From the Department of Interdisciplinary Life Sciences
the University of Veterinary Medicine Vienna
(Head: Univ.-Prof. Dott. Leonida Fusani, MPhil PhD)

Automated identification of individual Spotted Bowerbirds (*Chlamydera maculata*)

Master's Thesis

Submitted by
Sarah Paola Juárez Jerez (12204530)

Supervised by
Univ.-Prof. Dott. Leonida Fusani, MPhil PhD
Dr. rer. nat. Cliodhna Quigley
Job Knoester, MSc

Declaration of originality

It is hereby confirmed that no resources or literature other than those cited in this thesis have been used. The essential parts of the work were carried out by the author. This thesis was written independently and has not been submitted or published, in whole or in part, elsewhere.

Vienna, September 2025

Table of Contents

List of Figures

List of Equations

List of Tables

Abstract

Automating the identification of individual animals in the wild remains a significant challenge in behavioural ecology. In this study, we developed a machine learning pipeline to identify individual Spotted Bowerbirds (*Chlamydera maculata*) from camera trap video footage recorded in natural field conditions. We trained a ResNet50 Convolutional Neural Network (CNN) on automatically detected and segmented instances of 16 individual Bowerbirds and achieved a mean F1-score of 0.98 on the validation set. The model's high accuracy provides empirical evidence for consistent, learnable visual differences between individuals of this species, which have not been quantitatively described before.

To assess the data efficiency of our approach, we evaluated model performance across a range of training set sizes. We found that a modest dataset of 350 instances per individual was sufficient to reach the target F1-score of ≥0.85. We also trained a ResNet18 CNN to classify four viewpoints: front, back, left side, and right side, with an F1-score of 0.8891. We observed that using viewpoint-specific datasets reduced the data requirement to approximately 150 instances per bird. These findings highlight how considering the visibility of key identifying features can improve model efficiency, i.e. note that the spotted patterns that give the species its name are on the birds' back strip and wings, and are better visible from certain viewpoints. We demonstrated that training on viewpoint-specific data not only enhanced performance but also reduced the number of required instances, demonstrating the value of task-specific design for scalable identification, especially when data is scarce.

Our approach enables scalable, automated individual identification, which can reduce the need for the time- and labour-intensive manual classification.

1. Introduction

The study of a species behaviour often relies on the study of individual subjects, which requires researchers to identify and track specific animals over time [1]. Traditional methods of individual identification include the use of bands, tags, or natural markings [2, 3].

Video recordings from camera traps have been increasingly used for non-invasive collection of behavioural data in ornithological studies [4, 5, 6, 7, 8]. However, processing the large volumes of video data generated by these methods is time- and labour-intensive [9]. Additionally, manually analysing videos to identify individuals from their natural or artificial markers can be challenging due to physical occlusions, lighting changes and poor image quality.

Machine learning techniques have been used to identify animals from images and videos [3]. For instance, Ferreira et al. [10] used automatically labelled images of birds from multiple species to train a Convolutional Neural Network (CNN) for individual identification. Their method relied on a Radio Frequency Identification (RFID)-based setup where tagged birds triggered the camera by landing on a perch near an RFID antenna. This setup automatically embedded the individual's identity into each captured image, eliminating the need for manual labelling. It also ensured consistent image framing, as birds had to approach the camera at a fixed distance to trigger recording. However, such controlled setup is impractical in the wild, and capturing still frames instead of videos limits the observation of animals' behaviour. Thus, a significant gap remains in the ability to identify individuals from video footage of animals exhibiting natural behaviours in their natural habitat. Developing robust automated methods applicable to such data is crucial for advancing behavioural research.

The Spotted Bowerbird (*Chlamydera maculata*) serves as an excellent model species for developing and validating automated individual identification methods in natural settings. Males build and decorate structures known as bowers, which are crucial for their overall reproductive success [11, 12]. These bowers act as natural focal points ideal for data collection through camera traps. Moreover, Bowerbirds' complex courtship displays and their interactions around the bower are individually variable and interesting from a behavioural perspective [11, 12, 13].

In this study, we created a pipeline for automated identification of individual Spotted Bowerbirds using camera trap video data collected in the wild, with four objectives: (i) to train a CNN to identify individual Spotted Bowerbirds from camera trap footage with high accuracy (F1-score ≥ 0.85); (ii) to evaluate the data efficiency of the approach by determining the minimal amount of training data required per individual to achieve this performance; (iii) to assess feature importance by training the classifier on four different viewpoints, i.e. frontal, lateral (left and right), and dorsal, to identify which features most impact the model's performance; and (iv) to test whether restricting training and validation data to only the most informative viewpoint could reduce the minimum number of frames required per individual, compared to the models trained on all viewpoints.

2. Methods

2.1. Data pre-processing

2.1.1. Dataset

The dataset was collected by Dr. Giovanni Spezie (former PhD candidate at the University of Veterinary Medicine Vienna, supervised by Prof. Leonida Fusani) in Taunton National Park (Scientific), Queensland, Australia, during the 2021 breeding season, between July and November. Data were gathered using 17 motion-triggered camera traps (Browning, Recon Force Edge) positioned to monitor active bowers owned by previously banded birds, with one camera trap per bower. Cameras were typically placed at 1.5 meters from the bower, secured to a tree trunk and on a tripod approximately 30 centimetres from the ground, angled to capture activity at the bower platform and avenue, as observed in Figure 1.

Figure 1. Camera (a) set up at a bower (b) site (Photo by J. Knoester).

Video recordings were triggered by motion with a 1 second delay between triggers. The dataset comprised 25,234 scored videos, with individual video durations ranging from 30 seconds to 2 minutes. The videos were in the MP4 format, with a resolution of 1920×1080 pixels and a frame rate of 30 frames per second (fps). A total of 32 uniquely banded individuals were recorded across all videos.

Ethical approval for this study was obtained from the Animal Ethics Committee of the Department of Agriculture and Fisheries (AEC reference number: CA 2021/04/1496); field activities at Taunton National Park (Scientific) were approved by Queensland Wildlife and Parks Service (P-PTUKI-100095367; P-PENPS-100095369), and banding of the birds was approved by the Australian Bird and Bat Banding Scheme under the R-class banding (authority numbers: 3374).

2.1.2. Video sampling and frame extraction

All recorded videos were visually inspected to document whether they contained visible birds, whether they were banded or unbanded, and to determine the identity of banded individuals. Videos were subsequently filtered to include only those featuring a single owner bowerbird. An owner bowerbird refers to a male bowerbird actively building and maintaining the bower structure being recorded, distinct from other bowerbird individuals that could occasionally be captured in video, e.g., females, juveniles, or transient males.

This filtering thereby excluded videos featuring multiple individuals (banded or unbanded) (Figure 2a), no birds at all (Figure 2b), single unbanded individuals (Figure 2c), individuals of different species (Figure 2d), and single banded non-owners. The resulting video bank consisted of 24,239 videos of 16 banded owner bowerbirds. It should be noted that, despite this filtering, some of the resulting videos still captured non-target individuals (Figure 2d). For practicality, these videos were not manually removed from the dataset.

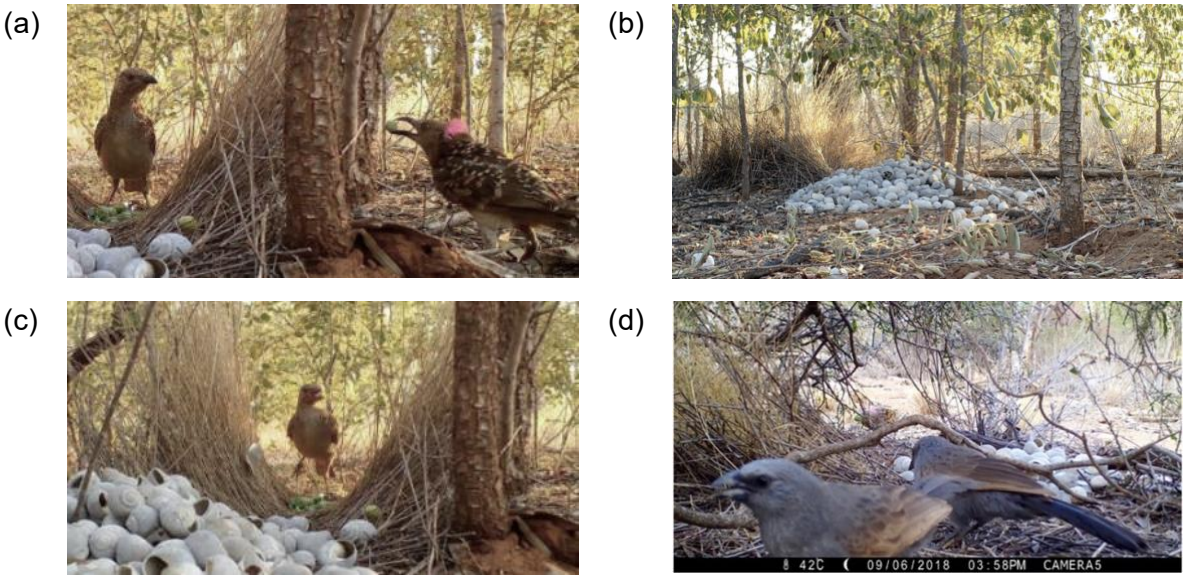


Figure 2. Examples of video recordings that fit the exclusion criteria.

For subsequent analysis, frames were extracted from the filtered videos. Before frame extraction, a 10% subset of the videos from each bird was held out for testing. Then, the remaining videos were processed using the OpenCV library to extract frames at intervals of 30 frames, or one frame every second, to obtain a representative sample of frames from each video.

2.1.3. Object detection

A multistage image processing pipeline was applied to the extracted raw frames aiming to standardise the size and position of the bird within the frames. First, a pre-trained YOLOv11 (Ultralytics) model was used to detect the birds in each frame (Figure *3*a). For frames in which a bird was detected, the frame was cropped to the bounding box with the highest confidence score, isolating the detected bird (Figure *3*b). Frames in which no bird was detected were excluded from further analysis.

(a) 

(b) 

Figure 3. Automated detection of individuals on a raw frame. (a) Raw frame. (b) Closeup view of the automatic detection within bounding box, predicted class label, e.g., "bird", and confidence score, e.g., 0.89.

YOLOv11 is a state-of-the-art single-pass CNN that offers an integrated framework for object detection and instance segmentation. It features fewer parameters compared to earlier iterations such as YOLOv8, which makes it more computationally efficient and faster [13]. Furthermore, YOLO models were pre-trained on the ImageNet dataset [14] and can generalise to novel object instances and imaging conditions without domain-specific retraining [13].

In this study, the performance of the pre-trained YOLOv11 model was qualitatively assessed on a randomly selected subset of the extracted frames. This subset comprised 200 randomly selected instances of each owner bowerbird. These instances were collected sequentially from randomly chosen videos until the target number of instances per bird was reached. Based on this assessment, it was determined that the predicted detections were sufficiently accurate for the research objectives, and the model was used for detection without retraining. The model's detection confidence threshold was empirically set to 0.8, after qualitative evaluation of the predicted bounding boxes at various confidence thresholds, e.g. 0.4-0.9, as no manually annotated bowerbird instances were available for quantitative evaluation.

2.1.4. Mask segmentation

Within the detected bounding box, the same YOLOv11 model was applied for mask segmentation, to generate a pixel-wise mask of each bird (Figure *4*).

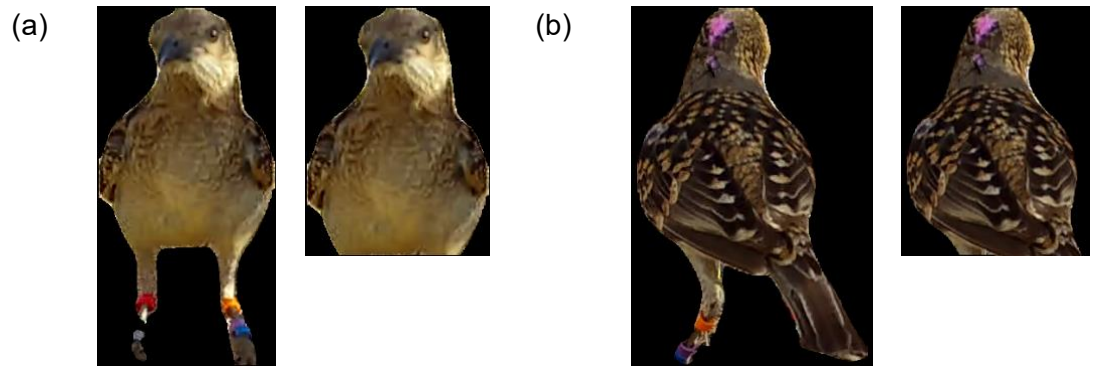


Figure 4. Examples of segmented instances before (left image in each pair) and after (right image in each pair) automatic leg removal. (a) Frontal view of an individual, (b) Dorsal view of the same individual.

YOLOv11's segmentation confidence threshold was set to 0.6 after qualitative evaluation of the predicted masks at various confidence thresholds, e.g. 0.4-0.9, as no manually annotated bowerbird instances were available for quantitative evaluation. This threshold minimised false positives, i.e. detection of portions of the background as birds, and false negatives, i.e. failing to detect the birds.

To remove noise from the instance segmentation, contiguous pixel regions below a certain threshold were discarded through connected component analysis. Frames where no pixel regions remained after this initial filtering were also discarded. Then, each mask was processed to digitally remove the leg bands, to prevent the classifier from overfitting the training data. This was done by iterating through each pixel row in the lower one-third portion of the mask, i.e. the fraction of the mask expected to contain the birds' legs, to identify narrow vertical structures, i.e. with a width less than or equal to 100 pixels. Pixel rows containing such structures were entirely cropped from the image.

The final dataset, consisting of the processed masks for each individual bird, was randomly split into training and validation subsets, with a 70:30 split (Scikit-learn library v1.3.0), and used to train the individual classifier model.

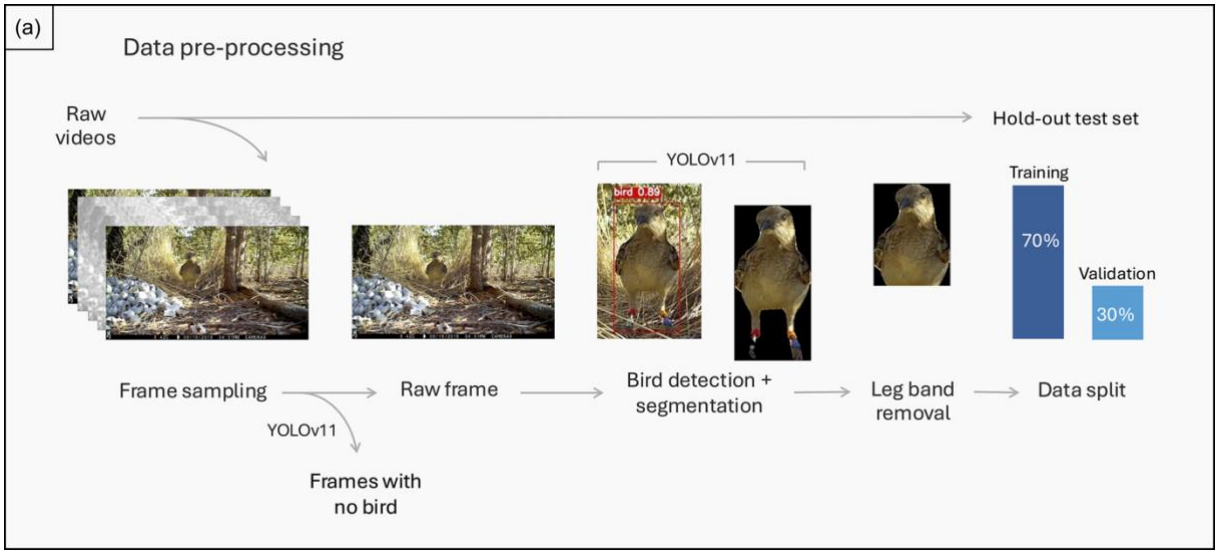
2.2. Individual bird classification

2.2.1. Establishing individual ID

Ground truth for individual identity was obtained from the scoring of all recorded videos, performed by other researchers as part of the larger project. To obtain enough training and validation data per class, only birds with ≥2,000 instances were included in the classification, resulting in 16 birds. The feasibility of visual discrimination between Spotted Bowerbird individuals by human observers was demonstrated in a bachelor's thesis by Reischle [16].

2.2.2. Individual classifier model configuration

We used a ResNet50 deep CNN as the individual classifier, with a transfer learning approach. Residual Networks (ResNets) are deep neural networks that have shown superior performance in classification tasks, as they able to learn residual functions with reference to the layer inputs, rather than learning unreferenced functions directly. This architecture makes the networks easier to optimise and allows for accuracy gains from the considerably increased network depth [15]. ResNet50 is a variant of the ResNet architecture that features 50 layers and has a robust capacity for extracting highly abstract and hierarchical features from images, and achieve better classification performance compared to shallower ResNets, while being less computationally expensive and more practical [15]. An overview of the data processing workflow and the ResNet50 model architecture is provided in Figure *5*b.



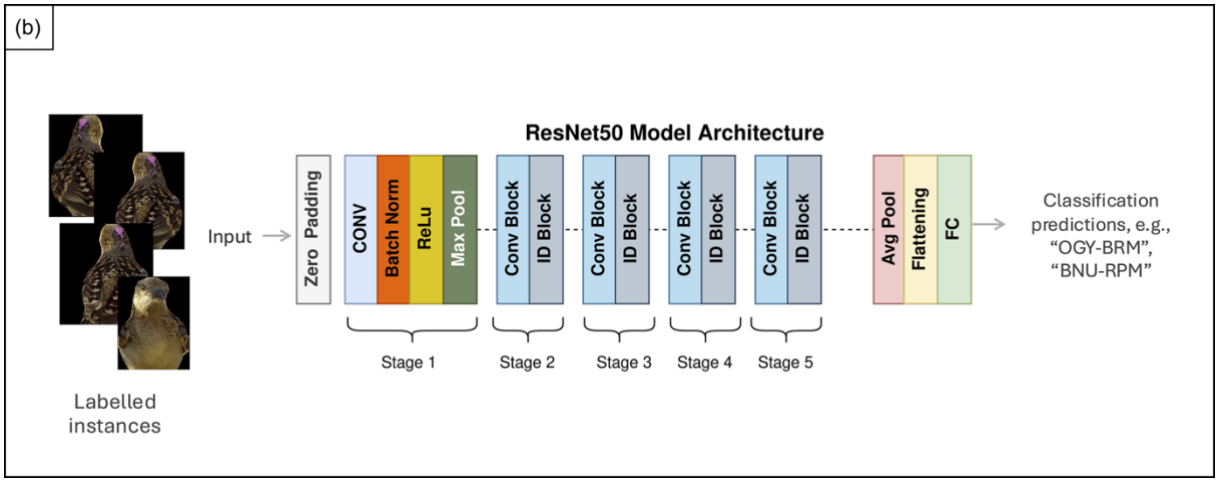


Figure 5. Data pre-processing for individual classification modelling (a) and individual classification pipeline using ResNet50 (b).

Our ResNet50 individual classifier was implemented in PyTorch (1.13.1) and Torchvision (0.14.1). Training was conducted on a computer node from the Life Science Compute Cluster (LiSC) with 8 CPU cores and 48 GB of RAM, equipped with an NVIDIA Tesla T4 GPU (NVIDIA Corporation). The system operated under Rocky Linux 9.5 (Blue Onyx), with CUDA support enabled for GPU acceleration.

The model was initialised with weights pre-trained on the ImageNet dataset. The final layer was replaced with a classification head to output 16 classes, corresponding to the individual birds, and used a softmax activation function for probability distribution across classes. Input images were resized to 512×512 pixels and normalized using the standard ImageNet mean and standard deviation values. During training, data augmentation was applied in the form of random horizontal flips (probability = 0.5). The model was trained using Stochastic Gradient Descent with a momentum of 0.9, an initial learning rate of 1×10−3, and a batch size of 32. The learning rate was reduced by a factor of 0.1 every 7 epochs.

The model was trained for a total of 10 epochs, with performance on the validation set monitored after each epoch to select the best performing model. The entire training process required 608 minutes (~10 hours). The selected model was then used for final evaluation on a test set, obtained from the held-out test videos, corresponding to 10% of the original videos.

2.2.3. Evaluation of the individual classifier

At inference, our ResNet-50 model produced, for each frame, a vector of class scores, and the final prediction for that frame corresponded to the class with the highest score. We evaluated the performance of our individual classifier, i.e., the accuracy of its predictions, through the F1-score metric (Equation *3*), which is calculated as the harmonic mean of precision and recall [16]. Precision (Equation *1*) is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP), and recall (Equation *2*), also known as sensitivity, is the ratio of true positives (TP) to the sum of true positives and false negatives (FN).

Equation 1. Precision.

$$Precision = \frac{TP}{TP + FP}$$

Equation 2. Recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Equation 3. F1-score.

$$\text{F1 score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

We established an F1-score performance cutoff of 0.85 as a study-specific benchmark, as no single cutoff can be applied across studies [17]. Classification performance was further visualised using a confusion matrix, which graphically shows the counts of true positives, true negatives, false positives, and false negatives across all classes.

2.2.3.1. Per-frame evaluation on the validation and test sets

For a general evaluation of the model's performance, the model's predictions were defined at the frame level, which meant each frame was an independent prediction instance. For the validation set, the model generated predictions for all validation frames, corresponding to 30%

of all frames randomly sampled from the non-test pool of videos of each bird. For the test set, the model generated predictions for all frames extracted from the held-out test videos of each bird. For both sets, we computed per-class precision, recall, and F1-score from the per-frame counts of TP, FP, and FN.

2.2.3.2. Per-video evaluation on the test set

To provide a more realistic estimate of the model's reliability in real-world conditions, we evaluated the individual classifier on the test set at the video level. First, we extracted all frames of each video in the held-out test set, i.e. 10% of all videos of a single bird, and used the ResNet50 individual classifier to generate per-frame predicitons. Then, a single identity was assigned to the entire video using majority voting, which is a decision fusion technique that returns the value that appears most often in a set [20]. In this context, the predicted identity for a video corresponded to the class most often predicted across all frames. The resulting video-level predictions were compared to the ground truth identities established during manual scoring, and classification performance was assessed using the F1-score metric, calculated across all videos for each bird.

To further examine whether classification performance varied across viewpoints, we carried out a viewpoint-specific analysis of the video-level predictions. We used the ResNet18 viewpoint classifier to determine the viewpoint visible on each frame of the held-out test videos. Then, we used majority voting within each video and within each viewpoint to generate viewpoint-specific identity predictions. These were compared against the ground truth labels to assess whether certain viewpoints resulted in more accurate identity classifications. This analysis aimed to inform future data collection and model training strategies.

2.3. Viewpoint classification

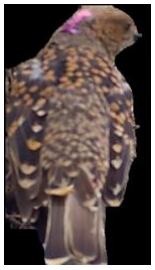
To assess which visual features contribute most to individual classification, we defined four viewpoints, i.e., front, back, left side, and right side, based on the morphological traits visible from each viewpoint. We first tested whether these viewpoints could be reliably distinguished through an inter-observer reliability (IOR) test involving two annotators. Then, one of the observers annotated a larger dataset, which was used to train and evaluate a viewpoint classification model.

2.3.1. Viewpoint labelling

To quantify the feasibility of reliably annotating these viewpoints, we measured the agreement between two independent observers on a randomly selected set of 400 instances. We calculated the percentage of raw agreement, accounting for agreement expected by chance by computing Cohen's Kappa coefficient [21], which is the most widely used measure for inter-rater reliability in categorical data involving two raters. We applied the unweighted version of this coefficient [22], as there is no inherent order to the viewpoint classes.

To prepare the training and validation data for the viewpoint classifier, we randomly sampled 3000 pre-processed instances, showing only cropped and masked birds, by selecting a uniform number of instances per individual. These instances were manually annotated based on the viewpoint labelling guide (Table *1*). The image labelling was performed in Label Studio, which is an open-source, web-based tool for custom data labelling (https://labelstud.io/). The labelled dataset was split into training, validation, and test sets, with a 70:20:10 split (Scikit-learn library v1.3.0).

Table 1. Viewpoint definition and labelling guide.

Viewpoint	Front	Back	Left side	Right side
Visible features	*Breast* *Belly*	*Tail* *Back* *Wings (at least partly visible)*	*Left wing* *Rump (partly visible)*	*Right wing* *Rump (partly visible)*
Hidden features	*Tail* *Rump* *Wings (mostly or entirely hidden)*	*Breast* *Belly*	*Right wing*	*Left wing*
Example				

2.3.2. Viewpoint classifier model configuration

We used a ResNet50 deep CNN as the viewpoint classifier, with a transfer learning approach. ResNet18 is an 18-layer variant of the ResNet architecture (described in Section 2.2.2.). The classifier was implemented in PyTorch (1.13.1) and Torchvision (0.14.1). Training was conducted on a computer node from the Life Science Compute Cluster (LiSC) with 8 CPU cores and 48 GB RAM, equipped with an NVIDIA Tesla T4 GPU (NVIDIA Corporation). The system operated under Rocky Linux 9.5 (Blue Onyx) with CUDA enabled for GPU acceleration.

The network was initialised with ImageNet-pretrained weights (IMAGENET1K_V1). The final layer was replaced with a classification head to output four classes, one for each viewpoint, using a softmax output for class probabilities. Input images were resized to 224×224 pixels and normalized using the standard ImageNet mean and standard deviation values. During training, data augmentation was applied in the form of random rotation (±7°) and colour jitter (brightness, contrast, and saturation ±0.15; hue ±0.05)

The model was trained using Stochastic Gradient Descent with a momentum of 0.9, an initial learning rate of 1×10−3, and a batch size of 32. The learning rate was reduced by a factor of 0.1 every 7 epochs. The model was trained for a total of 100 epochs, and performance on the validation set was monitored after each epoch.

2.3.3. Evaluation of the viewpoint classifier

The viewpoint classifier was applied to all extracted frames, across all individuals. From those predictions, 200 instances per viewpoint were randomly sampled for human validation, totalling 800 instances. These instances were independently labelled by two human annotators, and agreement between the annotators, and between the viewpoint classifier and each annotator was measured. The model was also tested for agreement with the human consensus, i.e. instances with identical labels from both annotators.

The performance of the classifier was assessed through the percentage of raw agreement and Cohen's Kappa coefficient [21] between each pair of raters (annotator–annotator, annotator–classifier, and human-consensus–machine).

2.4. Minimal data requirement

We compared the performance of fifteen individual classifiers trained and validated on increasingly large data subsets. Subsets contained 50-1000 instances per bird, with increments of 50 instances for subsets with up to 500 instances (50, 100, 150, 200, 250, 300, 350, 400, 450, 500), and subsequent increments of 100 instances for subsets with up to 1000 instances (600, 700, 800, 900, and 1000). These amounts refer to the total number of instances available for a single bird, before training, validation, and test split (70:20:10) (Scikit-learn library v1.3.0). Each subset was sampled randomly and independently. Thus, smaller subsets were not explicitly contained within the larger ones, but instances could overlap across subsets randomly. To isolate the impact of viewpoint on the data efficiency of individual classification, we conducted two experiments. In the first, subsets were created from the full dataset, regardless of viewpoint. In the second, subsets were created within each viewpoint class. The minimal data requirement was defined as the smallest subset size for which a model achieved an F1-score ≥ 0.85 on the validation set. For all experiments in this section, models were trained for 20 epochs (baseline model: 10 epochs) to account for slower learning on smaller datasets. However, performance plateaued after around 5 epochs, even for the smallest subsets.

2.4.1. Minimal data requirement in baseline dataset (all-viewpoints)

In this experiment, subsets were randomly created from the full baseline dataset. Thus, each subset contained a mixture of viewpoints, as present in the original data distribution. Additionally, contrary to the training of the baseline model, these subsets were created from the whole of the valid videos, without video hold-out for testing. Training, validation and test instances could therefore come from any of the videos.

2.4.2. Minimal data requirement across viewpoints

In this experiment, subsets were created from each viewpoint, i.e., front, back, left side, right side, and side view (left + right). For each viewpoint-specific dataset, subsets of 50–1000 instances per bird were sampled following the same increments and random-splitting described in section 2.4. Thus, each subset contained only images from a single viewpoint, and training, validation, and test instances were randomly sampled from that viewpoint’s pool of frames.

3. Results

3.1. Individual classification

3.1.1. Frame-level individual classification on validation and test sets

This section presents the performance of the individual classifier trained on the baseline dataset, consisting of 62,198 training instances, and 26,668 validation instances. The model was tested on all frames from the held-out test videos, totalling 222,227 instances, on a frame-level and on a video-level. The model achieved a mean F1-score of 0.968 on the validation set, and of 0.86 on the held-out test set (Table *2*).

Table 2. Dataset characteristics and performance (F1-score) of the individual classifier.

Bird individual ID	Train set	Validation set	Test set	F1-score		
				Validation	Test (Frame-level)	Test (Video-level)
BNU-RPM	9916	4250	33276	0.99	0.97	0.98
BNY-RPM	595	255	2686	0.92	0.83	0.89
BRG-YOM	12971	5560	48033	0.99	0.98	0.97
BRK-NOM	1298	557	3441	0.90	0.82	0.84
EYB-RPM	8326	3569	28112	0.99	0.97	0.98
GBM-ORY	6227	2670	18444	0.97	0.91	0.95
GBY-ORM	2831	1214	9356	0.98	0.92	0.94
OEB-RPM	3774	1618	16398	0.99	0.90	0.94
OGY-BRM	3917	1680	18931	0.99	0.91	0.94
ORB-UYM	1738	745	9564	0.97	0.85	0.89
OUB-RPM	1417	608	4323	0.97	0.83	0.87
OYR-BGM	1790	768	6726	0.96	0.89	0.92
RGY-BOM	589	253	1805	0.90	0.47	0.60
RYO-BGM	2874	1233	7509	0.98	0.82	0.92
YM-OBR	1903	816	11565	0.98	0.94	0.94
YRU-POM	1325	569	2058	0.98	0.84	0.87
Total	62,198	26,668	222,227	$\bar{x}$ = 0.98	$\bar{x}$ = 0.86	$\bar{x}$ = 0.90

Classes with fewer training instances, e.g., BNY-RPM, BRK-NOM and RGY-BOM, showed slightly lower F1-scores on the validation set (0.92, 0.90 and 0.90, respectively) and considerably lower scores on the test set. This performance drop was observed both at the frame-level (0.83, 0.82, and 0.47, respectively) and at the video-level (0.89, 0.84, and 0.60, respectively).

3.2. Viewpoint classification

3.2.1. Viewpoint labelling reliability

The inter-observer reliability analysis showed strong agreement between annotators, with 87.25% raw agreement and a Cohen's Kappa of 0.83, indicating almost perfect agreement [23].

The confusion matrix (Figure *6*) shows that most disagreements occurred between adjacent viewpoints, especially between the "back" and "left side" (19 times) or "right side" (11 times). Similarly, the front viewpoint was occasionally confused with "left side" (3 instances) and "right side" (3 instances). In contrast, disagreement was minimal for frontal and side views.

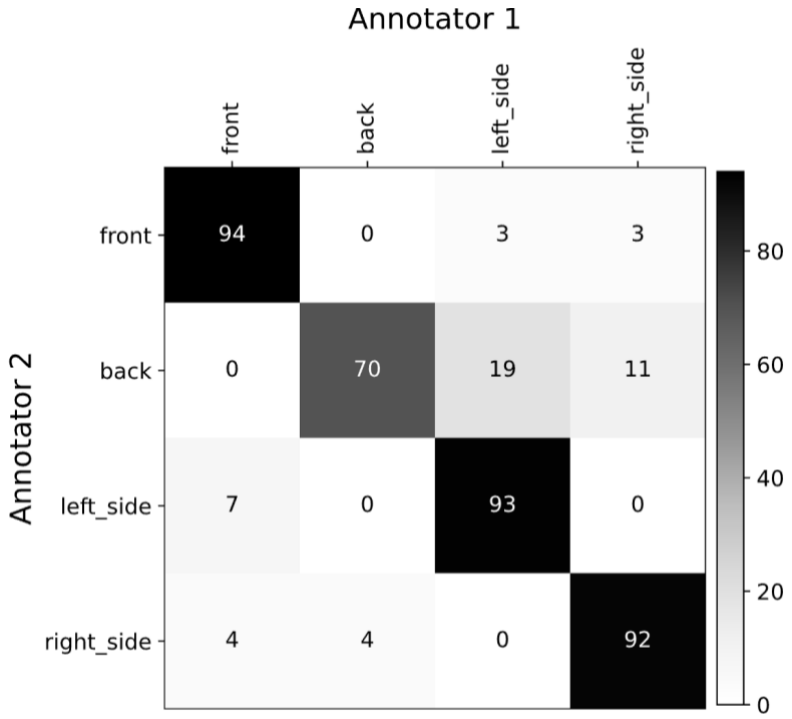


Figure 6. Confusion matrix illustrating viewpoint labelling agreement between annotators.

3.2.2. Viewpoint classifier performance

During viewpoint annotation, frames incorrectly detected, or mask were excluded, resulting in 2,946 labelled instances: back (921), front (569), left side (750), and right side (706), which were split into training, validation and test sets, with a 30:20:10 split (Table *3*).

Table 3. Viewpoint dataset characteristics and viewpoint classifier performance.

Viewpoint	Train set	Validation set	Test set	F1-score (test set)
Front	398	113	58	0.93
Back	644	84	93	0.86
Left side	525	150	75	0.88
Right side	494	141	71	0.88

The model achieved high classification performance on the test set. The front viewpoint was classified most accurately (0.93 F1 score), despite considerably fewer training instances, likely due to the more distinct features. The confusion matrix (Figure *7*) showed few misclassifications, mostly between adjacent viewpoints, such as back and side views, probably due to the similarity in the plumage appearance between the two.

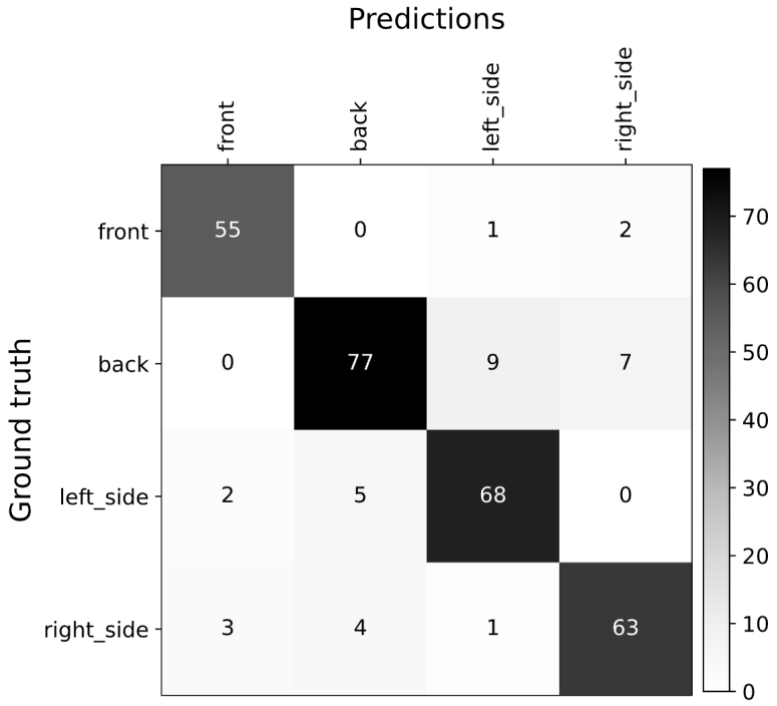


Figure 7. Confusion matrix illustrating ResNet18's viewpoint classification performance.

3.3 Minimal data requirement

3.3.1. Minimal data requirement in baseline dataset (all-viewpoints)

Figure *8* presents the model performance on the validation set across increasing subset sizes. There was a general trend of increasing F1-score with larger subset sizes, peaking at 0.89 with a subset of 350 instances. Beyond this point, the F1-score continued to improve but the rate of improvement diminished until reaching 0.95 F1-score at 900 instances.

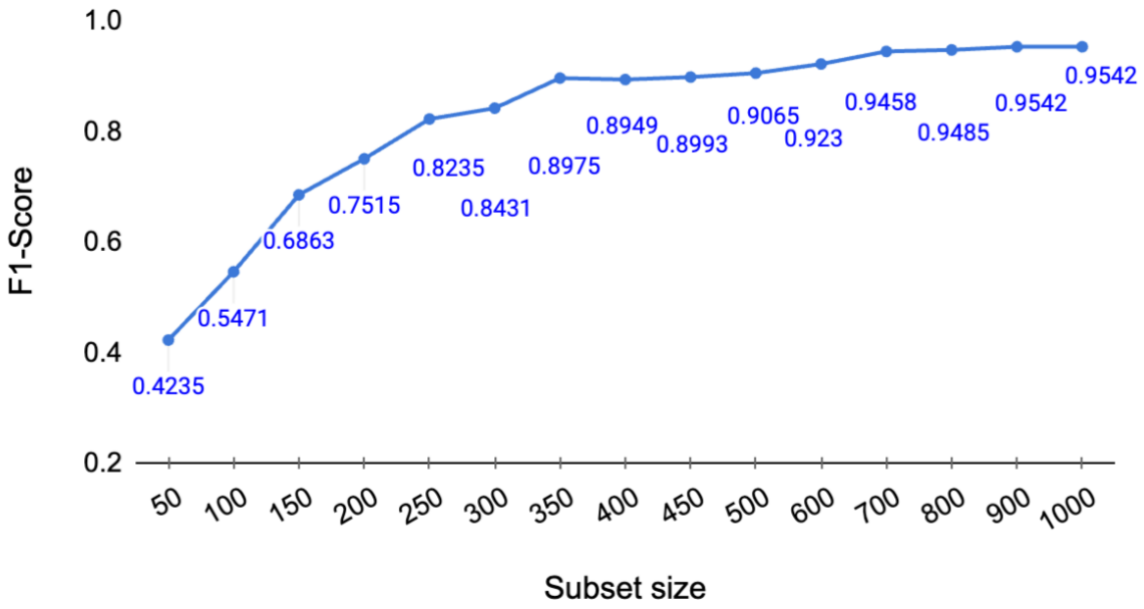


Figure 8. Validation classification performance (F1-score) across increasing subset sizes.

3.3.2. Minimal data requirement across viewpoints

Table *4* and Figure *9* compare the performance of models trained on 15 subsets with an increasing amount of training and validation data, when subsets were created within each viewpoint, against the model trained on all-viewpoints. The trend of increasing performance with more instances was observed across all models, regardless of viewpoint. Classifiers trained on instances of a single viewpoint required between 100 and 150 instances per bird to reach an F1-score of 0.85. In contrast, the classifier trained on data combining all viewpoints required significantly more data, about 350 instances per bird, to achieve the same level of accuracy. Furthermore, the classifiers trained on viewpoint-specific data achieved higher final F1-scores (above 0.97) when trained with larger datasets compared to the model trained on all viewpoints, which reached a maximum F1-score of 0.95.

Table 4. Individual classification performance (F1-score) across subset sizes, per viewpoint.

Subset size	Front	Back	Left side	Right side	Side view	All viewpoints
50	0.72	0.82	0.76	0.83	0.79	0.42
100	0.83	0.88	0.90	0.87	0.84	0.55
150	0.89	0.93	0.92	0.91	0.89	0.69
200	0.92	0.94	0.94	0.94	0.91	0.75
250	0.93	0.96	0.94	0.94	0.93	0.82
300	0.95	0.96	0.96	0.96	0.94	0.84
350	0.94	0.96	0.97	0.96	0.94	0.90
400	0.95	0.96	0.97	0.97	0.95	0.89
450	0.95	0.97	0.96	0.96	0.96	0.90
500	0.96	0.98	0.97	0.98	0.96	0.91
600	0.98	0.98	0.97	0.98	0.97	0.92
700	0.97	0.97	0.98	0.98	0.97	0.95
800	0.97	0.98	0.98	0.98	0.97	0.95
900	0.98	0.99	0.99	0.99	0.97	0.95
1000	0.98	0.99	0.99	0.99	0.98	0.95

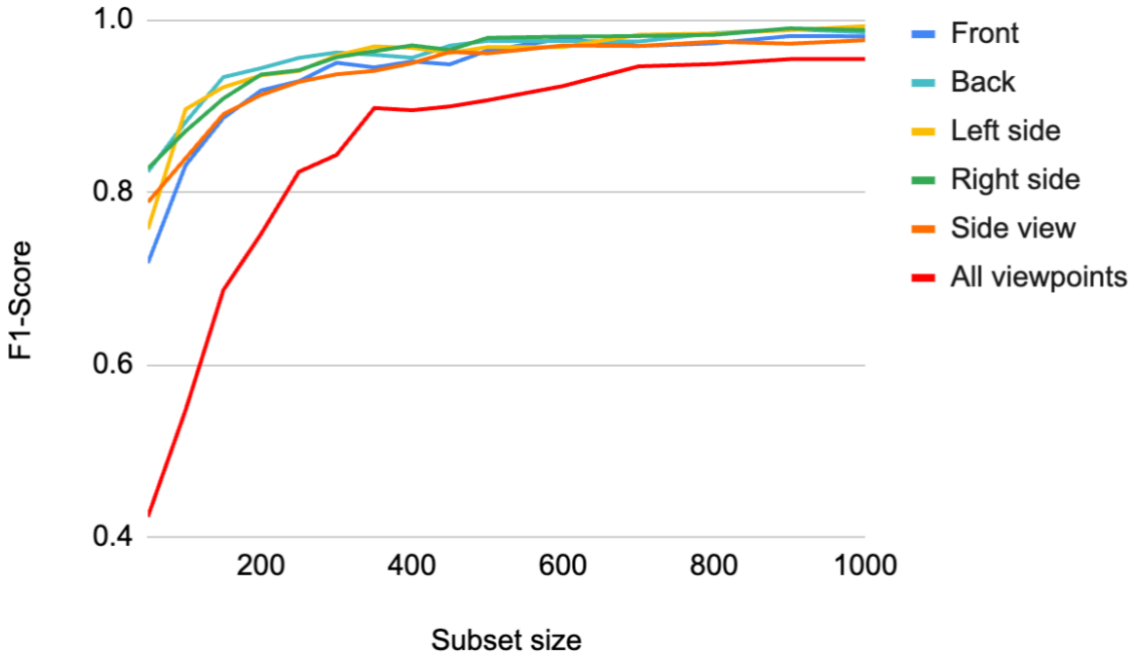


Figure 9. Individual classification performance (F1-score) on the validation set, across increasing subset sizes, by viewpoint.

Among the viewpoint-specific models, the classifier trained on the front viewpoint performed consistently worse than those trained on the left side, right side, and back viewpoints across all subset sizes.

3.3.3. Video-level individual classification performance on the test set, per viewpoint.

While the previous section shows results for models trained on different (and all) viewpoints, here we see results for the model trained on all viewpoints but tested on different (and all) viewpoints. Table *5* shows the video-level classification performance of the individual classifier on the held-out test set, i.e., completely unseen videos, across viewpoints. Each video was assigned a predicted label via majority voting of all frames in a video. Overall, the model achieved a high performance, with an average F1-score of 0.90 across all birds and viewpoints.

Table 5. Video-level individual classification performance on the test set, by viewpoint.

Bird_ID	Front	Back	Left side	Right side	All viewpoints
BNU-RPM	0.96	0.97	0.98	0.97	0.98
BNY-RPM	0.67	0.90	0.89	0.92	0.89
BRG-YOM	0.95	0.97	0.97	0.94	0.97
BRK-NOM	0.22	0.73	0.84	0.83	0.84
EYB-RPM	0.93	0.95	0.98	0.95	0.98
GBM-ORY	0.81	0.91	0.96	0.94	0.95
GBY-ORM	0.91	0.95	0.98	0.95	0.94
OEB-RPM	0.81	0.89	0.97	0.96	0.94
OGY-BRM	0.82	0.93	0.97	0.92	0.94
ORB-UYM	0.77	0.88	0.94	0.91	0.89
OUB-RPM	0.56	0.86	0.92	0.83	0.87
OYR-BGM	0.78	0.90	0.94	0.94	0.92
RGY-BOM	0.28	0.49	0.80	0.70	0.60
RYO-BGM	0.75	0.97	0.89	0.88	0.92
YM-OBR	0.92	0.90	0.96	0.94	0.94
YRU-POM	0.72	0.97	0.96	0.87	0.87
Average	0.74	0.89	0.93	0.90	0.90

Figure *10* illustrates the variation in classification performance across viewpoints. The left and right viewpoints achieved the highest and most consistent performance across individuals (mean F1-scores of 0.93 and 0.90, respectively). Meanwhile, the front viewpoint showed the greatest variability across individuals, and the lowest performance, e.g., BRK-NOM = 0.22, and RGY-BOM = 0.28. The model achieved considerably higher scores for the same two birds from the right and left side viewpoints (BRK-NOM = 0.84 and 0.83, respectively; and RGY-BOM = 0.80 and 0.70, respectively) and when predictions were aggregated across all viewpoints (0.84 and 0.60).

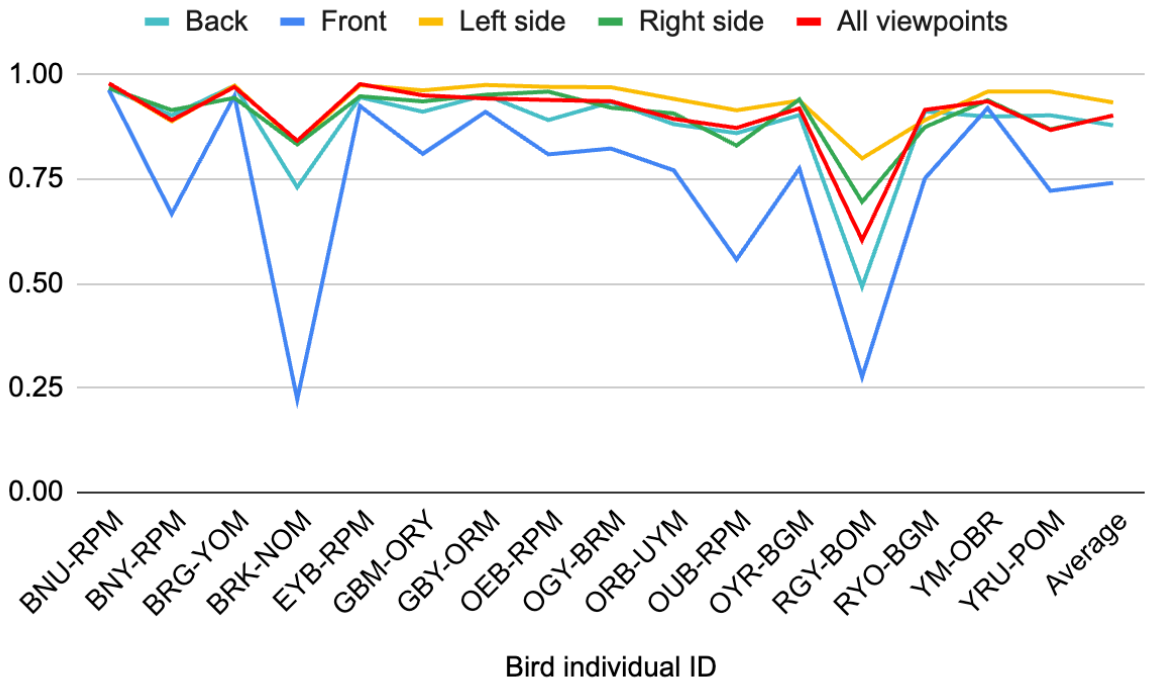


Figure 10. Video-level individual classification performance on the test set, per viewpoint.

4. Discussion

We developed and validated a machine learning pipeline to automatically identify individual Spotted Bowerbirds (*Chlamydera maculata*) from camera-trap video under natural field conditions. We evaluated the performance of our ResNet50 classifier at the frame level, achieving a mean F1-score of 0.98 on the validation set and 0.86 on the held-out test set. This high performance shows that deep learning models can reliably distinguish individual bowerbirds on single frames. To better approximate the model's real-world use, where predictions are needed for entire videos, we aggregated frame-level predictions using majority voting to assign a single identity per video. Under this video-level evaluation, the model achieved a mean F1-score of 0.90 across individuals, demonstrating that majority voting reduces the impact of occasional frame-level misclassifications and results in more robust identity predictions. A significant finding of this research is the empirical evidence provided by the model's high performance for the existence of sufficient, consistent, and learnable inter-individual variations in the visual appearance of Spotted Bowerbirds. To our knowledge, distinct inter-individual features in this species have not been quantitatively described in the literature.

We observed that individuals with fewer training samples tended to have slightly lower F1-scores, highlighting the importance of sufficient data availability per individual. Our analysis showed that an F1-score of 0.85 can be achieved with approximately 350 instances per individual (using a 70:30 training-validation split), equivalent to roughly 6 minutes of video per individual when extracting one frame per second. This amount of data can still be challenging to obtain in field studies for rarely seen or newly appearing subjects. Nonetheless, our model's ability to achieve this level of accuracy with relatively modest amounts of data per individual enhances the practicality and scalability of applying this method to a dataset collected in the wild. In cases where data is abundant, increasing the number of instances to 900-1,000 might be a better option, as the models trained on these subsets achieved a higher performance (0.9542 F1-score), marginally worse than the baseline model (0.98 F1-score).

To identify the most informative viewpoints, we trained the individual classifier separately on four viewpoints: front, side (left and right), and back. Human annotation of these viewpoints was highly consistent (87.25% raw agreement, Cohen's Kappa = 0.83). Most disagreements occurred between adjacent viewpoints, particularly between back and side, likely because

birds often appeared partially turned and did not fit cleanly into either category. In contrast, confusion between clearly distinct viewpoints (e.g., front vs. back, or left vs. right side) was minimal. The viewpoint classifier achieved high performance across all viewpoints (F1-scores between 0.86 and 0.93). Interestingly, the front viewpoint was classified most accurately (F1 = 0.93), despite having the fewest training instances, likely because of distinctive visual features such as the light-colored breast and belly. In comparison, the back viewpoint had the worst performance (F1-score = 0.86), which aligns with the observed human annotation confusion between back and side views, suggesting inherent visual ambiguity in this orientation. These results indicate that most classification errors occur when adjacent viewpoint features are visible, e.g., partially turned birds.

We used the viewpoint classifier to label all frames in the dataset and created viewpoint-specific subsets for training individual classifiers. These subsets included the original four viewpoints (front, back, left side, right side) as well as a combined “side view” category that merged left and right side instances. We then trained separate individual classifiers on each viewpoint-specific subset and compared their performance with a model trained on the full dataset containing all viewpoints. Models trained on a single viewpoint consistently required fewer training instances per bird to achieve the target performance of F1 ≥ 0.85. Specifically, viewpoint-specific models reached this threshold with only 100–150 instances per bird, whereas the all-viewpoints model required approximately 350 instances per bird. This finding suggests that including multiple viewpoints introduces additional visual variability, increasing the complexity of the classification task. Training on a single viewpoint removes this variability and allows the model to learn the discriminative features of each individual more efficiently, requiring less data. This finding is compatible with observations done on other species where identity recognition was focused on a small set of features. For example, in Greylag Geese, Kleindorfer et al. trained an algorithm using only the bill region, excluding plumage features because of their variability across conditions, and achieved reliable identity matching with only a few images per individual [24].

Finally, to further determine whether model performance could be improved by focusing on specific viewpoints, we evaluated the classifier trained on all viewpoints separately for each viewpoint, at the video-level. We observed that side and back views provided the most accurate predictions (F1 scores of 0.89, 0.93, and 0.90, respectively). Meanwhile, the front viewpoint showed the greatest variability across individuals, and the lowest performance (F1

score = 0.74), and the model consistently achieved higher scores when instances from all viewpoints were combined. For implementation, these results suggest that prioritising side and back viewpoints results in the highest accuracy when viewpoint information is available, but using all viewpoints could still be reliable when filtering by viewpoint is not feasible.

In the following section, we want to discuss the limitations of our data pre-processing pipeline. We used a multistage pre-processing pipeline that included object detection and mask segmentation on the extracted frames. Leg bands were automatically removed from the masked instances to prevent potential overfitting of the CNN to artificial markers. This process targeted narrow vertical structures within the lower third of the bird's mask. A limitation of this approach was the requirement for legs to be in a vertical position. In cases where oriented horizontally or tucked close to the body, they may not be detected as narrow structures and thus not removed. Additionally, this method in some cases led to removing portions of the tail (Figure *4*b) or resulted in inconsistent tail removal across images, which could be a limitation as the tail might contain potentially relevant identifying features. An alternative to our method could be to train a dedicated object detection model for the automated detection and segmentation of legs and leg bands. If trained on representative enough labelled data, such a model could localise leg regions regardless of their orientation, allowing for more robust and accurate leg removal.

Another limitation of our study relates to the continuous nature of camera trap recording in field conditions, as changes in lighting conditions throughout the day, and variations in the birds' posture can obscure or highlight different morphological features such as the plumage patterns and colours (Figure 11).

(a)  (b)  (c) 

Figure 11. Impact of lighting conditions and posture on plumage appearance. Frame closeups, illustrating how (a) and (b) lighting conditions and (c) changes in body posture can alter the apparent pattern and coloration of an individual's plumage.

An important consideration for both the viewpoint classification and the individual identification models is the use of horizontal flip as an augmentation technique, as it eliminates the distinction between left and right side view. This step was included as part of the augmentation pipeline to improve generalisation and was only retrospectively recognised as a factor that could affect the interpretability of viewpoint specific results. Left and right side viewpoints achieved similarly high F1-scores and showed nearly identical learning curves across subset sizes (Table 4, Figure 9), and video-level performance was highly consistent between the two viewpoints for most individuals (Figure 10). Both left and right models consistently outperformed the combined “side view” model trained on pooled left and right instances (Table 4, Figure 9), indicating that training the models separately provided an advantage despite the horizontal flipping. We speculate the mirrored instances may have helped the model learn individual specific features more robustly, regardless of the birds' orientation.

Finally, we want to discuss the potential implications of training and evaluating the model on data collected during a single breeding season. Birds' plumage features can undergo changes between breeding seasons due to moulting. Consequently, a model trained exclusively on data from one season may not maintain the same accuracy when applied to footage of the same individuals recorded in subsequent years without retraining or updating the dataset. Future work could test the model's performance on data from different breeding seasons for the same individuals to assess the model's reliability and long-term usefulness.

The development of an automated individual identification method for Spotted Bowerbirds from video footage holds significant implications for advancing behavioural research in this species. By automating the identification process, researchers can drastically reduce the time and labour currently required for manual video analysis. This capability opens avenues for investigating complex social interactions, such as identifying visiting females, auxiliary males, or rivals, and examining how individual behaviour, including courtship displays, might be modulated by the identity of the audience. Furthermore, automated individual identification facilitates the assessment of individual-level responses to environmental changes or experimental manipulations over extended periods. The methodology developed here is

potentially adaptable to other species, particularly those for which natural focal points suitable for camera trapping can be identified.

Our study demonstrates the feasibility of training a ResNet50 classifier with camera trap footage to identify individual Spotted Bowerbirds, empirically supporting the existence of consistent and learnable inter-individual variations in this species. A key finding was that an F1-score of ≥0.85 could be attained with a relatively modest dataset of 350 instances per individual (using a 70:30 training/validation split). We showed that viewpoint-specific training reduced the amount of data required per individual required to achieve the same performance to around 150 instances, regardless of the viewpoint showed. This data efficiency is particularly relevant for field studies where data accumulation for individual subjects can be limited.

Regarding the model's robustness, it is worth noting that some noise was present in the training data, which included a few non-target individuals, such as birds from other species, which were still detected by the object detector and were not manually filtered out. This level of noise is difficult to avoid in real-world datasets, especially when working with large-scale video collections. Nonetheless, the model achieved strong performance, indicating its robustness and ability to learn reliable identifying features despite such noise. The developed methodology holds considerable promise for enabling large-scale, automated analysis of individual behaviours of Bowerbird populations in the wild.

Acknowledgements

The computational results of this work have been achieved using the Life Science Compute Cluster (LiSC) of the University of Vienna. Data were collected thanks to a grant of the Austrian Science Fund (FWF: W1262-B29 [https://doi.org/10.55776/W1262]).

References

1. Vidal M, Wolf N, Rosenberg B, Harris BP, Mathis A. Perspectives on individual animal identification from biology and computer vision. Integr Comp Biol. 2021 Oct 4;61(3):900–16.
2. Cooper NW, Thomas MA, Marra PP. Vertical sexual habitat segregation in a wintering migratory songbird. Auk [Internet]. 2021 Jan 7; Available from: http://dx.doi.org/10.1093/ornithology/ukaa080
3. Li S, Li J, Tang H, Qian R, Lin W. ATRW: A benchmark for Amur Tiger Re-identification in the Wild [Internet]. arXiv [cs.CV]. 2019. Available from: http://arxiv.org/abs/1906.05586
4. O'Brien TG, Kinnaird MF. A picture is worth a thousand words: the application of camera trapping to the study of birds. Bird Conserv Int. 2008 Sep;18(S1):S144–62.
5. Rovero F, Zimmermann F, Berzi D, Meek DM. "Which camera trap type and how many do I need?" A review of camera features and study designs for a range of wildlife research applications. Hystrix, the Italian Journal of Mammalogy. 2013 May 1;24(2):48–156.
6. Rowcliffe JM, Kays R, Kranstauber B, Carbone C, Jansen PA. Quantifying levels of animal activity using camera trap data. Methods Ecol Evol. 2014 Nov;5(11):1170–9.
7. Caravaggi A, Banks PB, Burton AC, Finlay CMV, Haswell PM, Hayward MW, et al. A review of camera trapping for conservation behaviour research. Remote Sens Ecol Conserv. 2017 Sep;3(3):109–22.
8. Janisch J, Mitoyen C, Perinot E, Spezie G, Fusani L, Quigley C. Video recording and analysis of avian movements and behavior: Insights from courtship case studies. Integr Comp Biol. 2021 Oct 14;61(4):1378–93.
9. Harris G, Thompson R, Childs JL, Sanderson JG. Automatic storage and analysis of camera trap data. Bull Ecol Soc Am. 2010 Jul;91(3):352–60.
10. Ferreira AC, Silva LR, Renna F, Brandl HB, Renoult JP, Farine DR, et al. Deep learning-based methods for individual recognition in small birds. Methods Ecol Evol. 2020 Sep;11(9):1072–85.
11. Frith C and Frith D, The Bowerbirds: Ptilonorhynchidae. Oxford. Oxford University Press, 2004.
12. Borgia G, Pruett-Jones SG, Pruett-Jones MA. The evolution of bower-building and the assessment of male quality. Z Tierpsychol. 2010 Apr 26;67(1–4):225–36.

13. Spezie, G., & Fusani, L. (2023). Sneaky copulations by subordinate males suggest direct fitness benefits from male-male associations in spotted bowerbirds (Ptilonorhynchus maculatus). Ethology: Formerly Zeitschrift Für Tierpsychologie, 129(1), 55–61.
14. Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements [Internet]. arXiv [cs.CV]. 2024 [cited 2025 May 2]. Available from: http://arxiv.org/abs/2410.17725
15. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2009. p. 248–55.
16. Reischle, V. (2025). Can humans discriminate individual Spotted Bowerbirds (Chlamydera maculata) from photographs? [University of Vienna].
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition [Internet]. arXiv [cs.CV]. 2015. Available from: http://arxiv.org/abs/1512.03385
18. O'Reilly M, Duffin J, Ward T, Caulfield B. Mobile app to streamline the development of wearable sensor-based exercise biofeedback systems: System development and evaluation. JMIR Rehabil Assist Technol. 2017 Aug 21;4(2):e9.
19. Lipton ZC, Elkan C, Narayanaswamy B. Thresholding classifiers to maximize F1 score [Internet]. arXiv [stat.ML]. 2014. Available from: http://arxiv.org/abs/1402.1892
20. Aeeneh, S., Zlatanov, N., & Yu, J. (2023). New bounds on the accuracy of majority voting for multi-class classification. In arXiv [stat.ML]. arXiv. http://arxiv.org/abs/2309.09564
21. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.
22. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276–282.
23. Landis, J. & Koch, G. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159-74. PMID: 843571.
24. Kleindorfer S, Heger B, Tohl D, Frigerio D, Hemetsberger J, Fusani L, et al. Cues to individuality in Greylag Goose faces: algorithmic discrimination and behavioral field tests. J Ornithol. 2024 Jan;165(1):27–37.