Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)
Special Track on AI for Good (Projects)

Long-term Monitoring of Bird Flocks in the Wild

Kshitiz[1] , **Sonu Shreshtha**[1] , **Ramy Mounir**[3] , **Mayank Vatsa**[1] , **Richa Singh**[1] , **Saket Anand**[2] , **Sudeep Sarkar**[3] , **Sevaram Mali Parihar**[4]
[1]IIT Jodhpur, India
[2]IIIT Delhi, India
[3]University of South Florida, Tampa, Florida, USA
[4]Crane Conservationist, Khichan, India
{kshitiz.1, shreshtha.1, mvatsa, richa}@iitj.ac.in, {ramy, sarkar}@usf.edu, anands@iiitd.ac.in, sevarammaliparihar@gmail.com

Abstract

Monitoring and analysis of wildlife are key to conservation planning and conflict management. The widespread use of camera traps coupled with AI-based analysis tools serves as an excellent example of successful and non-invasive use of technology for design, planning, and evaluation of conservation policies. As opposed to the typical use of camera traps that capture still images or short videos, in this project, we propose to analyze longer term videos monitoring a large flock of birds. This project, which is part of the NSF-TIH Indo-US joint R&D partnership, focuses on solving challenges associated with the analysis of long-term videos captured at feeding grounds and nesting sites, among other such locations that host large flocks of migratory birds. We foresee that the objectives of this project would lead to datasets and benchmarking tools as well as novel algorithms that would be instrumental in developing automated video analysis tools that could in turn help understand individual and social behavior of birds. The first of the key outcomes of this research will include the curation of challenging, real-world datasets for benchmarking various image and video analytics algorithms for tasks such as counting, detection, segmentation, and tracking. Our recent efforts towards this outcome is a curated dataset of 812 high-resolution, point-annotated, images (4K - 32MP) of a flock of Demoiselle cranes (Anthropoides virgo) taken from their feeding site at Khichan, Rajasthan, India. The average number of birds in each image is about 207, with a maximum count of 1500. The benchmark experiments show that state-of-the-art vision techniques struggle with tasks such as segmentation, detection, localization, and density estimation for the proposed dataset. Over the execution of this open science research, we will be scaling this dataset for segmentation and tracking in videos, as well as developing novel techniques for video analytics for wildlife monitoring.

1 Problem Statement

Birds are essential components of numerous ecosystems and play pivotal roles in maintaining ecological balance [Şekercioğlu *et al.*, 2004]. In addition to being sensitive to alterations in habitat structure and composition, they serve as excellent indicators of habitat quality and biodiversity [Zakaria *et al.*, 2005]. Regular monitoring of wildlife is essential for conserving biodiversity and ensuring sustainable development. Due to their widespread distribution, rapid mobility, and enhanced sensitivity to environmental changes, birds have gained growing significance for monitoring. Migration of birds [Higuchi, 2012] is a vital indicator of biodiversity and can provide valuable insights into ecosystem changes as well as highlight potential climate change-related issues [Chen *et al.*, 2011]. By identifying major stopover and wintering sites of migratory birds and developing long-term monitoring tools, we can safeguard endangered species and track the state of the environment. Thus, precise and timely data on their composition, density, and distribution, along with reactions to environmental change and human-related activities, are required for the effective conservation and management of bird species. This necessitates the development of efficient bird counting and monitoring techniques that help provide insights into their interaction with environmental elements such as foraging and nesting preferences, which are essential for the creation of wildlife management plans and targeted conservation strategies [Nichols and Williams, 2006]. In addition, traditional methods of bird counting [Delany, 2010], which rely on human efforts, are not only labor-intensive, but also produce imprecise estimates of the density of birds.

The Demoiselle crane, which has the second-largest population of cranes in the world, is one of the winter-migrating birds that gather every year in Khichan village, near Jodhpur in India. In winter, more than 30,000 birds travel from Siberia to Jodhpur and spend 3-4 months in the Khichan village. Local communities feed the birds, and several ponds around the village provide drinking water and resting places for the cranes. Increased anthropogenic activities involving urbanization and development have resulted in the habitat destruction of the birds, leading to a decline in the population and diversity of bird species. The use of water from the ponds by the villagers for domestic purposes and cattle bathing, ex-

Figure 1: (a) Sample images of the Demoiselle crane from the Khichan village. (b) Sample images of different species of birds obtained from Bharatpur bird sanctuary.

cessive grazing by livestock, encroachment of habitats by residents for cultivation and developmental activities, and solid waste dumped in the area adjoining the village pond all result in a significant invasion of the cranes' habitat. Additionally, illegal hunting and trade use of pesticides and herbicides pose a major threat to the conservation of the population of cranes. Even though Demoiselle cranes are listed as "Least Concern" under version 3.1 of the International Union for Conservation of Nature (IUCN) Red List Categories [for Conservation of Nature (IUCN), 2010] and are listed in Appendix II of CITES, their shrinking and degrading habitats are still a threat to the species. Developing AI tools for monitoring these species will also enable the study of other endangered species in the future.

Current techniques in AI-based wildlife monitoring focus on short-term tracking and monitoring of birds and wildlife, primarily on an individual basis. This project aims to make two-fold contributions: (i) improve the AI technology to detect, recognize and analyze long-term videos of flocks of birds or herds of animals and (ii) improve the understanding of the behavior and activities of different wildlife, which will be instrumental towards the broader goal of preserving and protecting wildlife and to understand better the complex and interrelated factors that impact the health and well-being of these important species.

Our work is part of a larger project on Video Analytics for Wildlife conservation and protection under the Indo-US collaboration. The project involves creating a dataset of longer videos to monitor birds non-intrusively, allowing for a comprehensive and diverse analysis of the birds' behavior. By working closely with the local experts who have been looking after these birds for over two decades at the Khichan sites, we intend to study the migration patterns [Galtbalt *et al.*, 2022], social interactions, and feeding habits, which will be invaluable for understanding their ecological and evolutionary processes. This information can be used to predict the future impact on the species and help develop informed conservation and mitigation policies.

Our current contribution as part of this research, as shown in Fig.1, is to propose a novel bird monitoring dataset totaling around 812 annotated high-resolution images with point annotations. The dataset also includes 131 video samples ranging from 30 seconds to 5 minutes, captured under a variety of conditions, including different resolutions and viewpoints, and featuring multiple birds in a single frame. State-of-the-art vision algorithms struggle to automatically analyze images and videos of flocks of birds (both, on the ground and in-flight) and accurately predict tracks of individuals or groups, identify interactions, and other behavioral characteristics. This dataset presents a new challenge for researchers to develop algorithms that can accurately analyze and interpret the behaviors of birds when a flock is being captured in a single image or video. Subsequently, these techniques help to gain a greater understanding of their population.

2 Target SDGs and LNOB Principle: Societal Advantages

The 2030 Agenda for Sustainable Development of the United Nations has set 17 interconnected goals to promote prosperity and equality through the involvement of all individuals and sectors of society. Further, its dedication to the principle of Leaving No One Behind (LNOB) ensures that all members of society benefit from sustainable growth. Wildlife conservation is a critical component for sustainable development, as loss of biodiversity and habitat can adversely impact ecosystems and have the potential to directly impact humans via environmental degradation or human-wildlife conflicts. Similarly, the following Sustainable Development Goals (SDGs) are directly related to the research collaboration:

- **Goal 11: Sustainable Cities and Communities**: This goal aspires to make cities and human settlements more inclusive, safe, resilient, and sustainable. Monitoring migratory birds can provide useful insights into the effects of urbanization on wildlife and their ecosystems. For instance, the behavior and orientation of migratory birds can provide insights into mitigating the effects of urbanization on wildlife.
- **Goal 13: Climate Action**: This objective seeks to take immediate action to mitigate climate change and its consequences. Monitoring migratory birds can provide insights into the impacts of climate change on wildlife and their ecosystems.
- **Goal 15: Life on Land**: It aims to maintain, restore, and promote the sustainable use of terrestrial ecosystems, manage forests sustainably, and stop and reverse land degradation and biodiversity loss. Accordingly, wildlife conservation and monitoring can aid in the development of effective conservation policies and programs.
- **Goal 17: Partnerships to achieve the Goal**: To accomplish the SDGs, it is crucial to encourage partnerships and collaboration among all stakeholders. As part of a broader Indo-US collaboration, the NSF-TIH initiative in line with *target 17.16* aims to develop novel approaches for monitoring and understanding the behavior of migrating birds. The project seeks to encourage global sustainable development by combining academic expertise with the knowledge and engagement of the local community.

The United Nations SDGs recognize the significance of wildlife protection in attaining sustainable development, and monitoring migrating birds is an essential component of this effort. By inventing non-invasive ways for monitoring wildlife, we can reduce the need for direct intervention and ensure the natural recovery of wildlife populations. Using computer vision techniques to monitor migratory birds exemplifies how innovation and technology may help achieve the SDGs. Furthering these aims will necessitate a collaborative effort from all stakeholders, including NGOs, academics, and industry. Migrating birds are critical to the biological balance of many ecosystems. Changes in migratory bird populations indicate broader ecological changes; their habitat loss or degradation can be mitigated by sustainable practices that need a concerted effort from all stakeholders. By better understanding these effects, researchers and conservationists can develop strategies and policies in accordance with SDG targets such as *targets 15.7* and *15.9* to mitigate the negative impact of these factors on wildlife and their habitats. Additionally, in accordance with *target 1.a* it can help to preserve natural resources that are critical for the livelihoods of people, particularly those living in poverty.

Our research adheres to the LNOB principle by identifying individuals left behind and proposing solutions to address root causes. Non-invasive monitoring is critical for understanding the impact on the vulnerable wildlife populations. This information can lead to creating equitable conservation policies in line with *target 13.2* and practices that consider the various needs and views of all communities. It is crucial to recognize that conservation initiatives can impact human communities living near wildlife habitats, and addressing disparities within and between communities is essential for wildlife conservation and monitoring.

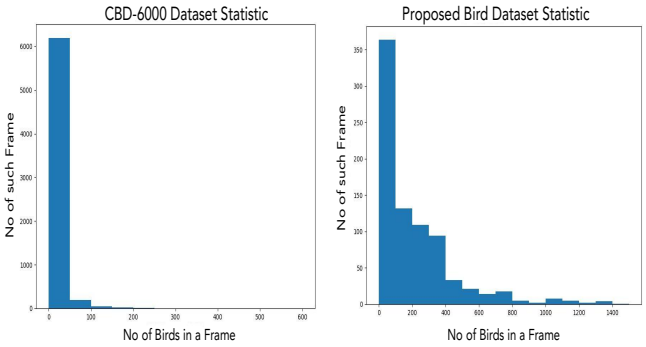

Figure 2: Plots depicting the distribution of the birds in the CBD-6000 and proposed birds dataset.

3 Proposed Birds Dataset

One of the significant contributions of our present research is the proposed dataset comprising images and videos of birds captured with different cameras and resolutions, providing a diverse range of data for training and evaluation. We collected the dataset from the village of Khichan through field trips from November to January with the assistance of Mr. Sevaram. The collected data consists of videos and images of birds in a range of settings, including on the ground and in flight, during the day and late evenings, with varying lighting situations. The details of the dataset are presented in Table 1. Traditional efforts to recognize and model birds in photos employ the use of datasets such as Caltech-UCSD Birds-200-2011 [Wah *et al.*, 2011] and NABirds V1 [Van Horn *et al.*, 2015]. These datasets primarily consist of images of a single bird species, which, while useful for research and development, are not entirely practical for real-world scenarios. The more recent CBD-6000 [Kim and Kim, 2020] dataset has around 6,477 images of seven different varieties of birds, with a maximum of 614 birds. After analyzing the distribution of the dataset in Fig.2, we can observe that in the case of the CBD-6000 dataset, the distribution of density of the bins is skewed towards the left, indicating that most of the images have a lower density of birds. The mean and median count are 9.18 and 2, respectively. On the other hand, the distribution of our dataset is more uniform, with a higher density of bins of up to 800. The mean and median count values, respectively, are 206.8 and 129.5.

In the proposed dataset, we have images of birds of varying resolutions, the highest being 6960 x 4640 and the lowest being 2400 x 1600, captured with SLR cameras as well as mobile cameras (including iPhone and Samsung-based smartphones) documenting and observing their behavior in their natural habitats. The annotations were performed with the VGG annotation tool (see Fig.3). In addition, we have approximately 100+ samples of video clips with a resolution of 1920 x 1080 at 30 frames per second, ranging from 30 sec-

Type	Resolution	No. of Samples
Image	6960 x 4640	**232**
Image	2400 x 1600	**445**
Image	6000 x 4000	**485**
Image	4032 x 3024	**109**
Image	4000 x 3000	**61**
Video(30fps)	1920x 1080	**102**
Video(25fps)	3840 x 2160	**29**

Table 1: Details of the proposed dataset consisting of images and videos of varying resolutions.

Figure 3: Point level annotations of the images using VGG Image Annotator.

onds to 5 minutes, as well as 29 video clips with a resolution of 3840 x 2160 at 25 frames per second. The diverse resolutions of the photos and movies will also allow for better image evaluation and comparison of the performance of the different sensors.

4 Experimental Framework and Baseline Techniques

We conducted experiments employing the ShanghaiTech B (SHB) [Zhang *et al.*, 2016] dataset and the proposed datasets to evaluate five benchmark approaches under different experimental settings. The proposed dataset has up to 1500 birds in a single frame, with varied lighting conditions and birds both on the ground and in flight. In the following sections, we will provide a comprehensive description of our dataset and evaluation procedure.

4.1 Dataset Protocol

Our experimental dataset includes two main parts: ShanghaiTech B and the proposed birds dataset. Part B of ShanghaiTech consists of 400 images for training and 316 for testing, with a total of 716 images taken on the streets of Shanghai. In addition, we created a curated dataset consisting of 304 annotated samples from the collected dataset for experimentation. This dataset has a 90-10 train-test split, which we maintain to ensure consistency across particular experimental settings.

4.2 Evaluation Protocol

We evaluated the performance of various state-of-the-art algorithms on the SHB dataset and the birds dataset using various evaluation frameworks. We first trained the models on the SHB dataset and tested them on the same dataset using the previously indicated split. We reported the results for the baseline models on the experiment's test set based on the Mean-Absolute Error (MAE) and Root Mean Squared Error (RMSE) (see Eq.1). In addition, we conducted experiments on the proposed dataset and benchmarked the findings using two protocols on the same dataset. For Protocol 1, we trained the model with 400 images from the SHB training dataset and evaluated it on the test split of the bird dataset. This approach allowed us to assess the model's ability to count birds based on the features learnt from trainin on the SHB dataset. For Protocol 2, we pre-trained the model on the SHB dataset, fine-tuned it on 90 percent of the birds dataset, and then evaluated its performance on the remaining 10 percent. This protocol allows the model to learn broad features that can be translated to bird counting during fine-tuning, allowing the model to learn more distinct relevant features.

4.3 Baseline Methods and Evaluation Metrics

Five models are used to benchmark the birds dataset. Each of the model is trained and tested on SHB dataset along with birds dataset using two protocols namely Protocol 1 and Protocol 2. For benchmarking, the following models are used.

- **Context-aware-crowd-counting (CAN)** [Liu *et al.*, 2019] is a deep learning architecture for crowd counting that encodes contextual information on an adaptive scale in order to accurately predict crowd density. It integrates information derived from numerous receptive field sizes and learns the significance of each feature at each image position, considering the possibility of rapid scale changes. In contrast with traditional crowd-counting methods, it combines multi-scale contextual information, allowing it to leverage the appropriate context at each image location.
- **DMNet**, or Distribution Matching for Crowd Counting [Wang *et al.*, 2020], is a method for crowd counting that employs Optimal Transport (OT) to quantify the similarity between predicted and ground truth density maps without the requirement for Gaussian smoothing. It avoids the use of Gaussian smoothing in annotations, which is known to degrade generalization performance, and instead use total variation loss in the model to increase OT computation stability.
- **CSRNet** or Dilated Convolutional Neural Networks for Understanding the Highly Congested Scenes [Li *et al.*, 2018] is intended to estimate crowd counts accurately and provide high-quality density maps for highly congested environments. It employs a convolutional neural network for feature extraction and a dilated CNN for the back-end, incorporating dilated kernels to increase reception fields and replace pooling procedures. The model focuses on developing a density map generator with pure convolutional layers as its foundation, allow-

ing for variable input image resolutions and collecting high-level features with broader receptive fields.

- **SGANet** [Wang and Breckon, 2022] employs Inception-v3 as its backbone and a segmentation map-guided attention layer to enhance the extraction of features for reliable density map estimation. It incorporates a novel curriculum loss technique that tackles the challenges created by highly dense regions in crowd counting. The curriculum learning technique is built on the image level and is distinguished by a novel curriculum loss function that takes into account the pixel-wise difficulty level based on a dynamic threshold when computing the density map loss.
- **CrowdFormer** or Weakly-supervised Crowd counting with Improved Generalizability [Savner and Kanhangad, 2023] is a weakly-supervised crowd counting method that extracts multi-scale characteristics with global context using a pyramid vision transformer. It combines features using an effective feature aggregation module and a simple regression head to estimate the crowd size. The suggested approach intends to increase generalizability in tasks involving crowd counting.

Evaluation Metrics: For evaluation purposes, we use the standard *MAE* (Mean Absolute Error) and *RMSE* (Root Mean Squared Error).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|\,,\ RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|^2} \tag{1}$$

The number of test images is represented by N. The actual number of individuals within the region of interest (ROI) in the ith image is denoted by x_i, and the estimated number of individuals in the ROI of the ith image is represented by $\hat{x}_i$.

5 Experimental Results

We perform experiments for various vision tasks using state-of-the-art methods for tasks such as detection and segmentation on the proposed dataset. We employ Mask R-CNN [He *et al.*, 2017] for instance segmentation, UNet [Ronneberger *et al.*, 2015] for semantic segmentation, DETR [Carion *et al.*, 2020] for both object detection and with ResNet-101 backbone for panoptic segmentation [Carion *et al.*, 2020]. We can observe from Fig. 5 that the aforementioned methods fail to perform their tasks on the dataset. In order to detect the birds, we have used the Megadetector toolkit [Beery *et al.*, 2019] that aims to detect the animals and people along with the vehicles in the camera trap images. The toolkit is trained on a diverse dataset, including the Caltech camera traps, Snapshot Serengeti, iNaturalist Dataset 2017, and many other datasets, which allowed for improved performance on publicly available camera trap datasets. However, the results on the manually collected dataset (see Fig.4) show that the model is not able to detect all of the birds in the images. Further, it also has difficulty in consistently detecting the birds in video files. The obtained results necessitate the development of computer vision techniques that are capable of detecting, localizing, and segmenting in the context of wildlife datasets,

Models	Backbone	SHB		Protocol-1		Protocol-2	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
CAN	CNN	11.16	16.10	110.18	162.49	10.10	16.46
CSRNet	CNN	8.67	14.09	72.72	116.60	7.20	10.13
SGANet	CNN	6.83	11.15	74.85	114.70	5.78	7.96
DMCount	CNN	8.76	16.24	106.97	164.23	7.37	13.13
CrowdFormer	Transformer	8.11	12.46	77.82	138.69	6.86	19.36

Table 2: Results of the different baseline models on the SHB dataset and birds dataset (Protocol-1 and Protocol-2). Protocol-1 entails training with the SHB dataset and testing on the proposed birds dataset. Protocol 2 includes training on the SHB dataset followed by fine-tuning on the birds dataset before testing it.

which are particularly difficult due to variable camera perspectives and crowded scenes. The results for the crowd-counting benchmarking experiments are provided in Table 2. We observe that Protocol-2, which involved fine-tuning a model on bird dataset after pre-training on the SHB dataset, resulted in a considerable performance gain over the former protocol. Protocol-1 consisted of testing the model solely on features learned from the SHB dataset, which is employed for human crowd counting. This increase in performance can be attributed to the fact that pre-training the model allows the model to learn relevant features that are then refined upon fine-tuning to improve results on the specific task of bird counting.

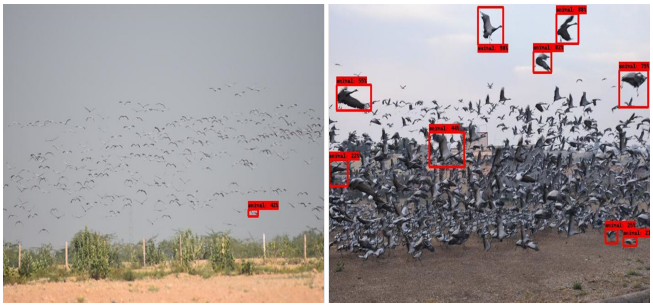

Figure 4: Sample predictions based on the use of the megadetector to detect the birds in the frames. Although the model was trained using a large and varied collection of camera traps and private datasets, it still struggled to detect all birds accurately.

6 Goals and Expected Results

In the context of wildlife monitoring and conservation, our project's objective is to create an annotated dataset for event comprehension in very long videos. This dataset will be the first in computer vision to consist of videos significantly longer than the standard 10-minute-long video datasets in vision. This will allow for the creation of novel computer vision algorithms for wildlife monitoring and conservation. In addition, there is potential for technology transfer in creating a new type of symbolic video camera trap that automatically creates labeled video data at the camera itself without requiring substantial external storage or delayed processing. This could also help provide more efficient and cost-effective methods of data collection and processing. The technical challenges of monitoring wildlife through continuous video include a wide variety of animals and movements,

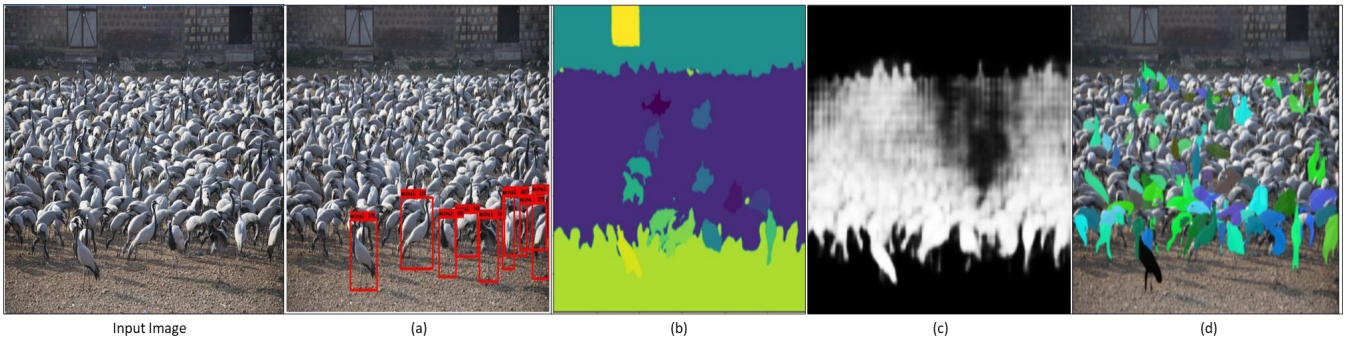

Figure 5: Results of the various vision tasks performed by the state of art methods. (a) DETR fails to perform object detection (b) DETR along with ResNet-101 backbone fails to perform pixel-level instance segmentation (c) UNet fails to perform pixel-level semantic segmentation (d) MaskRCNN fails to perform instance segmentation

extreme illumination, and weather variability in outdoor settings. To address these issues, the project will capture continuous footage of migratory bird behavior and annotate the spatial outlines of birds, other animals, and objects and their relationships in the scene. Furthermore, the video will also include temporal labels of observed activity at multiple scales.

A baseline algorithm will be built to segment and label long videos in a streaming manner, facilitating the analysis and comprehension of the events captured by the videos. These algorithms will, in turn, allow the development of online and continual learning approaches, leading to further improvements in tools for wildlife monitoring. The significance of this study rests in the potential for automated, non-invasive, long-term monitoring of wildlife.

The project also aims to create novel, unsupervised learning approaches grounded in cognitive science to recover spatio-temporal events in long videos. These spatio-temporal events may be atypical or anomalous interactions between birds recovered automatically without human intervention. State-of-the-art neural architectures for temporal segmentation, attention mapping, and spatial localization algorithms combined with cognitive science inspired design of self-supervised losses has the potential to automate this task without the need for arduous annotations on long videos necessary for training models. In addition to the Khichan dataset, we also intend to gather additional data from the Bharatpur bird sanctuary, which is a UNESCO world heritage site and is an ideal place for monitoring birds in their natural habitats. We plan to fine-tune the model using the pre-trained model from the Khichan challenging crane dataset on the birds dataset from the bird sanctuary. Beyond animal monitoring and conservation, the production of this annotated collection and the development of new algorithms have far-reaching ramifications. The techniques and technologies developed through this project can be used for surveillance and security where long-term video monitoring is required.

7 Long-term Impact on the SDGs

The wildlife bird conservation initiative has a direct and beneficial effect on a number of Sustainable Development Goals (SDGs). The project's main objective of developing techniques for wildlife conservation will enable sustainable ecotourism prospects in those regions, resulting in long-term poverty alleviation. Furthermore, by giving economic opportunities to local populations, the project can help reduce poverty and improve livelihoods in accordance with UN SDG 1. In addition, it will aid in the creation of sustainable urban development policies that encourage the coexistence of humans and wildlife, which are relevant to SDG 11. Monitoring migrating birds can provide insights into the impact of climate change on wildlife and their ecosystem, thereby assisting in the development of climate change mitigation and adaptation measures consistent with SDG 13. By promoting sustainable land use, the initiative can contribute to protecting terrestrial ecosystems, which is an intrinsic aspect of Objective 15. Additionally, partnerships and engagement between stakeholders enhance SDG 17 over the long run. Using frameworks like active learning [Norouzzadeh *et al.*, 2021] and self-supervised techniques [Pantazis *et al.*, 2021] to leverage the vast amount of available unlabeled data for predicting the behavior of birds is a long-term research objective. Computer vision algorithms inspired by cognitive science theories [Mounir *et al.*, 2022; Mounir *et al.*, 2023] can augment existing self-supervised and active learning methods for data-efficient learning from visual imagery. The ability to process long videos can aid in monitoring wildlife behavior and identifying major occurrences spanning multiple days and compile them as short clips highlighting the occurrences for further analyses by conservation biologists for assessing behavioral changes over longer periods as well as emerging trends in affected ecosystems. Overall, the project's success will have significant implications for both wildlife research and technological innovation in the field of computer vision.

8 Challenges and the Risk Associated

The nature of the data being collected poses several challenges. Collecting uncontrolled, long-term, outdoor videos of large flocks of birds demands robust algorithms that can deal with illumination variations, occlusions, and non-rigid shape deformation of individual birds, among other algorithmic challenges. Furthermore, annotating long videos is a de-

manding and expensive task, even with crowdsourced support. On the one hand, annotation of low-level tasks in videos is tedious but would be required for validation of the performance of the vision algorithms if not for training the algorithms. On the other hand, high-level annotation of *events of interest* would require experts on the species being monitored. We plan to deal with the first challenge of large-scale annotation by leveraging AI-augmented tools with crowdsourced annotations. For the latter, we plan to engage with local ornithologists as well as conservation biologists from organizations like the Wildlife Conservation Society (WCS).

The project relies on substantial data collection of birds for sufficiently long temporal periods that would capture events of interest. The current sites in Khichan, India, are well-suited for the task. However, there are risks associated with the data collection process, including possible damage or theft of camera devices, withdrawal of local support for maintaining data collection infrastructure, or in the worst case, the crane population abandoning the sites. We have plans for securing the data collection infrastructure as well as assurance from the locals for this study. For the latter two, we have also identified alternate sites, e.g., Bharatpur Bird Sanctuary, where we can collect similar quality and quantity of data.

9 Mechanism for Collaboration and Proposed Roadmap

The research project comprises multiple stakeholders, including local domain experts, AI experts from India and the United States, and researchers from IIT Jodhpur, IIIT-Delhi, and the University of South Florida (USF). The overall objective of the project is to leverage the advancements in AI to monitor birds in the wild and to develop tools that can aid in the understanding of their behavior, consequently assisting in developing strategies aimed at the conservation of biodiversity. The technological expertise of the USF team in building streaming video event segmentation and description is critical, as this technology is essential for monitoring events that involve large amounts of data (images/videos). IIIT-Delhi has collaborated with ecologists in the past to study snow petrels in Antarctica. In addition, IIT Jodhpur will assist in coordinating with the local population and help in the long-term monitoring of birds in the Khichan research region. It will also aid in developing camera trap technologies that, at the level of the camera itself, facilitate the creation of weakly labeled video datasets without considerable preprocessing.

The project comprises multiple phases and will last for two calendar years. The first step includes developing a diverse annotated collection of birds that will be released to the scientific community to help accelerate research on these organisms. The second phase focuses on developing computer vision algorithms for understanding events in wildlife monitoring images. This includes implementing various vision tasks such as crowd-counting, semantic segmentation, instance segmentation, and video summarization. These tasks will provide cues to analyze the behavior of wildlife, including different poses like heads-up and heads-down. Te final phase entails acquiring results, extracting insights about the birds' relationship with their habitat, and investigating the effects of climate change and its implications. To ensure successful collaboration across teams with diverse backgrounds, the project team will undertake regular meetings and discussion sessions with other local domain experts. The team plans to publish its findings and experiment results in reputable conferences and journals to benefit the research community.

To that end, a survey paper will be published detailing the status of efforts carried out concerning bird monitoring in the wild. The paper will also describe the existing datasets and the one manually created by the team, which will be made publicly available for researchers. A conference paper incorporating the development of novel computer vision techniques/algorithms for the event understanding of birds and their analysis and impact on their natural habitat, will be submitted. The resulting source code and findings of the work will be documented and open-sourced for the reproducibility of reported experiments. An online repository of the resources discovered by the team throughout their study, including conversations with topic experts, will be developed and hosted on public platforms such as the project website.

Through the research collaboration, we hope to leverage worldwide datasets and expertise, thereby facilitating the establishment of scientific standards and improving international cooperation through exchanges and co-development efforts. In addition to technical cooperation, this effort includes a partnership with locals who care for the migrating birds that visit Khichan annually. Mr. Sevaram, who has been instrumental in the care of these birds, will also work with the team to help conduct in-depth research on the nature and behavior of migratory birds. By working with all the stakeholders, we intend to contribute to research on monitoring bird behavior, migration patterns, and the development of new monitoring equipment, as well as create international cooperation and engage local communities.

A Curricula Vitae

Kshitiz
Indian Institute of Technology, Jodhpur
Kshitiz, is a final-year Computer Science undergraduate student at IIT Jodhpur. His areas of interest include machine learning, deep learning, and computer vision.

Sonu Shreshtha
Indian Institute of Technology, Jodhpur
Sonu Shreshtha is a 2nd-year Masters student pursuing specialization in Artificial Intelligence at the IIT Jodhpur. With over 7 years of industry experience, he specializes in computer vision, deep learning, image and video processing, and multimedia.

Ramy Mounir
University of South Florida, Tampa
Ramy Mounir is a Ph.D. candidate in the Computer Science and Engineering department at the University of South Florida (USF). He received his B.Eng and M.S. degrees in Mechanical Engineering from USF in 2015 and 2018, respectively. He graduated Summa Cum Laude and received the Outstanding Graduate Award in 2015. He is the recipient of outstanding reviewer awards from top-tier conferences (i.e.,

ECCV and ICLR) and the Early Innovation Award from Intel Corporation. His research interests include self-supervised learning of hierarchical representations of objects and events, and implementing perceptual and cognitive theories using computational deep learning methods and predictive models.

Mayank Vatsa
Indian Institute of Technology, Jodhpur
Mayank Vatsa, currently serves as a Professor and Dean (R&D) at IIT Jodhpur, India. He holds Masters and PhD in Computer Science from the West Virginia University, USA. He is a SwarnaJayanti Fellow, Fellow of IEEE and IAPR, and have published over 350 research papers. He has also been actively involved in the development of biometrics standards in India, serving as a member of the Indian Biometrics Standards Committee for e-gov applications, and the UIDAI's Biometrics Standard Subcommittee. Mayank is also a member of several international honor societies, and has held leadership positions in professional organizations such as Vice President (Publications) for IEEE Biometrics Council and Area Editor of Information Fusion (Elsevier). His research interests are focused on biometrics, computer vision, machine learning, deep learning, and image forensics.

Richa Singh
Indian Institute of Technology, Jodhpur
Richa Singh is a Professor and Head at the Department of CSE, IIT Jodhpur. She holds Ph.D. degree in computer science from West Virginia University in 2008. She is a Fellow of IEEE and IAPR, and has received several awards including the Kusum and Mohandas Pai Faculty Research Fellowship at the IIIT-Delhi, the FAST Award by the Department of Science and Technology, India, and several best paper and best poster awards in international conferences. She has also held leadership positions in professional organizations, such as Program Co-Chair of CVPR2022, General Chair of FG2021, and Vice President - Publications of the IEEE Biometrics Council, and an Associate Editor-in-Chief of Pattern Recognition.

Saket Anand
Indraprastha Institute of Information Technology, Delhi
Saket Anand is an Associate Professor at IIIT-Delhi. He completed his PhD in Electrical and Computer Engineering from Rutgers University in 2013. With over ten years of research experience in computer vision and machine learning, his research has been applied to various domains including wildlife conservation and road safety. His contributions to AI-based species segregation of camera trap images have been used by the Wildlife Institute of India (WII) for the All India Tiger Estimation in 2018-19 and 2022-23. Among other professional duties, he has served as Program Co-Chair for IEEE WACV 2022, Area Chair for IEEE WACV 2023 and and Associate Editor for the Pattern Recognition Journal. He is a member of the IEEE and has teaching experience in courses such as machine learning, computer vision, and deep learning.

Sudeep Sarkar
University of South Florida, Tampa
Sudeep Sarkar is a Distinguished University Professor, Chair of Computer Science and Engineering at the University of South Florida, Tampa, and Co-Director of the USF Institute for Artificial Intelligence + X. He holds an M.S. and Ph.D. in electrical engineering from The Ohio State University and B. Tech from the Indian Institute of Technology, Kanpur. With 35 years of experience, he has conducted and directed fundamental research in computer vision, predictive learning, biometrics, and artificial intelligence. His contributions include systems that recognize persons from gait biometrics, automated recognition of actions, activities, and events in a video, economic activity from satellite images, and extracting precise, medically relevant information from medical images. He has directed 22 Doctoral and 25 Master's students on these topics. He is a co-Editor-in-Chief of Pattern Recognition Letters, former President of the IEEE Biometrics Council, and a Fellow of several organizations, including the National Academy of Inventors (NAI), American Association for the Advancement of Science (AAAS), and the IEEE. He is also a member of the Academy of Science, Engineering, and Medicine of Florida.

Sevaram Mali Parihar
Crane Conservationist
Sevaram Mali Parihar, widely known as the Birdman of Khichan, has devoted most of his life to the conservation of demoiselle cranes. Despite leaving school in eighth grade, his tenacity and spirit have enabled him to become an exceptional conservationist. He has demonstrated remarkable courage and perseverance by disregarding the countless threats thrown at him in order to fight for the rights of the birds, including a court case to remove the high-voltage electrical poles, conserving their habitat. His selfless service has earned him various accolades, including the Sanctuary RBS Wildlife Service Award (2008) and the Green Warrior award (2018).

Ethical Statement

The project involves no potential harm to the birds or human entities involved in the study. The birds selection is strongly impacted by their status as Least Concern on the IUCN Red List. This assures that the study has no significant impact on the population. Furthermore, the data collection approaches used in the study are non-invasive and involve no physical treatment of the birds, such as tagging or radio collaring, avoiding any risk to the birds' well-being, such as disruption of their natural behavior and habitat. The dataset shall be devoid of any identifiable human entities, so respecting individuals' privacy and ensuring that the outcomes will not have a detrimental influence on any specific person or entity.

Acknowledgments

We thank Hemang Dahiya, Arsh Gupta, Neelabh Kumar Srivastava (IIIT-Delhi), and Ahmed Shahabaz (USF) for their assistance with point annotation for the proposed bird dataset, and the Bombay Natural History Society for the discussions. This research was supported by the US NSF grant IIS 1956050 and iHub Drishti, the TIH on CV, AR and VR. Saket Anand was partly supported by the Infosys Center for AI, IIIT-Delhi.

References

[Beery *et al.*, 2019] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*, 2019.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020*, pages 213–229, 2020.

[Chen *et al.*, 2011] I-Ching Chen, Jane K Hill, Ralf Ohlemüller, David B Roy, and Chris D Thomas. Rapid range shifts of species associated with high levels of climate warming. *Science*, 333(6045):1024–1026, 2011.

[Delany, 2010] Simon Delany. Guidance on waterbird monitoring methodology: field protocol for waterbird counting. *Wetlands international*, 25:4–5, 2010.

[for Conservation of Nature (IUCN), 2010] International Union for Conservation of Nature (IUCN). Iucn red list of threatened species. version 2010.2, 2010.

[Galtbalt *et al.*, 2022] Batbayar Galtbalt, Nyambayar Batbayar, Tuvshintugs Sukhbaatar, Bernd Vorneweg, Georg Heine, Uschi Müller, Martin Wikelski, and Marcel Klaassen. Differences in on-ground and aloft conditions explain seasonally different migration paths in demoiselle crane. *Movement Ecology*, 10(1):1–11, 2022.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Higuchi, 2012] Hiroyoshi Higuchi. Bird migration and the conservation of the global environment. *Journal of Ornithology*, 153(Suppl 1):3–14, 2012.

[Kim and Kim, 2020] Saehun Kim and Munchurl Kim. Learning of counting crowded birds of various scales via novel density activation maps. *IEEE Access*, 8:155296–155305, 2020.

[Li *et al.*, 2018] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[Liu *et al.*, 2019] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5099–5108, 2019.

[Mounir *et al.*, 2022] Ramy Mounir, Sathyanarayanan Aakur, and Sudeep Sarkar. Self-supervised temporal event segmentation inspired by cognitive theories. In *Advanced Methods and Deep Learning in Computer Vision*, pages 405–448. Elsevier, 2022.

[Mounir *et al.*, 2023] Ramy Mounir, Ahmed Shahabaz, Roman Gula, Jörn Theuerkauf, and Sudeep Sarkar. Towards automated ethogramming: Cognitively-inspired event segmentation for streaming wildlife video monitoring. *International journal of computer vision*, pages 1–31, 2023.

[Nichols and Williams, 2006] James D Nichols and Byron K Williams. Monitoring for conservation. *Trends in ecology & evolution*, 21(12):668–673, 2006.

[Norouzzadeh *et al.*, 2021] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *Methods in ecology and evolution*, 12(1):150–161, 2021.

[Pantazis *et al.*, 2021] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 10583–10592, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015*, pages 234–241, 2015.

[Savner and Kanhangad, 2023] Siddharth Singh Savner and Vivek Kanhangad. Crowdformer: Weakly-supervised crowd counting with improved generalizability. *Journal of Visual Communication and Image Representation*, page 103853, 2023.

[Şekercioğlu *et al.*, 2004] Çağan H Şekercioğlu, Gretchen C Daily, and Paul R Ehrlich. Ecosystem consequences of bird declines. *Proceedings of the National Academy of Sciences*, 101(52):18042–18047, 2004.

[Van Horn *et al.*, 2015] Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 595–604, 2015.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[Wang and Breckon, 2022] Qian Wang and Toby P Breckon. Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15233–15243, 2022.

[Wang *et al.*, 2020] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33:1595–1607, 2020.

[Zakaria *et al.*, 2005] Mohamed Zakaria, Puan Chong Leong, and Muhammad Ezhar Yusuf. Comparison of species composition in three forest types: Towards using bird as indicator of forest ecosystem health. *Journal of biological sciences*, 5(6):734–737, 2005.

[Zhang *et al.*, 2016] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.