BirdCollect: A Comprehensive Benchmark for Analyzing Dense Bird Flock Attributes

Kshitiz .[1], Sonu Shreshtha[1], Bikash Dutta[1], Muskan Dosi[1], Mayank Vatsa[1], Richa Singh[1], Saket Anand[2], Sudeep Sarkar[3], Sevaram Mali Parihar[4]

[1]Indian Institute of Technology, Jodhpur, India, [2]Indraprastha Institute of Information Technology Delhi, India, [3]University of South Florida, Tampa, Florida, USA [4]Crane Conservationist, Khichan, India
{kshitiz.1, shreshtha.1, d22cs051, dosi.1, mvatsa, richa}@iitj.ac.in, anands@iiitd.ac.in, sarkar@usf.edu, smaliparihar@gmail.com

Abstract

Automatic recognition of bird behavior from long-term, uncontrolled outdoor imagery can contribute to conservation efforts by enabling large-scale monitoring of bird populations. Current techniques in AI-based wildlife monitoring have focused on short-term tracking and monitoring birds individually rather than in species-rich flocks. We present *BirdCollect*, a comprehensive benchmark dataset for monitoring dense bird flock attributes. It includes a unique collection of more than 6,000 high-resolution images of Demoiselle Cranes (Anthropoides virgo) feeding and nesting in the vicinity of Khichan region of Rajasthan. Particularly, each image contains an average of 190 individual birds, illustrating the complex dynamics of densely populated bird flocks on a scale that has not previously been studied. In addition, a total of 433 distinct pictures captured at Keoladeo National Park, Bharatpur provide a comprehensive representation of 34 distinct bird species belonging to various taxonomic groups. These images offer details into the diversity and the behaviour of birds in vital natural ecosystem along the migratory flyways. Additionally, we provide a set of 2,500 point-annotated samples which serve as ground truth for benchmarking various computer vision tasks like crowd counting, density estimation, segmentation, and species classification. The benchmark performance for these tasks highlight the need for tailored approaches for specific wildlife applications, which include varied conditions including views, illumination, and resolutions. With around *46.2 GBs* in size encompassing data collected from two distinct nesting ground sets, it is the largest birds dataset containing detailed annotations, showcasing a substantial leap in bird research possibilities. The database is available at: https://iab-rubric.org/resources/wildlife-dataset/birdcollect

Introduction

Birds are vital components of our ecosystems worldwide, playing critical roles in pollination, pest control, seed dispersal, and other ecological processes. However, both migratory and non-migratory bird populations face escalating threats from anthropogenic pressures like habitat loss[1], climate change (Li, Liu, and Zhu 2022), and overexploitation. Habitat degradation and destruction are major threats to avian biodiversity globally. Consequently, North America has lost nearly 3 billion birds since 1970 (Rosenberg et al. 2019), with grassland-dwelling and aerial insectivores amongst the most affected. Globally, 1 in 8 bird species is now threatened with extinction[2]. Urgent conservation action is imperative to reverse these declines and prevent irreparable damage to ecosystems. Long-distance migratory birds that traverse continents are especially vulnerable, as they depend on networked habitats along their flyways for nesting and wintering. For instance, collision with man-made structures like power lines and wind turbines results in millions of migratory bird mortalities annually.

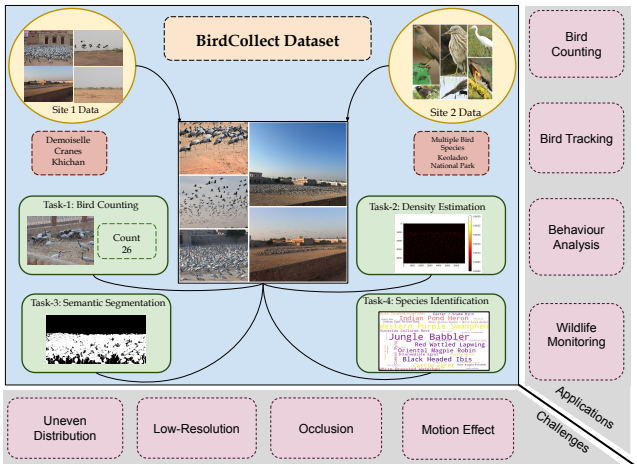

Figure 1: Visual description outlining the potential applications and challenges associated with our proposed dataset *BirdCollect*.

Quantitative long-term monitoring is essential to track climate change indicators like shifts in arrival time, nesting locations, and migration routes. Targeted conservation planning hinges on high-quality data on avian distribution, abundances, and behavior. AI and vision techniques through tasks like crowd-counting, density estimation, segmentation and species classification can automate bird monitoring from imagery. However, lack of labeled data for migratory birds limit the research progress in this field.

Our research endeavors to leverage advanced vision techniques for an enriched analysis of bird behavior. We present

[1]https://www.3billionbirds.org/

[2]https://www.stateofthebirds.org/2022/

Datasets	Size	Source		Annotation Type			Density	Diverse Conditions
		Curation	Annotation	Point Annotation	Segmentation Mask	Species Classes		
CBD-6000	$\sim$ 800 MB	Internet	Manual	✓	✗	✓	Low	✓
Penguin Dataset	28GB	Site Collection	Crowd Sourced	✓	✗	✓	High	✓
BirdSnap	-	Internet	Crowd Sourced	✗	✗	✓	Low	✗
NA Birds	512MB	Internet	Crowd Sourced	✗	✗	✓	Low	✗
BirdCollect (Ours)	**$\sim$ 46 GB**	**Site Collection**	**Manual**	✓	✓	✓	**Very High**	✓

Table 1: Comparison of *BirdCollect* with existing public birds dataset. Here, diverse conditions refers to the varying conditions of lighting, viewpoints and aerial/ground conditions. Further, the average number of birds per image is considerably larger than in any other public dataset, depicting true monitoring conditions.

a comprehensive and one of the largest bird datasets, named ***BirdCollect***, compiled from two prominent locales: Khichan village in Rajasthan and Keoladeo National Park (hereafter referred to as Bharatpur) in Rajasthan, India. In collaboration with Indian wildlife experts, we focus on Demoiselle cranes that gather in Rajasthan, India during winter months. The dataset encompasses high-resolution annotated images and videos, capturing crane flocks across diverse scenarios. Furthermore, a distinct subset within our dataset features high-quality images of 34 bird species, aiding species identification for protective measures. We further conduct extensive experiments to benchmark the performance of cutting-edge models on the proposed dataset, demonstrating its intricacies across multiple computer vision tasks.

Relevant Literature and Birds Datasets

Existing Datasets: The existing bird monitoring datasets predominantly contain images with low bird density and fewer birds per image (see Table 1). Among the publicly available datasets, most of the annotations are crowd-sourced, which can have unreliable annotations especially with high densities. In contrast, our dataset has manual supervised annotations. Other datasets like CBD-6000 (Kim and Kim 2020), NA birds [3], and BirdSnap (Berg et al. 2014) also comprise lower density internet images lacking real-world diversity. The Penguin dataset covers varying conditions with penguins. However, the dataset contains species information and labels for three different classes of penguins.

Literature: Recent advances in crowd counting employ density map regression or localization approaches. Density map based techniques (Bai et al. 2020) predict maps using feature extraction and regression head but rely substantially on human point annotations. Localization methods (Sam et al. 2020) on the other hand forecast individual locations but struggle with duplicate detections and noise. To mitigate the reliance on dense annotations, weakly-supervised transformer architectures such as (Liang et al. 2022) have gained traction by requiring fewer labels. Semantic segmentation has primarily utilized Convolutional Neural Networks (CNN) (Long, Shelhamer, and Darrell 2015). Transformers (Strudel et al. 2021) leverage context modules and self-attention to enhance per-pixel accuracy. Universal architectures like DETR (Carion et al. 2020) and MaskFormer (Cheng, Schwing, and Kirillov 2021) have inspired works (Cheng et al. 2022; Jain et al. 2023) that enable adaptability across segmentation tasks. Fine-grained recognition relies on object-part and attention-based methods. Object-part approaches (Zheng et al. 2017) extract features from discriminative regions. Attention techniques including (Bera et al. 2022) enhance features and localization via attention. Self-attention mechanisms, as used in models (Sun, He, and Peng 2022) enhance feature representation, while weakly supervised methods (Zhang et al. 2019) are able to identify and fuse parts using only image labels. With the emergence of denoising diffusion probabilistic models (DDPM) (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021), diffusion models have been applied to object detection (Chen et al. 2022) and segmentation (Gu et al. 2022).

[3] https://dl.allaboutbirds.org/nabirds

Benchmark Results and Analysis

As shown in the review section, there are limited datasets that provide an opportunity for analysing the activities of birds in high-density settings. Therefore, in this research, we present the *BirdCollect* dataset with the objective to prepare an annotated benchmark dataset for promoting design and development of algorithms for long term ethograming of birds. This is one of the largest datasets with detailed annotations available in the research community for bird monitoring and analysis. The proposed dataset aims to address the following key research questions:

RQ1: How can we accurately count individual birds in highly dense flocks?

RQ2: How can density be estimated robustly across diverse conditions with occlusion and variability?

RQ3: What techniques can effectively segment clustered, partially occluded birds?

RQ4: How can rare species be effectively classified from imbalanced datasets and limited training samples?

Figure 2: (a) Sample images of the Demoiselle crane from the Khichan village. (b) Sample images of different species of birds obtained from Bharatpur.

Dataset Collection

The proposed *BirdCollect* is a large-scale dataset for analyzing and tracking dense avian flocks utilizing computer vision techniques including crowd counting, semantic segmentation, and species classification. It comprises of images collected from two distinct nesting grounds associated with the conservational efforts. Site-1 is the Khichan village, an important wintering habitat for migratory birds like Demoiselle Cranes shows large bird densities in a migratory staging area. The images from Khichan village, renowned for hosting migratory birds from Central Asia, contain dense flocks of Demoiselle Cranes *(K, S1)*. While Site-2 is in Bharatpur, Rajasthan a UNESCO World Heritage Site (Arya and Syriac 2018) known for its rich diversity of resident and migratory bird species. Bharatpur provides variability in birds and environmental conditions like water, trees, and nature. The dataset curated from site-2, renowned for its large congregation of non-migratory resident breeding birds, includes 34 avian species *(B, S2)*. By collecting data from both sites, we aimed to analyze and compare the presence and diversity of birds across different habitats. The dataset comprises 6,986 high-resolution multi-scale images captured using multiple cameras across the two locations of Khichan and Bharatpur.

Data collection spanned from November to March to capture migratory species arriving in winter to their natural habitats, enabling detailed environmental study. This manual image capturing spanned over several months at the prime seasonal period for avian populations across both sites. The images were collected using multiple mobile camera (iPhones and Samsung) and DSLR cameras to capture varying resolutions. This enables evaluating image quality and model performance across sensors. Data from group *(K, S1)* was captured using both mobile and DSLR cameras, while data from group *(B, S2)* was obtained using DSLR cameras. Despite different original formats, all images were converted to standard JPEG to ensure consistency for public release of the dataset. Drone imagery was avoided due to non-invasive monitoring concerns, as demoiselle cranes are sensitive to their environment and could be frightened or injured. Moreover, gathering data for *(B, S2)* posed challenges due to factors like camouflage, low lighting, frequent motion, and obstructions caused by dense forests.

Dataset Statistics: Table 2 summarizes the characteristics of data gathered from these two distinct sites. For the Demoiselle Crane dataset *(K, S1)*, the labeled samples comprise 2,163 total images divided into training and test sets in a 70-30 ratio. The 1,473 image training set has a mean of 191 and median of 82 birds per image. Furthermore, the 690 test images have a mean of 188 birds and median of 69.5 per image. Similarly, the multi-species bird dataset *(B, S2)* consists of 433 images with 34 different species captured in their natural habitats.

Similar to existing crowd counting datasets, the Demoiselle crane dataset *(K, S1)* provides a challenging benchmark tailored to wildlife scenarios, specifically avian populations characterized by significant density and distribution variations and frequent occlusion in each image. This enables more robust analysis and benchmarking while accounting for key challenges in wildlife data that diverge from other domains based on viewpoint diversity, highly variable crowd counts, scale ranges, and image variety. As evident in Figure 3, the data exhibits a long-tailed distribution, which is prevalent in natural wildlife imagery. The differences in the inherent populations of bird species, along with varying ease of capture, lead to uneven image distribution across categories. As mentioned earlier, this data imbalance poses challenges for fine-grained classification within the constrained data context.

Annotation Process

The annotation process for our proposed dataset is categorized into three distinct computer vision tasks: Crowd counting and Density estimation, Semantic Segmentation and Species classification. As shown in Table 2, a total of 2596 images were annotated.

Crowd Counting: Quantifying the bird count within an image contributes to effective bird monitoring. Point

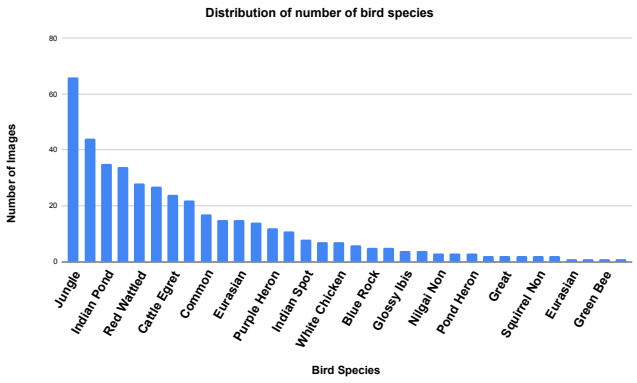

Figure 3: Distribution of bird species in the Bharatpur dataset *(B, S2)* closely resembles a long-tailed distribution.

Site Location	Camera Sensor	Resolution	# Samples
Site-1	Mobile Sensor	1920 x 1080	220
Site-1	Mobile Sensor	2400 x 1600	445
Site-1	Mobile Sensor	4000 x 3000	61
Site-1	Mobile Sensor	4032 x 3024	109
Site-1	DSLR	6000 x 4000	484
Site-1	DSLR	6960 x 4640	2907
Site-1	Mobile Sensor	3840 x 2160	2327
Site-2	DSLR - Bharatpur	5472 x 3648	433
# Total Samples			6986

Table 2: Details of the number of images of birds collected from two distinct sites, hereafter referred to as Site-1, Khichan village *(K, S1)* and Site-2, Bharatpur *(B, S1)*.

Site Location	Camera Sensor	Resolution	# Samples
Site-1	Mobile sensor	1920 x 1080	220
Site-1	Mobile Sensor	2400 x 1600	317
Site-1	Mobile Sensor	4000 x 3000	28
Site-1	DSLR	6960 x 4640	1598
Site-2	DSLR - Bharatpur	5472 x 3648	433
# Total Samples			2596

Table 3: Details of annotated samples corresponding to different resolution and camera sensors.

based annotations are used to localize a bird in perception, using https://www.robots.ox.ac.uk/~vgg/software/via/via_demo.html **Oxford VGG Image Annotator** tool. Unlike most avian datasets that are curated via internet and annotated using crowd-sourcing [4], our dataset has been annotated using manual supervision particularly because the scale and density of dataset makes it unreliable to trust the crowdsourced annotations. Counting in crowded scenes is challenging, particularly because they rely on large amounts of data annotation to achieve high performance thereby impeding the development of accurate models due to the accompanying cost and time constraints associated with the annotations. Further, annotating the wildlife bird datasets is even more challenging due to the natural variation including significant changes in scene illumination, strong object and location correlations, and diversity in bird species.

Semantic Segmentation of individual birds facilitates behavioral analysis and tracking. The complexity of labeling bird flocks encourages the adoption of advanced segmentation models, like Segment Anything (SAM)(Kirillov et al. 2023) for pseudo-labeling. However, directly applying SAM is hindered (Zhang et al. 2023) by background noise and extraneous objects. To mitigate this, we generate segmentation masks utilizing SAM guided by existing point annotations for crowd counting. Additionally, input images are partitioned into distinct regions to further refine pseudo-labels. For images in *(B, S2)* with fewer birds, CVAT is used to manually generate segmentation masks.

Species Classification The dataset contains 34 distinct bird species labeled by experts using species information per the IOC 13.1 taxonomy (Gill, Donsker, and Rasmussen 2023). This helps to ensure accuracy and reliability in identifying the comprehensive representation of avian classes.

Considering the scale and diversity of the dataset, along with complexity associated with the annotation process a set of data has been annotated using point labels under manual supervision. The remaining images stay unlabeled, presenting opportunities to explore unsupervised or weakly supervised learning methods.

Visual Quality Evaluation

We also evaluate the visual quality of the proposed dataset using the BRISQUE score (Mittal, Moorthy, and Bovik 2012). The dataset achieves an average BRISQUE score of 32.34 on a scale of 0 to 100, where lower the score better is the quality. In addition, the BRISQUE scores for the dataset obtained from two distinct sites are as follows: 31.39 (31.24 - training set; 33.55 - testing set) for *(K, S1)*, and 37.80 (37.55 - training set; 39.25 - testing set) for *(B, S2)*. This finding suggests a consistent and high level of image quality across several collection sites. These results imply that the dataset contains images of high visual quality and complexity, influenced by a variety of factors that affect image representation.

Experimental Setup

Implementation Details: The dataset contains images and annotations for birds collected manually from two sites over 4-5 months. Benchmark experiments were run on a multi-GPU Nvidia DGX A100 station with 2 80GB GPUs. We have also used the Megadetector toolkit (Beery, Morris, and Yang 2019) to initially crop out the birds in case of species classification task, as feeding the entire image makes it difficult for the model to learn meaningful features of the birds.

Baseline Methods

We perform benchmarking for the aforementioned tasks and the details are mentioned below.

Crowd counting and Density Estimation In the realm of crowd counting and analysis, several innovative approaches have been proposed. **Context-aware-crowd-counting (CAN)** (Liu, Salzmann, and Fua 2019) adaptively

[4]https://www.birds.cornell.edu/citizenscience/

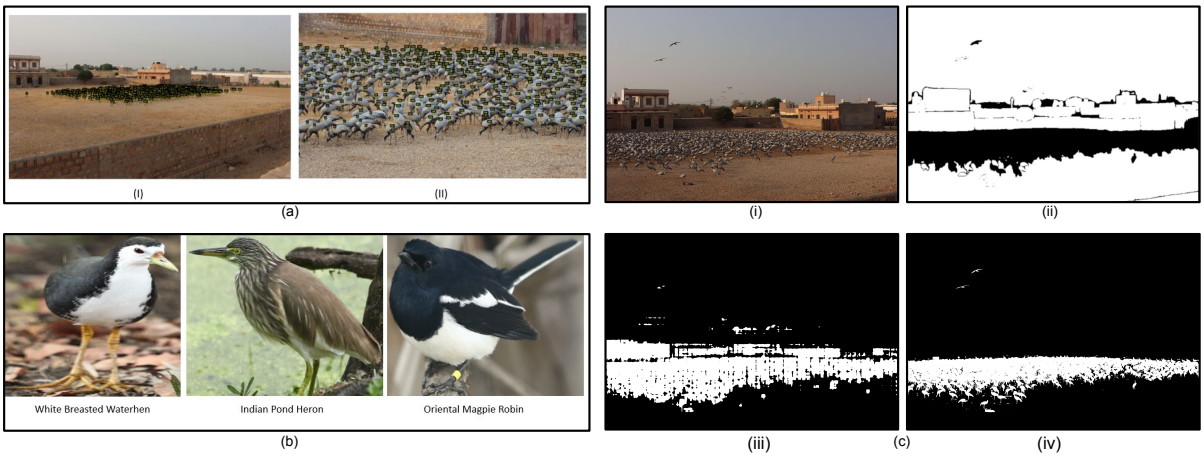

Figure 4: Sample images visualizing annotation process. (A): Point annotations for Crowd counting, (B): Species identification annotation using labels and (C): Segmentation masks (i) input image, (ii) SAM generated mask without prompts, (iii) SAM generated mask using point annotations as guidance. (iv) SAM mask when image is given in parts as input along with points.

encodes contextual information, prioritizing high-density regions. By incorporating multi-scale features, contrasts, and learned attention maps, it generates scale-aware density representations. Another notable method, **DM-Count** (Wang et al. 2020), introduces a novel solution by employing distribution matching through optimal transport to overcome the limitations associated with Gaussian smoothing. This eliminates the need for fixed kernel smoothing, enhancing accuracy. **P2PNet**, on the other hand, takes a unique approach by directly outputting 2D point coordinates for head locations. Utilizing Hungarian matching and regression, it achieves accurate counting without the requirement for intermediate density maps. **M-SFANet** (Thanasutives et al. 2021) addresses scale variation and background suppression challenges by integrating multi-scale features through ASPP and CAN modules, yielding high-resolution density and attention maps. Finally, **CrowdFormer** (Savner and Kanhangad 2023) uses pyramid vision transformers and patch embeddings to capture global context for weakly-supervised crowd counting, significantly enhancing prediction accuracy.

Diffuse-Denoise-Count (Ranasinghe et al. 2023) formulates crowd density map generation as a conditional denoising diffusion process (Ho, Jain, and Abbeel 2020). It models the diffusion process by adding Gaussian noise to the density maps and reversing the process for generation. Forward process $\mathcal{F}_q$ is formulated as :

$$\mathcal{F}_q : q\left(s_t \mid s_{t-1}\right) = \mathcal{N}\left(s_t \mid \sqrt{1-\eta_t} s_{t-1}, \eta_t \mathcal{I}\right) \quad (1)$$

Here $\mathcal{I}$ is the identity matrix. The sample s_0 is gradually transformed to a noisy sample s_t for $t \in \{1, \ldots, T\}$ upon addition of Gaussian noise. Further, the noise is sampled from a Gaussian distribution with a variance determined by the noise schedule $\eta_1, \ldots, \eta_j, \ldots, \eta_T$. s_t is computed by applying the forward transformation to s_0 and a noise vector $\nu \sim \mathcal{N}(0, \mathcal{I})$

$$s_t = \sqrt{\bar{\alpha}_t} s_0 + \left(1 - \overline{\alpha_t}\right) \nu \quad (2)$$

$\bar{\alpha}_t := \prod_{\tau=1}^{t} \alpha_\tau = \prod_{\tau=1}^{t} (1 - \eta_\tau)$ and η_τ. The conditional density map training objective $\mathcal{L}_{cwd}$ is a weighted sum of (i) $\mathcal{L}_{hybrid}$ or denoising loss and (ii) $\mathcal{L}_{vlb}$ or variational lower bound loss (Nichol and Dhariwal 2021).

$$\mathcal{L}_{cwd} = \mathcal{L}_{hybrid} + \gamma_{ct} \mathcal{L}_{ct} \quad (3)$$

Auxiliary $\mathcal{L}_{ct}$ loss computed from encoder-decoder features helps to learn the crowd-specific features.

Segmentation For the task of semantic segmentation, several notable methods have been proposed in the literature. **OneFormer** (Jain et al. 2023) excels at generating ground truths from panoptic annotations and adapting dynamically with multi-scale features, employing contrastive loss and CoordConv for spatial information. **Mask2former** (Cheng et al. 2022) utilizes masked attention for precise segmentation masks, featuring a multi-scale design that integrates hierarchical features. **SegFormer** (Xie et al. 2021) introduces a hierarchical transformer encoder for multi-scale features and an MLP decoder combining local and global attention for robust representations. **ClipSeg** (Lüddecke and Ecker 2022) distinguishes itself by supporting both zero-shot and one-shot segmentation through CLIP and a transformer decoder, leveraging CLIP's text-image embedding space for enhanced generalization. Finally, **Segment Anything (SAM)** (Kirillov et al. 2023) contributes efficient instance segmentation with point annotations, addressing real-world complexities while guiding accurate mask generation, offering potential for unsupervised learning and maintaining semantic-aware features.

Species Classification In fine-grained classification, innovative techniques have emerged. **Mutual-Channel (MC)** (Chang et al. 2020) leverages individual feature map channels with an MC Loss for distinct recognition. **Progressive Multi-Granularity (PMG)** (Du et al. 2020) sequentially captures multi-granularity representations. **HERBS** (Chou, Kao, and Lin 2023) enhances features with high-temperature refinement and background suppression.

Models	Backbone	MAE	RMSE
CAN	CNN	29.37	44.65
CSRNet	CNN	28.12	47.06
P2P-Net	CNN	145.06	296.41
M-SFANet	CNN	199.29	265.47
DMCount	CNN	27.44	58.28
CrowdFormer	Transformer	28.90	70.72
Diffuse Denoise-Count	DDPM	26.18	41.24

Table 4: Benchmarking results of the crowd counting techniques on the *(K, S1)* dataset

Model	*(K, S1)*	*(B, S2)*
Mask2Former	0.24	0.56
SegFormer	0.54	0.52
ClipSeg	0.6	0.77
OneFormer	0.76	0.94
Grounded Dino + SAM	0.49	0.92
Grounded SAM	0.63	0.93
SAM + Point Annotation	0.76	0.88

Table 5: Benchmarking the segmentation performance on the proposed dataset in terms of mIoU values.

FGVC-PIM (Chou, Lin, and Kao 2022) selects discriminative regions and fuses them via graph convolution. **ViT-NeT** (Kim, Nam, and Ko 2022) integrates ViT encoder and neural tree decoder for hierarchical decision-making in fine-grained classification.

Evaluation Protocol

BirdCollect dataset consists of two distinct subsets: *(K, S1)* and *(B, S2)*. We conduct studies for four different tasks, and to maintain consistency across these tasks, we define two different protocols, the details of which are listed below:

Protocol 1 - Unified Protocol: In the unified protocol, the 2596 annotated images, encompassing *(K, S1)* and *(B, S2)* datasets, are divided into a 70-30 ratio for training and testing. Accordingly, we have 1817 samples in the training set and 779 samples in the test set. We then keep the split consistent for performing the benchmarking experiment of the recent state of the art methods for (i) Crowd counting, (ii) Density estimation, and (iii) Segmentation respectively. For the crowd counting and density estimation task we utilize the point annotations as the ground truth and then predict the counting estimate along with the density maps. The semantic segmentation task aims to evaluate the mean-IoU scores scores across the segmentation masks.

Protocol 2 - BirdSpecies Protocol: The second protocol includes 433 images of 34 different bird species *(B, S2)* from Bharatpur. Further, we use samples from the *(K, S1)* dataset as the 35th class, resulting in a total sample count of 441. The addition of eight images introduces the **demoiselle crane** as a new class, creating a 35-class problem. This setup is designed to test classification within an imbalanced dataset, addressing the challenge of identifying species with limited samples. The benchmarking results are evaluated for accuracy in classifying the test samples into one of the 35 species using a 70-30 split.

Method	Backbone	%Acc
HERBS	Transformer	82.11
PMG	CNN	66.67
MC Loss	CNN	72.10
Vit-Net	Transformer	21.48
FGVC-PIM	Transformer	38.26

Table 6: Benchmarking results for species classification tasks on *(B, S2)* dataset

Evaluation Metrics

To evaluate the performance of the bird counting algorithms on the dataset, we utilize the standard metrics including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and Mean Intersection over Union (mIoU) for segmentation. In the case of species classification, we assess the effectiveness of the methods using classification metric of computing the overall accuracy by comparing the predicted class labels to the corresponding ground truth labels. The evaluation process involves rigorous testing and validation procedures to ensure the reliability.

Results and Analysis

In this section, we provide an overview of the baseline results obtained for different computer vision tasks. The state of the art models, protocols, and dataset details are discussed in the previous section.

For the unified protocol-1, we present the findings of crowd counting and density estimation experiments conducted on the *(K, S1)* dataset, as detailed in Table 4. We observe the variation of MAE ranging from 26.18 to 199.29, with the minimum MAE acchieved by Diffuse Denoise-Count(Ranasinghe et al. 2023). It is evident that state-of-the-art methodologies struggle when tasked with automatically analyzing images of bird flocks, both on ground and in-flight. These can be attributed to the variations in density and occlusion levels within the images. The method centered around diffusion models performs relatively better by generating high quality narrow kernel density maps that maintain pixel value distribution and enable accurate contour-based counting. Their stochastic map fusion also improves localization compared to single-output CNNs and Transformers. Furthermore, the *(B, S2)* dataset comprises several bird species, so counting birds lack meaningful interpretability.

Table 5 presents the outcomes of semantic segmentation in the context of both the *(K, S1)* and *(B, S2)* datasets. The distinctive characteristics of individual birds pose challenges for models to accurately segment dense flocks. Their distinct shapes, color, features and high degree of variability in crowded settings complicates the ability to distinguish the boundaries between them. We employ MIoU as a metric to assess segmentation model effectiveness. Upon examination, it's evident that OneFormer(Jain et al. 2023) and SAM(Kirillov et al. 2023), utilizing input point annotations, demonstrate the highest performance on the *(B, S2)* with mIoU of 0.76. Further, Results on *(B, S2)* dataset shows some gains in the performance with highest mIoU of 0.94 achieved by OneFormer, particularly because the samples

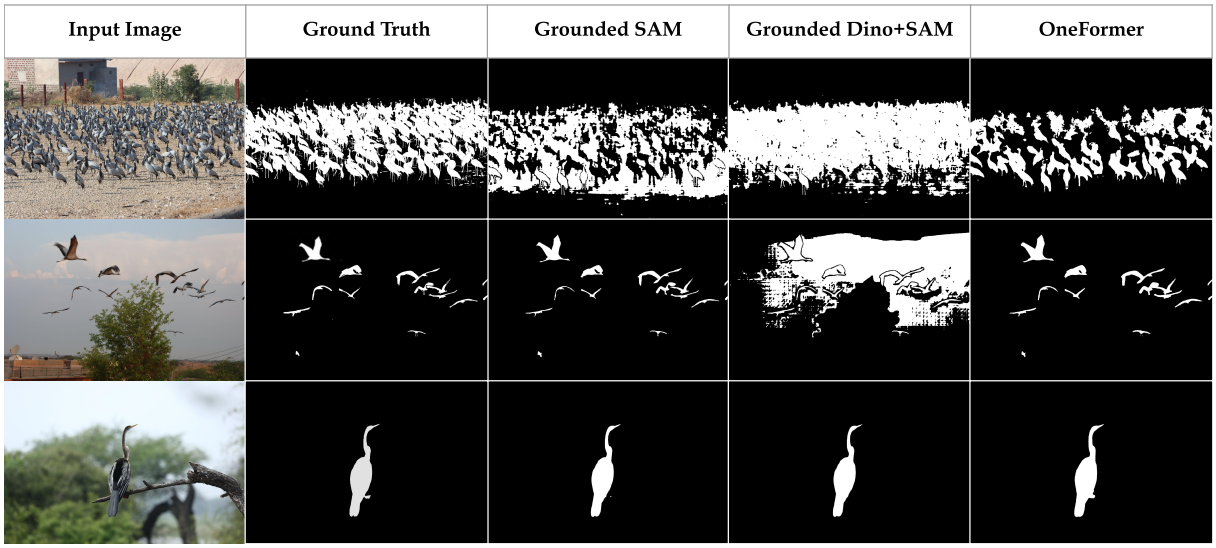

Figure 5: Visual sample representations of segmentation mask obtained from cutting-edge models. Column 1 displays input images, column 2 shows pseudo ground truth. Columns 3 to 5 depict the binary masks from various models. Rows 1-2 show *(K, S1)* subset, while the last row presents *(B, S2)* subset from the *BirdCollect* dataset.

have fewer birds though with significant variations. However, the challenges of density, flight pattern and diverse backgrounds makes the task challenging. Some examples of predicted segmentation masks by top-performing models are visualised in Figure 5.

For protocol-2, experiments performed (Table 6) for species classification on *(B, S2)* challenges the state-of-the-art models in the limited data regime. The variations in performance with highest accuracy of 82.11% depends on the ability to mitigate background noise and enhance multi-scale features to focus on discriminative regions. The challenge of classifying rare species within imbalanced datasets with constrained samples, as seen in our dataset, is mainly due to reliance of models on localization of subtle discriminative regions. Also, the model sensitivity to class imbalance often results in suboptimal feature suppression or refinement.

Across various protocols, it becomes evident that existing methods face challenges in analyzing high-density bird flocks under diverse conditions and identifying species within data exhibiting a long-tailed distribution. Thorough experiments highlighted the intricate nature of the proposed dataset, emphasizing its potential to propel advancements in deep learning models.

Conclusion and Broader Impact

The development of automated non-invasive monitoring (Kshitiz et al. 2023) technologies to track large bird flocks offers immense potential for enhancing bird behaviour analysis with minimal human intervention. Computer vision techniques including crowd counting and density estimation help facilitate examining migration patterns and spatial distribution of the avian populations. Further, the study of changes in flock density provides critical ecological insights, unveiling shifts in environmental factors influencing population health, such as climate and food chain ecosystem. In addition, semantic segmentation of flocks is crucial for unraveling behaviour dynamics and supporting effective ecological monitoring for conservation. However, the lack of comprehensive annotated large scale datasets capturing the intricacies of diverse habitats and behaviors have constrained the avian habitat research. Through extensive experiments, we demonstrate the complexity of the proposed dataset including diverse density and variable illumination conditions. These ecological insights can help drive data-driven conservation initiatives, safeguarding threatened migratory species. We believe curating and releasing benchmark datasets of this nature is crucial for pushing the boundaries of technologies to help develop long-term monitoring of flocks to protect vulnerable avian populations facing escalating anthropogenic threats worldwide.

Ethical Statement

Our non-invasive research methods, approved by institutional committee, ensure no harm occurs to the birds or the human participants during data collection. The birds considered for our study belong to the IUCN Least Concern bird species. The work aligns with several UN Sustainable Development Goals (SDGs) (SDG 13 and 15), involving providing insights into climate change and habitat destruction through bird monitoring. By furthering wildlife conservation and community well-being, we work in accordance with the 2030 Agenda for Sustainable Development.

Acknowledgments

We extend our thanks to Manoj Sharma, a naturalist at Keoladeo National Park, for his expert guidance on the details of Bharatpur bird species. We thank all the volunteers for participating in the annotation process. This research was sup-

ported by iHub Drishti, the Technology Innovation Hub on CV, AR and VR and US National Science Foundation grant IIS 1956050.

References

Arya, S.; and Syriac, E. K. 2018. Wetlands: The living waters-A review. *Agricultural Reviews*, 39(2): 122–129.

Bai, S.; He, Z.; Qiao, Y.; Hu, H.; Wu, W.; and Yan, J. 2020. Adaptive dilated network with self-correction supervision for counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4594–4603.

Beery, S.; Morris, D.; and Yang, S. 2019. Efficient pipeline for camera trap image review. *arXiv preprint arXiv:1907.06772*.

Bera, A.; Wharton, Z.; Liu, Y.; Bessis, N.; and Behera, A. 2022. Sr-gnn: Spatial relation-aware graph neural network for fine-grained image categorization. *IEEE Transactions on Image Processing*, 31: 6017–6031.

Berg, T.; Liu, J.; Woo Lee, S.; Alexander, M. L.; Jacobs, D. W.; and Belhumeur, P. N. 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2018.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Chang, D.; Ding, Y.; Xie, J.; Bhunia, A. K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; and Song, Y.-Z. 2020. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Transactions on Image Processing*, 29: 4683–4695.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2022. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Chou, P.-Y.; Kao, Y.-Y.; and Lin, C.-H. 2023. Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. *arXiv preprint arXiv:2303.06442*.

Chou, P.-Y.; Lin, C.-H.; and Kao, W.-C. 2022. A novel plug-in module for fine-grained visual classification. *arXiv preprint arXiv:2202.03822*.

Du, R.; Chang, D.; Bhunia, A. K.; Xie, J.; Ma, Z.; Song, Y.-Z.; and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 153–168. Springer.

Gill, F.; Donsker, D.; and Rasmussen, P. 2023. IOC World Bird List (v13.2). Eds. Doi: 10.14344/IOC.ML.13.1.

Gu, Z.; Chen, H.; Xu, Z.; Lan, J.; Meng, C.; and Wang, W. 2022. Diffusioninst: Diffusion model for instance segmentation. *arXiv preprint arXiv:2212.02773*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2989–2998.

Kim, S.; and Kim, M. 2020. Learning of counting crowded birds of various scales via novel density activation maps. *IEEE Access*, 8: 155296–155305.

Kim, S.; Nam, J.; and Ko, B. C. 2022. Vit-net: Interpretable vision transformers with neural tree decoder. In *International Conference on Machine Learning*, 11162–11172. PMLR.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kshitiz, S. S.; Mounir, R.; Vatsa, M.; Singh, R.; Anand, S.; Sarkar, S.; and Parihar, S. M. 2023. Long-term monitoring of bird flocks in the wild. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6344–6352.

Li, X.; Liu, Y.; and Zhu, Y. 2022. The Effects of Climate Change on Birds and Approaches to Response. In *IOP Conference Series: Earth and Environmental Science*, volume 1011, 012054. IOP Publishing.

Liang, D.; Chen, X.; Xu, W.; Zhou, Y.; and Bai, X. 2022. Transcrowd: weakly-supervised crowd counting with transformers. *Science China Information Sciences*, 65(6): 160104.

Liu, W.; Salzmann, M.; and Fua, P. 2019. Context-aware crowd counting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5099–5108.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Lüddecke, T.; and Ecker, A. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7086–7096.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Ranasinghe, Y.; Nair, N. G.; Bandara, W. G. C.; and Patel, V. M. 2023. Diffuse-Denoise-Count: Accurate Crowd-Counting with Diffusion Models. *arXiv preprint arXiv:2303.12790*.

Rosenberg, K. V.; Dokter, A. M.; Blancher, P. J.; Sauer, J. R.; Smith, A. C.; Smith, P. A.; Stanton, J. C.; Panjabi, A.; Helft, L.; Parr, M.; et al. 2019. Decline of the North American avifauna. *Science*, 366(6461): 120–124.

Sam, D. B.; Peri, S. V.; Sundararaman, M. N.; Kamath, A.; and Babu, R. V. 2020. Locate, size, and count: accurately resolving people in dense crowds via detection. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2739–2751.

Savner, S. S.; and Kanhangad, V. 2023. CrowdFormer: Weakly-supervised crowd counting with improved generalizability. *Journal of Visual Communication and Image Representation*, 94: 103853.

Strudel, R.; Garcia, R.; Laptev, I.; and Schmid, C. 2021. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7262–7272.

Sun, H.; He, X.; and Peng, Y. 2022. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5853–5861.

Thanasutives, P.; Fukui, K.-i.; Numao, M.; and Kijsirikul, B. 2021. Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In *2020 25th international conference on pattern recognition (ICPR)*, 2382–2389. IEEE.

Wang, B.; Liu, H.; Samaras, D.; and Nguyen, M. H. 2020. Distribution matching for crowd counting. *Advances in neural information processing systems*, 33: 1595–1607.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Zhang, C.; Liu, L.; Cui, Y.; Huang, G.; Lin, W.; Yang, Y.; and Hu, Y. 2023. A Comprehensive Survey on Segment Anything Model for Vision and Beyond. *arXiv preprint arXiv:2305.08196*.

Zhang, L.; Huang, S.; Liu, W.; and Tao, D. 2019. Learning a mixture of granularity-specific experts for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8331–8340.

Zheng, H.; Fu, J.; Mei, T.; and Luo, J. 2017. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, 5209–5217.