scientific data

OPEN

DATA DESCRIPTOR

Visual WetlandBirds Dataset: Bird Species Identification and Behavior Recognition in Videos

Javier Rodriguez-Juan[1] ✉, David Ortiz-Perez[1], Manuel Benavent-Lledo[1], David Mulero-Pérez[1], Pablo Ruiz-Ponce[1], Adrian Orihuela-Torres[2], Jose Garcia-Rodriguez[1] ✉ & Esther Sebastián-González[2,3]

The current biodiversity loss crisis makes animal monitoring a relevant field of study. In light of this, data collected through monitoring can provide essential insights, and information for decision-making aimed at preserving global biodiversity. Despite the importance of such data, there is a notable scarcity of datasets featuring videos of birds, and none of the existing datasets offer detailed annotations of bird behaviors in video format. In response to this gap, our study introduces the first fine-grained video dataset specifically designed for bird behavior detection and species classification. This dataset addresses the need for comprehensive bird video datasets and provides detailed data on bird actions, facilitating the development of deep learning models to recognize these, similar to the advancements made in human action recognition. The proposed dataset comprises 178 videos recorded in Spanish wetlands, capturing 13 different bird species performing 7 distinct behavior classes. In addition, we also present baseline results using state of the art models on two tasks: bird behavior recognition and species classification.

Background & Summary

Under the current scenario of global biodiversity loss, there is an urgent need for more precise and informed environmental management[1]. In this sense, data derived from animal monitoring plays a crucial role in informing environmental managers for species conservation[2,3]. Animal surveys provide important data on population sizes, distribution and trends over time, which are essential to assess the state of ecosystems and identify species at risk[4,5]. By using systematic monitoring data on animal and bird populations, scientists can detect early warning signs of environmental changes, such as habitat loss, climate change impacts, and pollution effects[6–8]. This information helps environmental managers develop targeted conservation strategies, prioritize resource allocation, and implement timely interventions to protect vulnerable species and their habitats[2,9,10]. Furthermore, bird surveys often serve as indicators of local ecological conditions, given birds' sensitivity to environmental changes, making them invaluable in the broader context of biodiversity conservation and ecosystem management[11]. However, monitoring birds, as any other animal, is highly resource-consuming. Thus, automated monitoring systems that are able to reduce the investment required for accurate population data are much needed.

The first step to create algorithms that detect species automatically is to create datasets with information on the species traits to train those algorithms. For example, a common way to classify species is by their vocalizations[12]. For this reason, organizations such as the Xeno-Canto Foundation (https://xeno-canto.org) compiled a large-scale online database[13] of bird sounds from more than 200,000 voice recordings and 10,000 species worldwide. This dataset was crowsourced and today it is still growing. The huge amount of data provided by this dataset has facilitated the organization of challenges to create bird-detection algorithms using acoustic data in understudied areas, such as those led by Cornell Lab (https://www.birds.cornell.edu). This is the case of BirdCLEF2023[14], or BirdCLEF2024[15], which used acoustic recordings of eastern African and Indian birds, respectively. While these datasets contain many short recordings from a wide variety of different birds, other authors have released datasets composed of fewer but longer recordings, which imitate a real wildlife scenario.

[1]Department of Computer Technology, University of Alicante, Alicante, 03690, Spain. [2]Department of Ecology, University of Alicante, Alicante, 03690, Spain. [3]'Ramón Margalef' Multidisciplinary Institute for the study of the Environment. University of Alicante, Alicante, 03690, Spain. ✉e-mail: jrodriguez@dtic.ua.es; jgarcia@dtic.ua.es

Name	Modality	Region	#Samples	#Species	Only birds
Xeno-canto[13]	Audio	Worldwide	+200,000	12115	✓
BirdCLEF2023[14]		Eastern Africa	16,900	264	✓
BirdCLEF2024[15]		India	24,460	942	✓
NIPS4BPlus[16]		Spain/France	687	61	✓
BirdVox-full-night[17]		USA	6	25	✓
Birds525[19]	Image	Worldwide	89,885	525	✓
CUB-200-2011[20]		Worldwide	11,788	200	✓
NABirds[21]		North America	48,000	400	✓
VB100[26]	Video	North America	1,416	100	✓
Animal Kingdom[27]		Worldwide	NA	NA	
WetlandBirds (*Proposed*)		Spain	178	13	✓

Table 1. Summary of reviewed bird datasets.

Name	Only birds	Species	Behaviors	Localization	Frame-level	#Minutes recorded
VB100[26]	✓	✓				755
Animal Kingdom[27]		✓	✓			NA
WetlandBirds (*Proposed*)	✓	✓	✓	✓	✓	58

Table 2. Features comparison between available birds video datasets.

Examples of this are NIPS4BPlus[16], which contains 687 recordings summing a total of 30 hours of recordings or BirdVox-full-night[17], which has 6 recordings of 10 hours each.

Although audio is a common way to classify bird species and the field of bioacoustics has increased tremendously in the latest years, another possible approach to identify species automatically is using images[18]. One of such bird image datasets is Birds525[19], which offers a collection of almost 90,000 images involving 525 different bird species. Another standard image dataset is CUB-200-2011[20], which provides 11,788 images from 200 different bird species. This dataset not only provides bird species, but also bounding boxes and part locations for each image. There are also datasets aimed at specific world regions like NABirds[21], which includes almost 50,000 images from the 400 most common birds seen in North America. This dataset provides a fine-grained classification of species as its annotations differentiate between male, female and juvenile birds. These datasets can be used to create algorithms for the automatic detection of the species based on image data.

However, another important source of animal ecology information that has been much less studied because of the technological challenges of its use are videos. Video recordings may offer information not only about which species are present in a specific place, but also about their behavior. Information about animal behavior may be very relevant to inform about individual and population responses to anthropogenic impacts and has therefore been linked to conservation biology and restoration success[22–25]. Besides its potential for animal monitoring and conservation, the number of databases on wildlife behavior are more limited. For example, the VB100 dataset[26], comprises 1416 clips of approximately 30 seconds. This dataset involves 100 different species from North American birds. The unique dataset comprised by annotated videos with birds behavior available in the literature is the Animal Kingdom dataset[27], which is not specifically aimed at birds and contains annotated videos from multiple animals. Specifically, it contains 30,000 video sequences of multi-label behaviors involving 6 different animal classes; however, the number of bird videos was not specified by the authors. Table 1 summarizes the main information of the datasets reviewed.

Due to the scarcity of datasets involving birds videos annotated with its behaviors, this study proposes the development of the first fine-grained behavior detection dataset for birds. Differently from Animal Kingdom, where a video is associated with the multiple behaviors happening, in our dataset, spatio-temporal behavior annotations are provided. This implies that videos are annotated per-frame, where the behavior happening and the location is annotated in each frame (*i.e.*, bounding box). Moreover, the identification of the bird species appearing in the video is also provided. The proposed dataset is composed by 178 videos recorded in Spanish wetlands, more specifically in the region of Alicante (southeastern Spain). The 178 videos expand to 858 behavior clips involving 13 different bird species. The average duration of each of the behavior clips is 19.84 seconds and the total duration of the dataset recorded is 58 minutes and 53 seconds. The annotation process involved several steps of data curation, with a technical team working alongside a group of professional ecologists. In comparison to other bird video datasets, ours is the first to offer annotations for species, behaviors, and localization. Furthermore, Visual WetlandBirds is the first dataset to provide frame-level annotations. A features' comparison of bird video datasets is presented in Table 2.

Table 3 reflects the different species collected for the dataset, distinguishing between their common and scientific names. The number of videos and minutes recorded for each species is also included.

Seven main behaviors were identified as key activities recorded in our dataset. These represent the main activities performed by waterbirds in nature[28]. In Table 4, these behaviors are specified alongside the number

Common name	Scientific name	Videos	Recorded minutes
Yellow-legged Gull	*Larus michahellis*	13	5.08
White wagtail	*Motacilla alba*	13	4.33
Squacco Heron	*Ardeola ralloides*	15	4.94
Northern shoveler	*Spatula clypeata*	14	3.49
Mallard	*Anas platyrhynchos*	10	2.94
Little-ringed plover	*Charadrius dubius*	10	1.93
Glossy ibis	*Plegadis falcinellus*	8	3.96
Gadwall	*Mareca strepera*	13	2.59
Eurasian moorhen	*Gallinula chloropus*	18	9.18
Eurasian magpie	*Pica pica*	16	5.95
Eurasian coot	*Fulica atra*	19	4.11
Black-winged stilt	*Himantopus himantopus*	14	3.55
Black-headed gull	*Chroicocephalus ridibundus*	15	6.84

Table 3. Statistics for each of the bird species.

Metric	Alert	Feeding	Flying	Preening	Resting	Swimming	Walking
Number of clips	124	271	46	58	122	78	159
Mean duration in frames	166	240	61	195	157	257	108

Table 4. Total number of clips per behavior and their mean duration.

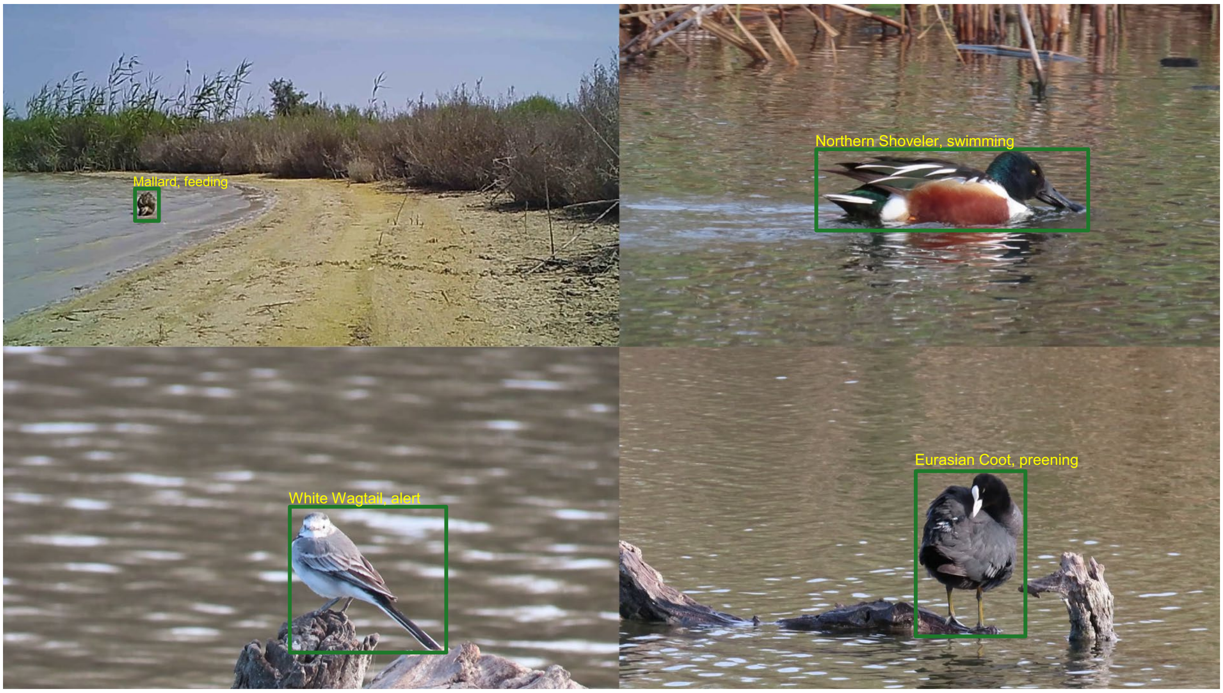


Fig. 1 Sample frames from the dataset.

of clips recorded per each of them and the mean duration of each behavior in frames. A clip is a piece of video where a bird is performing a specific behavior.

Figure 1 presents sample frames where only a single bird individual can be distinguished. However, this dataset contains not only videos with a single individual, but also videos where several birds appear together. This is the case for gregarious species, which are species that concentrate in an area for the purpose of different activities. Although the individuals of gregarious species often share the same behavior at the same time, it is also common that several behaviors can be seen in the same video at the same time. Figure 2 shows some sample frames where this phenomenon happens. Videos involving different birds and/or performing different activities sequentially were cut in clips where a unique individual is performing a unique behavior in order to get the statistics shown in Table 4.

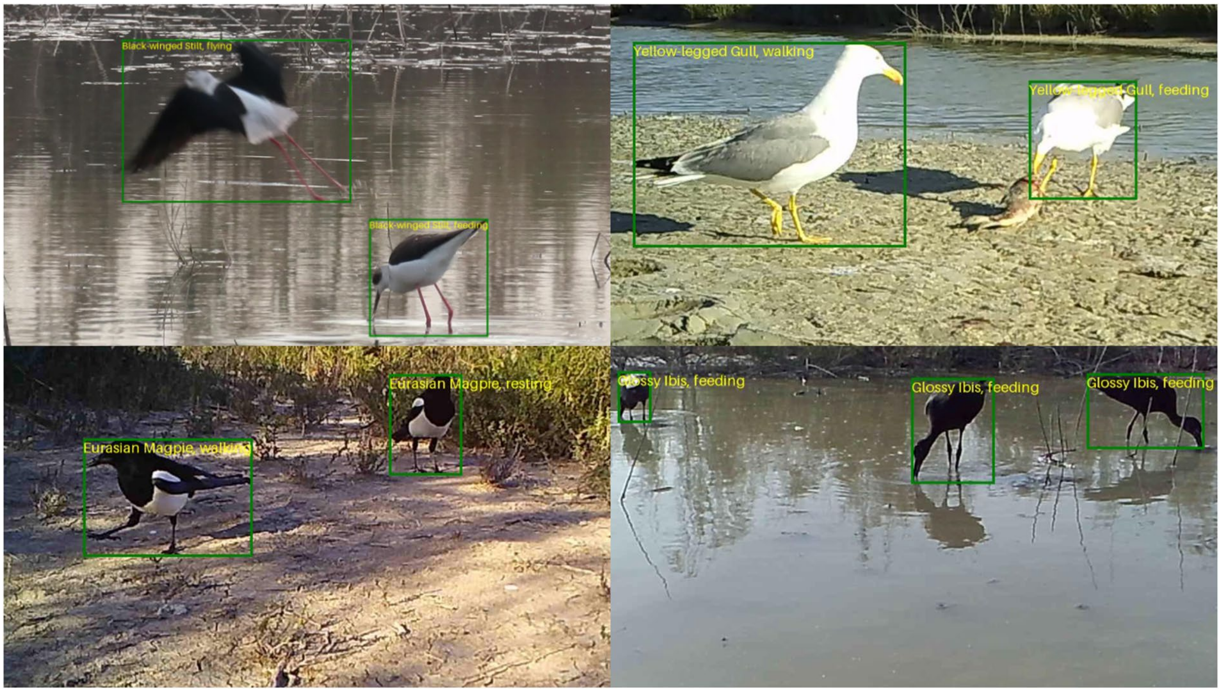


Fig. 2 Sample frames where gregarious birds appear performing different behaviors.

Between the seven behaviors proposed, it should be underlined the difference between the *Alert*, *Preening* and *Resting* behaviors. These action distinctions were established by the ecology team. We considered that the bird was *Resting* when it was standing without making any movement. The bird was performing the *Alert* behavior when it was moving its head from one side to another, moving and looking around for possible dangers. Finally, we considered that the bird was *Preening* when it was standing and cleaning its body feathers with its beak. The remaining behaviors are not explained because of their obvious meaning.

As it can be seen in Table 4, the number of clips per behavior is unbalanced between classes. This is because the recording of videos where some specific behaviors are happening is more uncommon, as happens with *Flying* or *Preening*, which represent the activities with the lowest number of clips in the dataset. These behaviors are difficult to record since they are performed with a lower frequency. In order to be able to collect more data on these less common behaviors, more hardware and human resources (*i.e.*, cameras and professional ecologists) are needed to cover a wider area of the wetlands. Furthermore, another technique like data augmentation[29] can generate synthetic data from the actual one. While allocating more human and hardware resources would ensure that the quality of the new data remains high, it is also costly as high-quality cameras and extra ecology professionals are expensive. On the counterpart, synthetic generation techniques are being highly used in current research as they provide a costless way of increasing the amount of data available for training. Although costless, synthetic data can decrease the quality of the dataset, so a trade-off between real and synthetic data would be necessary to limit the cost without compromising the quality of the videos. Although the unbalanced nature of the behaviors, no balancing technique over this data was applied in the released dataset in order to maximize the number of different environments captured, ensuring in this way the variability of contexts where the birds are recorded.

Additionally, Table 4 also shows the mean duration of the clips per behavior. It is worth noting the difference in the number of frames between *Flying*, that represents the minimum with 61 frames with respect to *Swimming*, which represents the absolute maximum with a value of 257 frames. This difference is explained in the nature of the behaviors, as swimming is naturally a slow behavior, which can be performed for a long time over the same area. However, flying is a fast behavior, and the bird quickly get outs of the camera focus, especially for videos obtained by camera traps, which cannot follow the bird while it is moving.

In order to collect the videos, we deployed a set of camera traps and high quality cameras in Alicante wetlands. The camera traps were able to automatically record videos based on the motion detected in the environment. We complemented the camera trap videos with recordings from high quality cameras. In these videos, a human is controlling the focus of the camera, obtaining better views and perspectives of the birds being recorded. Species recorded, behaviors identified and the camera deployment areas were described by professional ecologist based on their expertise. In Fig. 3 some video frame crops can be observed, where all the bird species developing the different behaviors available in the dataset can be seen.

After the data collection, a semi-automatic annotation method composed by an annotation tool and a deep learning model was used in order to get the videos annotated. After the annotation, a cross-validation was conducted to ensure the annotation quality. This method is deeply explained in the next section.

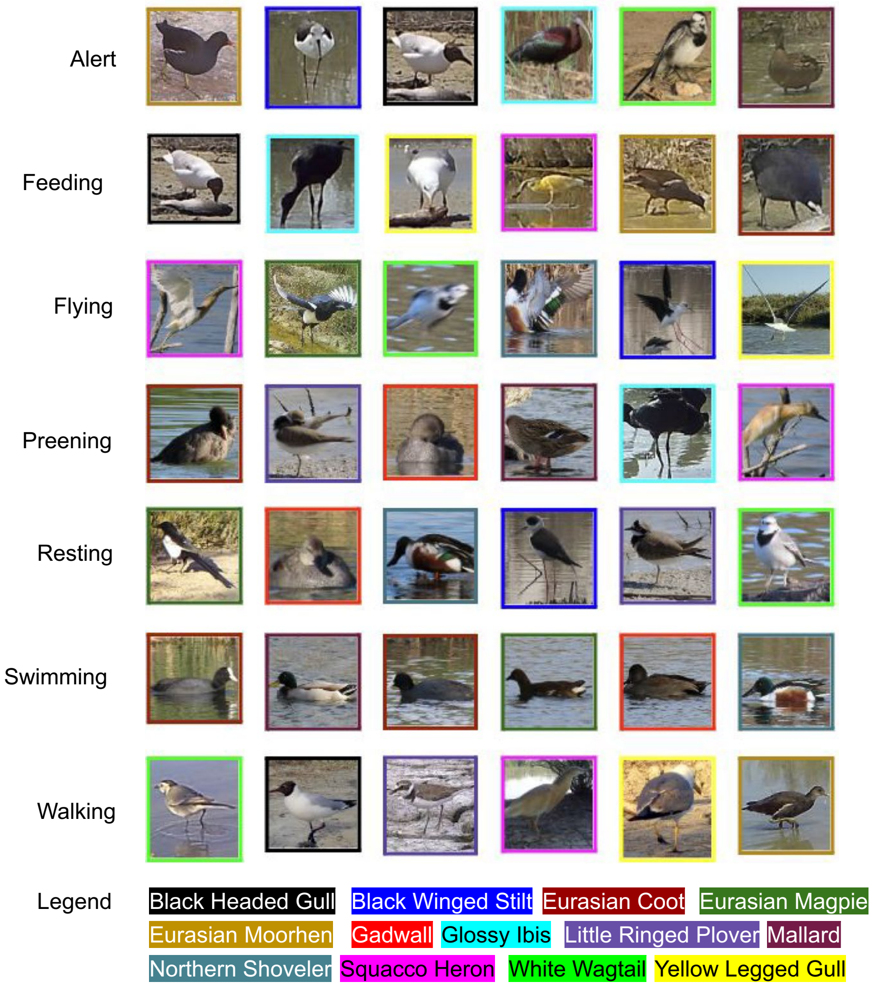


Fig. 3 Video frame crops of bird species performing the seven behaviors composing the dataset.

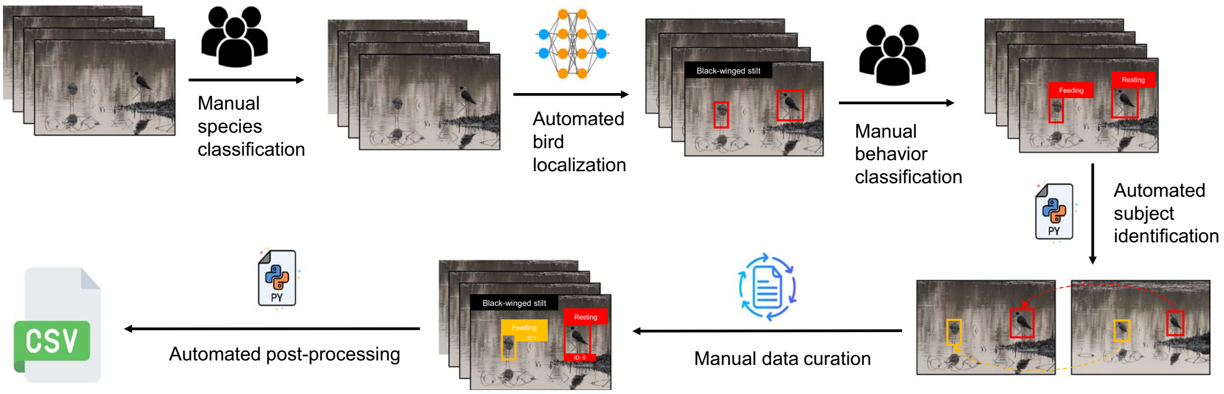

Manual species classification
Automated bird localization
Black-winged stilt
Manual behavior classification
Feeding
Resting
Automated subject identification
PY
Manual data curation
Black-winged stilt
Feeding
Resting
ID: 0
Automated post-processing
PY
CSV

Fig. 4 Visual representation of stages involved in the annotation process. Birds are first classified into species by annotators and localized using a YOLO model. Then, annotators recognize bird behaviors, subjects are identified using a Python script, and finally, the data is curated and post-processed.

In order to test the dataset for species and behavior identification, two baseline experimentation were carried out: one for the bird classification task, which involves the classification of the specie and the correct localization of the bird given input frames, and a second one for the behavior detection task, which involves the correct classification of the behavior being performed by one bird during a set of frames.

Methods

Data acquisition.

The acquisition of the data was conducted within Alicante wetlands, specifically within the wetlands of *La Mata Natural Park* and *El Hondo Natural Park* (sutheastern Spain). In these places, we deployed a collection of high-resolution cameras and camera traps in different areas of the wetlands. These areas were determined by the species expected to be recorded, as different species can be commonly seen in different wetland areas.

Camera traps are activated when movement is detected and thus can record for long periods of time without human intervention. The usage of automatic camera traps[30–32] is common in the monitoring of wildlife as it provides a low-cost approach to collect video and image data from the environment. However, the focus of this camera is fix and thus the videos of the same individual are often short. Manual cameras require the presence of a human while recording and are thus more time-consuming. Also, the presence of the cameraman may affect the animal's behavior. However, it permits manual changes of cameras' perspectives in order to correctly record the bird behavior.

Two models of camera traps were used: the Browning Strike Force Pro HD and the Bushnell Core HD, both featuring a sensor resolution of 24 megapixels, a shot speed of 0,21 seconds, and a field of view of 55°. For manual recordings, the Canon Powershot SX70 was employed, which has a sensor resolution of 20,3 megapixels and a shot speed of 5×10^{-4} seconds. As different camera models and capture settings were used, videos of different resolutions were obtained: 87 videos at 1920 × 1080px, 75 videos at 1296 × 720px, 14 videos at 1280 × 720px, 1 video at 960 × 540px, and 1 video at 3840 × 2160px.

The species selected were the most commonly found in Alicante wetlands, facilitating the recording of videos and providing valuable data to the natural parks where videos were recorded. In terms of behaviors, we identified the most representative ones of the selected species, in order to cover as much as possible the range of activities developed by the birds.

To ensure the generalization capabilities of models trained on this dataset, a variety of lighting and seasonal conditions, backgrounds, viewpoints, and video resolutions were considered. Regarding lighting conditions, the professionals responsible for the recordings were instructed to capture footage of birds at different times of day, thereby enhancing data variability. The dataset includes diverse lighting scenarios such as daylight, sunset, low-light, and backlight. Low-light and backlight scenes pose additional challenges for detection models, as they reduce the visibility of color features (often relevant for species identification) and make it more difficult to distinguish bird silhouettes from the background (e.g., top-right crop in Fig. 3). Regarding seasonal conditions, video recordings were conducted throughout the entire year to ensure representation of the environmental variability associated with the four seasons. However, due to Alicante's characteristically low annual precipitation, most of the videos in the dataset feature either sunny or cloudy weather conditions. While this may limit atmospheric diversity, it can also benefit the model training process by facilitating clearer visual identification of bird species, as the absence of rain-related distortions contributes to more interpretable video data.

To mitigate background bias in species detection, recordings were captured in a variety of natural contexts. Although the dataset was collected in wetland environments, it includes birds situated on water, the ground, grass, and tree branches (e.g., background differences between Alert crops in Fig. 3). Additionally, variations in lighting conditions affect water color, further contributing to background diversity (e.g., water color differences in Resting crops in Fig. 3). The dataset also includes a range of camera viewpoints. In some sequences, birds appear in the foreground, while in others, they are captured at greater distances, simulating real-world variability. Lastly, the inclusion of videos with different resolutions enhances the adaptability of models to real-world deployment settings, where camera quality may vary. For optimal performance in specific environments, it is recommended to fine-tune the models using data collected from the intended deployment context.

Data annotation.

Accurate annotation of the captured data is a determining factor in obtaining relevant results when training deep learning models on this data. To ensure annotation accuracy, the usage of annotation tools[33,34] is extended, as they provide a user-friendly interface that makes this process easy and accessible to non-technical staff.

There are many open-source annotation tools available on the market. CVAT (https://github.com/cvat-ai/cvat) is one of the most popular ones, as it provides annotation support for images and videos, including a variety of formats for exporting the data. VoTT (https://github.com/microsoft/VoTT) is also popular when annotating videos, as it offers multiple annotation shapes and integration with Microsoft services to easily upload data to Azure. Other simpler annotation tools are labelme (https://github.com/labelmeai/labelme) or LabelImg (https://github.com/HumanSignal/labelImg), which are aimed at annotating images and their capabilities are more limited. For our purpose, we decided to use CVAT because of the large number of exportable formats available, the great collaborative environment it offers, and its easy integration with semi-automatic and automatic annotation processes.

As the need for larger amounts of data to train deep learning models increased, researchers began to enhance annotation tools with automatic systems that could alleviate this task. Annotation tools integrate machine learning models[35] that can automatically infer what would otherwise be manually annotated. Common tasks performed by automated annotation tools are object detection[36] and semantic segmentation[37]. While the former predicts the bounding box and class of each object in the image, the latter predicts regions of interest associated with specific categories.

Although automated annotation systems have demonstrated strong performance, semi-automated annotation processes are ultimately used because they ensure the creation of highly accurate annotations while greatly reducing the amount of human intervention required. Semi-automated annotation studies are widely used in the medical field[38,39], where precision is a key factor throughout the design.

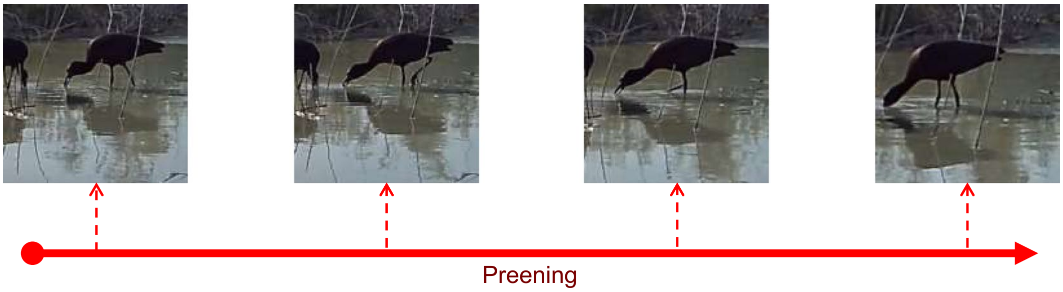


Fig. 5 Sample clip in which a bird performs the *Feeding* and *Walking* behavior simultaneously. In such cases, *Feeding* is prioritized by annotators due to its higher biological relevance.

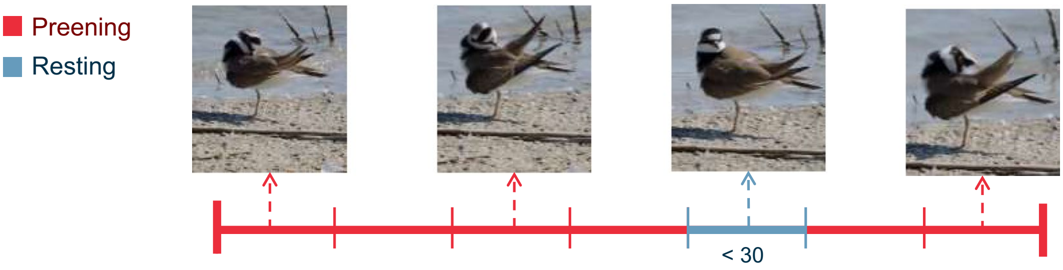


Fig. 6 Within this clip, a bird is doing the behavior *Preening*, but shortly interrupts the behavior and transitions to *Resting*. As *Resting* lasts for less than 30 frames, it is considered as part of the *Preening* behavior.

In this study, a semi-automated annotation approach was followed, based on CVAT and its possible integration with powerful computer vision models. Our approach consisted of five main steps: Species classification, bird localization, behavior classification, subject identification, data curation, and post-processing. Each of these stages is described in more detail below. Figure 4 shows this process.

1. **Manual species classification:** In this first step, the ecologists manually labeled each video with the main bird species that appeared. The main species is that of the bird in the focus of the camera. This way, annotations of birds that are different from the main species will not be included in the video annotations.
2. **Automated bird localization:** Then, an object detection model was used to predict the localization of the bounding boxes of the birds that appear in each of the video frames. YOLOv7[40] was chosen as the object detection model for ease of implementation, as it is already integrated into CVAT. Since the model provided by CVAT is trained on general purpose data, the class predicted by default for each bounding box is not be the bird species, but the class *bird*. To avoid manually changing all the bounding box classes, we used an option provided by CVAT to associate a user-defined class with the class detected by the model. In this way, the class *bird* was associated with the species appearing in the video.
3. **Manual behavior classification:** This manual stage had a twofold objective. First, they checked and corrected erroneous bounding boxes, and second, they annotated for each bounding box the behavior performed by each bird. To annotate the behaviors, CVAT bounding boxes *tags* were used.
4. **Automated subject identification:** When using automatic annotation models such as YOLOv7, CVAT does not support bounding box correspondence between frames. In other words, if a video shows two birds developing different behaviors, there is no relationship between the bounding boxes of adjacent frames, so it is not possible to analyze the birds' behaviors. This is not possible because the next frame will show two new bounding boxes whose relation to the one being analyzed is not known. To address this problem, the Euclidean distance[41] was used to correlate the bounding boxes of adjacent frames. The euclidean distance calculates the distance between the centers of the bounding boxes of adjacent frames and then correlates the bounding boxes with the minimum distance. The center of the bounding box was calculated as follows:

$$c=\left(\frac{x_{\min}+x_{\max}}{2},\frac{y_{\min}+y_{\max}}{2}\right) \tag{1}$$

Given the centre of the bounding boxes, the Euclidean distance was calculated as:

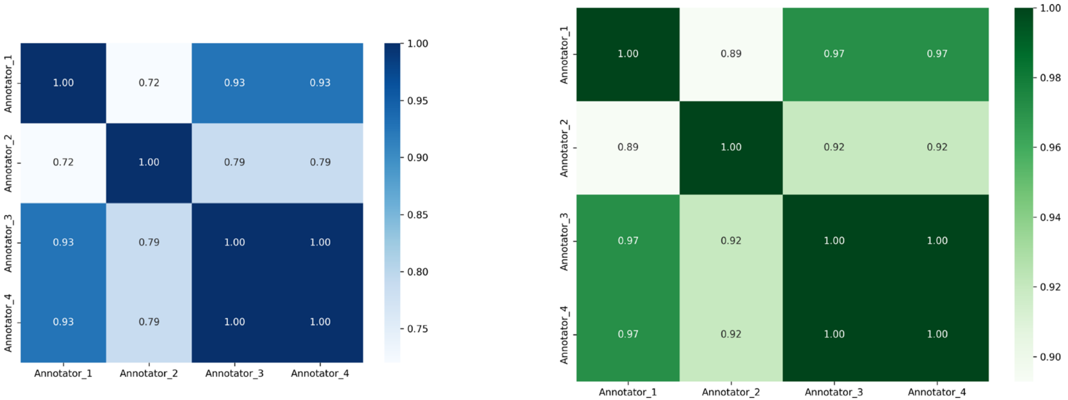
$$d(c_1,c_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2} \tag{2}$$

Metric	Cohen's Kappa	Fleiss' Kappa	Macro F1
Score	0,858	0,855	0,946

Table 5. Results of the inter-annotator agreement evaluation. Cohen's Kappa and Macro F1 values shown are the average of each of the pairwise values obtained.

Model	Precision	Recall	mAP50	mAP50-95
YOLOv9	0.835	0.759	0.801	0.556

Table 6. Results of the YOLO-based baseline developed for bird species classification.



(a) Pairwise Cohen's Kappa agreement matrix.

(b) Pairwise macro F1 score matrix.

Fig. 7 Inter-annotator agreement matrices for the dataset. (**a**) Cohen's Kappa. (**b**) Macro F1 Score.

$$\begin{aligned} \text{where } c_1 &= (x_1, y_1) \\ \text{and } c_2 &= (x_2, y_2) \end{aligned}$$

5. **Manual data curation:** After the labeling of species, bounding boxes, behaviors, and subjects, an overall review of all annotations was conducted to ensure the high quality of the data. To conduct the review, videos were assigned to all ecologists equally.
6. **Automated post-processing:** Once the annotations were complete, their format was adapted to make them easy to use and understand. To achieve this goal, the approach used in the AVA-Kinetics dataset[42] was followed. In this approach, a CSV file was used to contain annotations containing localized behaviors of multiple subjects. To export the data into the CSV format, the data was first exported from CVAT using the CVAT Video 1.1 format. Some Python scrips were then used to extract only the relevant information from the exported data and dump it into the output CSV file.

Annotation criteria.

The annotation process involved addressing several specific challenges identified during the annotation stage. Two primary issues emerged: the annotation of individual birds exhibiting multiple behaviors simultaneously, and the misclassification of minor sub-movements as the dominant behavior.

It is common for birds to perform more than one activity at the same time. However, in our annotation protocol, only a single behavior could be assigned to each bird per frame. In such cases, the behavior considered to be most biologically relevant was selected. In our dataset, this situation was associated with the behavior *Feeding*, which often co-occurred with locomotor behaviors such as *Walking* or *Swimming*. Based on input from ecological experts involved in the project, *Feeding* was prioritized due to its higher biological relevance. This behavior is closely linked to key ecological functions. Moreover, it serves as an indicator of habitat quality, as successful foraging reflects the availability of adequate food resources within the wetland environment. Moreover, *Feeding* can provide insights into species-specific foraging strategies and dietary preferences, which are valuable for ecological monitoring and conservation applications. In addition, *Feeding* behavior tends to be more behaviorally diverse and species-specific, thereby offering richer information for training models to distinguish fine-grained differences between species (a central contribution of the dataset). In contrast, behaviors such as *Walking* or *Swimming* are more ubiquitous and less discriminative across species. Finally, as *Feeding* occurred less frequently than other behaviors, prioritizing it in multi-behavior frames also contributed to improve class balance across the annotated data. Figure 5 shows an example of how *Feeding* is prioritized.

Animals often change among actions very fast, as a response to the changing environment. Thus, to consider a collection of movements of a bird as a behavior, this had to last a minimum of 30 frames, otherwise this

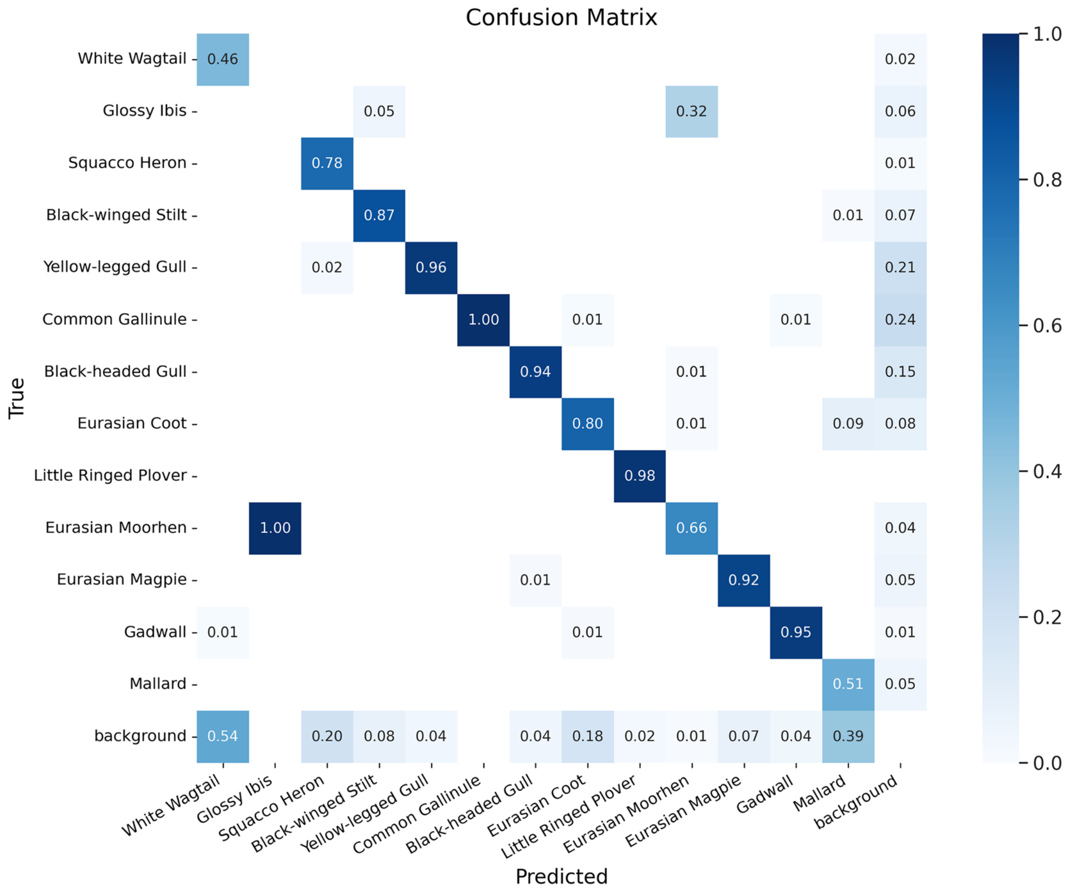


Fig. 8 Confusion matrix of species classification pipeline.

Model	Learning rate	Accuracy
MViT[47]	0.005	0.51
S3D[48]	0.005	0.29
SwinTransformer[49]	0.009	0.51
Video ResNet[50]	0.003	0.56
TimeSFormer[51]	0.001	0.49

Table 7. Results of the baseline models for behavior detection in terms of accuracy. The learning rate shown is the one that achieved the highest accuracies during hyperparameter tuning.

collection of movements was identified as a sub-movement of another main behavior, which is the one annotated for those frames. Moreover, this also facilitates the training of deep learning models, as very short behaviors are difficult to segment and classify due to the limited motion information they provide. In the annotation tool used (CVAT, as described in Data annotation subsection), a progress bar displaying the total number of frames and the current frame being viewed was provided. This feature made it easier for annotators to determine whether a behavior lasted at least 30 frames. Figure 6 shows an example of this annotation strategy.

Data Records

The dataset presented in this study is open access and accessible through Zenodo[43]. Within the Zenodo repository, there are five main elements:

- **Videos folder:** This folder contains the 178 videos that comprise the dataset. Videos are identified by their name, which is composed of a numeric value and the species that appears in the video. The format is the following "ID-VIDEO.SPECIES-NAME.mp4".
- **Bounding boxes CSV:** The *bounding_boxes.csv* file contains all the annotations of the dataset. It follows a format of 10 columns, ordered as follows: Global identifier of the row, video identifier, frame identifier within the video, activity identifier, subject identifier, species appearing in the video, and the four coordinates of the bounding box (top-left x-coordinate, top-left y-coordinate, bottom-right x-coordinate, and bottom-right

y-coordinate). Each of the CSV rows represents the information of one bounding box within one frame of a video.

- **Behavior identifiers CSV:** The *behavior_ID.csv* file contains a mapping of the seven behavior classes that make up the dataset and their numeric identifiers.
- **Species identifiers CSV:** The file *species_ID.csv* contains a mapping between the 13 different bird species and their numerical identifier.
- **Splits JSON:** The *splits.json* file contains the videos associated with each train, validation and test split.

Technical Validation

To ensure high quality recordings and accurate annotations, the entire process was carried out by expert ecologists. Ecologists used a semi-automated approach during the annotation process, as mentioned in the Data annotation section.

Firstly, video recordings were supervised by a group of experts who set up camera traps in strategic areas and also recorded some high-quality videos. For each video, these experts manually annotated the species appearing in the video. The same experts then manually corrected bounding box errors and annotated bird behavior, together with a number of collaborators with a background in ecology. Finally, a final stage of manual cross-checking of annotations was carried out by the experts and collaborators. The expertise of the annotators responsible for collecting and annotating the videos, together with the final cross-review process, ensures the quality and cleanliness of the data.

To qualitatively evaluate the quality of the annotations, we conducted an inter-annotator agreement assessment. This evaluation measures the consistency of the labeling criteria adopted by different annotators using three complementary metrics: Cohen's Kappa[44], Fleiss' Kappa[45], and the macro-averaged F1 score. Furthermore, as this dataset has been conceived mainly to be used in deep learning pipelines, baseline deep learning models trained on our dataset were developed. As mentioned previously, the purpose of this dataset is twofold, as it provides annotation data for performing bird species detection and behavior classification tasks. Thus, one baseline per each task was developed using PyTorch as coding platform.

Inter-annotator agreement assessment.

In order to evaluate the annotation consistency between annotators, three metrics were used in the assessment: Cohen's Kappa, Fleiss' Kappa, and the macro-averaged F1 score. Cohen's Kappa and Fleiss' Kappa are widely used metrics for assessing annotation quality in multi-annotator settings, as they quantify agreement beyond what would be expected by chance. While Cohen's Kappa measures the agreement between two annotators, Fleiss' Kappa generalizes this concept to more than two annotators, offering a single global measure of inter-annotator reliability. This is particularly relevant for our dataset, which was annotated by four individuals. In contrast, the macro-averaged F1 score measures the degree to which annotators consistently assign the same class labels, evaluating agreement on a per-class basis. We specifically use the macro version of the F1 score because it treats each class equally, thus mitigating the effects of class imbalance. Table 5 reports the results obtained using these metrics. Since both Cohen's Kappa and the macro F1 score are pairwise metrics, the table presents their average pairwise scores. For further information, Fig. 7 shows full pairwise agreement matrices.

The results reported in Table 5 indicate a high degree of annotation consistency. The average pairwise Cohen's Kappa (0,858) and the overall Fleiss' Kappa (0,855) suggest an excellent level of agreement among annotators. These results confirm that annotators followed a consistent set of criteria when labeling bird behaviors. Additionally, the macro-averaged F1 score of 0,946 highlights strong class-wise consistency, showing that annotators not only agreed in general but also consistently identified the same behavior categories across clips. This supports the reliability of the dataset for training and evaluating deep learning models.

Species classification.

As the dataset was primarily designed for training deep learning models, two baseline models were developed to evaluate its applicability. First, the baseline pipeline for species classification is introduced. This baseline is based on a YOLOv9[46] model trained over 50 epochs in the proposed dataset. YOLOv9 was selected due to its widespread use and strong reputation in object detection pipelines. The model is notable for its low inference times, making it suitable for real-time applications, while maintaining high accuracy across a wide range of scenarios.

Train, test and validation splits were generated from the full set of videos with a distribution 70-15-15. The splits were constructed using a stratified strategy based on the species and behaviors appearing in the videos. The distribution was computed by taking into account the number of frames that constitute each video (*e.g*., one video with 1,000 frames is equivalent to five videos with 200 frames). For efficient training, a downsampling of 10 is performed on the frames extracted from the videos. This can be done without affecting the performance of the model, as the difference between successive frames is minimal. The frames were extracted while maintaining the source FPS (Frames Per Second) of each video. During the training stage, a learning rate of 0.01 was used and a GeForce RTX 3090 GPU was used as the hardware platform. The test results from the baseline are shown next:

Table 6 shows the test results for species classification in terms of precision, recall, mAP50, and mAP50-95 metrics. mAP50-95 is a common object recognition metric that refers to the mAP (mean Average Precision) computed over 10 different IoU (Intersection over Union) thresholds, specifically from 0.50 to 0.95 in increments of 0.05. The results demonstrate that YOLOv9 achieves strong performance for the task, with a maximum precision of 0.835 and a high recall of 0.759. The mAP metrics, which evaluate the accuracy of bounding box localizations, also indicate robust performance, reaching 0.801 for mAP50 and 0.556 for mAP50-95. These are notable results, especially considering the challenges of achieving high mAP scores with stricter IoU thresholds.

To provide a more comprehensive understanding of the evaluation, the confusion matrix for the results is given. Figure 8 shows the confusion matrix, where it can be observed that the majority of the errors are due to the confusion of the ground truth class with the background class.

Behavior detection. Secondly, the behavior detection baseline is presented. In this baseline, four different video classification models were trained end-to-end to perform the behavior classification task. The trained models were Video MViT[47], Video S3D[48], Video SwinTransformer[49], Video ResNet[50], and TimeSFormer[51]. These models were selected due to their popularity for video classification tasks across a wide range of contexts, as well as their ease of use through the PyTorch and HuggingFace Transformers libraries, which facilitates the reproducibility of experiments. Moreover, the selected models are based on different architectures commonly used in computer vision: while Video S3D and Video ResNet rely on convolutional networks, Video MViT, Video SwinTransformer, and TimeSFormer are built upon the Transformer architecture as their fundamental building block. All model architectures and pretrained weights were extracted from PyTorch.

For the training, test and validation splits, the same distribution is used as for the species classification baseline. Input videos were downsampled with a downsample rate of 3, selecting the first frame as the representative of each set (*i.e.*, only the first frame of each set of 3 is kept). Regarding the training hyperparameters, a learning rate tuning was conducted using a uniform sampling strategy with minimum and maximum values of 0.0001 and 0.01, respectively. Similarly to the species classification baseline, training was performed on a GeForce RTX 3090 GPU. The results for each model are shown below:

From the Table 7 it can be concluded that the Video ResNet model is the one which learns better the complexity of the dataset, showcasing a maximum performance of 0.56. Conversely, the model with the lowest score is the S3D model, with an accuracy of 0.29. These results show the challenge posed by the dataset under study, which presents a limited amount of data. The limited amount of data available to train complex deep learning models demonstrates the need for more resources to capture more data. Furthermore, the development of new training strategies and deep learning architectures that fit the data needs should be explored in order to improve the baseline results obtained.

Usage Notes

Since the data annotations are provided in CSV format, it is recommended to use Python libraries such as Pandas, which is specifically designed to read and manage CSV data. In the official GitHub repository containing the code, there are usage examples of how to load and prepare the data to be fed into deep learning models. It is recommended to read the *dataset.py* script in the *behavior_detection* directory as an example.

Code availability

The data processing and experimentation code shown in the Technical Validation section is available on GitHub (https://github.com/3dperceptionlab/Visual-WetlandBirds). The GitHub repository is organized into two main directories. The *species_classification* directory contains all the code related to the species classification, and the *behavior_detection* directory contains the experiment with the behavior detection models proposed for the dataset.

Received: 28 January 2025; Accepted: 2 July 2025;
Published online: 11 July 2025

References

1. O'Riordan, T. *Environmental Science for Environmental Management* (Longman, 1995).
2. Nichols, J. D. & Williams, B. K. Monitoring for conservation. *Trends in ecology & evolution* **21**, 668–673, https://doi.org/10.1016/j.tree.2006.08.007 (2006).
3. Hays, G. C. *et al.* Translating marine animal tracking data into conservation policy and management. *Trends in ecology & evolution* **34**, 459–473, https://doi.org/10.1016/j.tree.2019.01.009 (2019).
4. Margules, C. & Usher, M. Criteria used in assessing wildlife conservation potential: a review. *Biological conservation* **21**, 79–109, https://doi.org/10.1016/0006-3207(81)90073-2 (1981).
5. Smallwood, K. S., Beyea, J. & Morrison, M. L. Using the best scientific data for endangered species conservation. *Environmental Management* **24**, 421–435, https://doi.org/10.1007/s002679900244 (1999).
6. Morrison, M. L. Bird populations as indicators of environmental change. In *Current Ornithology: Volume 3*, 429–451 (Springer, 1986).
7. Bonebrake, T. C., Christensen, J., Boggs, C. L. & Ehrlich, P. R. Population decline assessment, historical baselines, and conservation. *Conservation Letters* **3**, 371–378, https://doi.org/10.1111/j.1755-263X.2010.00139.x (2010).
8. Carvalho, S. B., Brito, J. C., Crespo, E. J. & Possingham, H. P. From climate change predictions to actions–conserving vulnerable animal groups in hotspots at a regional scale. *Global Change Biology* **16**, 3257–3270 (2010).
9. Joseph, L. N., Maloney, R. F. & Possingham, H. P. Optimal allocation of resources among threatened species: a project prioritization protocol. *Conservation biology* **23**, 328–338 (2009).
10. Nuttall, M. N. *et al.* Long-term monitoring of wildlife populations for protected area management in southeast asia. *Conservation Science and Practice* **4**, e614 (2022).
11. Fraixedas, S. *et al.* A state-of-the-art review on birds as indicators of biodiversity: Advances, challenges, and future directions. *Ecological Indicators* **118**, 106728 (2020).
12. Stastny, J., Munk, M. & Juranek, L. Automatic bird species recognition based on birds vocalization. *EURASIP Journal on Audio, Speech, and Music Processing* **2018**, 1–7, https://doi.org/10.1186/s13636-018-0143-7 (2018).
13. Vellinga, W.-P. & Planqué, R. The xeno-canto collection and its relation to sound recognition and classification. In *CLEF (Working Notes)*https://ceur-ws.org/Vol-1391/166-CR.pdf (2015).
14. Kahl, S. *et al.* Overview of birdclef 2023: Automated bird species identification in eastern africa. In *CLEF (Working Notes)*, 1934–1942 https://theses.hal.science/CIRAD/hal-04345437v1 (2023).
15. Miyaguchi, A., Cheung, A., Gustineli, M. & Kim, A. Transfer learning with pseudo multi-label birdcall classification for ds@gt birdclef 2024 Preprint at: https://arxiv.org/abs/2407.06291 (2024).

16. Morfi, V., Bas, Y., Pamuła, H., Glotin, H. & Stowell, D. Nips4bplus: a richly annotated birdsong audio dataset. *PeerJ Computer Science* **5**, e223 (2019).
17. Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S. & Bello, J. P. Birdvox-full-night: A dataset and benchmark for avian flight call detection. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 266–270 (IEEE, 2018).
18. Huang, Y.-P. & Basanta, H. Bird image retrieval and recognition using a deep learning platform. *IEEE Access* **7**, 66980–66989, https://doi.org/10.1109/ACCESS.2019.2918274 (2019).
19. M, C. Bird 525 species dataset Available on Hugging Face: A dataset containing bird species images. (2025).
20. Wah, C., Branson, S., Welinder, P., Perona, P. & Belongie, S. The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001, https://authors.library.caltech.edu/records/cvm3y-5hh21 (2011).
21. Van Horn, G. *et al.* Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 595–604, https://doi.org/10.1109/CVPR.2015.7298658 (2015).
22. Alados, C. L., Escos, J. M. & Emlen, J. Fractal structure of sequential behaviour patterns: an indicator of stress. *Animal Behaviour* **51**, 437–443 (1996).
23. Lindell, C. A. The value of animal behavior in evaluations of restoration success. *Restoration Ecology* **16**, 197–203 (2008).
24. Berger-Tal, O. *et al.* A systematic survey of the integration of animal behavior into conservation. *Conservation Biology* **30**, 744–753 (2016).
25. Goldenberg, S., Douglas-Hamilton, I., Daballen, D. & Wittemyer, G. Challenges of using behavior to monitor anthropogenic impacts on wildlife: a case study on illegal killing of african elephants. *Animal Conservation* **20**, 215–224 (2017).
26. Harvey, S. Deepbird: A deep learning pipeline for wildlife camera data analysis https://cs230.stanford.edu/projects_fall_2019/reports/26261732.pdf (2019).
27. Ng, X. L. *et al.* Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19023–19034 (2022).
28. Rose, P. *et al.* Evaluation of the time-activity budgets of captive ducks (anatidae) compared to wild counterparts. *Applied Animal Behaviour Science* **251**, 105626 (2022).
29. Mulero-Pérez, D., Ortiz-Perez, D., Benavent-Lledo, M., Garcia-Rodriguez, J. & Azorin-Lopez, J. Text-driven data augmentation tool for synthetic bird behavioural generation. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, 75–84 (Springer, 2024).
30. Fontúrbel, F. E., Orellana, J. I., Rodríguez-Gómez, G. B., Tabilo, C. A. & Castaño-Villa, G. J. Habitat disturbance can alter forest understory bird activity patterns: A regional-scale assessment with camera-traps. *Forest Ecology and Management* **479**, 118618 (2021).
31. Fontúrbel, F. E. *et al.* Sampling understory birds in different habitat types using point counts and camera traps. *Ecological Indicators* **119**, 106863, https://doi.org/10.1016/j.ecolind.2020.106863 (2020).
32. Murphy, A. J. *et al.* Using camera traps to examine distribution and occupancy trends of ground-dwelling rainforest birds in north-eastern madagascar. *Bird Conservation International* **28**, 567–580, https://doi.org/10.1017/S0959270917000107 (2018).
33. Arandjelovic, M., Stephens, C. & Diéguez González, P. *et al.* Highly precise community science annotations of video camera–trapped fauna in challenging environments. *Remote Sensing in Ecology and Conservation* **10**, 702–724, https://doi.org/10.1002/rse2.402 (2024).
34. Aljabri, M., AlAmir, M., AlGhamdi, M., Abdel-Mottaleb, M. & Collado-Mesa, F. Towards a better understanding of annotation tools for medical imaging: a survey. *Multimedia tools and applications* **81**, 25877–25911, https://doi.org/10.1007/s11042-022-12100-1 (2022).
35. Guillermo, M. *et al.* Implementation of automated annotation through mask rcnn object detection model in cvat using aws ec2 instance. In *2020 IEEE Region 10 Conference*, 708–713, https://doi.org/10.1109/TENCON50793.2020.9293906 (2020).
36. Kiyokawa, T., Tomochika, K., Takamatsu, J. & Ogasawara, T. Fully automated annotation with noise-masked visual markers for deep-learning-based object detection. *IEEE Robotics and Automation Letters* **4**, 1972–1977, https://doi.org/10.1109/LRA.2019.2899153 (2019).
37. Pavoni, G. *et al.* Taglab: Ai-assisted annotation for the fast and accurate semantic segmentation of coral reef orthoimages. *Journal of field robotics* **39**, 246–262, https://doi.org/10.1002/rob.22049 (2022).
38. Krenzer, A. *et al.* Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine* **21**, 33, https://doi.org/10.1186/s12938-022-01001-x (2022).
39. Li, H. *et al.* A semi-automated annotation algorithm based on weakly supervised learning for medical images. *Biocybernetics and Biomedical Engineering* **40**, 787–802, https://doi.org/10.1016/j.bbe.2020.03.005 (2020).
40. 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 https://arxiv.org/abs/2207.02696 (2022).
41. Sadli, R., Afkir, M., Hadid, A., Rivenq, A. & Taleb-Ahmed, A. Aggregated euclidean distances for a fast and robust real-time 3d-mot. *IEEE Sensors Journal* **21**, 21872–21884, https://doi.org/10.1109/JSEN.2021.3104390 (2021).
42. Li, A. *et al.* The ava-kinetics localized human actions video dataset https://arxiv.org/abs/2005.00214 (2020).
43. Rodriguez-Juan, J. *et al.* Visual wetlandbirds dataset: Bird species identification and behaviour recognition in videos, https://doi.org/10.5281/zenodo.15696105 (2025).
44. Kornblith, A. E. *et al.* Analyzing patient perspectives with large language models: a cross-sectional study of sentiment and thematic classification on exception from informed consent. *Scientific Reports* **15**, https://doi.org/10.1038/s41598-025-89996-w (2025).
45. Alzeer, H. M. *et al.* Validity and reliability of the arabic national nutrition plans checklist. *Scientific Reports* **15**, https://doi.org/10.1038/s41598-025-89928-8 (2025).
46. Wang, C.-Y. & Liao, H.-Y. M. YOLOv9: Learning what you want to learn using programmable gradient information https://arxiv.org/abs/2402.13616 (2024).
47. Li, Y. *et al.* Mvitv2: Improved multiscale vision transformers for classification and detection https://arxiv.org/abs/2112.01526 (2022).
48. Xie, S., Sun, C., Huang, J., Tu, Z. & Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification https://arxiv.org/abs/1712.04851 (2018).
49. Liu, Z. *et al.* Video swin transformer https://arxiv.org/abs/2106.13230 (2021).
50. Tran, D. *et al.* A closer look at spatiotemporal convolutions for action recognition https://arxiv.org/abs/1711.11248 (2018).
51. Bertasius, G., Wang, H. & Torresani, L. Is space-time attention all you need for video understanding? https://arxiv.org/abs/2102.05095 (2021).

Acknowledgements

We would like to thank “A way of making Europe” European Regional Development Fund (ERDF) and MCIN/AEI/10.13039/501100011033 for supporting this work under the “CHAN-TWIN” project (grant TED2021-130890B- C21 and HORIZON-MSCA-2021-SE-0 action number: 101086387, REMARKABLE, Rural Environmental Monitoring via ultra wide-ARea networKs And distriButed federated Learning. This work is part of the HELEADE project (TSI-100121-2024-24), funded by Spanish Ministry of Digital Processing and by the European Union NextGeneration EU. This work has also been supported by three Spanish national and two regional grants for PhD studies, FPU21/00414, FPU22/04200, FPU23/00532, CIACIF/2021/430 and CIACIF/2022/175.

Author contributions

J.R.J. conceived the manuscript, J.R.J., D.O.P., M.B.L., D.M.P. and P.R.P. conducted the experiments and analysed the results, A.O.T. and E.S.G. managed the data collection and annotation, J.G.R. led the project management. All authors contributed to the data annotation review and manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.R.-J. or J.G.-R.

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.