Zhu et al. *Zool. Res. Divers. Conserv.* 2025, 2(4): 211−222
https://doi.org/10.24272/j.issn.2097-3772.2025.018
https://cstr.cn/32249.14.zrdc.2025.018

ZRDC Zoological Research: Diversity and Conservation

Article Open Access

IAE-YOLOv9: Advancing automatic wildlife detection and its practical application in *Pantholops hodgsonii* monitoring for biodiversity conservation

Rui Zhu[1,2], Zhou-Yuan Li[3], Xin-Ming Lian[4], Qi-Hao Jiang[1,2], Jiang-Jian Xie[1,2,5,*], Wen-Ying Wang[6,*]

[1] *School of Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*
[2] *Multimodal Eco Data Intelligence Analysis Lab*, *Beijing Forestry University*, *Beijing* 100083, *China*
[3] *School of Grassland Science*, *Beijing Forestry University*, *Beijing* 100083, *China*
[4] *Key Laboratory of Adaptation and Evolution of Plateau Biota*, *Northwest Institute of Plateau Biology*, *Chinese Academy of Sciences*, *Xining*, *Qinghai* 810008, *China*
[5] *State Key Laboratory of Efficient Production of Forest Resources*, *Beijing Forestry University*, *Beijing* 100083, *China*
[6] *School of Life Science*, *Qinghai Normal University*, *Xining*, *Qinghai* 810000, *China*

ABSTRACT

Pantholops hodgsonii functions as a keystone taxon within plateau ecosystems, yet its harsh high-altitude habitat imposes substantial limitations on conventional field monitoring efforts. This study presents an advanced IAE-YOLOv9 architecture, incorporating core modules—involution operator, alterable kernel convolution (AKConv), and element multi-scale attention (EMA) mechanism —to enhance detection performance under complex field conditions. Notably, involution expands the effective receptive field, thereby strengthening detection of small-bodied targets. AKConv enables adaptive kernel reconfiguration, increasing responsiveness to scale variation. EMA facilitates refined object localization and classification by aggregating features across multiple spatial scales. When coupled with infrared-triggered camera traps, the system supports real-time surveillance, accurate detection, and automated counting of *P. hodgsonii*. On the self-constructed Sichuan-Gansu Wildlife Infrared Dataset (SGWID), IAE-YOLOv9 achieved 94.8% precision (95% confidence interval (CI): 94.2%–95.1%), 95.2% mean average precision (mAP@0.5) (95% CI: 94.4%–95.5%), and 91.2% recall (95% CI: 90.8%–91.5%), exceeding baseline detectors. Field deployment along migration corridors yielded 89.81% detection accuracy (95% CI: 88.34%–91.12%) and 88.28% counting accuracy (95% CI: 87.45%–89.12%) on the Hoh Xil Wildlife Camera Trap Dataset (HXWCTD), with high concordance to manual annotations. This integrative framework enables robust wildlife surveillance, migratory tracking, and conservation planning under high-altitude conditions, highlighting the effectiveness of combining infrared imaging with deep learning for ecological monitoring in extreme environments.

Keywords: *Pantholops hodgsonii*; Infrared-triggered camera; YOLO; Automatic counting

INTRODUCTION

Amid the accelerating erosion of global biodiversity, migratory species remain among the most vulnerable taxa due to their dependence on ecological connectivity and habitat integrity. Sustained survival requires access to distant breeding and foraging grounds, making these species acutely sensitive to environmental disturbance and habitat fragmentation (Bauer & Hoye, 2014). The Convention on the Conservation of Migratory Species (CMS) highlights the urgency of restoring critical habitats, establishing ecological corridors, strengthening cross-border monitoring frameworks, and implementing conservation strategies tailored to population dynamics (Lyster, 1989). Despite targeted efforts, the State of the World's Migratory Species report indicates that 44% of assessed migratory species continue to exhibit population declines, 51% of key biodiversity areas remain inadequately protected, and 58% of long-term monitoring sites experience unsustainable anthropogenic pressures. These findings

Received: 12 October 2025; Accepted: 25 November 2025; Online: 15 December 2025

Foundation items: This work was jointly supported by the National Key Research and Development Program of China (2023YFF1304301& 2023YFF1304305) and Beijing Forestry University Science and Technology Innovation Program (2024XY-G002).

*Corresponding authors, E-mail: shyneforce@bjfu.edu.cn; wangwy0106@163.com

emphasize the imperative for expanded surveillance systems and precise population assessments to inform conservation planning and accelerate ecosystem recovery (UN Environment Programme World Conservation Monitoring Centre, 2024).

Pantholops hodgsonii represents a flagship high-altitude migratory species endemic to the western Qinghai-Xizang Plateau, with core populations concentrated in the Hoh Xil National Nature Reserve (Leslie & Schaller, 2008). Currently classified as Near Threatened by the International Union for Conservation of Nature (IUCN) (Dharaiya et al., 2016), this species follows a distinct seasonal migration in which females undertake long-distance movements each summer to reach the calving grounds near Zhuonai Lake (Buho et al., 2011). Multiple environmental stressors, including climate instability, landscape fragmentation, and progressive grassland degradation, have intensified pressures on both habitat integrity and the persistence of endemic plateau species (Zhang et al., 2019). Extreme weather disrupts migratory timing and overwintering success, while infrastructure expansion interferes with traditional migratory routes. Concurrently, declining vegetation productivity undermines the ecological carrying capacity of core rangelands. These converging threats constrain migration, reduce geographic distribution, and elevate extinction risk.

Accurate characterization of population dynamics and effective conservation management require continuous, real-time monitoring coupled with reliable automated detection and counting of *P. hodgsonii* along migratory corridors. Infrared-triggered cameras provide clear advantages over traditional manual field surveys by improving sampling efficiency, reducing operational costs, and minimizing disturbance to wildlife (Li et al., 2014). However, extended deployment of these systems generates massive image datasets that make manual inspection impractical. This data burden underscores the need for scalable, automated analytical approaches capable of processing large volumes of image data with high accuracy and speed (Swann et al., 2004). Recent advances in deep learning have transformed automated wildlife monitoring, with detection algorithms broadly classified into two-stage and one-stage frameworks. Two-stage approaches, such as Fast R-CNN and RetinaNet, first generate candidate regions and subsequently perform classification and localization. These methods typically achieve strong detection accuracy but incur substantial computational overhead, limiting suitability for real-time applications. For instance, Vecvanags et al. (2022) reported average precision values of 0.4073 and 0.4364 for Fast R-CNN and RetinaNet in datasets involving wild boar and Père David's deer. Ukwuoma et al. (2022) introduced multi-scale attention mechanisms and feature pyramid networks to improve small-object detection, resulting in an approximate 5.0% gain in average precision on a wildlife dataset. Simões et al. (2023) combined MegaDetector with Faster R-CNN and Inception-ResNet-v2 to enhance species detection and counting in infrared-triggered camera videos.

To overcome the computational constraints inherent to two-stage detectors, one-stage models, such as YOLO and its derivatives, directly predict object categories and spatial locations, enabling faster inference and improved real-time performance. Applications in ecological monitoring have shown strong detection capability under variable illumination conditions. For example, Tan et al. (2022) reported a mean average precision (mAP@0.5) of 0.98 for image-based species detection and 88% accuracy for video classification using YOLOv5 across daytime and nighttime scenarios. Similarly, Bakana et al. (2024) further optimized performance through the lightweight WildARe-YOLO model, increasing inference speed by 17.65% and reducing computational cost by 50.92%.

Despite these advancements, integration of infrared-triggered camera networks with deep learning for systematic monitoring of migratory species remains limited. To address this gap, this study established a transect-based infrared camera monitoring system along migration corridors of *P. hodgsonii* in the Hoh Xil region. An improved YOLOv9 framework (Wang et al., 2023), designated IAE-YOLOv9, was developed by incorporating specialized feature enhancement modules, an automated counting component, and an interactive user interface. This integrated system successfully enables end-to-end automation of detection and counting during seasonal migration, providing a robust technical foundation for quantitative population density estimation and analysis of migratory dynamics.

MATERIALS AND METHODS

Dataset construction

This study employed two curated wildlife image datasets: the Sichuan-Gansu Wildlife Infrared Dataset (SGWID) and Hoh Xil Wildlife Camera Trap Dataset (HXWCTD). Both datasets were compiled using standardized data processing workflows and were used to evaluate model generalization, species-specific detection, and automatic counting of migratory taxa. Geographic sampling locations are shown in Figure 1A.

SGWID: SGWID comprises two data sources. The first includes infrared-triggered camera images of 11 wildlife species from the publicly available LoTE-Animal dataset (Liu et al., 2023), collected in Wolong National Nature Reserve, Sichuan Province, China, between 2009 and 2022. The second component includes images and video footage of four additional species captured between August 2022 and June 2023 during field deployments in Yanchiwan, Gansu Province. All data underwent uniform preprocessing steps, as detailed in Appendix A, Section A1. Representative images of each species are presented in Appendix A, Section A1, Supplementary Figure S1, and corresponding image counts are provided in Supplementary Table S1.

HXWCTD: The HXWCTD dataset was collected between July 2024 and March 2025 using a network of 12 infrared-triggered cameras positioned along the migratory corridor spanning from the Sonam Dargye Conservation Station to the Wu-Bei underpass of Qinghai-Xizang railway, near Hoh Xil Nature Reserve Area in Qumalai County, Qinghai Province (Figure 1B). Cameras were installed at intervals of 2–3 km along a designated transect and mounted 1.5 m above ground level at an angle of approximately 15° to optimize field-of-view for monitoring *P. hodgsonii.* Each unit was solar powered, with panels oriented 20° southwest to maintain energy efficiency. Camera placement was informed by terrain features, known wildlife trails, and historical observation records. Environmental data, including lighting conditions (daylight or infrared) and ambient temperature, were also recorded. Field observations and expert consultations confirmed the ecological relevance of this corridor for seasonal migration of *P. hodgsonii.*

Across the full monitoring period, 3 123 valid images and

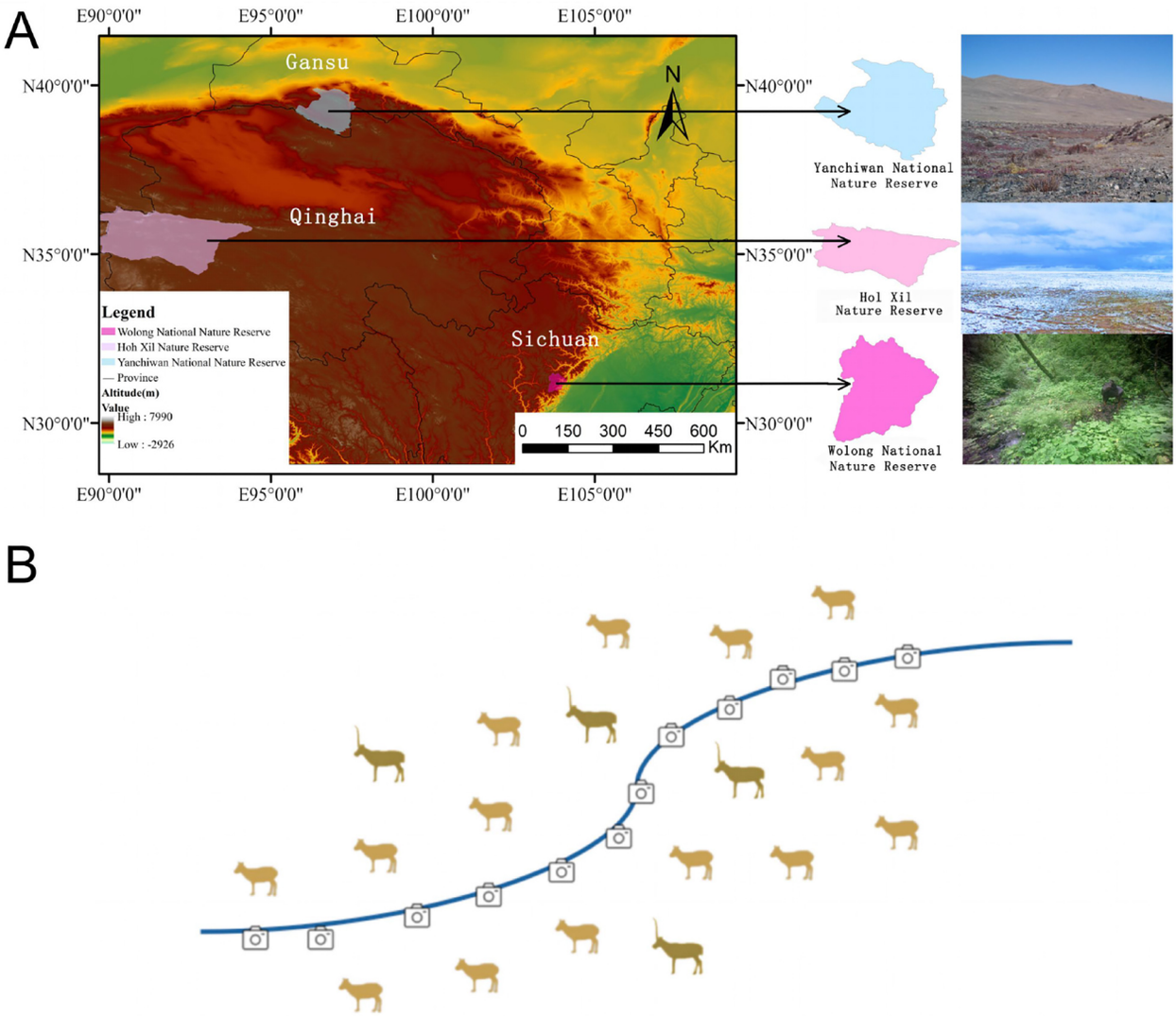

Figure 1 Nature reserves and infrared camera deployment

A: Geographic location and landscape overview of nature reserves associated with SGWID and HXWCTD. B: Schematic illustration of infrared-triggered camera placement and migration monitoring strategy.

3 123 ten-second video clips containing wildlife activity were collected. Of these, 1 067 images featured *P. hodgsonii* (Appendix A, Section A2, Supplementary Figure S2.1). Additional detected taxa included *Lepus oiostolus*, *Bos mutus*, *Procapra picticaudata*, *Equus kiang*, *Lynx lynx*, *Corvus corax tibetanus*, *Vulpes ferrilata*, *Canis lupus*, and *Buteo hemilasius* (Appendix A, Section A2, Supplementary Figure S2.2).

Annotation protocol and quality control: A unified, instance-level annotation protocol was developed for single-class object labeling across all images. Each individual was annotated when fully or partially visible and when at least one diagnostic feature, such as head, limbs, or horn morphology, was clearly discernible. Frames containing only shadows, severe motion blur, or indistinct silhouettes were excluded from analysis. Ambiguous or occluded cases were independently annotated in a blinded manner by two annotators using a standardized guideline and the LabelImg tool. Discrepancies were resolved through consensus discussion or, if necessary, majority voting adjudicated by a senior reviewer.

To assess inter-annotator consistency, a stratified random sampling was implemented based on camera location and lighting condition (day vs. night). Five overlapping subsets (S1–S5), each comprising 20% of the dataset, were independently labeled by two annotators. Inter-rater agreement was quantified using Cohen's κ, yielding κ=0.92±0.02, with a 95% confidence interval (CI) of [0.89, 0.96] (Warrens, 2015), indicating high concordance in instance-level detection and bounding-box placement (Figure 2).

IAE-YOLOv9 object detection model

The overall structure of the proposed IAE-YOLOV9 model is shown in Figure 3. This enhanced framework builds upon the original YOLOv9 architecture by sequentially integrating three key modules to improve detection accuracy, particularly for small objects under complex environmental conditions.

First, the involution operator (Li et al., 2021) was incorporated into the backbone network to expand the effective receptive field. Based on spatial anisotropy and channel-wise parameter sharing, the involution operator captures long-range dependencies and contextual cues more efficiently than standard convolution, thereby enhancing the ability of the model to detect small-object features.

Second, the traditional convolution layers in the backbone were replaced with the alterable kernel convolution (AKConv)

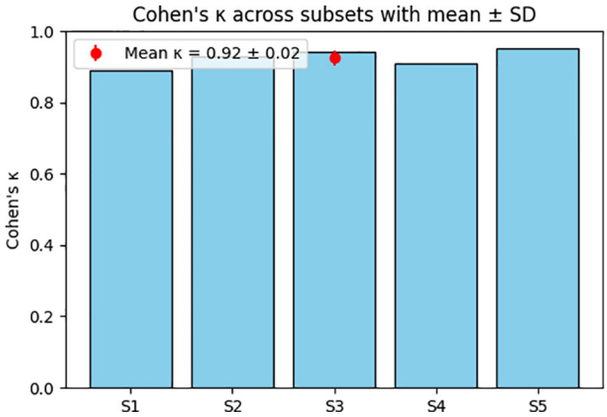

Figure 2 Cohen's κ across stratified subsets with overall mean and 95% confidence interval

module (Zhang et al., 2024). AKConv enables dynamic reconfiguration of kernel parameters and sampling patterns, allowing adaptive feature extraction across varying object scales and spatial resolutions while maintaining computational efficiency.

Third, an Element Multi-Scale Attention (EMA) module (Ouyang et al., 2023) was introduced before the detection head. EMA preserves per-channel information while selectively enhancing salient features through multi-scale fusion, achieving a balance between information processing and computational efficiency.

Implementation details for each module are provided in Appendix A: the YOLOv9 base module (Section A3), involution operator (Section A4), AKConv architecture (Section A5), and EMA mechanism (Section A6).

Training and evaluation of IAE-YOLOv9 models

Experimental setup and hyperparameter configuration

Model training and evaluation were performed using two platforms: a high-performance computing server and an NVIDIA Jetson Nano edge device. Detailed hardware and software specifications for both environments are summarized in Table 1.

Hyperparameter configurations for both model training and evaluation are listed in Table 2. The setup was optimized to ensure stable convergence and to mitigate overfitting during training. Specifically, the Adam optimizer was selected for its adaptive learning rate capabilities. A low dropout rate was applied to improve generalization, and both weight decay and momentum terms were introduced to enable training stability. The learning rate schedule was defined by an initial learning rate and a final learning rate factor, allowing fine control over convergence behavior and robustness during optimization.

Data augmentation

To enhance generalization and robustness, a range of data augmentation strategies were applied during training, including random flipping, scaling, cropping, and brightness adjustments. These augmentations served to artificially expand the diversity of the training dataset, enabling improved resilience to real-world variability and reducing susceptibility to overfitting.

Evaluation metrics

Model performance was assessed using multiple evaluation metrics, including precision, recall, mAP@0.5, frames per second (FPS), mean absolute error (MAE), and mean absolute percentage error (MAPE). The calculation formulas for these metrics are as follows:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{1}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{2}$$

$$mAP(0.5)=\frac{\sum_{i=1}^{k} AP_i}{k} \tag{3}$$

$$\text{AP}_i=\sum_{j=1}^{n-1}(r_{j+1}-r_j)P_{\text{inter}}(r_{j+1}) \tag{4}$$

$$\text{FPS}=\frac{1}{\text{Inference time per frame}} \tag{5}$$

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i-y_i| \tag{6}$$

$$MAPE=\frac{100\%}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i-y_i|}{y_i},\ y_i>0 \tag{7}$$

In the definitions of detection outcomes, a sample is recorded as a True Positive (TP) when wildlife is present in an image captured by an infrared-triggered camera and correctly identified by the model. A False Negative (FN) occurs when wildlife is present but not detected. A False Positive (FP) refers to the case where wildlife is absent but incorrectly predicted to be present. A True Negative (TN) denotes the correct identification of absence when no wildlife is present. In this study, mAP at an Intersection over Union (IoU) threshold of 0.5 was selected as the primary metric for evaluating wildlife detection performance (Appendix A, Section A7). Although COCO-style mAP@0.5:0.95 and per-class average precision (AP) are widely used for multi-species detection benchmarks, they are less applicable in this context given the single-species focus on *P. hodgsonii.* A broader evaluation across multiple IoU thresholds is provided in Appendix A, Section A8. Higher metric values indicate stronger detection capability of the model. In Eq. (3), AP_i denotes the average precision of the i-th category, and k represents the total number of object categories. In Eq. (4), r_j and r_{j+1} denote two adjacent recall levels on the Precision–Recall curve, $P_{inter}(r_{j+1})$ represents the interpolated precision at recall level r_{j+1}, and n is the number of sampled recall points used to calculate AP.

Additional performance indicators included the number of trainable parameters and computational complexity measured in Floating Point Operations (FLOPs). Parameter count reflects the scale of trainable parameters, with an excessive number likely to increase overfitting and training cost. FLOPs quantify the computation required for a single forward pass, with higher values implying increased hardware demands and reduced deployment efficiency on resource-constrained devices. Inference speed was assessed using FPS, defined as the number of images processed per second. Higher FPS indicates faster inference, which is essential for real-time monitoring, especially in edge-computing environments. For counting performance, MAE and MAPE were used. MAE represents the average absolute difference per image between predicted and manual counts, while MAPE captures the average percentage difference, reflecting both absolute and relative accuracy. Here, $\hat{y}_i$ and y_i denote the predicted and manually annotated counts for the i-th image, respectively, and N denotes the total number of evaluated images.

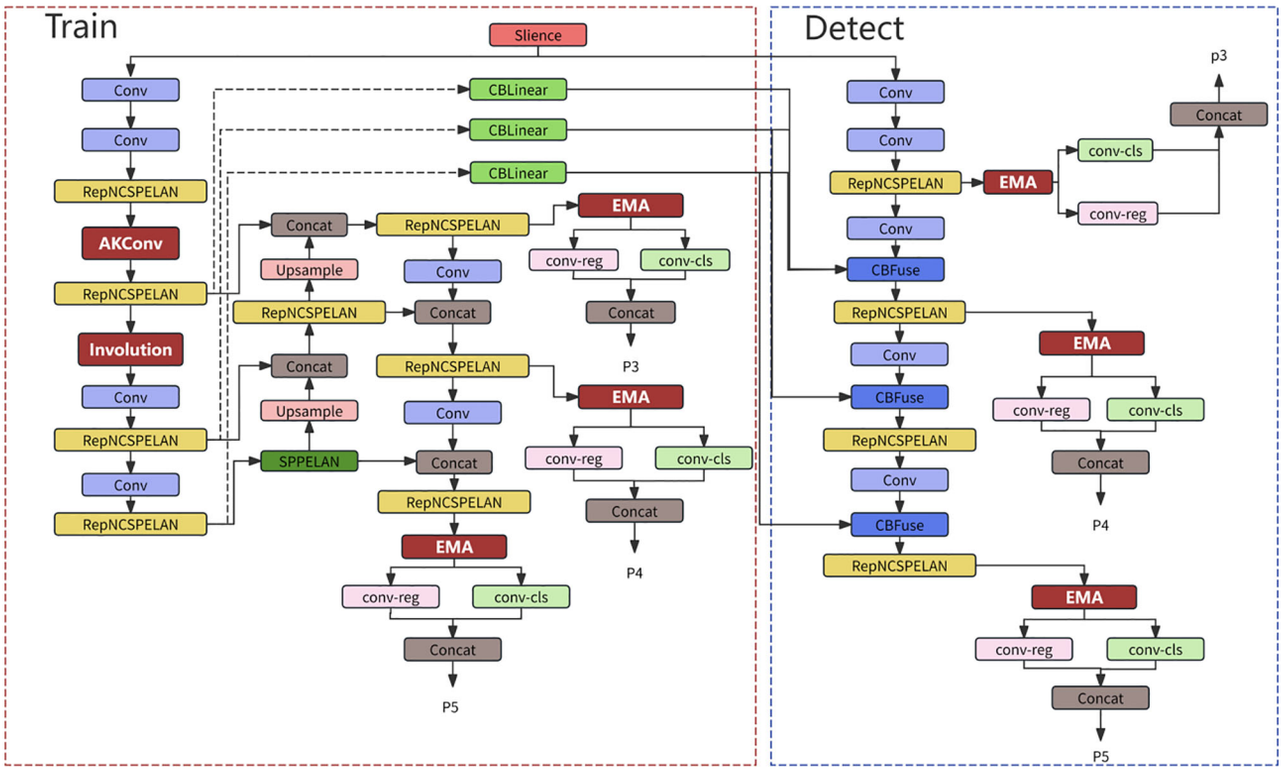

Figure 3 IAE-YOLOv9 model structure

Conv, Conv-reg and Conv-cls: Convolution; RepNCSPELAN: Feature extraction-fusion module; Upsample: Upsampling; Concat and CBFuse: Feature fusion; SPPELAN: Spatial pyramid pooling structure; EMA: EMA-Attention attention mechanism; Involution: Involution convolutional network operator; AKConv: AKConv lightweight architecture.

Table 1 Configuration parameters of the experimental setup

Name	Server	Jetson Nano
CPU	Intel(R) Core(TM) i9-9900KF CPU 3.60 GHz	ARM Cortex-A57 MPCore
GPU	NVIDIA GeForce GTX 1080 Ti	128-core NVIDIA Maxwell GPU
Operating system	Windows 10	Ubuntu 16.04
Programming language	Python 3.8	Python 3.8
Memory	27GB	4GB

Table 2 Configuration of experimental hyperparameters

Hyperparameter	Value
Dropout	0.005
Workers	8
Epoch	100
Batch-size	16
Momentum	0.937
Optimizer	Adam
Weight-decay	0.0005
lr0 (Initial learning rate)	0.01
Lrf (Final learning rate factor)	0.01

Inference and post-processing parameters

To ensure consistency and reproducibility across all experimental stages, a unified post-processing configuration was applied during training, validation, and testing. A confidence threshold of 0.5 was used as default for all inference tasks unless otherwise stated. Standard Greedy Non-Maximum Suppression (NMS) was implemented with an IoU threshold of 0.45. Given the single-species detection focus on *P. hodgsonii*, class-agnostic NMS was enabled. These settings were applied uniformly across ablation experiments, model comparisons, SGWID evaluations, and Jetson Nano deployment.

EXPERIMENTAL EVALUATION OF IAE-YOLOV9 ON THE SGWID DATASET

To evaluate the feasibility, generalizability, and adaptability of the proposed IAE-YOLOv9 model in wildlife monitoring applications, systematic training and testing were conducted using the SGWID dataset. A structured series of experiments was designed to assess model performance. Specifically, ablation studies were first performed to isolate the functional contributions of the involution, AKConv, and EMA modules. These experiments examined both individual effects and combinatorial interactions to clarify how each component contributed to performance enhancement. Subsequently, comparative evaluations were conducted against several representative object detection frameworks, including SSD (Liu et al., 2016), Faster R-CNN (Ren et al., 2017), YOLOv5s (Bochkovskiy et al., 2020), YOLOv7 (Wang et al., 2023), YOLOv9 (Wang et al., 2023), and YOLOv10-X (Wang et al., 2024). These comparisons enabled systematic benchmarking of IAE-YOLOv9 in terms of accuracy, robustness, and cross-architecture consistency. Confusion matrices were generated separately for both YOLOv9 and IAE-YOLOv9 on the test set,

enabling direct visual comparison of detection accuracy and inter-class prediction discrepancies.

All confidence intervals (95% CI) were computed using image-level bootstrap resampling with 1 000 iterations. For inter-model comparisons, paired bootstrap tests were applied to per-image metrics to determine statistical significance.

Analysis of ablation experiment results

To evaluate the functional contributions of the involution, AKConv, and EMA modules within the IAE-YOLOv9 architecture, ablation experiments were conducted using the SGWID dataset (Table 3). Each module was introduced independently into the YOLOv9 baseline to quantify its individual impact on detection performance, increasing precision by 1.4%, 0.4%, and 0.9%, respectively. When all three modules were combined, their interaction produced a non-additive enhancement in detection performance, resulting in a 2.2% increase in mAP@0.5, reaching 95.2% (95%CI: 94.4–95.5%), while maintaining real-time inference capability. These findings indicate that the proposed architecture delivers both high accuracy and strong applicability.

The performance gains observed from full integration can be attributed to the complementary inductive biases introduced by the three modules. Involution applies spatially adaptive kernels that suppress background interference and sharpen boundaries of small objects. AKConv utilizes anisotropic kernel structures to more effectively capture elongated or direction-sensitive morphological features, thereby improving localization precision. EMA introduces dynamic multi-scale attention that aligns fine-grained spatial details with coarse semantic context, enhancing discriminability under conditions of occlusion or crowding. By targeting orthogonal dimensions —spatial adaptivity, shape modeling, and scale alignment —these modules collectively reduce classification ambiguity, boundary shifts, and duplicate/missed detections. This synergistic interaction explains the observed non-linear improvement (Table 3) and demonstrates that the integration reflects a fusion of distinct inductive biases rather than a simple aggregation of components. √ indicates that the module is integrated; — indicates that the module is not integrated.

Performance comparison with benchmark models

Using the SGWID dataset, systematic comparative experiments were conducted to evaluate the performance advantages of IAE-YOLOv9 relative to several widely used object detection models (Table 4). Results demonstrated that IAE-YOLOv9 consistently outperformed SSD, Faster R-CNN, YOLOv5s, YOLOv7, YOLOv9, and YOLOv10-X across key performance metrics. Specifically, IAE-YOLOv9 achieved a precision of 94.8% (95% CI: 94.2%–95.1%), mAP@0.5 of 95.2% (95% CI: 94.4%–95.5%), and recall rate of 91.2% (95% CI: 90.8%–91.5%), corresponding to gains of up to 9.4% in mAP@0.5 compared to SSD and improvements of 3.5% in precision and 5.8% in recall compared to the baseline YOLOv9 model.

To assess the statistical reliability of these improvements, paired bootstrap significance tests (B=1 000) were performed using per-image AP differences between IAE-YOLOv9 and each baseline model on the SGWID dataset (Good, 2005). The performance advantages of IAE-YOLOv9 over YOLOv9, YOLOv7, and YOLOv10-X in mAP@0.5 were all statistically significant (P<0.05), with corresponding precision and recall gains also reaching significance. These findings confirm that the superior performance of IAE-YOLOv9 is not attributable to random fluctuations within the test set but instead reflects consistent and meaningful enhancements in detection accuracy over existing object detectors.

Traditional models such as SSD and Faster R-CNN exhibited missed detection rates and suboptimal computational efficiency when applied to complex backgrounds and heterogeneous targets. Similarly, certain lightweight models showed limited sensitivity to small-object features. In contrast, by integrating the involution operator, lightweight AKConv module, and the EMA mechanism, IAE-YOLOv9 effectively enhanced feature representation and global information perception, achieving superior detection accuracy and stronger generalization capability.

Regarding inference speed, IAE-YOLOv9 maintained a throughput of 24.8 FPS on the server platform, comparable to YOLOv9 and exceeding that of Faster R-CNN, YOLOv7, and YOLOv10-X, suggesting an acceptable computational profile

Table 3 Ablation experiment results for the IAE-YOLOv9 model

Experiment No.	Involution	AKConv	EMA	**Precision** (95% CI)	**mAP@0.5** (95% CI)	**Recall** (95% CI)
1	—	—	—	91.3% (90.6–92.8)	93.0% (92.1–93.7)	85.3% (84.4–85.9)
2	√	—	—	92.7% (91.4–93.4)	93.1% (92.2–94.0)	87.2% (85.9–88.3)
3	—	√	—	91.7% (90.9–92.9)	92.1% (91.5–92.2)	85.8% (84.4–86.1)
4	—	—	√	92.2% (90.9–92.7)	92.3% (91.3–93.2)	86.6% (85.2–87.1)
5	√	√	—	92.0% (91.1–92.3)	92.7% (91.9–93.2)	86.0% (85.3–86.4)
6	√	—	√	93.7% (93.1–94.2)	94.6% (94.2–94.7)	89.7% (89.1–90.1)
7	—	√	√	92.6% (92.0–93.1)	93.7% (92.7–94.1)	87.9% (87.1–88.4)
8	√	√	√	94.8% (94.2–95.1)	95.2% (94.4–95.5)	91.2% (90.8–91.5)

Table 4 Performance comparison of different object detection models on the SGWID dataset

Method	Parameter	**Precision** (95% CI)	**mAP@0.5** (95% CI)	**Recall** (95% CI)	FLOPs	FPS
SSD	34.5M	82.7% (80.1–82.9)	85.8% (84.9–86.1)	79.7% (78.9–81.4)	74.3G	36.5
Faster R-CNN	41M	85.3% (84.8–85.9)	88.3% (87.5–88.9)	82.2% (81.5–82.9)	251.4G	10.8
YOLOv5s	7.2M	91.8% (91.1–92.3)	92.5% (91.9–93.1)	84.9% (84.4–85.2)	16.4G	165
YOLOv7	36.49M	92.3% (91.9–92.5)	94.2% (93.9–94.5)	90.2% (89.6–90.3)	315.9G	8.6
YOLOv9	25.3M	91.3% (90.8–91.8)	93.0% (93.3–94.1)	85.3% (84.9–85.7)	102.7G	26.4
YOLOv10-X	63M	92.4% (91.8–92.8)	94.1% (93.8–94.3)	89.6% (89.2–89.9)	162.4G	16.7
IAE-YOLOv9	27.3M	94.8% (94.2–95.1)	95.2% (94.4–95.5)	91.2% (90.8–91.5)	109.3G	24.8

for practical deployment (Table 4). The confusion matrix analyses for IAE-YOLOv9 and YOLOv9 on the test dataset (Appendix A, Section A9) provided further validation through visualization of class-wise prediction accuracy and misclassification patterns.

To evaluate edge performance, IAE-YOLOv9 was deployed on the Jetson Nano using an optimized pipeline incorporating TensorRT-based INT8 quantization with calibration, layer fusion, kernel auto-tuning, and asynchronous preprocessing-inference scheduling. Under this configuration, the model achieved a stable throughput of 2.78 FPS at the input resolution used in this study, with a relative drop in mAP@0.5 of no more than 1.23% compared to the full-precision model. This configuration supports preliminary recognition and counting of *P. hodgsonii* on edge platforms, enabling rapid triage of camera-trap data and reducing manual annotation burden. Further details on cross-platform inference performance are provided in Appendix A, Section A10.

Visualization of detection results

To intuitively assess the effectiveness of the proposed model, representative detection results were visualized to highlight the performance improvements achieved by IAE-YOLOv9. As shown in Figure 4A, the optimized model demonstrated superior accuracy in small and distant targets. In the case of *Macaca thibetana*, IAE-YOLOv9 achieved precise localization of distant individuals, with bounding boxes closely aligned to object contours and confidence scores markedly enhanced. For *Hystrix brachyura* and *Rhinopithecus roxellana*, IAE-YOLOv9 significantly reduced missed detections compared to YOLOv9, particularly for small and distant individuals, confirming its enhanced reliability and robustness in small-object detection tasks.

Figure 4B presents detection comparisons between IAE-YOLOv9 and YOLOv9 under complex environmental conditions. Natural scenarios involving variable lighting, object occlusions, and complex backgrounds present substantial challenges for object detection. The optimized IAE-YOLOv9 model demonstrated excellent performance in such scenarios. Under dim lighting, the confidence level for detecting *Capricornis sumatraensis* increased from 0.65 to 0.82. In the complex settings where *H. brachyura* appeared against a rock-like background, confidence increased from 0.88 to 0.92. For *Ailuropoda melanoleuca* in a nighttime environment, confidence increased from 0.80 to 0.93. These improvements highlight the enhanced detection reliability of IAE-YOLOv9 under challenging field conditions.

Figure 4C illustrates multi-scale detection scenarios involving overlapping and variably sized individuals. When detecting *M. thibetana*, IAE-YOLOv9 accurately identified small individuals even when partially occluded by larger ones, with bounding boxes conforming tightly to object contours. When detecting *Pseudois nayaur,* IAE-YOLOv9 successfully identified partially exposed individuals positioned close to the camera. In addition, the model effectively distinguished between multiple individuals without generating duplicate boxes. In contrast, YOLOv9 exhibited several failure modes, including misaligned bounding boxes for small objects, insufficient coverage for larger ones, and reduced detection accuracy across all scenarios. These results demonstrate the superior precision, adaptability, and robustness of IAE-YOLOv9 in diverse and complex visual environments.

CASE STUDY OF *P. HODGSONII* MONITORING

Building upon the previously validated accuracy and robustness of the IAE-YOLOv9 framework in object detection tasks, this study further integrated an automated counting module and an interactive user interface. A focused case study was conducted on *P. hodgsonii*, a national first-class protected species, to evaluate practical applicability in real-world ecological monitoring. This application aimed to assess utility for key conservation objectives, including population density estimation and analysis of migration dynamics along established movement corridors.

Detection model and interface design for *P. hodgsonii*

Detection of *P. hodgsonii* was performed using the IAE-YOLOv9 model, fine-tuned via transfer learning on 1 067 infrared-triggered camera images. The model achieved an identification accuracy of 89.81% (95% CI: 88.34%–91.12%) and mAP@0.5 of 72.67% (95% CI: 71.98%–73.82%). Post-processing involved application of non-maximum suppression (NMS) with an IoU threshold of 0.45 to remove overlapping bounding boxes corresponding to the same individual, based on the heuristic that any two boxes with IoU above this threshold represent a duplicate detection. An automatic counting module was implemented by aggregating the number of predicted bounding boxes per image. All reported counts refer to image-level outputs, with no attempt made to estimate unique individuals across frames or camera stations.

To facilitate user interaction, a graphical interface was developed to support batch image input and visualized outputs. The interface displays relevant metadata for each detected *P. hodgsonii*, including category, count, confidence score, detection time, and bounding box coordinates, enabling intuitive interpretation of detection results (Appendix A, Section A11).

Quantitative comparison between model detections and manual counts of *P. hodgsonii*

To rigorously assess detection and counting accuracy, manually annotated results were used as benchmark ground truth. Manual counts were produced by trained annotators who independently reviewed each infrared-triggered camera image. Individuals were identified based on full or partial visibility of key morphological features—such as head, limbs, horn shape, and posture. In cases of ambiguity due to occlusion or poor image quality, annotations were cross-validated by at least two annotators, with final counts determined by consensus or majority vote to reduce subjectivity.

Model-generated counts were then quantitatively compared to manual annotations on an image-by-image basis. Three metrics were used for evaluation: absolute error (AE), representing deviation per image; relative error (RE), reflecting proportional deviation from the reference count; and accuracy, representing the proportion of correct predictions. The corresponding formulas are as follows:

$$\mathrm{AE} = \left| N_{\mathrm{model}} - N_{\mathrm{manual}} \right| \tag{8}$$

$$RE = \frac{N_{\mathrm{model}} - N_{\mathrm{manual}}}{N_{\mathrm{manual}}} \times 100\% \tag{9}$$

where N_{model} is the number identified by the model and N_{manual} is the number counted manually.

All detection results were generated using a confidence

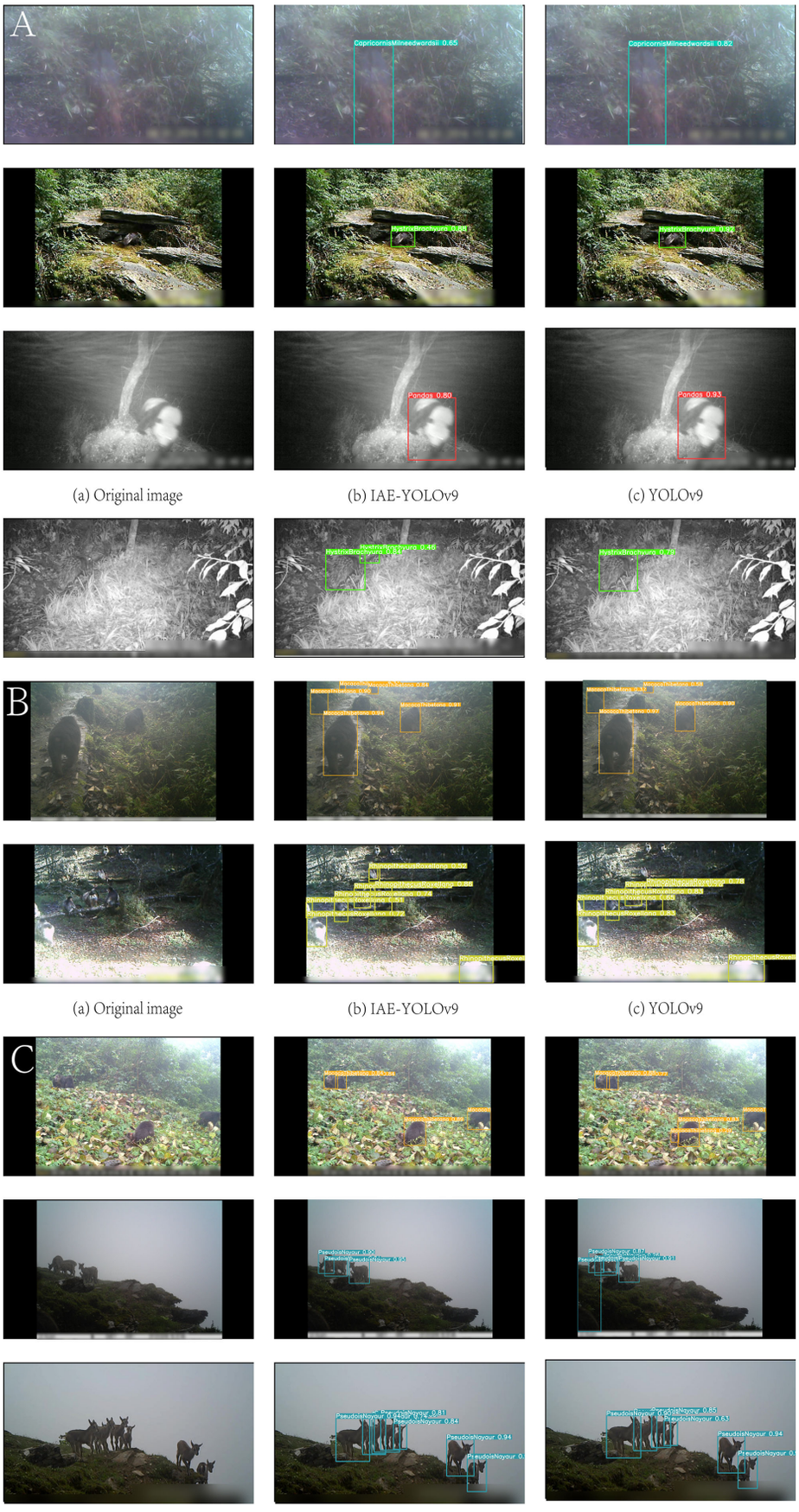

Figure 4 Comparative multi-scene and small object detection performance of IAE-YOLOv9 and YOLOv9

A: Performance comparison in open field environments. B: Performance comparison in visually complex natural scenes. C: Performance comparison in multi-object field scenarios.

threshold of 0.5, a standard setting in object detection workflows. Additional information regarding threshold selection and the corresponding precision-recall trade-offs is provided in Appendix A, Section A12.

From August 2024 to March 2025, 12 infrared-triggered cameras were deployed to collect monitoring images of *P. hodgsonii*. Both manual annotation and automated detection were used to quantify individuals in the captured images (Figure 5). To prevent overcounting within the same field of view, a 30 min temporal window was applied for within-camera deduplication (Ahmad et al., 2024). Consecutive detections of the same species within this interval were treated as a single trigger event, and only the maximum count within each interval was retained. Based on this protocol, the manual count totaled 1 826 individuals, while the model produced a count of 1 612, yielding a model-based counting accuracy of 88.28% (95% CI: 87.45%–89.12%). In addition, to explore how specific scene conditions influence counting accuracy, Appendix A, Section A13 provides a detailed analysis of typical error types and their relationships with key scene factors. Building on this understanding of error sources, a bootstrap paired significance test (B=1 000) was conducted using per-image counting errors. Results confirmed that the observed counting accuracy was statistically significant (P<0.05), indicating consistent agreement between model predictions and manual annotations.

In addition to overall counts, error distributions were analyzed at the image level (Figure 5B). The histogram of absolute error revealed high concordance between model predictions and manual annotations, with most discrepancies limited to 0–1 individual per image. The histogram of relative error (RE%) further confirmed that deviations were generally minor, clustering tightly around 0%. Outlier RE% values were observed predominantly in images containing very few individuals, where a single discrepancy resulted in proportionally large errors. These findings indicate that the model not only delivers strong overall accuracy but also maintains reliable performance across individual images.

To assess temporal dynamics, detection and counting data were aggregated in 10 day intervals throughout the monitoring period. Model-based results were compared directly with manual annotations to evaluate temporal consistency. Trends in population fluctuations showed strong convergence across both methods (Figure 5A, B). These findings not only validate the stability and reliability of the model in migratory species monitoring but also provide support for the analysis of relative migration patterns and future density estimation. Therefore, the proposed approach offers a strong foundation for long-term ecological monitoring and species conservation management, with significant practical value in identifying critical migration periods and habitat utilization patterns.

The analysis also revealed distinct seasonal trends. From early August to late October, the number of detected individuals increased markedly, corresponding to the known post-calving return migration of *P. hodgsonii*. This pattern reflects seasonal habitat use and migration rhythms across life cycle stages. From late October to early January, detection rates plateaued, followed by a renewed increase after January —potentially driven by localized movement or environmental cues. The consistency between model-based and manual detection patterns throughout these periods validates the capacity of the model to effectively capture biologically meaningful migration dynamics and highlights its practical value for long-term monitoring and adaptive conservation management of *P. hodgsonii*.

DISCUSSION

Methodological enhancements for automated wildlife detection

As a keystone species, *P. hodgsonii* plays a critical role in maintaining biodiversity and ecological balance within the fragile alpine ecosystem of the Sanjiangyuan region. Effective monitoring of this migratory ungulate is essential for evidence-based conservation but remains constrained by two key challenges, notably the high-altitude and harsh habitat conditions, which restrict manual surveys, and the analysis of massive image datasets, which demands substantial labor and resources (Brookfield & Allen, 1989). Although infrared-triggered camera traps coupled with deep learning algorithms have gained traction in ecological research, their application in migratory species monitoring remains underexplored. To overcome these limitations, this study established a transect-based surveillance network using infrared-triggered cameras deployed along migration routes of *P. hodgsonii* and incorporated an enhanced detection framework based on the IAE-YOLOv9 architecture. This integrated approach enabled automated detection and dynamic tracking of *P. hodgsonii*, substantially improving both the accuracy and efficiency of migration data acquisition. Quantitative benchmarks demonstrated that IAE-YOLOv9 achieved superior performance compared to the baseline YOLOv9 model in wildlife detection. The clear improvements in precision, mAP0.5, and recall can be attributed to the synergistic integration of three novel architectural modules.

Involution convolutional operator: This module facilitates position-specific channel-wise interactions, thereby enhancing feature extraction in visually complex scenes, such as infrared images containing dense vegetation or rocky terrain. Enhanced discrimination of background and target features supports more reliable detection performance. This effect aligns with the findings of Wang et al. (2022), who reported notable accuracy improvements when embedding involution into YOLOv5 fusion layers.

AKConv: The adoption of asymmetric convolutional kernels reduces computational redundancy and heightens sensitivity to small, low-contrast objects, particularly relevant for distant or partially obscured antelopes. Consistent with results from Pei et al. (2024), incorporation of AKConv led to an observed 1.7-fold increase in small-object detection rates.

EMA: By enhancing global semantic representation and enabling more coherent multi-scale feature fusion, EMA improves both boundary localization and class discrimination, particularly in scenes with object overlap or dense clustering. Comparable enhancements have been reported by Hao et al. (2023) in livestock head-and-leg detection tasks.

Evaluation across both datasets revealed that model performance was suboptimal when trained on limited data, whereas training on larger datasets yielded not only improved accuracy but also greater performance stability. These findings underscore the importance of long-term monitoring to expand the image repository of *P. hodgsonii*, which is expected to further improve species identification accuracy. Model optimization was also achieved by fine-tuning hyperparameters, including adjustment to training epochs and convolutional layers. To assess generalizability on the

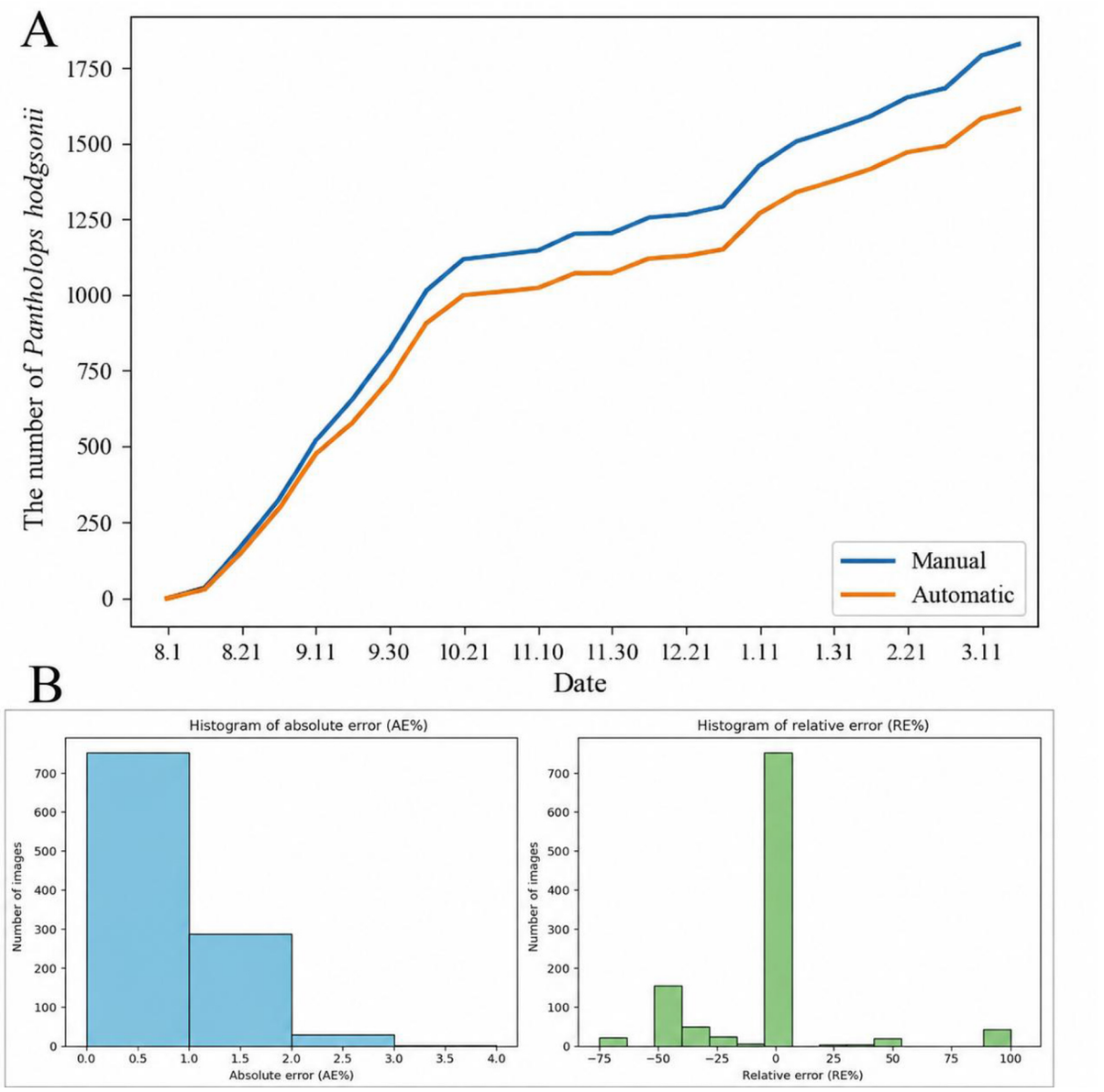

Figure 5 Population trend of *P. hodgsoni* based on manual and model-based count comparison with error analysis

A: Migration dynamics of *P. hodgsonii* derived from manual counts and automated model-based estimates. B: Error distribution of model-based counts for *P. hodgsonii* on the HXWCTD dataset.

HXWCTD dataset, transfer learning was employed (Weiss et al., 2016). When trained on this dataset, the IAE-YOLOv9 model achieved an identification accuracy of 89.81% (95% CI: 88.34%–91.12%) for *P. hodgsonii*. Integration of an automated counting module yielded a counting accuracy of 88.28% when benchmarked against manual annotations.

Applications and limitations of deep learning in monitoring *P. hodgsonii* migration

This study successfully applied the IAE-YOLOv9 model for automated detection and counting of the nationally protected species *P. hodgsonii*, achieving a counting accuracy of 88.28% relative to manual annotations. The model identified 1 612 individuals, compared to 1 826 recorded manually. Time-series analysis of detection data revealed distinct seasonal trends, with a marked increase in individual detections from early August to late October, corresponding to the post-calving return migration, followed by a plateau through to early January and a subsequent rise after mid-January. These fluctuations captured the annual migratory dynamics and localized activity rhythms of the species with high temporal resolution. However, it should be noted that all analyses were based on image-level counts without inter-camera deduplication or individual re-identification. As a result, the findings represent relative activity trends rather than estimates of absolute population size or migration magnitude.

The observed patterns aligned broadly with the migration patterns of *P. hodgsonii* derived from satellite tracking studies (Buho et al., 2011), underscoring the reliability of the automated approach in capturing biologically relevant signals. Furthermore, the user interface developed in this study facilitates real-time visualization and retrieval of *P. hodgsonii* occurrence data, streamlining access to quantitative information for future ecological analysis and demographic modeling. This platform provides an efficient, scalable tool for long-term monitoring and supports applied conservation initiatives targeting *P. hodgsonii*.

Given the conservation sensitivity of *P. hodgsonii* and its vulnerability to poaching, strict data protection protocols were

implemented. Geospatial coordinates and sensitive habitat information were encrypted and restricted to authorized institutions, while public visualizations were masked to prevent location disclosure. Furthermore, privacy safeguards were applied to image data to avoid incidental capture of individuals without consent, thereby mitigating potential ethical concerns. Detailed procedures concerning data security, ethical compliance, and community engagement are described in the final sections of the paper.

CONCLUSION

The integration of infrared-triggered cameras and deep learning offers substantial ecological utility for the monitoring of migratory species. This study presents an automated monitoring framework for migratory species that integrates species detection, individual counting, and a user interface to facilitate real-time identification and migration analysis of target taxa. The proposed system provides scalable and efficient technical support for long-term ecological surveillance and the study of migration dynamics in the field.

To enhance the ecological inference of such monitoring efforts, future research should incorporate image-level detection data into population density estimation frameworks, such as Random Encounter Models (REM) (Kavčić et al., 2021) and Spatially Explicit Capture-Recapture (SECR) (Efford & Fewster, 2013). Integrating these approaches will enable the translation of relative activity patterns into more accurate estimates of abundance, thereby improving the utility of automated monitoring data for conservation planning and population management.

SCIENTIFIC FIELD SURVEY PERMISSION INFORMATION

All field activity and data collection were conducted under permits issued by the appropriate regulatory authorities. All research activities complied with applicable ethical standards and data governance requirements.

DATA AVAILABILITY

After publication, a de-identified subset of the Hoh Xil Wildlife Camera Trap Dataset (HXWCTD) will be made publicly available. This subset contains 120 images (10 per camera from 12 infrared-triggered cameras) in COCO format, including bounding-box annotations, class labels, and camera IDs. All image metadata, such as geographic coordinates and timestamps, have been masked to protect sensitive habitats along the migration corridor. Access to the complete dataset will be reviewed and approved by the 21st Century Center for Science and Technology, Ministry of Science and Technology. Applicants will be required to sign a *Confidentiality Agreement* and/or *Data Use Agreement*; review time may vary depending on the project scope and approval workflow.

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

REFERENCES

Ahmad F, Mori T, Rehan M, et al. 2024. Applying a random encounter model to estimate the Asiatic black bear (*Ursus thibetanus*) density from camera traps in the Hindu Raj Mountains, Pakistan. *Biology*, **13**(5): 341.

Bakana SR, Zhang YF, Twala B. 2024. WildARe-YOLO: a lightweight and efficient wild animal recognition model. *Ecological Informatics*, **80**: 102541.

Bauer S, Hoye BJ. 2014. Migratory animals couple biodiversity and ecosystem functioning worldwide. *Science*, **344**(6179): 1242552.

Bochkovskiy A, Wang CY, Liao HYM. 2020. YOLOv4: optimal speed and accuracy of object detection. *arXiv*, doi: https://doi.org/10.48550/arXiv.2004.10934.

Brookfield H, Allen B. 1989. High-altitude occupation and environment. *Mountain Research and Development*, **9**(3): 201–209.

Buho H, Jiang Z, Liu C, et al. 2011. Preliminary study on migration pattern of the Tibetan antelope (*Pantholops hodgsonii*) based on satellite tracking. *Advances in Space Research*, **48**(1): 43–48.

Dharaiya N, Bargali HS, Sharp T. 2016. Melursus ursinus. *The IUCN Red List of Threatened Species 2016*: e. T13143A45033815.

Efford MG, Fewster RM. 2013. Estimating population size by spatially explicit capture-recapture. *Oikos*, **122**(6): 918–928.

Good P. 2005. Permutation, Parametric, and Bootstrap Tests of Hypotheses. New York: Springer.

Hao WL, Ren C, Han M, et al. 2023. Cattle body detection based on YOLOv5-EMA for precision livestock farming. *Animals*, **13**(22): 3535.

Kavčić K, Palencia P, Apollonio M, et al. 2021. Random encounter model to estimate density of mountain-dwelling ungulate. *European Journal of Wildlife Research*, **67**(5): 87.

Leslie Jr DM, Schaller GB. 2008. *Pantholops hodgsonii* (Artiodactyla: Bovidae). *Mammalian Species*, (817): 1–13.

Li D, Hu J, Wang CH, et al. 2021. Involution: inverting the inherence of convolution for visual recognition. *In*: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 12316–12325.

Li S, Wang DJ, Xiao ZS, et al. 2014. Camera-trapping in wildlife research and conservation in China: review and outlook. *Biodiversity Science*, **22**(6): 685–695. (in Chinese)

Liu D, Hou J, Huang SL, et al. 2023. LOTE-animal: a long time-span dataset for endangered animal behavior understanding. *In*: Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris: IEEE, 20007–20018.

Liu W, Anguelov D, Erhan D, et al. 2016. SSD: single shot MultiBox detector. *In*: Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 21–37.

Lyster S. 1989. The convention on the conservation of migratory species of wild animals (the Bonn Convention). *Natural Resources Journal*, **29**(4): 979.

Ouyang DL, He S, Zhang GZ, et al. 2023. Efficient multi-scale attention module with cross-spatial learning. *In*: Proceedings of the ICASSP 2023-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 1–5.

Pei YP, Jia GQ, Hui BF, et al. 2024. YOLOv8-based brown bear recognition algorithm in Qinghai-Tibet Plateau. *In*: Proceedings of the SPIE 13399, 9th International Workshop on Pattern Recognition. Xiamen: SPIE, 1339903.

Ren SQ, He KM, Girshick R, et al. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(6): 1137–1149.

Simões F, Bouveyron C, Precioso F. 2023. DeepWILD: wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics*, **75**: 102095.

Swann DE, Hass CC, Dalton DC, et al. 2004. Infrared-triggered cameras for detecting wildlife: an evaluation and review. *Wildlife Society Bulletin*, **32**(2): 357–365.

Tan MY, Chao WT, Cheng JK, et al. 2022. Animal detection and classification from camera trap images using different mainstream object detection architectures. *Animals*, **12**(15): 1976.

Ukwuoma CC, Qin ZG, Yussif SB, et al. 2022. Animal species detection and classification framework based on modified multi-scale attention mechanism and feature pyramid network. *Scientific African*, **16**: e01151.

UN Environment Programme World Conservation Monitoring Centre. 2024. State of the world's migratory species. Cambridge: UN Environment

Programme World Conservation Monitoring Centre.
Vecvanags A, Aktas K, Pavlovs I, et al. 2022. Ungulate detection and species classification from camera trap images using RetinaNet and Faster R-CNN. *Entropy*, **24**(3): 353.
Wang A, Chen H, Liu LH, et al. 2024. YOLOv10: real-time end-to-end object detection. *In*: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc. , 3429.
Wang CY, Bochkovskiy A, Liao HYM. 2023. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *In*: Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver: IEEE, 7464–7475.
Wang FF, He PF, Zhang TJ, et al. 2022. Wildlife detection algorithm based on INV-YOLOv5m. *In*: Proceedings of 2022 Global Reliability and Prognostics and Health Management (PHM-Yantai). Yantai: IEEE, 1–6.
Warrens MJ. 2015. Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy*, **5**(4): 1000197.
Weiss K, Khoshgoftaar TM, Wang DD. 2016. A survey of transfer learning. *Journal of Big Data*, **3**(1): 9.
Zhang X, Song YZ, Song TT, et al. 2024. LDConv: linear deformable convolution for improving convolutional neural networks. *Image and Vision Computing*, **149**: 105190.
Zhang YJ, Zhu YX, Li JX, et al. 2019. Current status and future directions of the Tibetan Plateau ecosystem research. *Science Bulletin*, **64**(7): 428–430.