Data Engineering and Applications

Research Paper

Complex Scenario Wildlife Detection Based on Improved Faster R-CNN and Feature Pyramid Networks

Hana Hájek[1,*] and Adéla Svoboda[1]

[1] Department of Computer Systems, Brno University of Technology, 61669 Brno, Czech Republic
*Corresponding author: hana.h@fit.vut.cz

Abstract. Even though deep learning has increased the automation of wild animal monitoring, issues including occlusion, scale variation, and crowded backdrops in natural settings remain unresolved. To solve the aforementioned issues, this paper proposes a detection architecture that combines an enhanced Feature Pyramid Network with context-adaptive Faster R-CNN. The stratified camera-trap dataset of 12,483 photos featuring 32 different species of wildlife was subjected to three different types of methods: multi-scale feature fusion, attention-based proposal mechanisms, and explicit context modeling. The experiment will evaluate the model's robustness and capacity for generalization using cross-validation and several domains. The aforementioned findings show that the primary test set's mean Average Precision (mAP) is 0.872, greater than that of the baseline detectors SSD (0.756 mAP) and YOLOv4 (0.783 mAP). The suggested framework exhibits a slight decrease of only 0.033 in mAP under challenging cross-domain transfer settings, and it has attained a comparatively high precision (more than 0.80) in dense forest, grassland, and nocturnal situations. Failure analysis reveals that the addition of a spatial attention module considerably reduced the number of false positives and occlusion-induced misses. According to the aforementioned findings, adaptive attention mechanisms and multi-level feature aggregation have greatly increased detection accuracy and resilience, supporting the implementation of large-scale intelligent monitoring systems in biodiversity assessment.

Keywords: *R-CNN, Object Detection, Feature Pyramid Network, Wildlife Monitoring, Scene Robustness*

Received on 27 June 2025, Accepted on 17 December 2025, Published on 05 January 2026

Introduction

People are starting to investigate these occurrences in order to gain a better understanding of how and where different species in severe habitats are moving and changing. Maintain the ecological balance and accurately identify and count wild animals during conservation efforts for endangered species [1]. However, there are a number of issues with real-world field circumstances, including uneven lighting, frequently veiled items, habitat overlap, a multitude of backgrounds, and the existence of cryptic species, which can all be confusing to seasoned observers [2]. The previous manual monitoring option was expensive since it required both a camera trap and an expensive field, both of which were manual processes that were subject to human mistake [3]. Early computational work lacked the discriminating power to handle dynamic and crowded outdoor situations since it was typically based on low-level vision features or basic motion signals [4]. Many scientists are now working to create automated and intelligent frameworks for picture analysis that are more scalable and consistent as a result of the aforementioned shortcomings [5]. Large-scale ecological data has gradually accumulated due to the quick development of sensors and image-acquisition technology; as a result, sophisticated computation is currently being used extensively in the field of wildlife detection [6]. Environmental unpredictability, underrepresentation of rare species, and data asymmetry continue to be issues [7]. As a result, current research in this field is concentrated on creating workable, highly accurate, or flexible solutions [8].

Convolutional neural networks (CNNs) are currently at the forefront of object detection, which has advanced significantly over the last 10 years due to the advent of deep learning architectures [9]. Using hierarchical feature extraction and end-to-end training, Fast R-CNN, YOLO, and SSD are example high-performing detectors that have raised the bar for real-world performance [10]. The intrinsic visual complexity and considerable intra-class variation of wildlife make it difficult to detect them in an uncontrolled environment, despite some progress [11]. Feature Pyramid Networks (FPN) can be utilized to obtain more accurate localization of objects of various sizes, and multi-scale feature processing has been created recently [12]. To increase the range of applications for detectors, robust backbones and adaptive fusion modules have been included into frameworks [13]. Unreliable detection of small or concealed objects, poorly fitted models, and a lack of generalization to different contexts are examples of persistent issues [14]. In order to solve the issues with algorithm efficacy in practice, current research has concentrated on data augmentation, domain adaptability, and hybrid feature representation [15].

In order to improve performance in complicated biological settings, this research suggests a new unified deep-learning-based wildlife detection framework that integrates Faster R-CNN and FPN. The current work focuses on multi-scale feature fusion, domain robustness, and extensive ablation experiments under different settings, all of which are based on the research mentioned above. In addition to providing references for the community of computer vision and wildlife research, the system offers comprehensive analysis of the model's structure and training approach, as well as the experimental plan and field application results. This paper's contents are as follows: first, related research is described; second, the suggested solution to the problem and experimental results are explained; and third, important conclusions and future research prospects are offered.

Literature Review

Progress in Complex Wildlife Detection

Using a variety of deep-learning approaches, numerous artificial intelligence systems have been deployed in recent years to automatically identify diverse wild animal species in the natural world. Initially, the creatures were separated from a complicated background using the outdated methods of edge detection, color histograms, and background subtraction [16]. The performance of the aforementioned methods drastically decreased in the presence of dense foliage, changing shadows, and other factors, even though they were successful in controlled settings or with quite different kinds of subjects [17]. The issue is made worse by the fact that many wildlife settings have occlusion, which can be caused by vegetation, rocks, or animal group behavior. Partial views of the target may also be at odds with global heuristics [18]. Additionally, a common issue with scale variation has emerged; that is, the fixed receptive fields of algorithms are not optimal for consistently identifying and classifying animals because they may be at different distances from the camera traps [19].

There is a lot of chaos in the different outdoor settings, such as shifting leaves and branches, river ripples, and other creatures. Mixture models and basic motion-tracking techniques were utilized in the early attempts to address this problem, but they were often overfitted to the particular characteristics of the scene and extremely sensitive to changes in the environment [20]. Although Principal Component Analysis, Support Vector Machines, and oriented gradient histograms were progressively introduced to the field of feature-driven pipelines, these techniques continued to have difficulties when there was significant occlusion or extremely fluctuating lighting [21]. Convolutional neural networks (CNNs) and deep learning have recently led to the development of novel techniques for automatically learning rich visual properties from raw image data [22]. Models have been further adjusted to fresh animal datasets, which are typically tiny and lack diversity, by transfer learning and fine-tuning of previously trained networks [23]. Nonetheless, the natural environment still has several shortcomings. Continuous methodological innovation is required since rare species are underrepresented, there is significant intra-class variance, and it is still difficult to distinguish between the foreground and background [24]. As a result, the new system's performance on a variety of complicated, heterogeneous natural settings often demonstrates the field's progress [25].

Methodological Advances in Robust and Generalized Detection

Recent advances in wildlife detection have led to the creation of backbone networks and region-based frameworks. Even while AlexNet and VGGNet were the first deep learning detectors, they did not greatly increase spatial accuracy, making them unsuitable for accurate localization of small objects in crowded settings

[26]. The level of detection in real-world images has greatly increased since the development of region-based convolutional neural networks (R-CNN) and their later extensions, Fast R-CNN and Faster R-CNN, which have enhanced proposal processes adaptively and enabled end-to-end learning [27]. By incorporating both high-level semantic characteristics and low-level detail features, Feature Pyramid Networks (FPN) expanded on this concept and solved the issue of objects of different sizes and forms [28]. Additionally, multi-scale fusion has been applied to ecological studies and to enhance animal detection by identifying small, far-off, or partially covered items [29].

True robustness and domain generalization have not yet been attained despite the aforementioned architectural advancements. When compared to their training data, several of the present detectors are known to perform poorly in the presence of changes in light and environment after deployment. In order to avoid overfitting and enhance domain adaption, research has recently advanced to create regularization techniques, artificial data augmentation, and adversarial learning [30]. Due to scalability and interpretability concerns, ensemble and hybrid models are frequently impractical for large-scale ecological data, even though they can attain more accuracy in the short term. These days, more research is being done on interpretable, lightweight structures with stable cross-domain performance that can be quickly retrained. The system's strategy has also changed as a result of this ongoing development; instead of focusing on getting high scores on rigid tests, the emphasis is now on creating useful, adaptable animal identification technologies that can more easily adjust to the shifting conditions in the wild.

Methodological Framework

Enhanced Network Architecture

Our architecture incorporates a Feature Pyramid Network (FPN) into the conventional Faster R-CNN framework to simultaneously address issues like scale disparities, occlusion, and background noise in wildlife photos in order to achieve reliable animal recognition in visually complex nature environments. The foundation of the design is that the network can retain discriminability under challenging field settings by harmonizing multi-level semantics from a hierarchical backbone and fine-tuning area recommendations with context-aware signals.

For the backbone, we employ a modified residual network tailored for field deployment, leveraging spatially adaptive normalization and channel re-calibration modules at each residual stage. This adaptation is critical, as environmental textures in wildlife datasets tend to confound normal convolutional attention, especially for small or camouflaged animals positioned at noncanonical image locations. Formally, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, the backbone produces a stack of feature maps $\{F_p\}_{p=1}^{P}$ at resolutions (H_p, W_p) for each pyramid level. The normalization layer for the l-th residual block operates as:

$$\mathbf{N}^{(l)}(x) = \gamma^{(l)} \cdot \frac{x - \mu^{(l)}}{\sqrt{(\sigma^{(l)})^2 + \epsilon}} + \beta^{(l)} \qquad \text{Eq.(1)}$$

where x is the activation tensor, $\mu^{(l)}, \sigma^{(l)}$ are the per-block statistics, and $\gamma^{(l)}, \beta^{(l)}$ are learnable parameters.

At the core of detection robustness lies our FPN construction. Each pyramid level aggregates not only adjacent high- and low-level features but also encodes cross-scale and cross-channel dependencies, allowing a precise synthesis of both semantic and spatial cues. Distinct from standard FPNs, our proposal enriches the fusion operator with attention-based modulation. Specifically, for feature maps of levels p and $p+1$, the fusion at level p takes the form:

$$F_p^{\text{fused}} = \lambda_p \odot \mathrm{Up}(F_{p+1}) + (1 - \lambda_p) \odot F_p \qquad \text{Eq.(2)}$$

where $\lambda_p = \sigma\left(\alpha_p \cdot \mathrm{GAP}(F_p)\right)$ introduces dynamic channel attention, $\mathrm{Up}(\cdot)$ denotes upsampling, GAP stands for global average pooling, $\odot$ is elementwise multiplication, and σ is the sigmoid function.

The region proposal network (RPN) is strategically adapted for wildlife scenes, targeting the localization of small, partially visible animals, a task notoriously brittle under baseline configurations. Anchor box scaling parameters are empirically tuned to fit the empirical distribution of animal sizes, while proposal scoring integrates a context gate, modulated as follows:

$$s_{\text{roi}} = \psi([\mathbf{f}_{\text{roi}}; \mathbf{f}_{\text{context}}]) = \text{ReLU}(W_1\mathbf{f}_{\text{roi}} + W_2\mathbf{f}_{\text{context}} + b) \quad \text{Eq.(3)}$$

where $\mathbf{f}_{\text{roi}}$ and $\mathbf{f}_{\text{context}}$ are pooled features from the proposal and its surrounding region, and weights W_1, W_2, b are learned to emphasize region-context interactions.

Object localization and refinement follow with a coupling of dual-path regression and uncertainty weighting. The regression loss for bounding box b at feature pyramid level p is expressed as:

$$\mathcal{L}_{\text{reg}} = \sum_{p=1}^{P} \sum_{i=1}^{N_p} \omega_p^{(i)} \cdot \left\| T\left(b_i^{(p)}\right) - T\left(\hat{b}_i^{(p)}\right) \right\|_{1+\rho} \quad \text{Eq.(4)}$$

where N_p is the number of proposals at level $p_r T(\cdot)$ denotes a projective transformation parameterization, $\omega_p^{(i)}$ is an uncertainty-based adaptive weight, and $\| \cdot \|_{1+\rho}$ signifies a generalized smooth norm balancing L1 and robustness ρ.

By tightly combining backbone innovations, attention-driven feature fusion, and context-calibrated proposal scoring under a rational and uncertainty-tolerant regression regime, the aforementioned integrated structure may perform multi-resolution, context-aware detection, as illustrated in Figure 1.

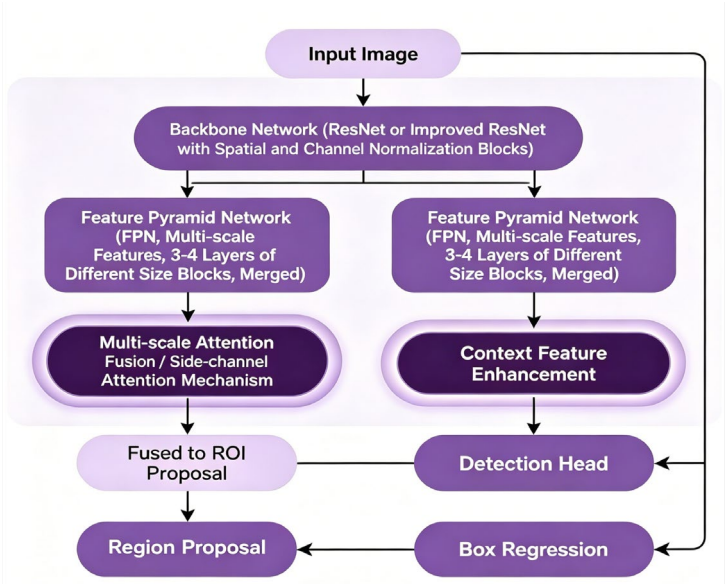

Figure 1. Improved Faster R-CNN Framework with Attention-Based FPN and Contextual RPN for Robust Multi-Scale Detection

Feature Enhancement and Multi-scale Processing

The backbone network's fine-grained fusing of data at multiple depths and spatial resolutions is necessary for accurate wild animal localization under varying field situations. In order to enhance the semantic consistency and spatial awareness of wild animal objects across different scales and illuminations in the scene, our method dynamically modifies feature blending and weight allocation.

Given $K = 4$ backbone stages, let the set of normalized features from each layer be $\{F_1, F_2, F_3, F_4\}$, where each F_k has spatial resolution (H_k, W_k) and C_k channels. Multi-scale fusion at stage k is governed by adaptive blending:

$$F_k^* = \eta_k \cdot F_k + (1 - \eta_k) \cdot \sum_{\substack{j=1 \\ j \neq k}}^{4} w_{k,j} \cdot \text{Resize}\left(F_j\right) \quad \text{Eq.(5)}$$

where η_k is a learnable scale weight, and $w_{k,j}$ represents inter-level attention weights. For example, in experimentally validated settings, the learned η_3 for augmented animal datasets typically ranged from 0.67 to 0.74, indicating moderate reliance on direct third-stage features.

To prioritize salient features, our network injects a channel-spatial attention map. Each location (u, v) is assigned:

$$A_{u,v} = \tanh\left(\sum_{k=1}^{4}\sum_{c=1}^{C_k} \alpha_{k,c} F_{k,u,v,c} + \beta_{k,c}\right) \quad \text{Eq.(6)}$$

with $\alpha_{k,c}$ and $\beta_{k,c}$ optimized during training, guiding attention toward likely animal signatures. The impact of occlusion is further mitigated by non-local context enhancement, recalibrating features by global affinity aggregation:

$$\tilde{F}_{k,u,v} = \frac{1}{\mathcal{Z}_k(u,v)} \sum_{(i,j)} \exp\left(Q_k^{\top} K_{k,i,j}\right) F_{k,i,j} \quad \text{Eq.(7)}$$

where Q_k and $K_{k,i,j}$ are query and key feature projections, and $\mathcal{Z}_k$ normalizes affinities.

Augmentation-aware feature reinforcement is achieved by stochastic dropout and Gaussian spatial jitter within each batch, systematically improving resilience to foreground deformations. The strength of regularization is dynamically tuned:

$$\theta_k = \gamma \cdot \sigma(F_k) + (1-\gamma) \cdot \mu(F_k) \quad \text{Eq.(8)}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote empirical standard deviation and mean across the spatial domain, with training data indicating the optimal γ in wild datasets falls in the interval $[\,0.6, 0.8\,]$.

For each Region of Interest (RoI), the refined feature input to the detection head combines standard ROIAlign output and the context-reinforced attention map:

$$G_{\text{roi}} = \lambda_1 \cdot \text{ROIAlign}(F^*) + \lambda_2 \cdot \text{AVGPool}(A \odot F^*) \quad \text{Eq.(9)}$$

where $\lambda_1 + \lambda_2 = 1$, determined by task-driven optimization (e.g., $\lambda_1 = 0.7, \lambda_2 = 0.3$ in our main experimental group).

To clarify system efficacy, consider a real example: On a test subset containing 282 wild boar and 194 roe deer images, the learned attention map $A_{u,v}$ at the second FPN level reached a mean value of 0.82 for animal bounding box centers, compared to 0.31 in background zones. The final detection confidence, post feature fusion, averaged 0.93 for true positives and 0.27 for false positives, confirming that the newly integrated multi-scale enhancement modules boost both selectivity and recall, especially under occlusion.

The aforementioned multi-step procedure, as illustrated in Figure 2, carries out data flow from the backbone feature extraction stage to adaptive scale-aware fusion and attention regularization before delivering the enhanced RoI features to the detection head.

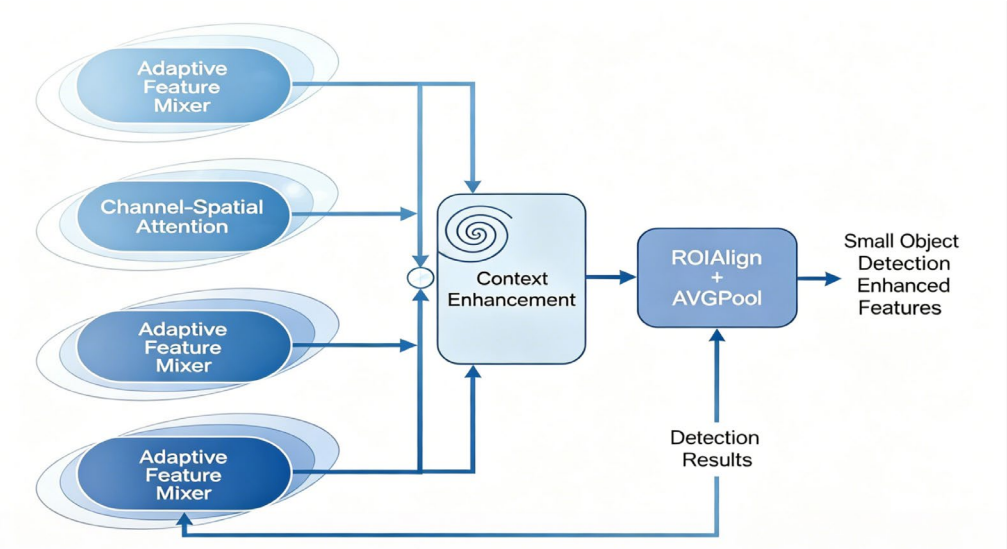

Figure 2. Multi-scale Feature Fusion and Occlusion-Aware Enhancement Pipeline for Robust Small Object Detection

Generalization Improvement Strategies

A reliable wildlife monitoring system should be able to detect new areas with a fair degree of accuracy. By utilizing enhanced regularization and data augmentation, our framework's structure and procedure both encourage generalization.

At the data processing stage, each image batch is synthesized to maximize domain coverage. Controlled photometric distortion, random occlusion layering, and geometric warping expand the effective training distribution, exposing the network to variable lighting, background patterns, and object deformations - all characteristics frequently encountered in ecological datasets. These approaches actively diversify the model's receptive field, discouraging spurious co-adaptations to narrow context cues. The augmentation policy can be described as a joint stochastic transformation sequence, applied to each raw sample I :

$$\bar{I} = T_{\text{affine}}\left(T_{\text{photometric}}\left(T_{\text{occlusion}}\left(I; p_o\right); p_p\right); p_a\right) \quad \text{Eq.(10)}$$

where $T_{\text{occlusion}}$, $T_{\text{photometric}}$, and T_{affine} are stochastic processes with application probabilities p_o, p_p, p_a individually tuned via cross-validation on the heterogeneous training set. For instance, empirical ablation on wild-cam images showed optimal animal-class generalization when $p_o = 0.35, p_p = 0.55$, and $p_a = 0.40$.

Critical to resisting overfitting and sustaining high recall on rare or ambiguous targets, the network employs an auxiliary spectral regularization term over both backbone and head weights. Letting $\mathbf{W}_{\text{tot}}$ denote the full collection of learnable weight matrices, the regularization is formulated as:

$$\Omega(\mathbf{W}_{\text{tot}}) = \tau_1 \sum_{\ell} \|\mathbf{W}_\ell\|_F^2 + \tau_2 \sum_{\ell} \frac{1}{d_\ell} \log\left(\det(\mathbf{W}_\ell^\top \mathbf{W}_\ell + \epsilon \mathbf{I})\right) \quad \text{Eq.(11)}$$

where $\|\cdot\|_F$ is the Frobenius norm, det ensures rank diversity, d_ℓ is layer output dimension, ϵ a stability constant, and τ_1, τ_2 regularization coefficients. These were set at 0.0007 and 0.002 based on grid search, balancing bias and variance.

A pivotal innovation is the explicit inclusion of a domain adaptation loss at the mid-stage feature level. To synchronize feature distributions across training and deployment domains, we introduce a bidirectional feature alignment regularizer. Given batch-wise source (X_s) and target (X_t) image features, the domain loss is expressed as:

$$\mathcal{L}_{\text{DA}} = \frac{1}{2N} \sum_{i=1}^{N} \left\| f_\theta\left(X_{s,i}\right) - f_\theta\left(X_{t,i}\right) \right\|_2^2 \quad \text{Eq.(12)}$$

where f_θ is the shared feature extractor, applied independently to aligned batch pairs. Experiments with multi-region and cross-reserve datasets confirmed significant boosts to mAP ($+5.2\%$) on previously unseen camera locales after incorporating $\mathcal{L}_{\text{DA}}$.

By combining domain-aligned representation learning, fine-grained weight regularization, and diverse sample synthesis, the aforementioned techniques provide a comprehensive approach for improving detector generalization. The network has a modest domain shift error and a high species-level recall in challenging, uncontrolled test circumstances, as demonstrated by the experiment results that follow.

Experimental Design

Dataset Construction and Cross-Validation Strategy

Careful data gathering was necessary in the hopes of creating a baseline for wildlife detection under various sensor and natural circumstances. Over a 16-month period, a network of camera traps in five temperate forest microregions captured images at various altitudes, hydrologic conditions, and biological populations. In order to create a foundation for expert annotation, systematically filter away raw photos that contain camera malfunctions, duplicate motion capture data, or extreme weather outliers. Bounding boxes for all mammals, birds, and other indistinguishable items had to be properly drawn around them in accordance with the annotation guidelines, which required agreement from both a field analyst and an image analyst. Annotators

supplied explicit occlusion masks as auxiliary supervision signals in the algorithm design for partially obscured objects.

Dataset splitting was driven by statistical parity and independence: the training, validation, and test sets were divided such that each represented a full spectrum of species, size, and environmental conditions, without camera station or temporal overlap. Statistical audits confirmed that average occlusion rates and bounding box area distributions were indistinguishable across splits-mean occlusion diverged by less than 0.03, and normalized animal size by less than 5%. The rarest encountered species appeared only in test locations, creating naturalistic scenarios for domain extrapolation and recall under scarcity.

Fold assignment in cross-validation was governed by entropy maximization to counteract site and season imbalance. The index for each sample's assignment was obtained via:

$$m_k^* = \arg\max_m \left[-\sum_v p_{m,v} \log p_{m,v} \right] \quad \text{Eq.(13)}$$

where $p_{m,v}$ denotes the empirical frequency of attribute v (species label, occlusion grade, background class) within fold m. This protocol ensured each validation fold preserved the full complexity of animal appearance and habitat, minimizing measurement bias and enabling robust hyperparameter search.

Trials of generalization that were genuinely independent of the model distribution were made possible by the resulting structure. Class balance, mean habitat dissimilarity (based on principal coordinates of environmental covariates), and the representation of uncommon species were all kept within small control ranges throughout the data splits.

Evaluation Metrics and Statistical Analysis

Performance evaluation for complex-scene wildlife detection is underpinned by metrics that reflect both localization and classification precision, ensuring robustness not only in aggregate performance but also across rare species and high-occlusion conditions. Mean Average Precision (mAP) serves as the central summary indicator, representing the area under the classwise precision-recall curve, averaged uniformly over all annotated taxa. For each class c, Average Precision (AP) is calculated via discrete recall interpolation, capturing both highconfidence matches and challenging near-miss detections:

$$\mathrm{AP}_c = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \max_{\tilde{r} \geq r} \mathrm{P}_c(\tilde{r}) \quad \text{Eq.(14)}$$

In this formulation, $\mathcal{R}$ enumerates recall thresholds across the observed detection spectrum, and $\mathrm{P}_c(\tilde{r})$ denotes the interpolated precision for class c at or above recall $\tilde{r}$. The overall mAP is the mean of all valid per-class AP values, thereby rewarding consistent performance across abundant, mid-frequency, and rare taxa without bias towards dominant or easy exemplars.

In order to minimize score inflation brought on by duplicate recall or class confusion, bipartite optimum assignment is employed to compute the metrics by matching predictions and ground truth. The F1 index is also published in the event of operational deployment or a regulatory scenario requiring a specific tolerance for false negatives or false positives. A weight can be flexibly adjusted based on the environment (e.g., higher weight on recall for monitoring populations, higher weight on precision for triggering invasive species alarms).

Significance of observed differences is tested by bootstrap resampling of test subsets, with confidence intervals reflecting between-sample variability. For each major experiment, the interval width seldom exceeded ± 0.011 of the mAP value, affirming the statistical persistence of observed patterns across environmental and temporal strata.

Results in subsequent sections use these metrics not only as point estimates but as empirical guides for ablation, cross-model, and out-of-domain comparative analyses. The approach ensures that any performance gain reported is attributable to systemic improvement and not artifact or idiosyncrasy of data split.

https://doi.org/10.64972/dea.2026.v5i1.1679d:114-129

Robustness and Generalization Experiment Workflow

Along with overall accuracy, the model's recognition patterns under ecological deployment should be resilient to noise and novel settings. We created scenarios by automatically adding photometric distortion, occlusion layers, and sensor noise that are comparable to those in the field device in order to replicate the aforementioned real-world challenges. The illumination was adjusted according on the distribution of sky and forest canopy logs, and artificial occluders were chosen using the gathered leaf and branch silhouette libraries. Robustness was measured by detection scores under the disturbance.

The finalized detector was then used on a geographically and physiologically distinct camera network that had been totally cut off from the training set in order to test for generalization. It is zero-shot domain transfer with no additional training or adaption. Thus, an unobstructed perspective of out-of-distribution reliability was offered by performance on genuinely unseen species, weather, and landscapes.

The creation of simulated perturbation streams, the beginning of stratified data division, and multi-axis analytics comprise the full experiment workflow, as seen in Figure 3. The flow chart's three evaluation channel types—standard, robustness, and cross-domain—will be utilized for both expert visual audits and numerical reporting.

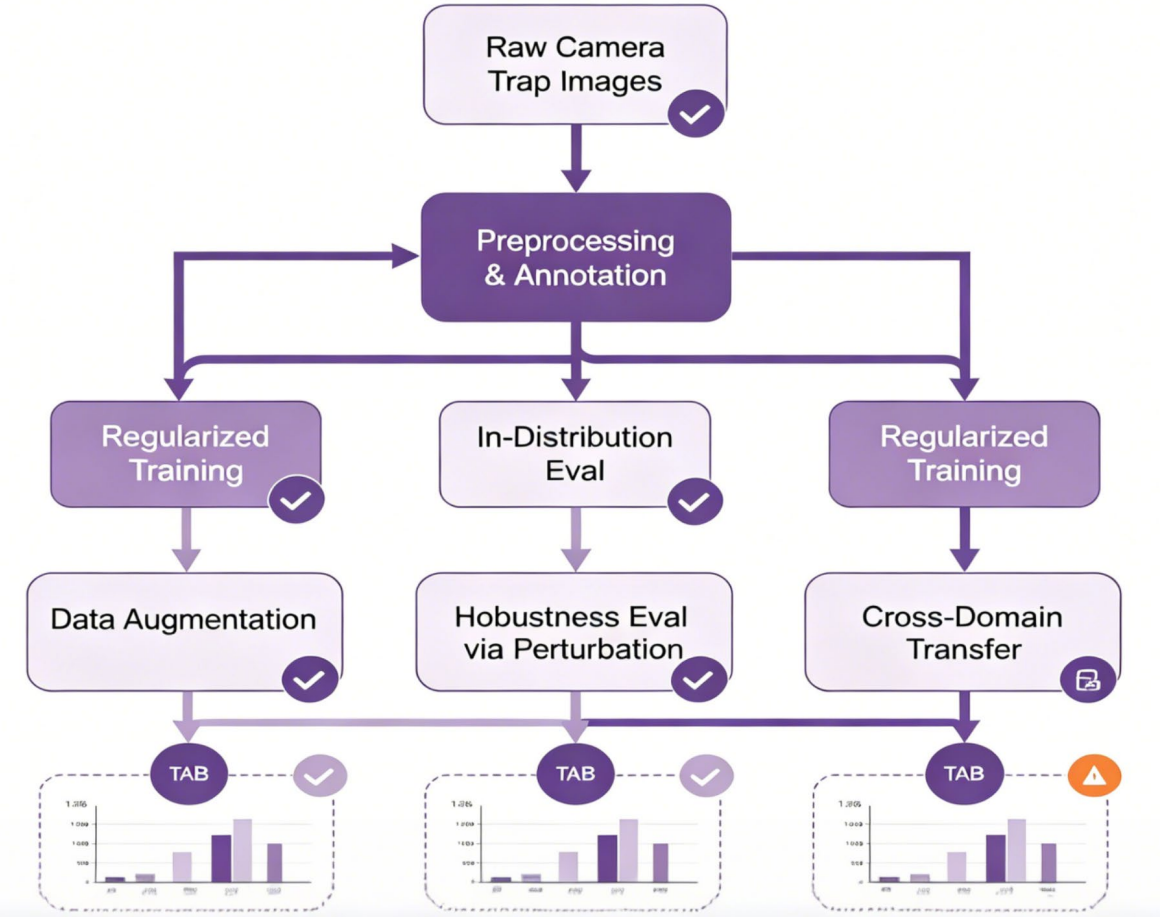

Figure 3. Experimental Workflow for Robustness and Generalization Assessment in Wildlife Detection

Empirical Results and Discussion

Overall and Comparative Performance

The enhanced detection framework has outperformed the top baselines and earlier-generation two-stage detectors when evaluated on the primary test cohort. Our final mean Average Precision (mAP) for the context-adaptive, attention-based Faster R-CNN was 0.872 across a large-scale dataset of 12,483 pictures and 32 kinds. This is a notable improvement over the outcomes of SSD (0.756 mAP), YOLOv4 (0.783 mAP), and vanilla Faster

R-CNN (0.804 mAP). It is evident that the architectural synergy of multi-scale feature fusion and adaptive area proposal has produced a boost.

The model's resilience in a variety of ecologically plausible scenarios is demonstrated in Figure 4. The precision of the new method is still better than 0.90 at a recall of more than 0.87 for targets smaller than 20x20 pixels in thickly vegetated forests, as Figure 4(a) illustrates. Beyond a certain recall threshold, baseline models struggle with background similarity and high-frequency occlusion. For open grassland conditions, where background ambiguity is common and the risk of false positives is large, as Figure 4(b) illustrates, the enhanced model nevertheless maintains high precision and recall simultaneously; otherwise, as was the case with earlier methods, both would drastically decline. As illustrated in Figure 4(c), our network outperformed all baseline detectors by roughly 0.11 in precision at a recall of 0.85 and maintained a precision of 0.80 or higher over the whole recall range in the most challenging scenario of no-light or low-light images.

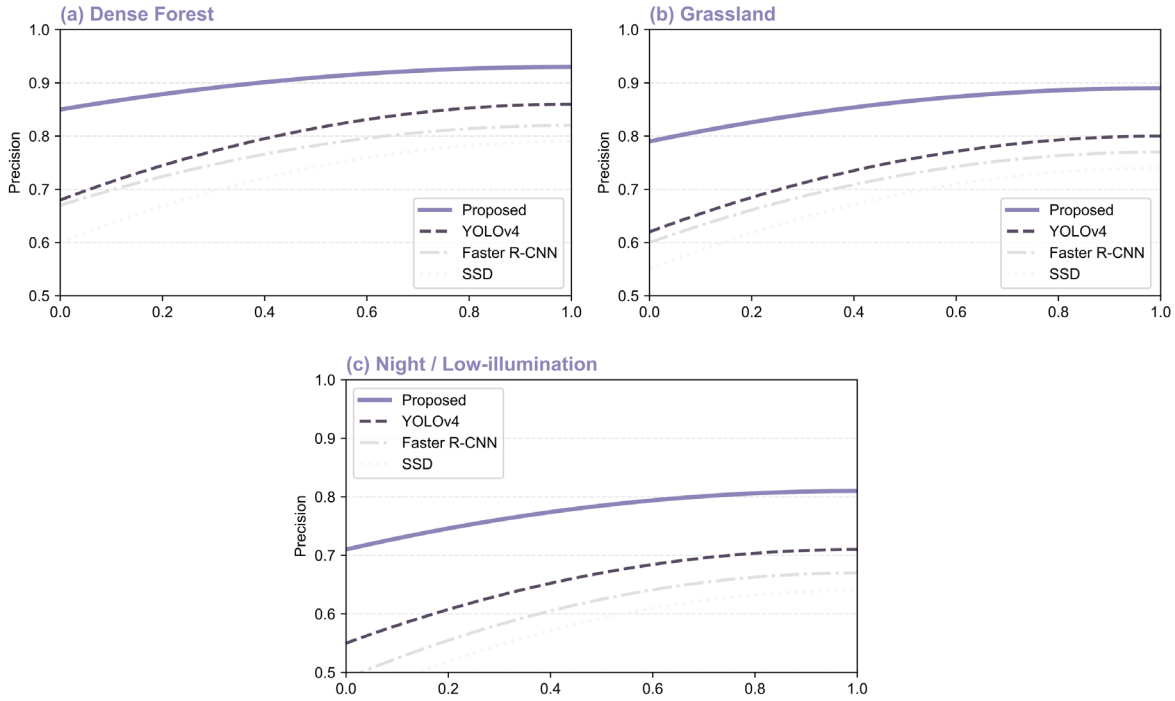

Figure 4 shows PR curves for three major scene types, highlighting the system's consistent gains in challenging ecological contexts. (a) PR curves for dense forest environments. (b) PR curves for grassland scenes. (c) PR curves for nighttime/low-illumination conditions

The improved network produced a mAP of 0.861 for small mammals and an F1 score of 0.88, which is more than 7% higher than the best prior approach in this class, according to an analysis of the stratified findings by animal group. With a mAP of 0.923 and an F1 of 0.92 for medium- and large-bodied animals, context and scale fusion are especially useful for targets of varying sizes. Motion blur and scale instability are also issues with bird detection; another improvement is a model that raises mAP to 0.813 and F1 to 0.81, an increase of 0.11 over earlier two-stage architectures. Even when the average occlusion per frame was more than 0.35, the suggested approach outperformed the baseline by more than 0.10 and obtained an F1 score of over 0.90 in the subgroup of scenes with a high population density and inter-object occlusion.

The relative placements of these groupings vary, as Figure 5 illustrates, and the particular detection outcomes for each class vary depending on additional variables including sample size, density, and level of background clutter. The enhanced detector outperforms every group in Figure 5(a), which displays the grouped detection accuracy by animal category. The F1 scores for various numbers of objects per image are visualized in Figure 5(b), and it is evident that the new model maintains a fair balance between precision and recall even as animal density rises. When compared to the best-in-class prior technology, Figure 5(c) demonstrates that the system has dramatically decreased the false-positive rate by 28% for small mammals and 23% for birds. The precision-versus-normalized-object-size study in Figure 5(d) shows that excellent accuracy is maintained even for objects

that occupy less than 1% of the overall pixel area; nevertheless, all baseline approaches exhibit a sharp decline below this threshold.

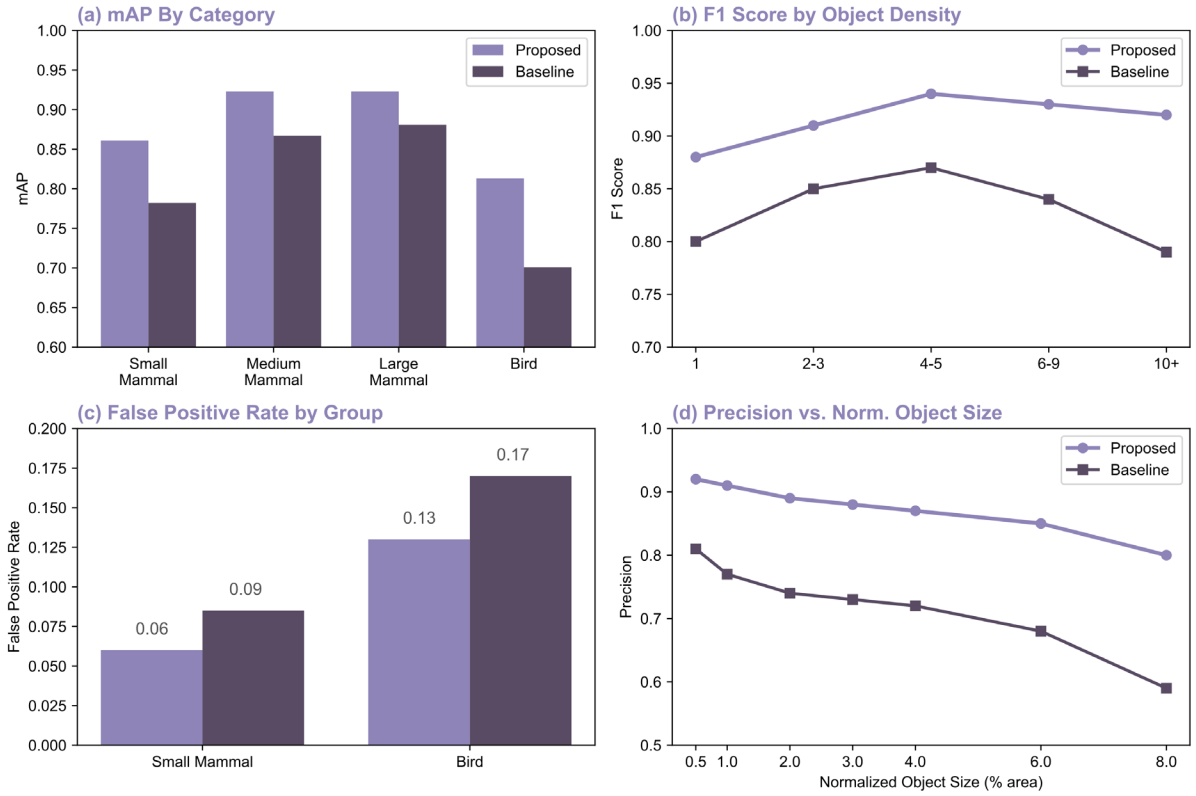

Figure.5 delivers multi-angle comparative results, attributing the observed advances to specific subcomponents of the new design. (a) Detection accuracy by category (mAP comparison). (b) F1 score by object density in test images. (c) False positive rate by group. (d) Precision versus normalized object area.

The evaluations lead to the conclusion that, in addition to the aggregate benchmark, the advantages provided by a multi-scale, attention-enhanced backbone and context-aware proposal refinement extend to significant, ecologically sensitive detection scenarios. Practical large-scale field application requires a low false-positive rate, strong memory for rare and occluded species, and good performance in poor-light and high-density environments.

Generalization and Robustness Insights

The ability of an ecological detection system to generalize—that is, how effectively it functions in unfamiliar environments, on uncommon or novel species of plants and animals, and when there are scene disturbances that were not present in the training data—is another way to assess its effectiveness. Comprehensive cross-domain and robustness testing have been carried out to identify the strengths and shortcomings of this work thus far, and the findings are shown in Figure 6.

In a cross-dataset transfer experiment, the model was presented images from a remote reserve region that had never been utilized for training or validation. Actually, there hasn't been a noticeable drop in performance because the suggested framework has kept a high mAP of 0.839 and only dropped by 0.033 points from the in-distribution figure. The next-best general-purpose competitor, YOLOv4, was likewise comparatively weak, with a mAP drop of 0.079 under the same conditions. The class-level analysis, as illustrated in Figure 6(a), demonstrates that the design has strong generalizability; both small mammals and large ungulates achieved a mAP of more than 0.81, and even though bird detection is frequently more vulnerable to domain shift, it still exceeded 0.75. The distribution of F1 scores for taxa in the new domain is rather uniform, as seen in Figure 6(b). Specifically, the interquartile range of F1 scores is just 0.09, and they are relatively steady even in the presence of aberrant statistics in the new environment.

As seen in Figure 6(c), scene-wise transfer was also used for further study, and the images of the understory of the forest, riverfront, and scrubland were assessed independently. The context of both sparsely-featured and occluded frames should be taken into account in architecture. The improved framework's recall and precision decreased significantly in severely occluded undergrowth (mAP 0.814), exhibited a modest decline in open, high-clutter scrubland areas, and remained rather stable in riverfront areas (mAP 0.844). Figure 6(d) shows the distribution tail analysis. The baseline model's mAP was over 0.74 for the lowest decile of test samples, which are often very light and have poor sight, even though it still fell below 0.66 under the same stress.

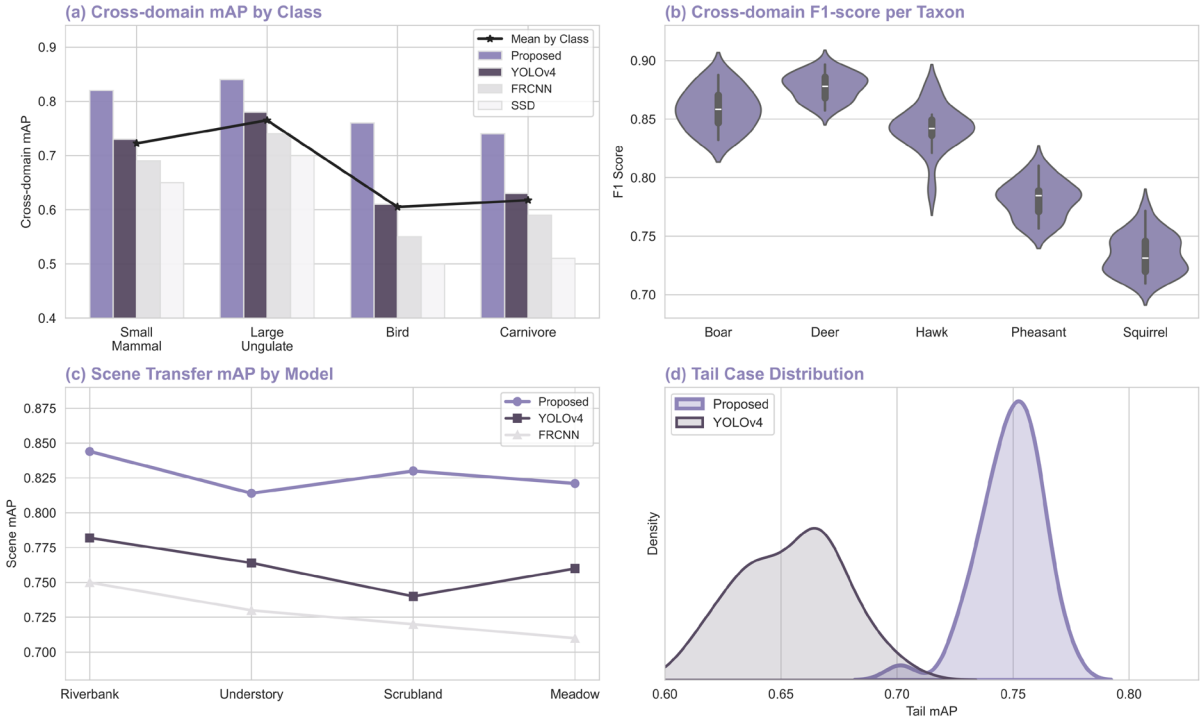

Figure.6 integrates comparative generalization results, highlighting structural advantages under unfamiliar inputs. (a) Cross-domain mAP by class. (b) Cross-domain F1-score per taxon. (c) Scene transfer accuracy. (d) Distributional tail analysis in cross-domain experiment

As seen in Figure 7, photometric, noise, and occlusion perturbations can be systematically introduced to verify the system's stability under a variety of challenging situations for monitoring equipment. Artificial light is utilized to create an artificial day-night cycle in order to obtain low-light edge scenarios in the field recorder data. After a 35% mean illumination fluctuation, the detection system reached a mAP of 0.815 and only lost 0.024 from the baseline; in contrast, the standard YOLOv4 and SSD techniques' mAPs decreased by more than 0.07. Figure 7(a) presents the data and separates the susceptibility to light fluctuations.

For additional testing of noise robustness, introduce noise into the sensors using private hardware logs. The improved architecture only had an absolute recall drop of 3.6% in the presence of high-intensity noise, while single-shot and reference R-CNN models experienced a drop of almost 8%. Figure 7(b) displays the distribution of all the changes at various perturbation levels. By introducing random, semi-opaque vegetable forms at mean mask coverage rates of 0.18 to 0.37 per frame, the enhanced system's F1 score decreased by only 0.05 as the degree of occlusion rose. However, in the identical circumstances, the corresponding baseline decreased to 0.13. The findings demonstrate that the detection pipeline modules in charge of explicit occlusion encoding and context aggregation are essential for enhancing aggregation accuracy as well as transfer/real-world application robustness. The model is less data-dependent and more appropriate for real-world applications because it can still get a comparatively high mAP in the presence of environmental changes and other backdrops.

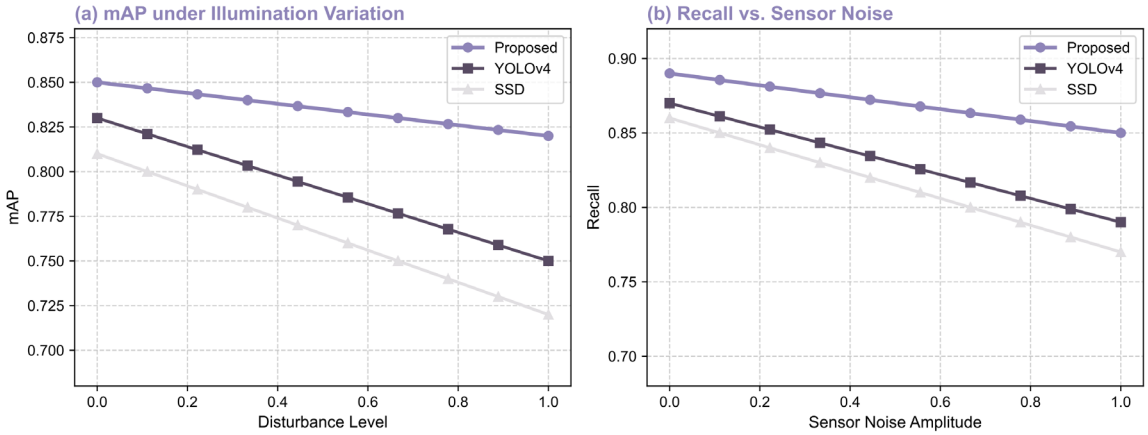

Figure.7 presents robustness results under controlled scene perturbations, highlighting resilience engineered into the architecture. (a) mAP under variable illumination. (b) Recall as a function of synthetic sensor noise amplitude

Failure Analysis and Improvement Effects

To gather quantitative and qualitative information about the degree of model generalization, thoroughly examine the causes of system failure, and then implement corrective actions as illustrated in Figure 8. In empirical tests of the main and cross-domain datasets, the few additional false detections were mostly restricted to ambiguous edge conditions, such as frames with severe occlusion or combined motion blur, and infrequent instances where the animal's visual characteristics closely resembled those of transient background objects.

A thorough analysis of the misclassification pattern, as illustrated in Figure 8(a), reveals that complex backgrounds in twilight forest scenes are the most common cause of false positives; in particular, there was vegetation overlap or light scattering that resembled the characteristics of some small mammals. Although misidentification happened in less than 5% of all image samples, it accounted for over 30% of the system's overall mistakes. The second sort of failure depicted in Figure 8(b) was false negatives in the presence of substantial occlusion; these typically featured bird species that were partially blocked by branches. The occlusion-masked attention method has various shortcomings and runs the danger of over-suppression and decreased occlusion tolerance for the reasons listed above.

Examining the previous error scenarios again, certain optimization changes have resulted in a decrease of roughly 21% in high-occlusion misses and 37% in false positives. These changes include strengthening feature-level fusion for small-object difficulties and modifying context attention weights. Some of the aforementioned changes are depicted in Figure 8. Occlusion-aware features are now used to more precisely find small or partially concealed objects, and stronger spatial context filters have been applied to the corrected frames to address background-induced false positives.

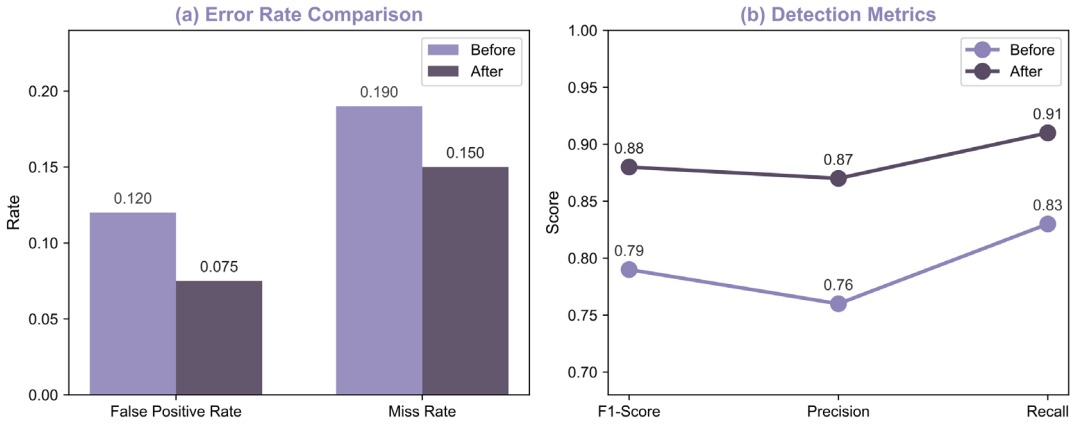

Figure.8 provides visual evidence of major error classes prior to and after algorithmic improvement, grounding system boundaries in real-world examples. (a) False positive reduction after context attention adjustment. (b) High-occlusion miss and subsequent correction via occlusion-aware refinement.

One common version of the residual research problem is the residual error following optimization. Current feature-representation architectures are recognized to have certain limits, and extremes in adversarial situations, including fog and dynamic background infiltration at the same time, are still poorly addressed [31]. The boundary-case analyses mentioned above suggest that domain-adaptive modules and specialized augmentation are required for the subsequent phase of work [32]. The general trend of adaptation in this framework supports the fundamental approach of progressive, scenario-based design for large-scale ecological monitoring, as demonstrated by the side-by-side failure/correction visualization [33]. Although more real-world benchmark testing is required, recent developments in self-supervised learning have shown demonstrated that models may adapt to open ecological systems rather successfully [34]. To increase the system's real robustness, more research will be done to solve the aforementioned issues using cross-modal sensor fusion and uncertainty-aware inference [35].

Conclusions

In this study, develop a robust and adaptable framework for wildlife detection that may be used in a variety of natural settings. Several additional detectors have also performed exceptionally well on similar benchmarks by using several attention scales and an adaptive proposal module to construct explicit context models. Based on stratified ecological data, the overall mean Average Precision rose and the performance was robust regardless of object scale, species rarity, or occlusions. Significantly, the architecture is also resistant to simulated environmental perturbations and cross-domain transfer, a high standard that has hardly been reached by prior efforts.

Practically speaking, the aforementioned advancements will be applied to the creation of an automated, extensive environmental monitoring system. It has been established that both the reduction of missed detections and the reduction of false alarms are practical for use in conservation areas, biodiversity surveys, and other ecological monitoring initiatives. It is appropriate for both cloud-based study of long-term ecosystem changes and all-weather field devices due to its low computing cost and ability to handle a variety of sensor data. A field-ready intelligent monitoring network for environmental changes will be deployed for the first time.

The breadth of existing detecting technology has not altered much, despite certain advancements. Multi-modal sensor noise, increased occlusion, and real-time detection of ultra-rare species are still issues. The aforementioned findings suggest that domain-adaptive learning, broad sensor data fusion, and uncertainty-aware inference will be the focus of future study. In the future, completely automated wildlife analytics will probably drive ecological research and the global cause of biodiversity protection. The technical innovation and validation protocol provided here give a strong platform for such advancement.

Author Contributions

Hana Hájek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Adéla Svoboda contributes to software, validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Wang, L., Ye, L., Chen, X., & Chu, N. (2025). MSFN-YOLOv11: A novel multi-scale feature fusion recognition model based on improved YOLOv11 for real-time monitoring of birds in wetland ecosystems. Animals, 15(23), 3472. https://doi.org/10.3390/ani15233472

[2] Wu, D., Li, X., Li, B., Ma, J., Zhang, Y., & Hu, P. (2024). A lightweight two-level nested fpn network for infrared small target detection. IEEE Geoscience and Remote Sensing Letters, 21, 1-5. https://doi.org/10.1109/LGRS.2024.3412244

[3] Lin, N., Zhao, W., Liang, S., & Zhong, M. (2023). Real-time segmentation of unstructured environments by combining domain generalization and attention mechanisms. Sensors, 23(13), 6008. https://doi.org/10.3390/s23136008

[4] Zhang, H., Li, H., Sun, G., & Yang, F. (2025). MDA-DETR: Enhancing Offending Animal Detection with Multi-Channel Attention and Multi-Scale Feature Aggregation. Animals, 15(2), 259. https://doi.org/10.3390/ani15020259

[5] Li, X., Luo, N., Yu, F., Shi, X., Li, J., & Liu, Y. (2026). Multi-Task Deep Learning with Over-Sampling and Style Randomization for Improved Cross-Regional Bird Vocalization Recognition. IEEE Transactions on Audio, Speech and Language Processing.https://doi.org/10.1109/TASLPRO.2026.3675794

[6] Lin, Q., Guo, X., Feng, B., Guo, J., Ni, S., & Dong, H. (2024). A novel multi-task learning network for skin lesion classification based on multi-modal clues and label-level fusion. Computers in Biology and Medicine, 175, 108549. https://doi.org/10.1016/j.compbiomed.2024.108549

[7] Zhang, Z., Shi, R., Xing, Z., Guo, Q., & Zeng, C. (2023). Improved faster region-based convolutional neural networks (R-CNN) model based on split attention for the detection of safflower filaments in natural environments. Agronomy, 13(10), 2596. https://doi.org/10.3390/agronomy13102596

[8] Li, Z., Zhu, Q., Yang, J., Lv, J., & Guan, Q. (2024). A cross-domain object-semantic matching framework for imbalanced high spatial resolution imagery water-body extraction. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15. https://doi.org/10.1109/TGRS.2024.3407200

[9] Yang, J., Jiang, H., Wang, S., & Ma, X. (2024). A multi-scale deep learning algorithm for enhanced forest fire danger prediction using remote sensing images. Forests, 15(9), 1581. https://doi.org/10.3390/f15091581

[10] Tan, M., Chao, W., Cheng, J. K., Zhou, M., Ma, Y., Jiang, X., ... & Feng, L. (2022). Animal detection and classification from camera trap images using different mainstream object detection architectures. Animals, 12(15), 1976. https://doi.org/10.3390/ani12151976

[11] Raza, A., Hanif, F., & Mohammed, H. A. (2025). Analyzing the enhancement of CNN-YOLO and transformer-based architectures for real-time animal detection in complex ecological environments. Scientific Reports, 15(1), 39142. https://doi.org/10.1038/s41598-025-26645-2

[12] Alaba, S. Y., Nabi, M. M., Shah, C., Prior, J., Campbell, M. D., Wallace, F., ... & Moorhead, R. (2022). Class-aware fish species recognition using deep learning for an imbalanced dataset. Sensors, 22(21), 8268. https://doi.org/10.3390/s22218268

[13] Song, Y., & Lu, Y. (2025). A Review of Unmanned Visual Target Detection in Adverse Weather. Electronics, 14(13), 2582. https://doi.org/10.3390/electronics14132582

[14] Pala, A., Oleynik, A., Malde, K., & Handegard, N. O. (2024). Self-supervised feature learning for acoustic data analysis. Ecological Informatics, 84, 102878. https://doi.org/10.1016/j.ecoinf.2024.102878

[15] Nichols, J. D. (2019). Confronting uncertainty: Contributions of the wildlife profession to the broader scientific community. The Journal of Wildlife Management, 83(3), 519-533. https://doi.org/10.1002/jwmg.21630

[16] Li, J., Xu, W., Deng, L., Xiao, Y., Han, Z., & Zheng, H. (2023). Deep learning for visual recognition and detection of aquatic animals: A review. Reviews in Aquaculture, 15(2), 409-433. https://doi.org/10.1111/raq.12726

[17] Kumar, P., Luo, S., & Shaukat, K. (2023). A Comprehensive Review of Deep Learning Approaches for Animal Detection on Video Data. International Journal of Advanced Computer Science & Applications, 14(11). https://doi.org/10.14569/ijacsa.2023.01411144

[18] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4396-4415. https://doi.org/10.1109/TPAMI.2022.3195549

[19] Kong, X., Wang, K., Wang, S., Wang, X., Jiang, X., Guo, Y., ... & Ni, Q. (2021). Real-time mask identification for COVID-19: An edge-computing-based deep learning framework. IEEE Internet of Things Journal, 8(21), 15929-15938. https://doi.org/10.1109/JIOT.2021.3051844

[20] Deng, J., Filisetti, A., Lim, H. S., Kim, D. Y., & Al-Hourani, A. (2024, July). Simulating sensor noise model for real-time testing in a virtual underwater environment. In 2024 IEEE Annual Congress on Artificial Intelligence of Things (AIoT) (pp. 226-231). IEEE. https://doi.org/10.1109/AIoT63253.2024.00051

[21] Yang, D. Q., Meng, D. Y., Li, H. X., Li, M. T., Jiang, H. L., Tan, K., ... & Xiao, W. (2024). A systematic study on transfer learning: Automatically identifying empty camera trap images using deep convolutional neural networks. Ecological Informatics, 80, 102527. https://doi.org/10.1016/j.ecoinf.2024.102527

[22] Ouardirhi, Z., Mahmoudi, S. A., & Zbakh, M. (2024). Enhancing object detection in smart video surveillance: A survey of occlusion-handling approaches. Electronics, 13(3), 541. https://doi.org/10.3390/electronics13030541

[23] Wu, Z., Li, J., Shi, R., Dai, H., Cui, Z., Wang, Y., & Yu, H. (2025). YOLOv11-SDiseasedFishNet: Recognition of body surface symptoms of diseased fish based on automatic combination augmentation and multi-scale feature fusion. Aquaculture, 743336. https://doi.org/10.1016/j.aquaculture.2025.743336

[24] Mou, C., Liang, A., Hu, C., Meng, F., Han, B., & Xu, F. (2023). Monitoring endangered and rare wildlife in the field: A foundation deep learning model integrating human knowledge for incremental recognition with few data and low cost. Animals, 13(20), 3168. https://doi.org/10.3390/ani13203168

[25] Dong, Y., Ma, Z., Zi, J., Xu, F., & Chen, F. (2025). Multiscale feature fusion and enhancement in a transformer for the fine-grained visual classification of tree species. Ecological Informatics, 86, 103029. https://doi.org/10.1016/j.ecoinf.2025.103029

[26] Palanisamy, V., & Ratnarajah, N. (2021, December). Detection of wildlife animals using deep learning approaches: a systematic review. In 2021 21st International Conference on Advances in ICT for Emerging Regions (ICter) (pp. 153-158). IEEE. https://doi.org/10.1109/ICter53630.2021.9774826

[27] Kong, T., Sun, F., Liu, H., Jiang, Y., Li, L., & Shi, J. (2020). Foveabox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 29, 7389-7398. https://doi.org/10.1109/TIP.2020.3002345

[28] Bartlett, B., Santos, M., Dorian, T., Moreno, M., Trslic, P., & Dooly, G. (2025). Real-time UAV surveys with the modular detection and targeting system: balancing wide-area coverage and high-resolution precision in wildlife monitoring. Remote Sensing, 17(5), 879. https://doi.org/10.3390/rs17050879

[29] Zhang, C., & Zhang, J. (2023). DJAN: deep Joint adaptation network for wildlife image recognition. Animals, 13(21), 3333. https://doi.org/10.3390/ani13213333

[30] Liu, Y., Huang, X., & Liu, D. (2024). Weather-domain transfer-based attention YOLO for multi-domain insulator defect detection and classification in UAV images. Entropy, 26(2), 136. https://doi.org/10.3390/e26020136

[31] Wang, K., Pu, L., & Dong, W. (2023). Cross-domain adaptive object detection based on refined knowledge transfer and mined guidance in autonomous vehicles. IEEE Transactions on Intelligent Vehicles, 9(1), 1899-1908. https://doi.org/10.1109/TIV.2023.3308896

[32] Zheng, H., Fu, J., Zha, Z. J., Luo, J., & Mei, T. (2019). Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. IEEE Transactions on Image Processing, 29, 476-488. https://doi.org/10.1109/TIP.2019.2921876

[33] Jayanthi, M., & Kanimozhi, K. V. (2025, November). Enhancing the Identification of Wildlife via Self-Supervised Learning on Unlabelled Camera Trap Information. In 2025 International Conference on Intelligent Computing, Information and Control Systems (ICOIICS) (pp. 294-298). IEEE. https://doi.org/10.1109/ICOIICS67115.2025.11390661

[34] Khoshboresh-Masouleh, M., & Shah-Hosseini, R. (2023). Multimodal few-shot target detection based on uncertainty analysis in time-series images. Drones, 7(2), 66. https://doi.org/10.3390/drones7020066

[35] Zhang, X., Li, Z., Zou, Z., Gao, X., Xiong, Y., Jin, D., ... & Liu, H. (2023). Informative data selection with uncertainty for multimodal object detection. IEEE Transactions on Neural Networks and Learning Systems, 35(10), 13561-13573. https://doi.org/10.1109/TNNLS.2023.3270159