dougherty, E. R. & Merenlender, A. M. Keeping pace with climate change in global terrestrial protected areas. *Science advances* **6**, eaay0814 (2020).
12. Parmesan, C. Ecological and evolutionary responses to recent climate change. *Annu. Rev. Ecol. Evol. Syst.* **37**, 637–669 (2006).
13. Walther, G.-R. *et al*. Ecological responses to recent climate change. *Nature* **416**, 389–395 (2002).
14. Gigante, D. *et al*. Habitat conservation in italy: the state of the art in the light of the first european red list of terrestrial and freshwater habitats. *Rendiconti lincei. Scienze fisiche e naturali* **29**, 251–265 (2018).
15. Angiolini, C. *et al*. Detecting the imprints of past clear-cutting on riparian forest plant communities along a mediterranean river. *River Research and Applications* **39**, 1616–1628 (2023).
16. Lanzas, M. *et al*. Detecting management gaps for biodiversity conservation: An integrated assessment. *Journal of Environmental Management* **354**, 120247 (2024).
17. Kettunen, M. *et al*. Assessment of the natura 2000 co-financing arrangements of the eu financing instrument. *Institute for European Environmental Policy (IEEP), Brussels, Belgium* (2011).
18. Angelini, F. *et al*. Robotic monitoring of habitats: The natural intelligence approach. *IEEE Access* **11**, 72575–72591, https://doi.org/10.1109/ACCESS.2023.3294276 (2023).
19. Collins, A. C. Harnessing innovations in ai and robotics for environmental conservation: A comprehensive overview. (2024).
20. de Simone, L. *et al*. One small step for a robot, one giant leap for habitat monitoring: A structural survey of eu forest habitats with robotically-mounted mobile laser scanning (rmls). *Ecological Indicators* **160**, 111882 (2024).
21. Rolnick, D. *et al*. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)* **55**, 1–96 (2022).
22. Zou, Z., Chen, K., Shi, Z., Guo, Y. & Ye, J. Object detection in 20 years: A survey. *Proceedings of the IEEE* **111**, 257–276 (2023).
23. Minaee, S. *et al*. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence* **44**, 3523–3542 (2021).
24. Haug, S. & Ostermann, J. A crop/weed field image dataset for the evaluation of computer vision based precision agriculture tasks. In *Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part IV 13*, 105–116 (Springer, 2015).
25. Singh, D. *et al*. Plantdoc: A dataset for visual plant disease detection. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, 249–253 (2020).
26. Moupojou, E. *et al*. Fieldplant: A dataset of field plant images for plant disease detection and classification with deep learning. *IEEE Access* **11**, 35398–35410 (2023).
27. Olsen, A. *et al*. Deepweeds: A multiclass weed species image dataset for deep learning. *Scientific reports* **9**, 2058 (2019).
28. Gallo, I. *et al*. Deep object detection of crop weeds: Performance of yolov7 on a real case dataset from uav images. *Remote Sensing* **15**, 539 (2023).
29. Angelini, F. *et al*. Robotic monitoring of dunes: a dataset from the eu habitats 2110 and 2120 in sardinia (italy). *Scientific Data* **11**, 238 (2024).
30. Angelini, F., Pollayil, M. J., Bonini, F., Gigante, D. & Garabini, M. Robotic monitoring of grasslands: a dataset from the eu natura2000 habitat 6210* in the central apennines (italy). *Scientific Data* **10**, 418 (2023).
31. Angelini, F. *et al*. Robotic monitoring of alpine screes: a dataset from the eu natura2000 habitat 8110 in the italian alps. *Scientific Data* **10**, 855 (2023).
32. Pollayil, M. J. *et al*. Robotic monitoring of forests: a dataset from the eu habitat 9210* in the tuscan apennines (central italy). *Scientific Data* **10**, 845 (2023).
33. Dalle Fratte, M., Caccianiga, M., Ricotta, C. & Cerabolini, B. E. Identifying typical and early warning species by the combination of functional-based diagnostic species and dark diversity. *Biodiversity and Conservation* **31**, 1735–1753 (2022).
34. Hutter, M. *et al*. Anymal-toward legged robots for harsh environments. *Advanced Robotics* **31**, 918–931 (2017).
35. Ercole, S. *et al*. Manuali per il monitoraggio di specie e habitat di interesse comunitario (direttiva 92/43/cee) in italia: specie vegetali. *MANUALI E LINEE GUIDA* **140**, 1–292 (2016).
36. Tordoni, E. *et al*. Disentangling native and alien plant diversity in coastal sand dune ecosystems worldwide. *Journal of Vegetation Science* **32**, e12861 (2021).
37. Viciani, D. *et al*. A first checklist of the alien-dominated vegetation in italy. *Plant Sociology* **57**, 29–54 (2020).
38. Acosta, A. & Ercole, S. Gli habitat delle coste sabbiose italiane: ecologia e problematiche di conservazione. *ISPRA, Serie Rapporti* **215**, 115 (2015).
39. Bagella, S., Bulai, I. M., Malavasi, M. & Orrù, G. A theoretical model of plant species competition: The case of invasive carpobrotus sp. pl. and native mediterranean coastal species. *Ecological Informatics* **87**, 103070 (2025).
40. Pignatti, S., Guarino, R. & La Rosa, M. 2017–2019. flora d'italia (edizione ii). *Edagricole-Edizioni Agricole di New Business Media srl, Bologna*.
41. Biondi, E. *et al*. *Manuale Italiano di interpretazione degli habitat della Direttiva 92/43/CEE* (Società Botanica Italiana, Ministero dell'Ambiente e della tutela del territorio e del mare, DPN, 2009).
42. Gigante, D. *et al*. A methodological protocol for annex i habitats monitoring: the contribution of vegetation science. *Plant Sociology* **53**, 77–87 (2016).
43. Biagioli, M.*et al*. (eds.) *Orchidee d'Italia* (GIROS APS. Il Castello, Cornaredo (MI), 2024), 3rd edn.
44. Pignatti, S.*Valori di bioindicazione delle piante vascolari della flora d'Italia* (Dipt. di Botanica ed Ecologia dell'Università Camerino, 2005).
45. Suárez, J. R. O. & Villalba, C. Morfología y reproducción en dos poblaciones de" asphodelus albus" miller (" liliaceae"). In *Anales del Jardín Botánico de Madrid*, vol. 48, 189–200 (Real Jardín Botánico, 1990).
46. Lifante, Z. D. & Valdés, B.*Revisión del género Asphodelus L.(Asphodelaceae) en el Mediterráneo occidental* (Éd. des Conservatoire et Jardin Botaniques, 1996).
47. Biondi, E. *et al*. New and validated syntaxa for the checklist of italian vegetation. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology* **148**, 318–332 (2014).
48. Tesei, G. *et al*. Restoration strategies for grasslands colonized by asphodel-dominant communities. *Grassland science* **66**, 54–63 (2020).
49. Allegrezza, M. *et al*. The edge communities of asphodelus macrocarpus subsp. macrocarpus: the different ecological aspects and a new case study in the central apennines. *Plant Sociology* **52**, 19–40 (2015).
50. Landolt, E. *et al*. *Flora indicativa: Okologische Zeigerwerte und biologische Kennzeichen zur Flora der Schweiz und der Alpen* (Haupt, 2010).

51. Valle, B. *et al*. Glacial biodiversity of the southernmost glaciers of the european alps (clapier and peirabroc, italy). *Journal of Mountain Science* **19**, 2139–2159 (2022).
52. Fanfarillo, E. *et al*. Drivers and patterns of community completeness suggest that tuscan fagus sylvatica forests can naturally have a low plant diversity. *Forest Ecosystems* 100276 (2024).
53. Brunet, J. & von Oheimb, G. Colonization of secondary woodlands by anemone nemorosa. *Nordic Journal of Botany* **18**, 369–377 (1998).
54. Aubin, I., Ouellette, M.-H., Legendre, P., Messier, C. & Bouchard, A. Comparison of two plant functional approaches to evaluate natural restoration along an old-field–deciduous forest chronosequence. *Journal of Vegetation Science* **20**, 185–198 (2009).
55. Chytrý, M. *et al*. Floraveg.eu - an online database of european vegetation, habitats and flora. *Applied Vegetation Science* **27**, e12798, https://doi.org/10.1111/avsc.12798 (2024).
56. Angiolini, C. *et al*. Assessing the conservation status of eu forest habitats: The case of quercus suber woodlands. *Forest Ecology and Management* **496**, 119432 (2021).
57. Angelini, F., Di Lorenzo, G. & Garabini, M. Habitats labelled data, https://doi.org/10.5281/zenodo.11504938 Dataset (2024).
58. Angelini, F., Di Lorenzo, G. & Garabini, M. Habitats raw data 2023. https://doi.org/10.5281/zenodo.15050728 Data set (2025).
59. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7464–7475 (2023).
60. Jocher, G. *et al*. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo* (2022).
61. Wang, C.-Y., Yeh, I.-H. & Mark Liao, H.-Y. Yolov9: Learning what you want to learn using programmable gradient information. In *European conference on computer vision*, 1–21 (Springer, 2024).
62. Wang, A. *et al*. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* **37**, 107984–108011 (2024).
63. Li, Y., Qi, F. & Wan, Y. Improvements on bicubic image interpolation. In *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, vol. 1, 1316–1320 (IEEE, 2019).
64. Reyer, C. P. *et al*. A plant's perspective of extremes: terrestrial plant responses to changing climatic variability. *Global change biology* **19**, 75–89 (2013).
65. Rowcliffe, J. M., Field, J., Turvey, S. T. & Carbone, C. Estimating animal density using camera traps without the need for individual recognition. *Journal of Applied Ecology* 1228–1236 (2008).
66. Wearn, O. R. & Glover-Kapfer, P. Snap happy: camera traps are an effective sampling tool when compared with alternative methods. *Royal Society open science* **6**, 181748 (2019).
67. Norouzzadeh, M. S. *et al*. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* **115**, E5716–E5725 (2018).
68. Schneider, S., Taylor, G. W., Linquist, S. & Kremer, S. C. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution* **10**, 461–470 (2019).
69. Pignatti, S., Guarino, R. & La Rosa, M.*Flora d'Italia, 2 Edizione* (Edagricole di New Business Media, Bologna, 2017–2019).
70. Bruno, F. & Covarelli, G. I pascoli ei pratipascoli della valsorda (appennino umbro). *Notiziario Fitosociologico* **5**, 47–64 (1968).
71. Gigante, D. *et al*. 6210 formazioni erbose secche seminaturali e facies coperte da cespugli su substrato calcareo (festuco-brometalia). In *Manuali per il monitoraggio di specie e habitat di interesse comunitario (Direttiva 92/43/CEE) in Italia: habitat. ISPRA, Serie Manuali e Linee Guida, 142/2016.*, 140–141 (Angelini P., Casella L., Grignetti A., Genovesi P., ISPRA, 2016).
72. Gobbi, M. *et al*. Vanishing permanent glaciers: climate change is threatening a european union habitat (code 8340) and its poorly known biodiversity. *Biodiversity and Conservation* **30**, 2267–2276 (2021).
73. Bonari, G. *et al*. Shedding light on typical species: implications for habitat monitoring. *Plant Sociology* **58**, 157–166 (2021).
74. Erschbamer, B. & Caccianiga, M. S. Glacier forelands: Lessons of plant population and community development. *Progress in Botany* **78**, 259–284 (2017).
75. WFO. World flora online. http://www.worldfloraonline.org [Online; accessed 16-November-2022] (2022).
76. Zitti, S., Frattaroli, A. R., Carli, E., Cutini, M. *et al*. 9210* faggeti degli appennini con taxus e ilex. In *Manuali per il monitoraggio di specie e habitat di interesse comunitario (Direttiva 92/43/CEE) in Italia: habitat.*, 234–235 (ISPRA, Serie Manuali e Linee Guida, 142/2016., 2016).
77. Kazemi, M., Gupta, K. & Mehrandezh, M. Path-planning for visual servoing: A review and issues. *Visual Servoing via Advanced Numerical Methods* 189–207 (2010).
78. Macario Barros, A., Michel, M., Moline, Y., Corre, G. & Carrel, F. A comprehensive survey of visual slam algorithms. *Robotics* **11**, 24 (2022).
79. Bai, S., Shan, T., Chen, F., Liu, L. & Englot, B. Information-driven path planning. *Current Robotics Reports* **2**, 177–188 (2021).
80. Buslaev, A. *et al*. Albumentations: fast and flexible image augmentations. *Information* **11**, 125 (2020).
81. Romao, C. Interpretation manual of european union habitats. (1996).

## Acknowledgements

This research is partially supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101016970 (Natural Intelligence), and in part by the Italian Ministry of Education and Research (MIUR) in the framework of the "FoReLab" (Future-Oriented Research Lab) Project (Departments of Excellence). The Authors would like to thank Jacopo Cioni, Giuseppe Alfonso, Mathew J. Pollayil, and Andrea Timperi for their help on the field.

## Author contributions

All Authors conceived the proposed method and the experiments and built the dataset. G.D.L. wrote the manuscript. All Authors conducted the experiments. G.D.L., A.D., G.R., F.B., A.G., L.D.S., B.V., and M.S.B. conducted the labeling phase. A.D., G.R., M.C.C., and S.B. provided the methodological framework for the monitoring of habitats 2110 and 2120 and identified the plant species on the field. A.G., F.B., and D.G. provided the methodological framework for the monitoring of habitats 6210* and identified the plant species on the field. L.D.S., E.F., T.F., and C.A. provided the methodological framework for the monitoring of habitats 9210* and identified the plant species on the field. B.V., M.S.B., and M.C. provided the methodological framework for the monitoring of habitats 8110 and 8120 and identified the plant species on the field. F.A. and M.G. supervised the robotic and algorithmic pipeline. All authors reviewed and approved the dataset and the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to G.. .

**Reprints and permissions information** is available at www.nature.com/reprints.

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