IUCN Red List Threatened Bird Classification from Citizen-Science Images using Fine-Tuned CNN Features and a Hierarchical Lightweight Classifier

Takeshi Nishikawa

Foundation for Computational Science (FOCUS), Computational Science Center Building 1F, 7–1–28 Minatojima-Minamimachi, Chuo-ku, Kobe 650–0047, Japan,
nishikawa@j-focus.or.jp

Abstract

To prevent bird strikes at wind farms, an approaching protected bird must be identified so that the turbine can be halted before a collision. We present a bird-species classification module intended for that real-time, low-power edge setting: a hierarchical system assigning IUCN Red List categories to in-the-wild citizen-science bird images, combining a ResNet50 feature extractor fine-tuned on a bird-domain corpus with lightweight LightGBM classifiers in a five-stage cascade following the taxonomic hierarchy. We evaluate on 84 species × 10 images (840) plus 840 held-out non-bird images, so that every stage is scored against its own denominator, and all paired tests resample species rather than images. Fine-tuning improves Stage 4 top-1 accuracy and macro recall from 0.273 to 0.404 (+13.1 pt; Holm-adjusted $p = 0.015$), whereas neither the top-5 gain nor the end-to-end IUCN gain survives correction. Training a flat 71-class Stage 4 and a genus-level hierarchical variant to verified convergence under identical caps, we detect no difference, and a pre-specified 5 pt non-inferiority test is inconclusive. Collapsing the order and species stages into a single 74-class classifier likewise changes no accuracy metric significantly, while raising alert precision from 0.765 to 0.929 at the same per-object cost, so the two stages should be merged and four stages suffice. A controlled down-sampling study finds the cascade degraded but usable at 128×128 and broken at 32×32, where an apparent rise in protected-species alert recall is an artefact of precision collapsing to 0.10. That alert, the operational output, is limited at native resolution too (recall 0.25–0.33). On the target device, a Jetson Orin Nano, accuracy is unchanged ($p = 1.000$) at 226 ms per object; the cost lies in the Stage 4 classifier and JPEG decoding rather than the backbone, and tuning both—Stage 4 truncated to 500 rounds at equal accuracy, the crop supplied in memory as the camera does—gives 112 ms on the CPU and 13 ms with the GPU, a usable speed for the application.

Keywords: Fine-grained visual categorization, Transfer learning, Hierarchical classification, Convolutional neural network, IUCN Red List, Citizen science, Biodiversity monitoring, Edge inference

1 Introduction

1.1 Application background: preventing bird strikes at wind farms

Wind power is a mainstay of renewable energy, but bird strikes—in which birds, raptors in particular, are killed by colliding with the rotating blades—are a serious obstacle to reconciling the siting of generation facilities with ecosystem conservation. One promising way to reduce the damage is

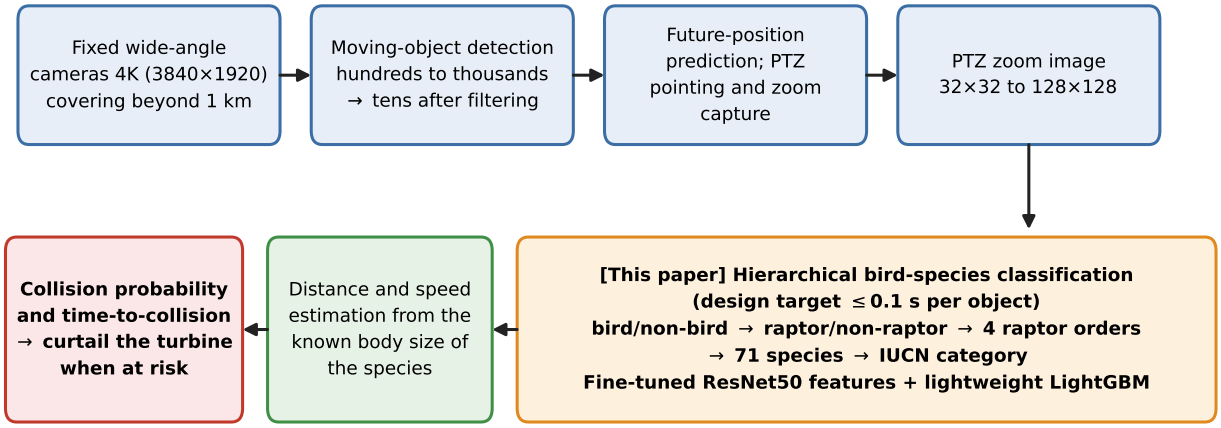

Figure 1: Overview of the wind-turbine bird-strike prevention system and the scope of this paper.

active control: detecting a protected bird as it approaches a turbine and temporarily halting the turbine when a collision with the rotor-swept area is predicted.

To realize this control, we are building the real-time system shown in Fig. 1. Multiple fixed wide-angle cameras sharing a field of view (imaging in 4K, 3840×1920, over ranges that include distances beyond 1 km) detect flying objects, their future positions are predicted, and a PTZ (pan–tilt–zoom) camera is pointed at the predicted position to acquire a magnified image. The bird species is identified from that magnified image, the distance and speed are estimated using the known body size of the identified species, and the collision probability with the rotor-swept area and the time to collision are computed so that the turbine can be stopped when at risk. The scope of this paper is the stage of this pipeline that **identifies the bird species and its IUCN Red List category from the PTZ magnified image**. As a prototype of this system we have built a sensor unit that integrates three 4K fixed wide-angle cameras and one PTZ camera (Fig. 2), and we have carried out vehicle-mounted field deployment tests. Targeting the direction of Enoshima Island from the Enoden parking center shown in Fig. 3, we detected and tracked flying objects with the 4K fixed wide-angle camera stage. An example of the distribution of the estimated distance and ground speed of detected Black Kites is shown in Fig. 4. Individuals at a median distance of about 1.25 km were captured, and their distance and ground speed (median about 4 m/s) could be estimated, showing that the motion information required for the subsequent computation of collision probability and time to collision is already available at the fixed-camera stage. The median estimated distance of about 1.25 km is consistent with the extent of Enoshima Island (about 700–1400 m from the origin) indicated by the concentric arcs in Fig. 3.

1.2 Problem setting and technical constraints

The input to the present task is a single natural image containing a bird as the main subject, and the output is its species name and IUCN Red List category (in particular CR/EN/VU). In operation, the input is a low-resolution image of a bird in flight, cropped after zooming with the

(a) Laboratory prototype (b) Field deployment mounted in a vehicle

Figure 2: Prototype sensor unit (three 4K fixed wide-angle cameras and one PTZ camera). (a) Laboratory prototype, (b) field deployment mounted in a vehicle.

PTZ camera; depending on the size of the magnified image obtained after zooming, resolutions of 32×32, 128×128 or 224×224 pixels are used selectively. That is, 32×32 is used when a distant bird still appears small even after zooming, 224×224 when it appears sufficiently large, and 128×128 in intermediate cases (in every case the crop is resized to 224×224 pixels before being fed to the feature extractor). A distant bird is, on the fixed camera, no more than a featureless grayscale blob, and because the pixel count is so small it is difficult to apply existing CNNs or recent Transformers directly for identification. The system therefore separates moving-object detection (fixed cameras) from species identification (PTZ magnified images) and concentrates the computational resources and the discriminative processing on the latter.

The final targets for issuing a turbine-stop command are, domestically, the four raptor species regarded as particularly important for conservation in Japan (Golden Eagle, White-tailed Eagle, Steller's Sea Eagle and Mountain Hawk-Eagle) and, internationally, the rare species corresponding to the CR (Critically Endangered) and EN (Endangered) categories of the IUCN Red List. The primary purpose of the system is to pass information to the turbine-stop decision, via the estimation of collision probability and time to collision, when the approach of one of these protected species is detected. Detecting these protected species with high recall among all 71 species is therefore the practical requirement.

The classification stage is subject to the following technical constraints. First, **real-time operation**. A raptor flying at 60 m/s has about 50 seconds (about 3 km) of margin between the first detection point and the rotor-swept area; the several hundred to several thousand moving-object candidates arising in one image must be reduced to a few tens using speed and track history, and each object must then be classified within 0.1 s. Second, **low power consumption and edge execution**. Because the cameras and PC are installed near the turbine, the classifier has to run on a low-power edge device. We target a Jetson Orin Nano 8GB (six Arm Cortex-A78AE

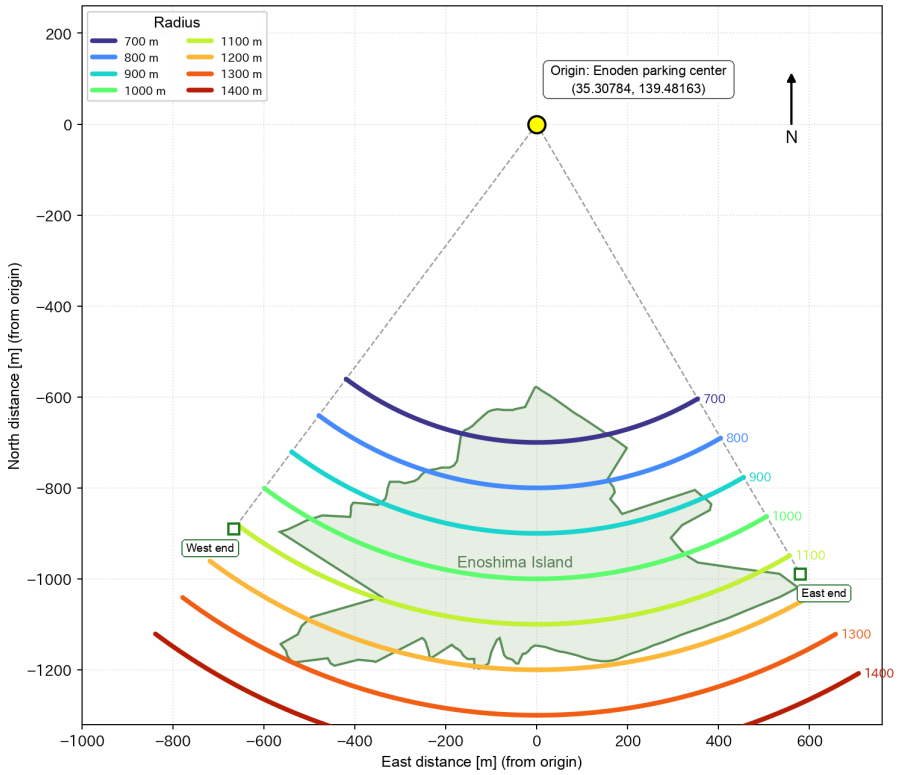

Figure 3: Concentric distance arcs (700–1400 m at 100 m intervals) overlaid on the shape of Enoshima Island, centered at the Enoden parking center (35.30784, 139.48163). Distances are drawn on top of the island shape.

cores and an Ampere GPU sharing 8 GB), run in its 15 W mode as it would be in service, and the measurements of Section 4.3 are made on that device; per camera, four cores are allocated to moving-object detection, future-position prediction/PTZ control and bird-species classification. Under these power and compute constraints it is difficult to deploy a large Transformer. These constraints make the design choices described below—"early rejection by a hierarchical cascade" and "lightweight classifiers"—a necessity rather than a preference.

The evaluation in this paper is carried out not on the operational PTZ magnified images themselves but on ordinary images from the citizen-science platform iNaturalist [1], used as a proxy, because real PTZ images of threatened species are currently scarce. Since iNaturalist images include diverse backgrounds, partial occlusion and a variety of flying and perched postures, they can be considered to act as a partial proxy for the factors of variation present in PTZ magnified images. Evaluation on real operational images is left as future work and is discussed in Section 7.3.

1.3 Contributions

In a preliminary study, a configuration using classical image features (HOG/HSV/LBP) with a decision-tree classifier reached only 5 % top-5 accuracy (the fraction of images for which the correct species is among the five highest-confidence candidates; the formal definition is given in Section 5.3) on an independent evaluation set of 720 images uniformly distributed over 72 species, and replacing the features with ImageNet-pretrained ResNet50 [3] features improved this only to 0.43, which is insufficient for the practical automatic detection of threatened (CR/EN/VU) species. Starting from this observation, this work makes the following four contributions.

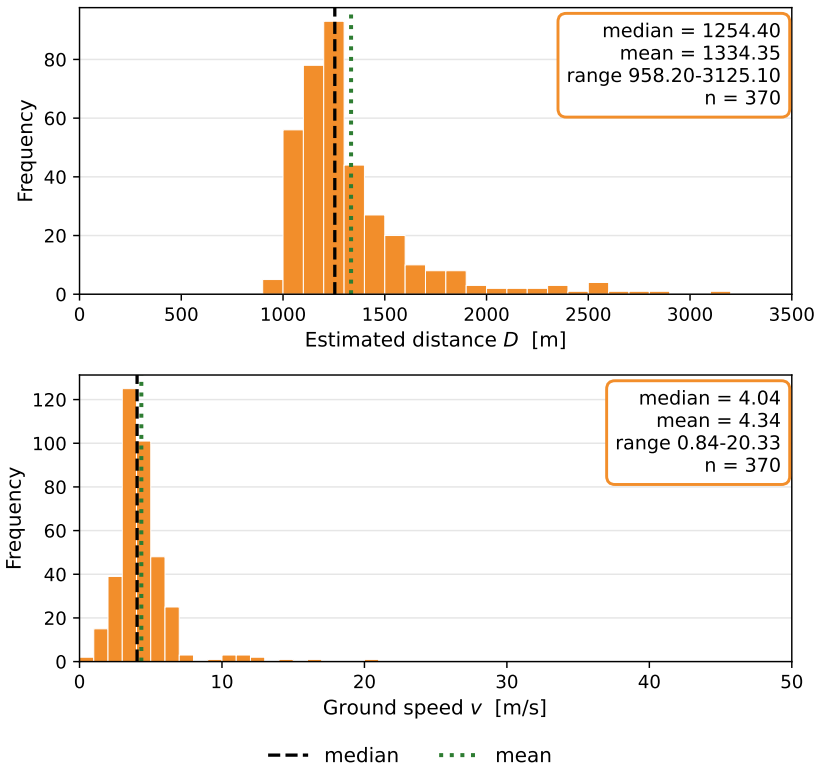

Figure 4: Distributions of the estimated distance (top) and ground speed (bottom) of flying objects (Black Kites) detected and tracked by the 4K fixed wide-angle camera (Enoshima 4K sample; 370 observations from 3 tracks). Dashed and dotted lines denote the median and the mean.

(1) A hierarchical pipeline consistent with real-time and edge constraints: We propose a configuration that cascades four LightGBM [5] classifiers along the taxonomic hierarchy and deterministically assigns the IUCN category in the final stage (Section 3). Early rejection at the lower stages acts as a reduction of computation under the real-time and low-power constraints.

(2) Restricted transfer learning (fine-tuning): When transferring ResNet50 to the bird-image domain we adopt a restricted setting in which only the head layer and the last residual block are updated, suppressing both overfitting and training cost (Section 4). This design is consistent with recent findings on which layers should be updated [9, 10].

(3) A convergence-controlled test of the added hierarchy, and an evaluation protocol that supports it: We train a flat 71-class Stage 4 and a genus-level hierarchical variant from identical features, with the same seed and the same 40,000-round budget, and verify that each stops by early stopping rather than by exhausting its budget. Under that control we detect no difference between them, and a pre-specified non-inferiority test is inconclusive, so we claim neither an advantage nor an equivalence for the hierarchy (Section 6). Supporting this, we evaluate every stage against its own denominator, add a held-out non-bird set so that the bird/non-bird stage is a real accuracy, and report species-cluster bootstrap intervals, Holm-corrected tests, per-category precision, and the structural recall ceilings of the ground truth (Section 5).

(4) A controlled measurement of the resolution gap between the evaluation domain and the operational one: Because the deployed classifier receives a PTZ crop of only tens of pixels, we re-extract features from the same ground truth reduced to 128×128 and 32×32 and re-run the unchanged models, which locates the operating boundary of the design (Section 6.6):

degraded but usable at 128, broken at 32. The same experiment shows the protected-species alert recall *rising* at 32×32 purely because precision collapses, which is a concrete reason to report that pair jointly.

2 Related work and positioning

2.1 Fine-grained visual categorization and bird benchmarks

Bird image classification is a representative instance of fine-grained visual categorization (FGVC), and standard benchmarks such as CUB-200-2011 [20], NABirds [21] and Birdsnap [16] are well established. Since the introduction of deep learning, bird species classification has progressed from part-based CNNs with pose normalization [17], through Bilinear CNNs that use second-order statistics of local features [18], to attention mechanisms that progressively zoom into discriminative regions [19]. More recently, Vision-Transformer-based methods have achieved high accuracy: TransFG [13] selects discriminative patch tokens and reports 91.7 % on CUB and 90.8 % on NABirds, FFVT [14] attains comparable accuracy at lower cost by aggregating tokens across layers, and DCAL [15] reports 92.0 % on CUB with dual cross-attention. These methods presuppose end-to-end large-scale training and relatively heavy inference. This work stands on a different axis from the accuracy race: it aims at a lightweight configuration that can be operated under the real-time, low-power and edge constraints of Section 1.2. In other words, it presents an alternative to giant models that is easy to implement, fits small data and consumes little power.

2.2 Citizen-science data and observation bias

iNaturalist [1] and eBird [2] provide bird images on the scale of tens of millions. The iNaturalist dataset [22] and iNat2021 [23] share the difficulties of the present task in that they exhibit a strong long-tailed distribution and contain visually nearly identical species. Here **observation-density bias** refers to the long-tailed bias by which observations and postings of common species dominate the platform while records of rare species are scarce. To cope with this bias, Mac Aodha et al. [24] jointly modeled geographic and temporal occurrence priors together with photographer bias, and Kang et al. [25] presented a framework that decouples representation learning from classifier learning and retrains only the classifier. This work demonstrates that, under the constraint that the rarer a threatened species is the less training data it inherently has, the combination of restricted transfer learning and lightweight classifiers fits such long-tailed, small-data regimes.

2.3 Camera traps and wildlife monitoring

In passive sensing of wildlife, MegaDetector [26] has become the de facto standard preprocessing step for removing empty images, and Wildlife Insights [27] provides a cloud platform for species identification and data sharing. Norouzzadeh et al. [28] demonstrated the application of deep learning to large-scale camera-trap images, and Beery et al. [29] showed a domain shift in which accuracy degrades markedly at unseen locations even when it is high at training locations. The early-rejection cascade of fixed cameras plus PTZ magnification proposed here addresses the same practical requirement of efficiently narrowing down a large number of candidates.

2.4 Machine learning applied to threatened species and the IUCN Red List

Research applying machine learning to the estimation of IUCN conservation status has concentrated mainly on predicting extinction risk from geographic distribution records. IUCNN [35] predicts

conservation status from distribution records with a deep network, and Borgelt et al. [36] predicted that many Data Deficient species are likely to be threatened. Prospects for ML/CV in conservation are surveyed by Tuia et al. [37], and WildlifeDatasets [38] has been released as a foundation for individual re-identification. In contrast, research that **directly identifies and detects IUCN threatened species from images** is comparatively sparse, and this work fills that gap with a practical system. BioCLIP [34] has been proposed as a foundation model for biological taxonomy, but its zero-shot performance is premised on assumptions different from those of edge inference.

2.5 Bird detection around wind-power facilities

The prior work closest to the present application is the image-based detection and species classification of birds around wind-power facilities by Yoshihashi et al. [30, 31]. They installed a fixed camera with a telephoto lens (5616×3744 pixels, field of view about 27°×19°) around a 2 MW, 80 m-diameter turbine in the Kinki region of Japan and, under conditions in which a bird with a 1 m wingspan appears as about 20 pixels at a distance of 580 m, constructed a dataset of about 32,000 time-lapse images (2 s intervals) containing more than 60,000 bird bounding boxes [30]. Their method extracts moving-object candidates by background subtraction and compares AdaBoost (Haar-like/HOG features) with convolutional neural networks, evaluating bird/non-bird detection and binary hawk-versus-crow species classification. The reported performance is a true-positive rate of 0.98 at a false-positive rate of 0.05 for bird detection and 0.85 at 0.1 for species classification, showing that shallow learning (AdaBoost) rivals deep learning for detection while deep learning is superior for species classification and for generalization to unseen locations. They cite "many threatened species belong to the hawks" as their reason for focusing on conservation, sharing the problem awareness of this work. Their dataset and code used to be publicly available (as noted in Section 7, they are currently unreachable), but they remain one of the few real-environment benchmarks in this field. In the earlier EWEA 2015 paper [31], the same imaging system was used to analyze detection and classification performance per image resolution (15–20, 21–30, 31–50 pixels), quantifying the tendency for identification to become harder at lower resolutions; the down-sampling study of Section 6.6 is the counterpart measurement for our cascade. Relatedly, Takeki et al. [32] showed, within the same framework, a method that detects birds appearing small in wide-field landscape images by integrating deep features at multiple scales; this is the prior work corresponding to our fixed wide-angle camera stage (detection of distant, small objects). Outside Japan, McClure et al. [33] reported a proof of concept for a system that automatically monitors eagle flights at a wind-power facility.

One point to note here is the increasing size of wind turbines. The 2 MW, 80 m-diameter turbine targeted by Yoshihashi et al. belongs to the generation installed in the 2000s; in recent years these have been replaced by large turbines of the 4 MW, 150 m-diameter class. As the rotor diameter grows (rotor-swept radius from 40 m to the 75 m class) the blade-tip speed also increases, so completing a stop before a collision requires starting detection and classification farther away—that is, at a stage where the bird appears even smaller. This lengthening of the detection range is difficult to address with post-hoc monitoring based on time-lapse imaging, and it strongly motivates our configuration, which captures distant, minute moving objects, actively magnifies them with a PTZ camera and classifies them in real time.

The positioning of this work with respect to these prior studies is as follows. First, whereas Yoshihashi et al. target **time-lapse images from a fixed camera** (0.5 fps), this work combines detection by fixed wide-angle cameras with active zoom capture by a PTZ camera to **track an incoming individual and obtain magnified images**. Second, whereas the species classification of the prior work is **binary (hawk versus crow)** or of relatively coarse granularity, this work goes

as far as identifying 71 Accipitriformes species and assigning an IUCN Red List category. Third, this work presupposes real-time operation **within 0.1 s per object on a low-power edge device** and a connection to a turbine-stop decision through collision prediction, incorporating the computational constraints explicitly into the design. Fourth, it introduces early rejection through a cascade along the taxonomic hierarchy, so that protected species (the four domestic raptors; internationally CR/EN species) are selectively picked up with high recall.

2.6 Hierarchical classification and transfer-learning strategies

The use of hierarchical (taxonomic) labels has been an active area. Bertinetto et al. [39] proposed losses that reduce the severity of mistakes using a hierarchy, Chang et al. [40] a framework that scans labels from coarse to fine and improves accuracy at both granularities simultaneously, and Khrulkov et al. [41] hyperbolic representations that embed tree structures with low distortion. Whereas these incorporate the hierarchy into the **loss or the embedding**, this work structures the hierarchy explicitly as a **cascade of classifiers** and asks, under a controlled and verified training budget, whether that refinement buys anything once the features have been adapted. HD-CNN by Yan et al. [43] is representative of methods that embed the hierarchy inside the network, which is again a contrasting approach (for a general survey of hierarchical classification see Silla et al. [42]). As for transfer learning, Yosinski et al. [8] quantified the transition from generic low layers to task-specific high layers, and He et al. [4] discussed the limited effect of ImageNet pre-training. Lee et al. [9] further showed that the layers that should be updated differ according to the type of distribution shift (surgical fine-tuning), and Kumar et al. [10] showed that full fine-tuning with an untrained head distorts pretrained features. As parameter-efficient adaptation, BiT [11], which uses large-scale pre-training with a simple transfer rule, and Visual Prompt Tuning [12], which freezes the backbone and adds learnable tokens in the input space, have also been proposed. Our restricted fine-tuning, which updates only the head and layer4, is directly supported by these findings and is well suited to edge deployment in that it requires no additional trainable modules and can be applied to existing lightweight backbone implementations as they are.

3 Configuration of the proposed system

3.1 A five-stage cascade along the taxonomic hierarchy

The proposed system consists of the five-stage cascade shown in Table 1. The basic design policy is to make the classification stages correspond to the taxonomic hierarchy of birds (class → order → family/genus → species). Stage 1 (bird/non-bird) and Stage 2 (raptor/non-raptor) are LightGBM binary classifiers, Stage 3 (four raptor orders) and Stage 4 (identification of 71 Accipitriformes species) are multi-class classifiers, and Stage 5 (IUCN Red List category) is a deterministic lookup from the `conservation_statuses` field of the iNaturalist API [45].

There are three reasons for adopting this hierarchical configuration. First, **reduction of the classifier-side computation by early rejection**. Rejecting non-birds and non-raptors with lightweight binary decisions at the lower stages restricts the heavy species identification to the candidates that pass. We are careful not to overstate this: as Section 4.3 shows, the backbone forward pass runs once per candidate *before* Stage 1 and dominates the per-object cost, so the cascade reduces the classifier-side work rather than delivering the 0.1 s budget by itself. Second, **divide and conquer over difficulty and data volume**. The number of classes, the difficulty of discrimination and the number of available training images differ greatly from stage to stage, so a classifier suited to each stage can be optimized independently. Third, **flexibility with respect**

Table 1: The five-stage hierarchical cascade configuration.

Stage	Decision	Input	Output classes	Implementation
1	bird/non-bird	all images	2 classes	LightGBM binary
2	raptor/non-raptor	Stage 1 = bird	2 classes	LightGBM binary
3	four raptor orders	Stage 2 = raptor	4 classes	LightGBM multi-class
4	species identification	Stage 3 = Accipitriformes	71 classes	LightGBM multi-class
5	IUCN Red List category	species from Stage 4	CR/EN/VU/NT/LC etc.	deterministic lookup

to adding classes. When a protected species is added, only the later stages need to be retrained; the feature extractor does not.

In Stage 5, the `ancestor_ids` field of the iNaturalist API is used to identify the class Aves (taxon_id = 3) and the four raptor orders, namely Accipitriformes (taxon_id = 71261), Falconiformes (67570), Strigiformes (19350) and Cathartiformes (559244), and the species output by Stage 4 is mapped to a Red List category. Of the 71 Stage 4 classes, 3 are CR, 5 EN, 4 VU, 6 NT and 53 LC. Of these, the ones used directly in the turbine-stop decision are the four domestically important raptors (Golden Eagle, White-tailed Eagle, Steller's Sea Eagle and Mountain Hawk-Eagle) and the CR/EN species that are subject to international stop measures; the output of Stage 5 is finally passed to turbine control as the decision of "whether or not this is one of the protected species".

3.2 Rationale for the classifier and the feature extractor

We explain the reasons for choosing ResNet50 [3] as the feature extractor and LightGBM [5] as the classifier, relating them to the constraints of Section 1.2.

Why ResNet50: It is standard and easy to reproduce, and stable 2048-dimensional features can be obtained from public implementations. Its parameter count and inference cost are smaller than those of ViT-based state-of-the-art models, and there is abundant knowledge on quantization and lightweighting for edge deployment (Section 4.3). In settings such as ours, where the number of classes is limited and low power consumption is required, the advantages of implementation simplicity and lightness outweigh those of the latest giant models.

Why LightGBM: For pre-extracted tabular features (2048 dimensions), gradient-boosted decision trees are less prone to overfitting on small data and train quickly, which suits independent per-stage optimization and partial retraining when classes are added. Compared with an end-to-end neural classification head, they are loosely coupled to feature extraction, so it is easy to keep the feature extractor fixed and replace only the later stages.

3.3 Inference pipeline

All five stages are driven by a single feature set (denoted full_y below), so that no stage-to-stage difference can be attributed to a change of features. Noisy labels contained in the training data (misidentifications) and duplicates caused by variant Japanese and English names (four pairs, e.g. Red-tailed Hawk) have been consolidated from 75 classes to 71 classes by a normalization procedure (the consolidation was based on the formal scientific and vernacular names rather than on the internal program-level label strings).

4 Transfer learning (fine-tuning) of the features and implementation

4.1 Policy for feature learning

In this study we obtain the feature extractor by transfer-learning (fine-tuning) an ImageNet-pretrained ResNet50 to the bird-image domain. Rather than retraining the whole network, we adopt a restricted setting in which only some layers close to the output (the classification head that discriminates species and the last convolutional block) are updated while the lower layers close to the input are frozen. This design follows the transfer-learning finding that low layers of a convolutional network learn generic features such as edges and textures while high layers learn task-specific ones [8, 4, 9], and it aims to suppress both overfitting on limited bird data—which include rare species—and training cost. Our method is transfer learning without any mechanism that explicitly corrects a distribution shift of the input data, and we consistently refer to it as transfer learning (fine-tuning) below.

4.2 Training setup

The feature extractor is ResNet50 [3] (pretrained with ImageNet1K_V2 [4]). The final classification layer (the fully connected layer that outputs probabilities over 1000 classes) is removed from ResNet50, and the 2048-dimensional vector produced by the global average pooling layer immediately before it (which aggregates each feature map into a single value over the spatial dimensions) is used as the feature representation summarizing one image. Input images are resized so that the shorter side is 256 pixels, center-cropped to 224×224 pixels, and normalized with the ImageNet per-channel mean and standard deviation. In transfer learning, in order to avoid overfitting and to keep the computational cost low, only the classification head and the last residual block (layer4) are made trainable, and all other layers are frozen. The training data consist of 151,539 images with taxonomic labels (131,539 birds plus 20,000 non-birds) over 895 training labels. We distinguish labels from species throughout: the 895 labels comprise 894 bird labels plus one non-bird class, and because the bird labels include near-duplicate name variants of the same taxon they correspond to 804 distinct bird species. The consolidation of such variants is applied downstream, at Stage 4, where 75 labels become 71 species. We use the AdamW optimizer with a head learning rate of 4×10^{-4} and a layer4 learning rate of 4×10^{-5} (linear scaling [7] with respect to an effective batch size of 512), at most 20 epochs, and early stopping (patience = 3) based on top-1 accuracy (the fraction of images for which the highest-confidence candidate matches the ground truth) on a validation split holding out 10 % of the training data. This "validation split" is a set separated from within the training data and is distinct from the independent evaluation set (GT) of Section 5.1; the relationship between the two is shown in Fig. 6. Computation used mixed precision with bfloat16 autocast, and parallelization used Distributed Data Parallel [6] over four H200 NVL GPUs. Validation top-1 accuracy reached its best value of 0.8482 at epoch 19; the run used the full 20-epoch allowance without triggering early stopping, and took 1,156 s of wall-clock time (Fig. 5).

4.3 Training and inference cost, and edge deployment

The training cost is 13 minutes on four GPUs, as stated above. The inference cost splits into two layers depending on the deployment form.

Reference (server CPU, ResNet50 features): For the ResNet50 feature extraction plus five-stage LightGBM inference used in the evaluation of this paper, ResNet50 feature extraction

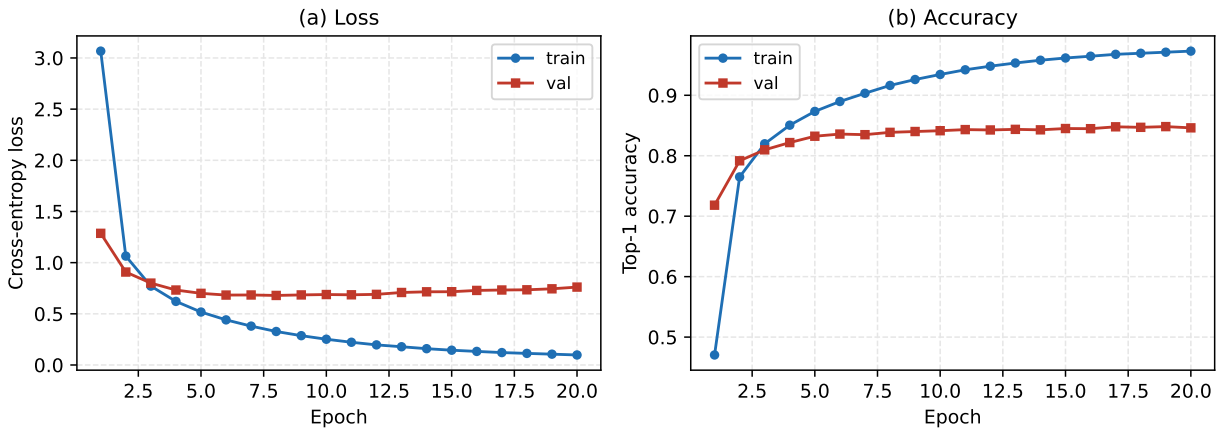

Figure 5: ResNet50 fine-tuning learning curves. (a) Loss, (b) top-1 accuracy.

Table 2: Measured per-object latency on the Jetson Orin Nano, over the 1,680 ground-truth images (mean, with the median in parentheses). Preprocessing and the LightGBM cascade run on the CPU in every column; the GPU column replaces only the feature extractor, whose TensorRT figure includes the host–device copies. The single-thread column uses a 300-image subset.

Component	6 threads	4 threads	1 thread	GPU (FP16)
Preprocessing (CPU)	73.2 (73.5)	73.1	65.1	73.2
ResNet50 features	104.9 (99.8)	132.7	451.3	**3.2**
LightGBM cascade	48.1 (3.1)	47.0	54.9	48.1
End-to-end per object	**226.2** (190.5)	252.8	571.4	**124.5**
Throughput [images/s]	4.03	3.75	1.71	—

is dominant on a server CPU (Intel Xeon Platinum 8480+, ONNX Runtime, batch = 1), taking about 70 ms per image.

Measured on the target edge device: We ran the unmodified pipeline on the device it is intended for, a Jetson Orin Nano 8GB (six Arm Cortex-A78AE cores and an Ampere GPU sharing 8 GB, in its 15 W mode), over the same 1,680 ground-truth images. The fine-tuned ResNet50 was exported to ONNX with the classification head removed, so the graph is the 2048-d feature extractor used everywhere else in this paper; preprocessing was reimplemented with PIL because torchvision is unavailable on the device. Two checks confirm the port computes the same function: the device features agree with the server features to a minimum cosine similarity of 0.99994 over all 1,680 images, and the cascade decisions agree on 99.8–99.9% of images at every stage. Feature extraction ran on the CPU through ONNX Runtime and on the GPU through TensorRT 8.5.2. File reads are excluded from the per-object figures, since in operation the crop arrives from the PTZ camera in memory. Table 2 reports the result.

Two things follow, and both revise what we had assumed.

The dominant cost is not the backbone once the cascade is deep. We previously reasoned that ResNet50 dominates and that early rejection removes only cheap downstream work. On the device this holds for the median candidate but not for the ones that matter: the LightGBM cascade costs 3.1 ms at the median, yet 48.1 ms on average and 245 ms at the 95th percentile, because the 312 of 1,680 images that reach Stage 4 pay for a 71-class model of 12,894 boosting rounds—one round being a single round of tree addition, which in a multi-class stage adds as many trees as there are classes. At that point the classifier exceeds the 104.9 ms backbone. The cascade's

Table 3: Effect of the tuning, measured on the same 1,680 images. Accuracy is Stage 4 top-1; the paired species-cluster test against the server run gives $p = 1.000$ for every configuration listed. The GPU rows keep preprocessing and the cascade on the CPU and replace only the feature extractor.

Configuration	Pre.	ResNet50	LightGBM	End-to-end	top-1
As measured (12,894 rounds, JPEG)	73.2	104.9	48.1	226.2	.4038
+ Stage 4 at 500 rounds	94.6	101.9	8.6	205.1	.4000
+ in-memory PTZ crop	2.1	100.5	9.2	**111.9**	.4000
+ GPU FP16, batch 1	2.1	3.1	9.2	**14.5**	.4000
+ GPU FP16, batch 16	2.1	1.7	9.2	**13.1**	.4000

early rejection is thus doing more work than we credited it with—it is what keeps the median low—while Stage 4 itself, not the backbone, sets the tail.

Most of the preprocessing is JPEG decoding. Of the 73 ms, decoding the citizen-science photograph accounts for 49.9 ms, resize and crop for 14.5 ms and normalization for 1.9 ms. The median image here is 2.8 Mpx, whereas an operational PTZ crop arrives from the camera as pixels, already cropped to the subject. The evaluation corpus therefore charges this stage for work the deployed system does not do.

Tuning against the measurement

The measurement names three places to act, and we acted on each, in every case keeping the requirement that the device must compute the same function as the server. Table 3 gives the result.

Stage 4 boosting rounds. Gradient boosting adds trees sequentially, so a model can be truncated at prediction time and the first k trees are exactly the model that training with k rounds would have produced. Sweeping k on the device shows the accuracy is flat well below the 12,894 rounds the early stopping selected: at 500 rounds Stage 4 top-1 is 0.4038 against 0.4038 for the full model, with the paired species-cluster test at $p = 1.000$, while the cascade cost falls from 45.2 ms to 5.3 ms. The protected-species alert is in fact slightly cleaner (precision 0.813 against 0.765), since a shorter model fires less confidently on the false positives. Truncation also shrinks the model from 733 MB to 123 MB and its load time from 14.1 s to 0.73 s, which matters on an 8 GB device that has to start unattended. We keep 500 rounds; 250 is also indistinguishable ($p = 0.727$) if the tail matters more than the mean.

Preprocessing under operational conditions. Feeding the cascade the 224×224 crop already in memory—the paper's own resize-and-crop applied beforehand, so the pixels are identical to those behind every accuracy number here—costs 2.1 ms instead of 94.6 ms, and the features still match the server to a minimum cosine of 0.999945, giving the same 0.4000 top-1. We also tried faster resize paths and rejected them: OpenCV's `INTER_LINEAR` is 22% faster but does not low-pass on downscale as PIL and torchvision do, which moves the features to a minimum cosine of 0.512, and Pillow's reduced-scale JPEG `draft` decode is 45% faster at cosine 0.925. Neither is usable with a classifier trained on the other convention.

The backbone on the GPU. TensorRT FP16 runs the extractor in 3.10 ms at batch 1, and 1.72 ms per image at batch 16—the regime the system actually sees, since one frame yields tens of candidates. Quantizing the backbone to INT8 for the CPU does not work here: dynamic quantization produces `ConvInteger`, which this ONNX Runtime build does not implement on aarch64, and static quantization runs 3.5× faster (28.6 ms) but destroys the predictions, agreeing with the server on only 53% of Stage 2 decisions.

Tuning brings the CPU-only pipeline from 226 ms to 112 ms per object and the GPU-assisted

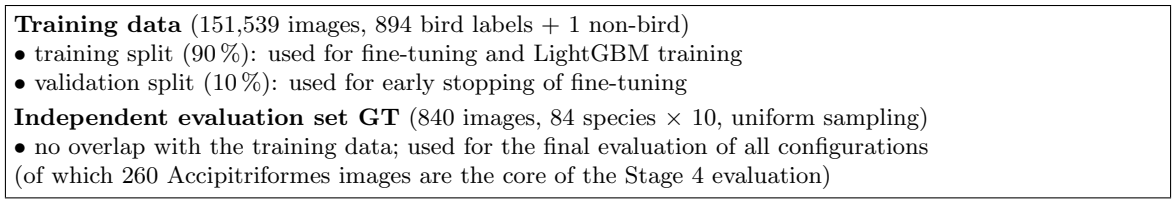

Figure 6: Relationship among the training data, the internal 10 % validation split, and the independent ground-truth (GT) set. The validation split is 10 % of the training data and is independent of the GT.

one to 13 ms, at unchanged accuracy. The CPU-only figure does not reach the 0.1 s design target, but it is comfortably inside what the application needs: the 0.1 s figure is a per-object target, and the deployment reduces several hundred to several thousand moving-object candidates to a few tens before classification (Section 1.2), so a few tens of objects at 112 ms is a few seconds against the roughly 50 s a raptor takes to cover the 3 km to the rotor. The device therefore runs this cascade at a usable speed without a GPU, and with the GPU it does so with an order of magnitude to spare.

Quantization and distillation of the backbone remain unmeasured as a training exercise: the 3–9 ms figure quoted above is a server-CPU projection for a lightweight backbone, and we did not train one. After tuning, that path is also no longer where the remaining time is—the backbone is 1.7–3.1 ms on the GPU, and what is left on the CPU is the 9.2 ms cascade.

5 Evaluation experiments

5.1 Evaluation dataset

With the composition shown in Table 4, we constructed an independent evaluation set (GT) of 84 species × 10 images each = 840 images that are not used in training. Sampling the same number of images for every species removes the *species-frequency* component of observation-density bias; the photographer, geographic, posture and background biases of the platform remain, and we do not claim otherwise. Of the GT, Accipitriformes account for 26 species × 10 = 260 images, which are the central object of evaluation for species identification (Stage 4). The relationship among the training data, the validation split and the GT is shown in Fig. 6.

The bird GT alone cannot score Stage 1: it contains no non-bird images, so bird/non-bird "accuracy" computed on it is only a bird recall. We therefore collected an additional **840 held-out non-bird images** from iNaturalist, mirroring the composition of the 20,000 non-bird training images (plants 49.8 %, insects 30.1 %, fungi, reptiles, mammals, fishes, molluscs, amphibians, arachnids and other animals). Aves was excluded explicitly at query time, every candidate was rejected if its content MD5 matched any training image, and the resulting set was inspected visually. All Stage 1 numbers below are computed on the union of the two sets (n = 1,680), and we additionally report the non-bird rejection rate separately.

Representative input images (examples of the CR/EN/VU/LC categories taken from the GT) are shown in Fig. 7. As stated above, this evaluation uses ordinary citizen-science images as a proxy, and the domain differs from that of the operational PTZ magnified images. There are, however, three reasons why this proxy evaluation is appropriate for validating the claims of this work. First, the focus of the evaluation is the relative comparison between configurations (the effect of transfer learning and of hierarchical refinement), which is a criterion that is relatively insensitive to the domain of the evaluation data. Second, citizen-science images qualitatively subsume the

Table 4: Dataset configuration.

Use	Images	Classes	Source
Training (full_y)	151,539	894 bird labels + 1 non-bird	iNaturalist + existing data
of which birds	131,539		
of which non-birds	20,000	10 iconic taxa	iNaturalist
Validation split	10 %	same as above	held out for early stopping
GT birds (independent)	840	84 species (26 Accipitri.)	iNaturalist, 10 per species
GT non-birds (held out)	840	607 taxa in 10 iconic taxa	iNaturalist, MD5-disjoint

appearance variation of PTZ magnified images. Specifically, citizen-science images widely include (a) **diverse backgrounds** such as sky, woodland, water surfaces and urban areas, (b) **partial occlusion** by branches, leaves or other individuals, (c) **diverse postures** such as perching, flying, soaring and flapping, as well as varied shooting angles, (d) **variation in resolution and scale** due to differences in shooting distance and zoom, and (e) **variation in lighting** such as front lighting, back lighting and twilight. These are qualitatively common to the factors of variation that inevitably arise in PTZ tracking (background diversity, clipping and occlusion during tracking, changes in flight posture, resolution differences depending on the zoom factor, and changes in outdoor lighting). The ability to discriminate species from the morphology and plumage of the bird itself, independently of the background, is therefore required in both domains, and robustness to the appearance of a species acquired on citizen-science images can be considered partially transferable to PTZ magnified images. Third, operational PTZ magnified images are also resized to the same 224×224 pixels and fed to the same feature extractor as in training and evaluation, so the tendencies of appearance-based discrimination performance can be extrapolated to a certain extent from this proxy evaluation. Evaluation on real images specialized to low resolution and flight posture remains future work, as stated in Section 7.3.

5.2 Compared configurations and naming convention

We validate the effect of the proposed method through combinations of the features (ImageNet pre-training vs. transfer learning) and the classification structure (flat vs. hierarchical). A naming convention is defined so that the content of each configuration can be read off from its name, and the definition of each label is given in Table 5. The prefix denotes the feature type (IN = ImageNet pre-training, FT = fine-tuning) and the suffix the classification structure (Flat = flat 71 classes, Hier = hierarchical cascade). The suffix Hier is a variant in which the flat 71-way classification of Stage 4 in the basic cascade of Table 1 is decomposed into two stages, a genus-level classifier and a within-genus species classifier: it adds a genus-level candidate restriction to species identification within Accipitriformes.

Classifier hyper-parameters. Every LightGBM stage uses a learning rate of 0.05, 63 leaves, unlimited depth, feature and bagging fractions of 0.8 with bagging every 5 rounds, and L_2 regularization of 0.1 for the binary stages and 0.2 for the multi-class ones. The only setting that varies is the minimum number of samples per leaf, which has to follow the size of the problem each classifier sees: 20 for the binary Stages 1–2, 10 for the four-way Stage 3, 3 for the flat 71-class Stage 4, and, in the hierarchical variant, 5 for the genus classifier and 2–3 for the within-genus classifiers, some of which separate only two or three species.

Convergence control. The comparison between Flat and Hier is meaningful only if neither arm is handicapped by its training budget. Both arms are trained from the same features with the same seed and, since the hierarchical arm trains one genus classifier plus fourteen within-genus classifiers rather than a single model, the same *per-classifier* cap of 40,000 boosting rounds and

Figure 7: Examples of input images used for evaluation (from iNaturalist). Badges at the upper left indicate the IUCN category (CR/EN/VU/LC). iNaturalist does not own these photographs; each is reproduced under the licence chosen by its observer, and the per-photograph observation ID, photographer attribution and licence are listed in the released manifest.

Table 5: Naming convention for the evaluated configurations.

Configuration	Features	Classification structure	Role
IN-Base	ImageNet pre-training	flat 71 classes	baseline
FT-Flat	fine-tuning	flat 71 classes	main proposed configuration ($\star$)
FT-Hier	fine-tuning	hierarchical cascade	test of the added hierarchy

the same early-stopping patience of 100 rounds. We verify that each stopped by early stopping rather than by exhausting its cap: the flat Stage 4 stopped at iteration 12,894 and the hierarchical Stage 4a at 2,451. A first attempt with a 10,000-round cap was discarded precisely because the flat arm reached the cap while still improving; that run and the converged run agree on every headline metric to the reported precision, which is itself evidence that the comparison is stable.

Pairwise comparisons are paired and two-sided, but they cannot use McNemar's exact test [44] as stated, because that test treats images as independent and the ground truth has ten images per species. We therefore use a cluster-robust analogue: the statistic is the net discordance $\sum_i (b_i - c_i)$ summed over species, and the null distribution is obtained by flipping the sign of whole species (exact enumeration when the number of informative species permits, Monte Carlo otherwise). With one image per species this reduces to McNemar's test. Non-bird images are clustered by their own taxon (607 taxa), not by iconic-taxon folder. Because nine tests are reported, we apply the Holm–Bonferroni correction across the family and quote both the raw and the adjusted p; significance is judged at $\alpha = 0.05$ on the adjusted value. Table 8 also lists the image-level p for reference, to make the size of the difference visible.

5.3 Evaluation metrics and error analysis

The evaluation metrics are top-1 accuracy for Stages 1–3, top-1 and top-5 accuracy plus macro-averaged recall for Stage 4, and IUCN accuracy together with per-category recall *and precision* for CR/EN/VU for Stage 5.

Denominators. Each stage is scored only on the samples it is defined on, and against that stage's own denominator: Stage 1 on all 1,680 GT images, Stage 2 on the 820 birds whose species appears in the taxon table, Stage 3 on the 370 true raptors, Stage 4 on the 260 Accipitriformes images, and Stage 5 on all 1,680. This is worth stating explicitly because the natural implementation mistake—counting a stage's hits only on its eligible samples while dividing by the size of the whole GT—silently deflates the lower stages; on this data it would report Stage 3 as 0.36 rather than 0.82.

Uncertainty and class balance. The GT contains 10 images per species, so image-level independence is not tenable; every headline number is accompanied by a 95% confidence interval from a species-level cluster bootstrap (10,000 resamples of species with replacement). For Stage 5 we also report the trivial always-OTHERS baseline, since the CR/EN/VU classes are a small minority and overall accuracy alone would look strong without any discriminative ability. For each IUCN category we report the *structural ceiling* of recall, i.e. the fraction of that category's images that Stage 4 could identify at all: 20 of the 60 VU images belong to species outside Accipitriformes, so VU recall cannot exceed 40/60 by construction.

The operational decision. Because the system exists to decide whether to alert on a protected species, we additionally report the precision and recall of that binary decision over the four domestically protected raptors.

Each classifier outputs a confidence (score) over all candidate species for an input image. **Top-1 accuracy** is the fraction of images for which the single highest-confidence candidate matches the correct species, which corresponds to ordinary accuracy. **Top-5 accuracy** is the fraction of images for which the correct species is contained among the five highest-confidence candidates. In this task, top-1 corresponds to automatic determination (immediately presenting a single candidate) and top-5 to assistance in narrowing down candidates (presenting the top five species) when an expert makes the final determination; both are therefore practical metrics. In addition, to analyze the quality of misclassifications, we tabulate the distribution of the **rank of the correct species** (the position at which the correct species appears when the class probabilities output by the model are sorted in descending order). A rank of 1 corresponds to a top-1 hit and a rank within 5 to a top-5 hit. In particular, a rank of 6 or worse (i.e. outside the top 5) represents a state in which the correct candidate does not appear anywhere near the top of the prediction. We track its frequency independently, because it distinguishes an improvement that promotes near-misses into first place from one that rescues species which were absent from the candidate list altogether.

6 Results and discussion

6.1 Overall performance comparison

A performance comparison of the three configurations is shown in Table 6. Fine-tuning improves the raptor-related stages: Stage 3 rises from 0.7865 to 0.8189, Stage 4 top-1 from 0.2731 to 0.4038 (+13.1 pt) and macro recall by the same margin, while top-5 rises more modestly from 0.5462 to 0.5962 (+5.0 pt). Two results deserve to be stated plainly. First, **fine-tuning makes Stage 1 worse**, not better: bird/non-bird accuracy falls from 0.9726 to 0.9577 and the non-bird rejection rate from 0.9583 to 0.9274. Adapting the backbone to fine-grained bird discrimination evidently

Table 6: Overall performance comparison (bold = best in each row). Brackets give 95% species-cluster bootstrap intervals. Each stage uses its own denominator n; Stages 1 and 5 are scored on the 840 bird plus 840 non-bird GT images. Top-1 accuracy is the hit rate of the single highest-confidence candidate; top-5 accuracy is the fraction of images whose correct species is among the top five candidates (definitions in Section 5.3).

Metric	n	IN-Base	**FT-Flat**	FT-Hier
Stage 1 (bird/non-bird)	1680	**0.9726** [.964, .981]	0.9577 [.946, .968]	0.9577
non-bird rejection rate	840	**0.9583**	0.9274	0.9274
Stage 2 (raptor/non-raptor)	840	0.9286 [.906, .950]	**0.9333** [.913, .952]	0.9333
Stage 3 (four raptor orders)	370	0.7865 [.684, .881]	**0.8189** [.738, .892]	0.8189
Stage 4 top-1	260	0.2731 [.177, .377]	**0.4038** [.296, .512]	0.4038
Stage 4 top-5	260	0.5462 [.446, .639]	0.5962 [.485, .704]	**0.6115** [.508, .712]
Stage 4 macro recall	260	0.2731 [.193, .350]	**0.4038** [.320, .488]	0.4038
End-to-end IUCN accuracy	1680	0.9369 [.908, .963]	0.9452 [.920, .968]	**0.9506** [.926, .973]
(always-OTHERS baseline)	1680	0.9167	0.9167	0.9167
CR recall (ceiling 30/30)	30	14/30 = 0.467	15/30 = 0.500	**17/30 = 0.567**
EN recall (ceiling 50/50)	50	20/50 = 0.400	34/50 = 0.680	**35/50 = 0.700**
VU recall (**ceiling 40/60**)	60	25/60 = 0.417	**26/60 = 0.433**	22/60 = 0.367
CR / EN / VU precision	—	.667 / .645 / .581	.625 / .756 / .591	**.680 / .796 / .667**
Protected-species alert precision	—	0.625	0.765	**0.833**
Protected-species alert recall	40	0.250	**0.325**	0.250

Table 7: Rank distribution of the correct species (Accipitriformes, $n = 260$; bold = best).

Rank	IN-Base	**FT-Flat**	FT-Hier
1	71 (27.3 %)	**105 (40.4 %)**	105 (40.4 %)
2	30 (11.5 %)	23 (8.8 %)	27 (10.4 %)
3	16 (6.2 %)	13 (5.0 %)	12 (4.6 %)
4	14 (5.4 %)	12 (4.6 %)	8 (3.1 %)
5	11 (4.2 %)	2 (0.8 %)	7 (2.7 %)
≥ 6	118 (45.4 %)	105 (40.4 %)	**101 (38.8 %)**

costs a little of the generic bird/not-bird separability that ImageNet features provide, which is a sensible reason to keep Stage 1 on unadapted features in a deployed system. The penalty is small at native resolution but widens to 20 points once the input is degraded to the operational pixel count (Section 6.6), so this is not a marginal design choice. Second, the Stage 5 accuracies (0.937–0.951) sit only a few points above the trivial always-OTHERS baseline of 0.9167, so overall IUCN accuracy is a weak summary and the per-category numbers below carry the real information.

6.2 Rank distribution and statistical significance

The distribution of the rank of the correct species over the 260 Accipitriformes cases is shown in Fig. 8 and Table 7. Fine-tuning acts almost entirely on rank 1: its frequency rises from 27.3 % to 40.4 %, while the rank-6-or-worse mass falls only from 45.4 % to 40.4 %. In other words, the adapted features mostly promote already-near-miss cases into first place rather than rescuing species that were far outside the candidate list, which is exactly the pattern that makes the top-1 gain significant and the top-5 gain marginal.

The paired comparisons are shown in Table 8. Only the top-1 improvement from fine-tuning survives both the clustering correction and the multiplicity correction ($p_{\text{Holm}} = 0.015$ for FT-Flat, 0.019 for FT-Hier). **Neither the top-5 nor the IUCN gain does**: top-5 reaches $p_{\text{Holm}} = 0.39$ and IUCN 0.83 for FT-Flat. We therefore rest the fine-tuning claim on top-1 and macro recall alone and report the others as trends. The contrast with the image-level column is instructive:

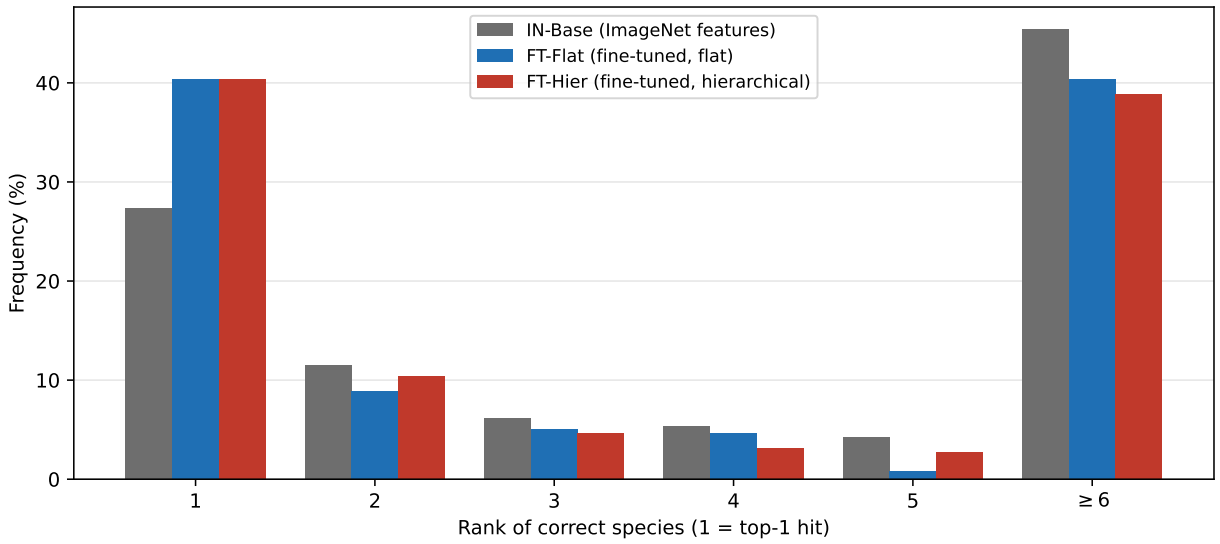

Figure 8: Distribution of the rank of the correct species over the 260 Accipitriformes cases.

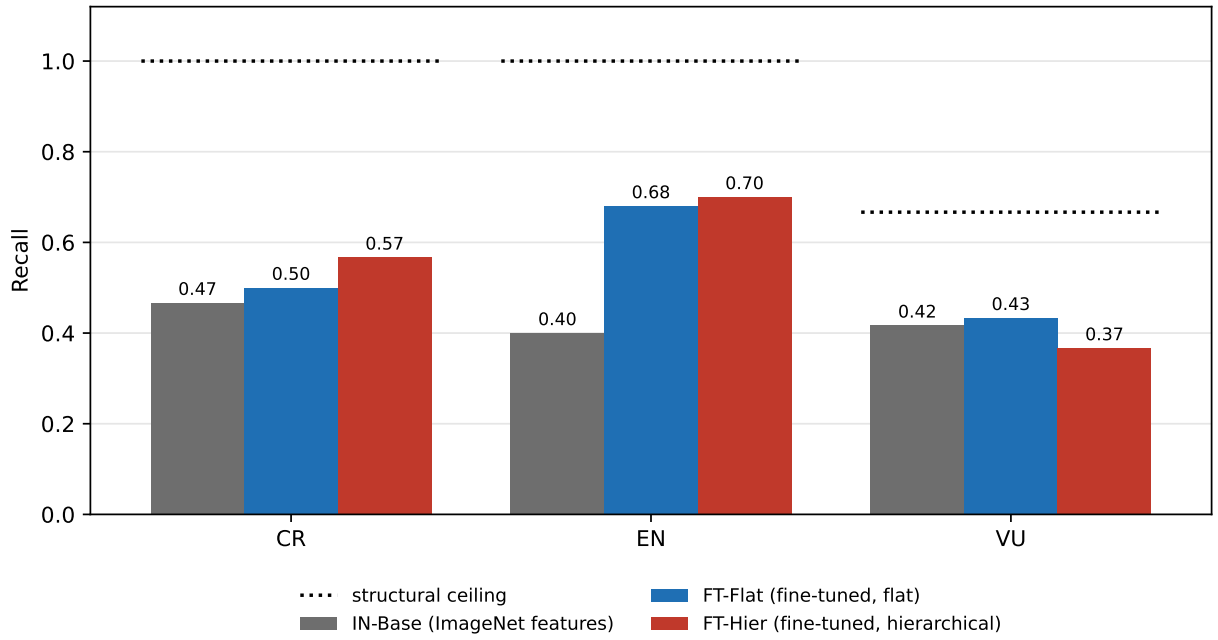

Figure 9: Recall per IUCN category (CR/EN/VU). Dotted lines mark the structural ceiling: 20 of the 60 VU images are species outside Accipitriformes, so VU recall cannot exceed 40/60.

treating the 840 images as independent would have reported the top-1 result as $p = 8.2 \times 10^{-6}$ rather than 1.7×10^{-3}, and would have made the FT-Hier IUCN gain look significant (0.015) when the species-level test puts it at 0.26.

Table 8: Paired comparisons. p is the species-cluster sign-flip test and is the value we interpret; p_{img} is the image-level McNemar p, shown only to expose how much image-level independence would inflate significance. Holm correction is applied across all nine tests. Stage 4 tests use $n = 260$, IUCN tests $n = 1{,}680$.

Comparison	Metric	b	c	p	p_{Holm}	p_{img}	Conclusion
IN-Base → FT-Flat	top-1	12	46	1.7×10^{-3}	**0.015**	8.2×10^{-6}	significant
IN-Base → FT-Flat	top-5	16	29	0.065	0.390	0.072	not significant
IN-Base → FT-Flat	IUCN	26	40	0.226	0.834	0.109	not significant
IN-Base → FT-Hier	top-1	9	43	2.4×10^{-3}	**0.019**	2.0×10^{-6}	significant
IN-Base → FT-Hier	top-5	23	40	0.193	0.834	0.043	not significant
IN-Base → FT-Hier	IUCN	15	38	0.036	0.255	2.2×10^{-3}	not significant
FT-Flat → FT-Hier	top-1	12	12	1.000	1.000	1.000	no difference
FT-Flat → FT-Hier	top-5	17	21	0.864	1.000	0.627	no difference
FT-Flat → FT-Hier	IUCN	9	18	0.167	0.834	0.122	no difference

6.3 Does the genus-level hierarchy help?

With both arms converged (Section 5.2), no difference between FT-Flat and FT-Hier is detectable: top-1 is identical at 0.4038 with $b = c = 12$ ($p = 1.00$), and top-5 differs by +1.5 pt in favour of the hierarchy ($p = 0.86$, $p_{\text{Holm}} = 1.00$).

The important point is what may and may not be concluded from this. A non-significant test is not evidence of equivalence, so we pre-specified an absolute margin of $\delta = 5$ pt on Stage 4 top-5 and tested non-inferiority against it using the species-cluster bootstrap. The observed difference is +1.5 pt with a 95% interval of $[-5.4, +10.4]$ pt. Because the lower bound falls outside $-\delta$, **non-inferiority is not established**, and since the interval is also wider than $+\delta$, neither is equivalence. The honest summary is that this experiment has too little resolution, at $n = 260$ with ten images per species, to separate the two designs in either direction.

We therefore make no design claim about hierarchical refinement in either direction. What the data support is narrower: under a fixed feature extractor and equal, verified per-classifier training caps, the two Stage 4 designs gave point estimates close enough that no difference was detected—but, as above, neither interchangeability nor equivalence was established. A practitioner choosing between them today is choosing under acknowledged uncertainty, and may reasonably weigh engineering factors: the hierarchical variant gave higher CR/EN recall and alert precision here, the flat variant is simpler to train and to retrain when classes change.

6.4 Merging Stage 3 and Stage 4

Section 6.3 asked whether adding a level to the hierarchy helps. The opposite question is whether one of the existing levels earns its place. Stage 3 decides the raptor order before Stage 4 decides the species, which makes an order-level mistake terminal: an Accipitriformes bird routed to Falconiformes is never offered to species identification at all, and Stage 3 is right on 0.8189 of raptors. We therefore trained a single classifier over the 20,238 raptor training samples with 74 classes—the 71 species, plus one class each for Cathartiformes, Falconiformes and Strigiformes—and put it in place of both stages. Everything else, including Stages 1 and 2, is unchanged, and the features, seed, caps and hyper-parameters are those of the separate stages. To remove the library as a variable we also retrained Stage 3 and Stage 4 in the same environment; the retrained Stage 4 stopped at iteration 12,894, exactly as the shipped model did.

Table 9 gives the result. No metric separates the two designs: the paired species-cluster tests return $p_{\text{Holm}} = 1.000$ for Stage 4 top-1 and top-5 and 0.81 for end-to-end IUCN. What does change is the balance between precision and recall on the alert. The merged model raises alert precision

Table 9: Merging Stage 3 and Stage 4, over the same 1,680 ground-truth images. Accuracy differences are not significant (paired species-cluster tests, $p_{\text{Holm}} = 1.000$ for top-1 and top-5, 0.81 for IUCN). Latency is measured on the Jetson Orin Nano under the operational preprocessing of Section 4.3; model size is the classifier stages the cascade must hold.

	Stage 3 → Stage 4	Merged (74 classes)		
Metric	500 rounds	3,385 (all)	500	250
Stage 3 raptor order	.8189	.8108	.7865	.7703
Stage 4 top-1	**.4038**	.3923	.3885	.3769
Stage 4 top-5	**.5962**	.5731	.5615	.5538
End-to-end IUCN	.9452	.9500	**.9512**	.9500
Alert precision	.765	.923	**.929**	.923
Alert recall	**.325**	.300	**.325**	.300
(tp / fp)	13 / 4	12 / 1	**13 / 1**	12 / 1
End-to-end latency [ms]	111.9	126.7	**109.9**	107.9
LightGBM p95 [ms]	33.0	100.5	33.1	22.4
Classifier size [MB]	126	380	134	83
Reached Stage 4 (of 370)	312	262	243	234

from 0.765 to 0.929, cutting false positives from four to one, and at 500 boosting rounds it does so while keeping recall at 0.325—the same recall as the two-stage cascade. End-to-end IUCN accuracy rises from 0.9452 to 0.9512.

The mechanism is visible in where candidates leave the cascade. With the merged model, 137 of the 370 raptors are assigned to one of the three other orders, against 68 under Stage 3, and Stage 4 is reached by 243 instead of 312. The three order classes act as a sink: a bird that is not an Accipitriformes has somewhere to go other than the nearest of 71 species. That also explains the small loss in top-1 (0.4038 to 0.3885, not significant), since some Accipitriformes are absorbed too. The training set makes this asymmetry plausible—each species contributes 150 to 372 samples while the three orders contribute 926, 2,381 and 2,371—so the sink classes are the well-populated ones.

On the device the merged cascade is not more expensive. At 500 rounds it costs 109.9 ms per object against 111.9 ms for the two-stage cascade, with the same LightGBM tail (p95 33.1 ms against 33.0 ms), and it replaces two models with one. Using the merged model at its full 3,385 rounds is the configuration to avoid: 126.7 ms with a p95 of 100.5 ms, for no accuracy gained. As in Section 4.3, the rounds beyond a few hundred buy nothing here.

Stage 3 was in the design because it mirrors a real taxonomic level, the raptor order, and that was a reasonable thing to build. The measurement, however, finds no reason in either accuracy or speed to keep it as its own stage. The two designs are statistically indistinguishable on every accuracy metric, and on the two that matter operationally—alert precision and end-to-end IUCN accuracy— the merged model is ahead. It costs the same per object (109.9 against 111.9 ms) and halves the number of classifiers the device must hold. We therefore conclude that **Stage 3 and Stage 4 should be merged**. The flexibility argument for keeping them apart—that the species model can be retrained without touching the order model—survives in principle, but retraining one 74-class model covers the same ground, so it is thin grounds for a second stage. A deployment that merges them should still truncate the model to a few hundred rounds rather than run it out, for the reason given in Section 4.3.

6.5 Recall by threatened-species category

Recall per IUCN category is shown in Fig. 9 and Table 6. The largest effect is in the EN (Endangered) category, which rises from 20/50 (40.0 %) to 34/50 (68.0 %) with FT-Flat and 35/50 (70.0 %) with FT-Hier; the species concerned include the Bateleur, the Steppe Eagle (*Aquila nipalensis*) and the Martial Eagle. CR rises from 14/30 to 15/30 (FT-Flat) and 17/30 (FT-Hier), and VU from 25/60 to 26/60 (FT-Flat) while FT-Hier falls to 22/60. Precision moves in the same direction as recall for EN (0.645 → 0.756 → 0.796), so the EN gain is not bought by over-predicting the category.

Two caveats belong next to these numbers. VU recall is bounded above by 40/60 because 20 of the 60 VU images are species outside Accipitriformes that Stage 4 cannot name at all; measured against its reachable subset, FT-Flat identifies 26 of 40. And the operational quantity—the alert on one of the four domestically protected raptors—remains weak: recall is 0.325 for FT-Flat and 0.250 for FT-Hier at precisions of 0.765 and 0.833. A system that misses two thirds of the individuals it exists to protect is not yet deployable on that criterion alone, and we regard raising this recall, rather than the aggregate IUCN accuracy, as the primary target for further work.

6.6 Sensitivity to input resolution

The evaluation above uses citizen-science photographs, whereas the operational input is a PTZ crop carrying on the order of 32×32 to 128×128 pixels (Section 1.2). To measure what that costs, we ran a controlled down-sampling study: the 1,680 ground-truth images were reduced to $S \times S$ and enlarged back to 224×224 (bicubic in both directions, applied after the usual resize-and-crop), and features were re-extracted at $S = 128$ and $S = 32$. **Nothing was re-trained.** The backbone and all LightGBM stages are the ones behind every other table in this paper, and the $S = 224$ column is the native-resolution result reproduced by the same evaluation script, so the differences isolate input resolution. Results are in Table 10.

At 128×128 the cascade degrades but keeps working. Stage 4 top-1 falls from 0.404 to 0.292 for FT-Flat and 0.300 for FT-Hier, a drop that is significant in every arm ($p_{\text{Holm}} = 0.014$, 0.011 and 0.009 for FT-Flat, FT-Hier and IN-Base). The top-5 loss is smaller and, in the fine-tuned arms, not significant ($p_{\text{Holm}} = 0.14$ for FT-Flat and 0.22 for FT-Hier). Read operationally, 128×128 still supports the narrow-the-candidates use of the system but no longer supports automatic determination.

At 32×32 the system fails. Stage 4 top-1 collapses to 0.065 (FT-Flat) and top-5 to 0.219, both significant against the native condition ($p_{\text{Holm}} = 9.0 \times 10^{-5}$ and 5.8×10^{-4}), and end-to-end IUCN accuracy of 0.918 is indistinguishable from the always-OTHERS baseline of 0.9167: at this resolution the pipeline has essentially no discriminative ability left. Two further observations matter more than the headline numbers.

First, **the apparent rise of the protected-species alert recall at 32×32 is not an improvement**. Recall goes up (0.325 → 0.500 for FT-Flat) only because precision collapses from 0.765 to 0.099: the confusion counts move from 13 true positives against 4 false positives to 20 against 183. The classifier is firing the alert almost indiscriminately, and a recall figure read on its own would misrepresent that as progress. This is the clearest case in our data where the pair of precision and recall must be read together. Operationally the consequence is blunt: at that resolution the alert carries no information a turbine controller could act on, and an operator who tuned the system on recall alone would conclude the opposite.

Second, **the benefit of fine-tuning does not survive to 32×32**. FT-Flat reaches 0.065 top-1 against IN-Base's 0.069 — the ordering reverses, though the gap is negligible — whereas at

Table 10: Sensitivity to input resolution. Each ground-truth image is reduced to $S \times S$ and enlarged back to the network input (bicubic both ways); no model is re-trained, so the columns differ only in input resolution. $S = 224$ is the native condition of Table 6. Stages 1–3 are shared between FT-Flat and FT-Hier, so those columns coincide. Recall rows for CR/EN/VU are hit counts out of n. The cascade stays usable at 128 but not at 32, where end-to-end accuracy falls to the always-OTHERS baseline (.917) and the alert fires almost indiscriminately, so its higher recall comes with a precision of .099.

		IN-Base			FT-Flat			FT-Hier		
Metric	n	224	128	32	224	128	32	224	128	32
Stage 1 (bird/non-bird)	1680	.973	.965	.827	.958	.951	.626	.958	.951	.626
non-bird rejection	840	.958	.952	.692	.927	.911	.261	.927	.911	.261
Stage 2 (raptor or not)	840	.929	.879	.745	.933	.918	.756	.933	.918	.756
Stage 3 (raptor order)	370	.786	.749	.595	.819	.797	.635	.819	.797	.635
Stage 4 top-1	260	.273	.165	.069	.404	.292	.065	.404	.300	.062
Stage 4 top-5	260	.546	.388	.188	.596	.542	.219	.612	.573	.215
End-to-end IUCN	1680	.937	.927	.912	.945	.939	.918	.951	.934	.919
(always-OTHERS)	1680	.917	.917	.917	.917	.917	.917	.917	.917	.917
CR recall (of 30)	30	14	12	1	15	15	6	17	13	4
EN recall (of 50)	50	20	14	0	34	25	6	35	23	4
VU recall (of 60)	60	25	7	4	26	16	2	22	15	1
Alert precision	—	.625	.417	.125	.765	.714	.099	.833	.588	.101
Alert recall	40	.250	.125	.575	.325	.250	.500	.250	.250	.450

128×128 the fine-tuned advantage is preserved (0.292 vs. 0.165). The same pattern is far stronger at Stage 1: the small penalty that fine-tuning imposes on bird/non-bird accuracy at native resolution (0.9726 → 0.9577, Section 6.1) widens into a 20-point gap at 32×32 (0.827 → 0.626), with the non-bird rejection rate falling from 0.692 to 0.261. Adapting the backbone to fine-grained bird discrimination evidently buys species-level detail at the price of robustness to degraded input, which strengthens the design suggestion of Section 6.1: keep Stage 1 on unadapted features.

The practical consequence is a requirement on the sensor side rather than on the classifier. Under the present design the PTZ stage must deliver at least 128×128 pixels on the target, and 32×32 crops cannot be salvaged by the current cascade. This is consistent in direction with Yoshihashi et al. [31], who reported degradation across 15–20, 21–30 and 31–50 pixel bands; our measurement locates the practical breakdown of this particular pipeline at the 32-pixel scale. Whether super-resolution preprocessing, training on synthetically degraded images, or a backbone chosen for low-resolution robustness can move that boundary is left for future work.

6.7 Accuracy on the target device

The latency measurement of Section 4.3 is only meaningful if the device produces the same answers as the server, so we scored the device predictions with the same evaluation script and compared them against the server run (Table 11). No metric separates the two. The paired species-cluster tests give $p = 1.000$ for Stage 4 top-1, top-5 and end-to-end IUCN alike: on top-1 the discordant pairs are $b = 0$ against $c = 1$, meaning a single image that the device gets right and the server does not.

Four of the 1,680 cascade decisions differ. All four are borderline cases in which the correct species was already outside the top candidate on one side or the other, and in one of them—a Black Kite—the device is the one that is correct. This is the expected consequence of a relative difference of at most 1.1% in the features (Section 4.3) acting on an argmax; it does not accumulate into a measurable accuracy loss. Running the extractor in FP16 on the GPU perturbs the features less than the CPU–server gap does (cosine 0.9999980, relative L_2 0.20%, against 1.07%), so the FP16

Table 11: Accuracy of the identical pipeline on the server and on the Jetson Orin Nano, over the same 1,680 ground-truth images, scored with the same script. Recall rows are hit counts. The paired species-cluster sign-flip test gives $p = 1.000$ for every comparison it applies to.

Metric	n	Server (H200)	Orin Nano
Stage 1 (bird/non-bird)	1680	.9577	.9583
non-bird rejection	840	.9274	.9286
Stage 2 (raptor or not)	840	.9333	.9345
Stage 3 (raptor order)	370	.8189	.8189
Stage 4 top-1	260	.4038	.4077
Stage 4 top-5	260	.5962	.5923
Stage 4 macro recall	260	.4038	.4077
End-to-end IUCN accuracy	1680	.9452	.9452
CR recall (of 30)	30	15	15
EN recall (of 50)	50	34	34
VU recall (of 60)	60	26	26
Protected-species alert precision	—	.765	.813
Protected-species alert recall	40	.325	.325

deployment path inherits this conclusion.

7 Practical value and contribution to ecology

7.1 Inference cost and integration into bird-strike prevention

The proposed system requires no GPU at inference time and runs on a CPU. Stage 1 rejects non-birds at 0.9274 (FT features) to 0.9583 (ImageNet features). Stage 2's 0.9333 is an accuracy over both of its classes, not a non-raptor rejection rate, and should not be read as the latter; what matters operationally is that the two lower stages together keep most candidates away from the expensive multi-class stages, which is what holds the median cost down on the device. As Section 4.3 makes precise, however, the narrowing does not remove the backbone forward pass ahead of Stage 1, and it does not help the candidates that do reach Stage 4, where the classifier itself becomes the largest term. Measured on the Jetson Orin Nano the pipeline costs 226 ms per object as written, and 112 ms on the CPU alone once Stage 4 is truncated and the crop arrives in memory, or 13 ms with the backbone on the GPU (Section 4.3). The CPU-only figure exceeds the 0.1 s per-object target but is well inside the application's timing, since a few tens of candidates at 112 ms is a few seconds against roughly 50 s of approach. We therefore describe the integration into the wind-turbine system of Section 1.1—identify the approaching bird from the PTZ magnified image and pass an alert for a stop-command species to turbine control—as the intended deployment path rather than as a demonstrated capability, and note that on today's protected-species alert recall (0.25–0.33) such an alert would miss most individuals. The same configuration can also be deployed in the continuously operating environment of automatic cameras (camera traps) in protected areas.

Power in the field. What the sensor unit draws decides how long a survey can run away from mains power, and reaching a distant site means flying with the equipment: lithium-ion batteries taken into a cabin are capped at 100 Wh. We therefore power the unit from an Anker PowerHouse 90, which stays under that cap while providing an AC outlet. Measured consumption is 9 W for the four-port PoE+ switch (I-O DATA ETG-POE04), 7 W for each of the three fixed 4K cameras (HV-800E6A5), 25 W for the 4K PTZ camera (HV-PTZ810A) while its motors are driven and 10 W at rest, and 15 W for the Jetson Orin Nano 8GB. The system draws 70 W at peak and 55 W in normal operation, since the PTZ motors do not run continuously, which gives about one hour

of survey time on that battery. Mounted in a vehicle the limit disappears: a DC–AC converter on the 12 V accessory socket supplies the same AC mains, so runtime is bounded by fuel rather than by the battery. This is also why the classifier is held to the 15 W mode of Section 1.2: the compute shares a budget with three fixed cameras and a motorised PTZ camera, and raising its clocks would come directly out of the survey hour.

7.2 Contribution to ecology and conservation research

The significance of this system lies in the following three points. **(a) Improved survey efficiency through automatic identification:** Species identification by experts is time-constrained and rate-limits large-scale analysis. The proposed system identifies the correct species outright in 40 % of cases over 71 Accipitriformes species and places it in the top five in 60 %, which is useful for narrowing down candidates for an expert's final determination but not for unattended labelling. **(b) Improved detection of threatened species:** It raises EN recall from 40 % to 68–70 % at improved precision, which helps surface low-frequency species that are easily overlooked. CR recall (50–57 %) and the protected-species alert recall (25–33 %) remain the limiting quantities. **(c) Reproducibility:** The trained model weights (102 MB), the prediction files behind every table, and the evaluation script that produces the denominators, bootstrap intervals, cluster sign-flip tests, Holm correction and non-inferiority test are deposited as a single versioned archive with a DOI [46], so that each number here can be recomputed rather than taken on trust. The deposit contains the weights themselves and not only the code, and it is citable through a concept DOI that always resolves to the latest version. The evaluation script requires nothing beyond the Python standard library and is seeded, and the command documented in the archive regenerates the evaluation output behind every table in this paper byte for byte, so reproduction does not depend on the reader's environment. The inference path, which does not require a particular high-end GPU, supports broad adoption. It should be noted that in this field there are not a few cases in which published data resources become unreachable in later years. For example, the dataset of bird images around a wind turbine (`http://bird.nae-lab.org/dataset`) that had been released by Yoshihashi et al. [30], whose application background is closest to this work, is currently inaccessible. Such cases show that the sustained publication and preservation of trained models and data is itself indispensable from the standpoint of reproducibility. For this reason the weights of this work are deposited in a preservation repository under a DOI rather than being left on a code-hosting service alone.

7.3 Limitations and future work

This evaluation uses ordinary citizen-science images as a proxy for operational PTZ magnified imagery, and the two domains differ in several respects at once: resolution, flight posture, motion blur, and outdoor illumination under tracking. Of these we have now measured only the first. The controlled down-sampling study of Section 6.6 quantifies the resolution gap and shows it to be severe — usable at 128×128, broken at 32×32 — but it degrades ordinary photographs rather than reproducing the other three factors, and it therefore gives an upper bound on performance at those pixel counts rather than an estimate of field performance. An evaluation on real PTZ images remains the single most important outstanding task, and until it exists the numbers in this paper should be read as characterizing the method, not the deployed system. All GT images come from iNaturalist and thus share the photographic, geographic and tagging biases of that platform, so independent validation using the Macaulay Library or GBIF is also necessary.

The protected-species alert recall of 0.25–0.33 at native resolution and the CR recall of 0.50–

0.57 are the limiting numbers for the intended application; training data for CR species are scarce (fewer than 70 images per species), which makes data augmentation and few-shot learning the natural next steps. The resolution study adds a caution about how this quantity is read: at 32×32 the measured alert recall rises to 0.45–0.58 while precision falls to about 0.10, so recall must always be reported against its precision rather than tracked alone.

The edge measurement covers the deployed configuration but not every variant of it. Section 4.3 measures the pipeline on the target device and tunes it there, and Section 6.7 shows the accuracy is unchanged throughout. Two caveats attach to the tuned figures. The in-memory crop is our reconstruction of what the PTZ stage would hand over—the paper's own preprocessing applied ahead of time, so the pixels and hence the accuracy are identical—but it is not a measurement on real PTZ output, which we do not yet have. And the quantized or distilled backbone remains a projection: the 3–9 ms server-CPU figure is for a lightweight backbone we did not train, and INT8 quantization of the present backbone failed on this device, either unsupported or, when it ran, agreeing with the server on barely half of the Stage 2 decisions.

One design choice of this study also bounds what it can say. All five stages are trained from a single feature set, which keeps stage-to-stage differences attributable to the classifiers rather than to the features, but it means we cannot report what an Accipitriformes-weighted training subset would do for Stages 2–4; that trade-off is worth measuring separately.

8 Conclusion

This work proposed a five-stage hierarchical classification system that assigns IUCN Red List categories to birds in citizen-science images, intended as a candidate module for the bird-species classification function of a real-time, low-power edge system for wind-turbine bird-strike prevention. Evaluation used citizen-science images as a proxy for operational PTZ magnified imagery. On an independent evaluation set of 84 species × 10 images, augmented with 840 held-out non-bird images so that every stage could be scored against its own denominator, fine-tuning the backbone raised Stage 4 top-1 accuracy and macro recall from 0.273 to 0.404 (+13.1 pt; species-cluster test, $p_{\text{Holm}} = 0.015$) and EN recall from 20/50 to 34/50 at improved precision, while the top-5 gain (+5.0 pt) and the end-to-end IUCN gain did not survive clustering and multiplicity correction. Training a flat and a genus-level hierarchical Stage 4 to verified convergence under identical per-classifier caps, we found no detectable difference between them and, under a pre-specified 5 pt margin, could establish neither superiority nor equivalence. A controlled down-sampling study locates the operating boundary of the present design: at 128×128 the cascade degrades but still narrows candidates, while at 32×32 it loses discriminative ability altogether, and the alert recall that appears to rise there is an artefact of precision collapsing to 0.10. Delivering at least 128×128 pixels on the target is therefore a requirement on the PTZ stage rather than something the classifier can absorb. Replacing the order and species stages with one 74-class classifier changes no accuracy metric significantly while raising alert precision from 0.765 to 0.929 and end-to-end IUCN accuracy to 0.9512, at the same per-object cost and with one classifier instead of two: the taxonomic order stage, though a natural thing to build, earns its place in neither accuracy nor speed, and the two stages should be merged. Running the same pipeline on the target edge device, a Jetson Orin Nano, leaves accuracy unchanged—no metric separates it from the server ($p = 1.000$). It costs 226 ms per object as written, and the measurement relocates the bottleneck away from the backbone: the tail is set by the 12,894-round Stage 4 classifier and three quarters of the preprocessing is JPEG decoding the evaluation photographs. Truncating Stage 4 to 500 rounds at equal accuracy ($p = 1.000$) and supplying the crop in memory as the camera does brings the CPU-only pipeline

to 112 ms, and the backbone in FP16 on the GPU brings it to 13 ms. The CPU-only figure does not reach the 0.1 s per-object target but is a usable speed for the application, in which a few tens of candidates are classified against roughly 50 s of approach. Future work will prioritize evaluation on real PTZ magnified images, independent validation with the Macaulay Library or GBIF, and few-shot learning for CR species—with the recall of the protected-species alert, currently 0.25–0.33 and always to be read against its precision, as the primary quantity to improve.

Acknowledgments

This work was carried out using the computing environment of the Foundation for Computational Science. We thank the providers of the iNaturalist API and of the IUCN Red List data.

References

[1] iNaturalist contributors: iNaturalist (online), available at `https://www.inaturalist.org/` (accessed 2026-04-22).

[2] Sullivan, B. L., Aycrigg, J. L., Barry, J. H. et al.: The eBird enterprise: an integrated approach to development and application of citizen science, *Biological Conservation*, Vol. 169, pp. 31–40 (2014).

[3] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778 (2016).

[4] He, K., Girshick, R. and Dollár, P.: Rethinking ImageNet pre-training, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4918–4927 (2019).

[5] Ke, G., Meng, Q., Finley, T. et al.: LightGBM: a highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst. (NIPS)*, Vol. 30, pp. 3146–3154 (2017).

[6] Li, S., Zhao, Y., Varma, R. et al.: PyTorch distributed: experiences on accelerating data parallel training, *Proc. VLDB Endow.*, Vol. 13, pp. 3005–3018 (2020).

[7] Goyal, P., Dollár, P., Girshick, R. et al.: Accurate, large minibatch SGD: training ImageNet in 1 hour, arXiv:1706.02677 (2017).

[8] Yosinski, J., Clune, J., Bengio, Y. and Lipson, H.: How transferable are features in deep neural networks?, *Adv. Neural Inf. Process. Syst. (NIPS)*, Vol. 27, pp. 3320–3328 (2014).

[9] Lee, Y., Chen, A. S., Tajwar, F. et al.: Surgical fine-tuning improves adaptation to distribution shifts, *Proc. Int. Conf. Learn. Represent. (ICLR)* (2023).

[10] Kumar, A., Raghunathan, A., Jones, R., Ma, T. and Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution, *Proc. Int. Conf. Learn. Represent. (ICLR)* (2022).

[11] Kolesnikov, A., Beyer, L., Zhai, X. et al.: Big Transfer (BiT): general visual representation learning, *Proc. European Conf. Comput. Vis. (ECCV)*, pp. 491–507 (2020).

[12] Jia, M., Tang, L., Chen, B.-C. et al.: Visual prompt tuning, *Proc. European Conf. Comput. Vis. (ECCV)*, pp. 709–727 (2022).

[13] He, J., Chen, J.-N., Liu, S. et al.: TransFG: a transformer architecture for fine-grained recognition, *Proc. AAAI Conf. Artif. Intell.*, pp. 852–860 (2022).

[14] Wang, J., Yu, X. and Gao, Y.: Feature fusion vision transformer for fine-grained visual categorization, *Proc. British Machine Vision Conf. (BMVC)* (2021).

[15] Zhu, H., Ke, W., Li, D. et al.: Dual cross-attention learning for fine-grained visual categorization and object re-identification, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4692–4702 (2022).

[16] Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W. and Belhumeur, P. N.: Birdsnap: large-scale fine-grained visual categorization of birds, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2019–2026 (2014).

[17] Branson, S., Van Horn, G., Belongie, S. and Perona, P.: Bird species categorization using pose normalized deep convolutional nets, *Proc. British Machine Vision Conf. (BMVC)* (2014).

[18] Lin, T.-Y., RoyChowdhury, A. and Maji, S.: Bilinear CNN models for fine-grained visual recognition, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1449–1457 (2015).

[19] Fu, J., Zheng, H. and Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4438–4446 (2017).

[20] Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S.: The Caltech-UCSD Birds-200-2011 dataset, Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011).

[21] Van Horn, G., Branson, S., Farrell, R. et al.: Building a bird recognition app and large scale dataset with citizen scientists, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 595–604 (2015).

[22] Van Horn, G., Mac Aodha, O., Song, Y. et al.: The iNaturalist species classification and detection dataset, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8769–8778 (2018).

[23] Van Horn, G., Cole, E., Beery, S. et al.: Benchmarking representation learning for natural world image collections, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12884–12893 (2021).

[24] Mac Aodha, O., Cole, E. and Perona, P.: Presence-only geographical priors for fine-grained image classification, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 9596–9606 (2019).

[25] Kang, B., Xie, S., Rohrbach, M. et al.: Decoupling representation and classifier for long-tailed recognition, *Proc. Int. Conf. Learn. Represent. (ICLR)* (2020).

[26] Beery, S., Morris, D. and Yang, S.: Efficient pipeline for camera trap image review, arXiv:1907.06772 (2019).

[27] Ahumada, J. A., Fegraus, E., Birch, T. et al.: Wildlife Insights: a platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet, *Environmental Conservation*, Vol. 47, No. 1, pp. 1–6 (2020).

[28] Norouzzadeh, M. S., Nguyen, A., Kosmala, M. et al.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning, *Proc. Natl. Acad. Sci. (PNAS)*, Vol. 115, No. 25, pp. E5716–E5725 (2018).

[29] Beery, S., Van Horn, G. and Perona, P.: Recognition in terra incognita, *Proc. European Conf. Comput. Vis. (ECCV)*, pp. 456–473 (2018).

[30] Yoshihashi, R., Kawakami, R., Iida, M. and Naemura, T.: Bird detection and species classification with time-lapse images around a wind farm: dataset construction and evaluation, *Wind Energy*, Vol. 20, No. 12, pp. 1983–1995 (2017).

[31] Yoshihashi, R., Kawakami, R., Iida, M. and Naemura, T.: Construction of a bird image dataset for ecological investigations, *Proc. IEEE Int. Conf. Image Process. (ICIP)*, pp. 4248–4252 (2015).

[32] Takeki, A., Trinh, T. T., Yoshihashi, R., Kawakami, R., Iida, M. and Naemura, T.: Combining deep features for object detection at various scales: finding small birds in landscape images, *IPSJ Trans. Computer Vision and Applications*, Vol. 8, Art. 5 (2016).

[33] McClure, C. J. W., Martinson, L. and Allison, T. D.: Automated monitoring for birds in flight: proof of concept with eagles at a wind power facility, *Biological Conservation*, Vol. 224, pp. 26–33 (2018).

[34] Stevens, S., Wu, J., Thompson, M. J. et al.: BioCLIP: a vision foundation model for the tree of life, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19412–19424 (2024).

[35] Zizka, A., Andermann, T. and Silvestro, D.: IUCNN — deep learning approaches to approximate species' extinction risk, *Diversity and Distributions*, Vol. 28, No. 2, pp. 227–241 (2022).

[36] Borgelt, J., Dorber, M., Høiberg, M. A. and Verones, F.: More than half of data deficient species predicted to be threatened by extinction, *Communications Biology*, Vol. 5, Art. 679 (2022).

[37] Tuia, D., Kellenberger, B., Beery, S. et al.: Perspectives in machine learning for wildlife conservation, *Nature Communications*, Vol. 13, Art. 792 (2022).

[38] Čermák, V., Picek, L., Adam, L. and Papafitsoros, K.: WildlifeDatasets: an open-source toolkit for animal re-identification, *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 5953–5963 (2024).

[39] Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S. and Lord, N. A.: Making better mistakes: leveraging class hierarchies with deep networks, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12506–12515 (2020).

[40] Chang, D., Pang, K., Zheng, Y. et al.: Your "Flamingo" is my "Bird": fine-grained, or not, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11476–11485 (2021).

[41] Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I. and Lempitsky, V.: Hyperbolic image embeddings, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6418–6428 (2020).

[42] Silla, C. N. and Freitas, A. A.: A survey of hierarchical classification across different application domains, *Data Min. Knowl. Discov.*, Vol. 22, pp. 31–72 (2011).

[43] Yan, Z., Zhang, H., Piramuthu, R. et al.: HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition, *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 2740–2748 (2015).

[44] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages, *Psychometrika*, Vol. 12, pp. 153–157 (1947).

[45] iNaturalist API documentation (online), available at `https://api.inaturalist.org/v1/docs/` (accessed 2026-04-22).

[46] Nishikawa, T.: bs2026-v6-artifacts: model weights, predictions and evaluation code for IUCN Red List threatened bird classification from citizen-science images, Zenodo (2026), `https://doi.org/10.5281/zenodo.21894830`.