SmartWilds: Multimodal Wildlife Monitoring Dataset

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present the first release of SmartWilds, a multimodal wildlife monitoring dataset. SmartWilds is a synchronized collection of drone imagery, camera trap photographs and videos, and bioacoustic recordings collected during summer 2025 at The Wilds safari park in Ohio. This dataset supports multimodal AI research for comprehensive environmental monitoring, addressing critical needs in endangered species research, conservation ecology, and habitat management. Our pilot deployment captured four days of synchronized monitoring across three modalities in a 220-acre pasture containing Pere David's deer, Sichuan takin, Przewalski's horses, as well as species native to Ohio, including bald eagles, white-tailed deer, and coyotes. We provide a comparative analysis of sensor modality performance, demonstrating complementary strengths for landuse patterns, species detection, behavioral analysis, and habitat monitoring. This work establishes reproducible protocols for multimodal wildlife monitoring while contributing open datasets to advance conservation computer vision research. Future releases will include synchronized GPS tracking data from tagged individuals, citizen science data, and expanded temporal coverage across multiple seasons.

1 Introduction

Conservation biology requires comprehensive ecosystem monitoring to inform evidence-based management decisions, yet traditional approaches provide fragmented views of wildlife activity and habitat use. The integration of multiple sensing modalities, powered by edge AI and computer vision approaches offers transformative opportunities for automated environmental monitoring at unprecedented scales (Besson et al., 2022; Tuia et al., 2022; Pringle et al., 2025; Kline et al., 2025). Our research advances multimodal AI for wildlife monitoring by creating datasets and models that enable object detection and tracking across environmental conditions, support fine-grained species classification from multi-sensor data, and facilitate behavioral analysis from long video sequences.

We present a multimodal dataset from sensor deployments at The Wilds Conservation Center (The Wilds, 2025). Our dataset was curated to evaluate multi-sensor fusion techniques and benchmark machine learning approaches for conservation applications across visual, acoustic, and environmental data streams. This dataset poses challenges to existing computer vision methods through its combination of rare species detection, variable conditions, and the need for behavioral state recognition over extended time periods. This work supports research areas including endangered species monitoring, conservation and restoration ecology, and habitat assessment for both exotic endangered species and native biodiversity conservation. Our contributions include: (1) a synchronized multimodal ecological dataset with comprehensive metadata, (2) reproducible protocols for environmental monitoring sensor networks, and (3) support for sensor fusion to advance multimodal learning research in environmental contexts.

The rest of the paper is organized as follows: Section 2 reviews related works on multi-modal sensor fusion for ecological research; Section 3 describes the study site and sensor network design; Section

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

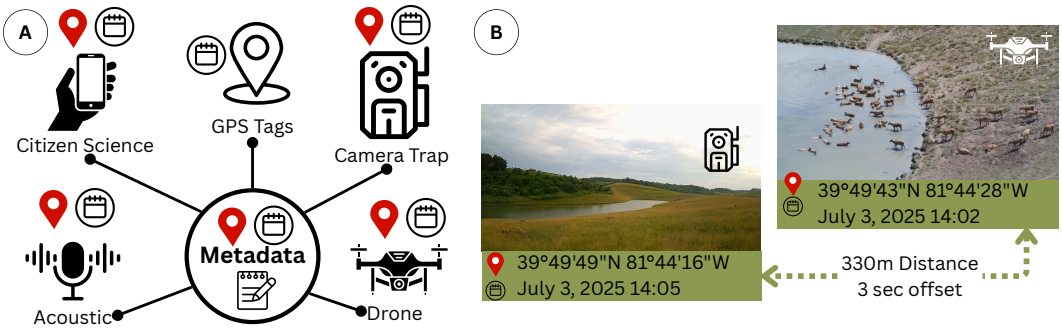

Figure 1: Representative images and data study design. GPS and time-stamp metadata allow for cross-referencing between modalities. (a) Diagram of dataset modalities, citizen science images, GPS tags, acoutic data, camera trap and drone images, joined via location and time-stamp metadata. (b) Example of multi-modal data cross-referencing using metadata. Camera trap view (TW02) of Pere David's deer synchronized with drone image of Pere David's deer wading in lake.

4 details the dataset, including camera trap, bioacoustic, drone, and metadata. Section 5 includes field observation, discusses the strengths and weaknesses of the sensing modalities and how the pilot study informed additional data collection; Section 6 includes details about future data releases and work in progress.

2 Related Work

Recent advances have introduced multimodal wildlife monitoring datasets addressing diverse conservation challenges (Smith and Pinter-Wollman, 2021; Buxton et al., 2018; Besson et al., 2022). MammAlps (Gabeff et al., 2025) combines multi-view video with synchronized audio data to analyze wild mammal behaviors in alpine environments, demonstrating the power of multimodal approaches for detailed behavioral analysis. BuckTales (Naik et al., 2024) provides multi-UAV tracking data for wild antelope identification and re-identification, advancing techniques for individual animal monitoring. The PanAf-FGBG dataset (Brookes et al., 2025) explores how environmental backgrounds impact wildlife behavior recognition, while YOLO-Behaviour (Chan et al., 2025) demonstrates automated behavior quantification frameworks. KABR (Kholiavchenko et al., 2024) focuses on Kenyan wildlife behavior recognition from drone footage, contributing to conservation AI in African ecosystems. However, most existing datasets focus on specific taxonomic groups or behavioral tasks rather than comprehensive ecosystem monitoring. Our work contributes synchronized multi-sensor data collection protocols designed specifically for conservation digital twin development, with planned integration of GPS tracking data for individual-level behavioral analysis.

3 Field Deployment and Data Collection

3.1 Study Site

The Wilds is a 10,000-acre conservation center in southeastern Ohio, home to endangered species conservation programs and native wildlife restoration efforts (The Wilds, 2025). Our pilot deployment focused on a single 220-acre enclosure containing a breeding population of GPS-tagged Pere David's deer (*Elaphurus davidianus*), chosen specifically to enable future integration with individual tracking data. Sensor sites were selected chronologically based on observed wildlife activity patterns and strategic coverage of diverse habitat types within the study pasture, illustrated in Fig. 2 and described in Table 3. Camera trap sites prioritized high deer activity areas, particularly around water sources, while bioacoustic monitors targeted diverse acoustic environments from open grasslands to woodland edges. Where possible, existing structures were utilized for sensor mounting to minimize environmental impact and maximize equipment protection.

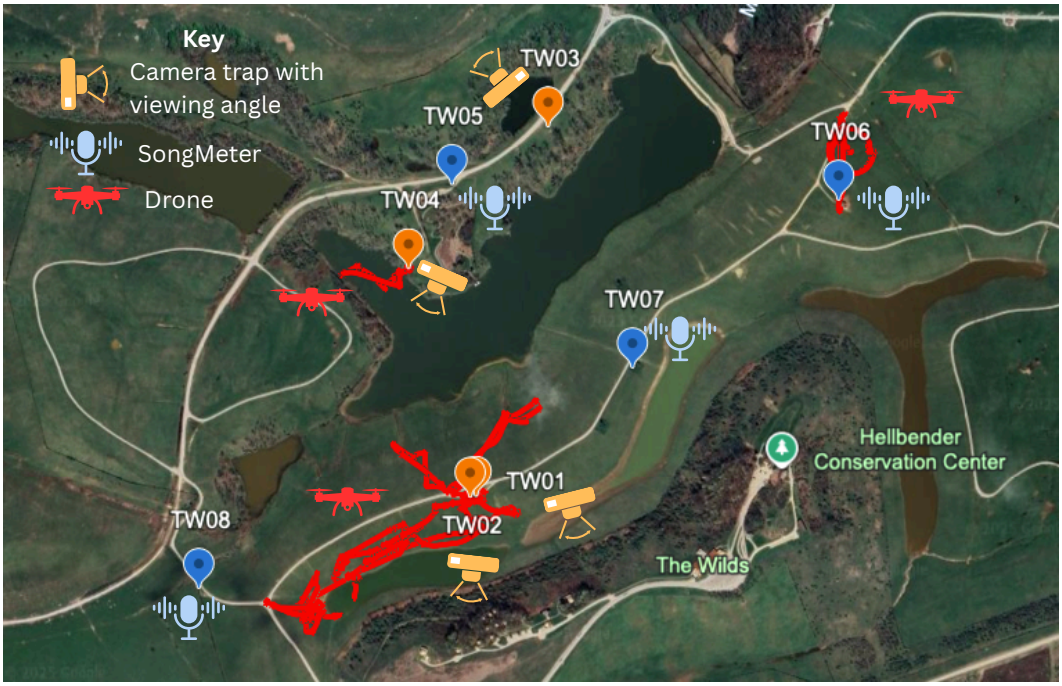

Figure 2: Map of sensor placements created with Google Earth. Camera trap locations in orange, bioacoustics sensors in blue, drone flight paths in red. See Table 3 for site details.

3.2 Sensor Network Design

The multimodal sensor network consisted of three complementary sensing technologies deployed for four days of continuous monitoring (June 30 - July 3 2025). Four camera traps, including GardePro T5NG and comparable trail camera models, were strategically positioned around lakes and wildlife congregation areas using motion-triggered photo/video hybrid mode to capture animal activity at key locations. See Table 3 for details on site selection. Four bioacoustic monitors (Song Meter Mini devices) recorded high-quality 48kHz, 16-bit mono audios. Multiple drone missions using Parrot ANAFI quadcopters provided flexible aerial coverage through systematic surveys and opportunistic behavioral tracking, with dedicated synchronization flights conducted within view of camera traps to enable precise cross-modal timestamp calibration. This integrated sensor array enabled comprehensive multimodal data collection across fixed monitoring locations and dynamic aerial observations.

4 Dataset

The multi-modal dataset is summarized in Table 4. The initial dataset release totals 101GB and over 20K files. The dataset is organized by sensor type and deployment location, with standardized metadata for each component.

4.1 Camera trap data

Camera trap data consists of motion-triggered images and videos organized by deployment site (TW01-TW04), providing comprehensive visual documentation specifically structured for object detection, tracking, and fine-grained classification of wildlife species. The systematic capture of animal behaviors and interactions across multiple camera locations enables localization and recognition tasks in naturalistic settings with varying environmental conditions and species compositions.

Table 1: The Wilds Multimodal Initial Release Dataset Summary. Initial release includes photos, videos, and acoustic data. Future releases will include GPS tag data, citizen science images, and weather and satellite data.

Modality	**Data Type**	**Total Files**	**GB**
Camera Traps	Visual monitoring (photos and videos)	20,014	49
Bioacoustic	Audio recordings	311	6
Drone	Aerial video	20 video files + metadata	46
Total	**All modalities**	**∼20K**	**101**

4.2 Bioacoustic data

Bioacoustic recordings from continuous and scheduled monitoring (TW05-TW08) were collected to enable multi-sensor fusion research that combines acoustic representations with visual data streams. Half of the monitors were configured to record 5 minutes every hour to capture ungulate vocalizations throughout the day, and half were configured to record bird song at dusk and dawn to capture local diversity.

4.3 Drone data

Drone mission data includes video footage with flight telemetry and detailed mission objectives, designed to support 3D modeling, temporal and behavioral reasoning from video sequences. The extended drone video captures are particularly valuable for recognition of complex behavioral states in long video sequences, including territorial displays, social interactions, and habitat use patterns that unfold over extended observation periods.

4.4 Metadata

All deployments include comprehensive metadata with GPS coordinates, habitat descriptions, technical sensor specifications, deployment timestamps, environmental conditions, and detailed field observations from researchers. This structured field note documentation, combined with visual and acoustic data streams, creates a rich foundation for multi-sensor fusion studies that integrate images, sounds, and contextual field observations. The standardized metadata framework supports human-in-the-loop and citizen-science annotation efforts by providing the contextual information necessary for active-learning pipelines that balance annotation cost and data quality.

5 Discussion

5.1 Field Observations

Field deployment revealed important insights about multimodal monitoring in conservation settings. Animal responses varied by sensor type. The deer initially showed curiosity toward drone flights but exhibited minimal behavioral disruption overall. Breeding season activity patterns were clearly observable, with territorial males vocalizing frequently and herds congregating around water sources during warm weather. Technical challenges included GPS signal limitations in remote areas affecting some sensor synchronization, weather impacts on acoustic recording quality, and the need for creative mounting solutions in areas lacking suitable structures. Despite these challenges, the sensor network successfully captured multimodal data across all target areas.

5.2 Comparison of Sensor Modalities

We summarize the relative performance each sensing modality across eight key performance dimensions relevant to conservation monitoring applications (Table 2). These metrics were selected based on established frameworks for conservation technology evaluation and practical deployment considerations in wildlife monitoring contexts. Spatial range and resolution determines monitoring coverage and detection capabilities across different habitat scales (Tuia et al., 2022; Pringle et al., 2025). Temporal range and resolution captures both short-term behavioral events and long-term ecological

Table 2: Comparative analysis of sensor modality performance across key conservation monitoring metrics. Metrics selected based on established frameworks for wildlife monitoring technology evaluation (Tuia et al., 2022; Besson et al., 2022) and practical deployment considerations in conservation settings. Performance rating: Poor, Moderate, Good.
*GPS tag data will be added in a future data release.

Metric	Camera Traps	Bioacoustics	Drones	GPS Tags*
Spatial Range	Fixed location, ∼30 m radius	Fixed location, ∼100 m radius	Mobile; battery-limited (∼2 km)	Entire home range
Spatial Resolution	High within field-of-view	Moderate directional	Sub-meter aerial resolution	∼1–10 m accuracy
Temporal Range	Weeks to months	Weeks to months	Hours per mission	Months to years
Temporal Resolution	Event-triggered; <1 s	Continuous or scheduled	30–60 fps video	Hourly locations
Species Detectability	Large ungulates, visible species	Cryptic/vocal species, birds	Large mammals, aerial view	Tagged individuals only
Behavior Detail	Limited to frame interactions	Vocalizations, acoustic behaviors	High detail: posture, interactions	Movement patterns only
Deployment Effort	Low–medium (site visits)	Low–medium (site visits)	High (active piloting)	Low once deployed
Data Volume	Moderate	Moderate–high	High	Low

patterns Besson et al. (2022). Species detectability captures different sensor modalities ability to sense specific taxonomic groups, especially more cryptic species such as birds or insects,(Smith and Pinter-Wollman, 2021). Behavioral detail is important for understanding complex social interactions of group-living animals and individual to group-level responses to environmental changes (Kline et al., 2025). Deployment effort and data volume captures practical considerations affecting scalability and cost of long-term monitoring efforts (Besson et al., 2022).

6 Future Directions

Building on pilot deployment insights, upcoming releases will address identified limitations and leverage demonstrated multimodal strengths. The pilot revealed minimal useful data from camera trap videos compared to drone footage, leading to modified protocols with co-located bioacoustic monitors and camera traps at three additional sites for direct detection capability comparison. Future releases will integrate synchronized GPS tracking from ear-tagged Pere David's deer with visual and acoustic data, enabling analysis of individual movement patterns and behaviors across the four additional weeks of planned data collection to capture seasonal variation.

The demonstrated complementary strengths across modalities—where camera traps excel at species identification, bioacoustic monitors provide continuous temporal coverage, and drones offer landscape-scale perspectives. These complementary strengths will inform machine learning research directions focused on multimodal fusion architectures. Development will prioritize real-time adaptive sampling through edge computing capabilities and AI-assisted management systems that leverage integrated sensor networks' superior performance over single-modality approaches. Extension to multiple habitat types, replication at additional conservation sites, and integration of citizen science observations will expand data collection while validating multimodal AI frameworks that can operate autonomously across diverse ecosystems and transform global environmental monitoring practices.

7 Data Availability Statement

The complete multimodal wildlife monitoring dataset will be made publicly available on Hugging Face upon publication of this manuscript under a CC0-1.0 license. Dataset cards with representative samples and comprehensive metadata are currently available for review. All code for data processing and analysis will be released alongside the dataset to ensure reproducibility.

References

Besson, M., Alison, J., Bjerge, K., Gorochowski, T. E., Høye, T. T., Jucker, T., Mann, H. M. R., and Clements, C. F. (2022). Towards the fully automated monitoring of ecological communities. *Ecology Letters*, 25(12):2753–2775.

Brookes, O., Kukushkin, M., Mirmehdi, M., et al. (2025). The panaf-fgbg dataset: Understanding the impact of backgrounds in wildlife behaviour recognition.

Buxton, R. T., Lendrum, P. E., Crooks, K. R., and Wittemyer, G. (2018). Pairing camera traps and acoustic recorders to monitor the ecological impact of human disturbance. *Global Ecology and Conservation*, 16:e00493.

Chan, A. H. H., Putra, P., Schupp, H., et al. (2025). Yolo-behaviour: A simple, flexible framework to automatically quantify animal behaviours from videos. *Methods in Ecology and Evolution*, 16:760–774.

Gabeff, V., Qi, H., Flaherty, B., et al. (2025). Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps.

Kholiavchenko, M., Kline, J., Ramirez, M., et al. (2024). Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos.

Kline, J., Afridi, S., Rolland, E. G., Maalouf, G., Laporte-Devylder, L., Stewart, C., Crofoot, M., Stewart, C. V., Rubenstein, D. I., and Berger-Wolf, T. (2025). Studying collective animal behaviour with drones and computer vision. *Methods in Ecology and Evolution*.

Naik, H., Yang, J., Das, D., Crofoot, M., Rathore, A., and Sridhar, V. H. (2024). Bucktales: A multi-uav dataset for multi-object tracking and re-identification of wild antelopes. *Advances in Neural Information Processing Systems*, 37:81992–82009.

Pringle, S., Dallimer, M., Goddard, M. A., Le Goff, L. E., Hart, E., Langdale, S. J., Fisher, J. C., Abad, S.-A., Ancrenaz, M., Angeoletto, F., Auat Cheein, F., Austen, G. E., Bailey, J., Baldock, K., Banin, L., Banks-Leite, C., Barau, A., Bashyal, R., Bates, A. J., Bicknell, J. E., Bielby, J., Bosilj, P., Bush, E., Butler, S., Carpenter, D., Clements, C. F., Cully, A., Davies, K., Deere, N. J., Dodd, M., Drinkwater, R., Driscoll, D., Dutilleux, G., Dyrmann, M., Edwards, D. P., Farhadinia, M. S., Faruk, A., Field, R., Fletcher, R. J., Foster, C., Fox, R., Francksen, R., Franco, A., Gainsbury, A., Gardner, C. J., Giogi, I., Griffiths, R. A., Hamaza, S., Hanheide, M., Hayward, M. W., Hedblom, M., Helgason, T., Heon, S. P., Hughes, K., Hunt, E., Ingram, D. J., Jackson-Mills, G., Jowett, K., Keitt, T., Kloepper, L., Kramer-Schadt, S., Labisko, J., Labrosse, F., Lawson, J., Lecomte, N., de Lima, R. F., Littlewood, N. A., Marshall, H., Masala, G. L., Maskell, L., Matechou, E., Mazzolai, B., McConnell, A., Melbourne, B., Miriyev, A., Nana, E., Ossola, A., Papworth, S., Parr, C., Payo-Payo, A., Perry, G., Pettorelli, N., Pillay, R., Potts, S. G., Prendergast-Miller, M., Qie, L., Rolley-Parnell, P., Rossiter, S. J., Rowcliffe, J. M., Rumble, H., Sadler, J. P., Sandom, C., Sanyal, A., Schrodt, F., Sethi, S. S., Shabrani, A., Siddall, R., Smith, S., Snep, R. P. H., Soulsbury, C. D., Stanley, M. C., Stephens, P. A., Stephenson, P. J., Struebig, M. J., Studley, M., Svátek, M., Tang, G., Taylor, N., Umbers, K., Ward, R., White, P., Whittingham, M. J., Wich, S., Williams, C. D., Yoh, N., Zaidi, S. A. R., Zmarz, A., Zwerts, J., and Davies, Z. G. (2025). Opportunities and challenges for monitoring terrestrial biodiversity in the robotics age. *Nature Ecology and Evolution*. Accepted: 2025-04-07.

Smith, J. E. and Pinter-Wollman, N. (2021). Observing the unwatchable: Integrating automated sensing, naturalistic observations and animal social network analysis in the age of big data. *Journal of Animal Ecology*, 90(1):62–75.

The Wilds (2025). The wilds: Conservation center. `https://www.thewilds.org/`. Accessed: 2025-09-09.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., Van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., Van Horn, G., Crofoot, M. C., Stewart, C. V., and Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):792.

A Supplemental Material

Table 3: Sensor deployment sites with selection rationale and habitat characteristics.
CT = camera trap. SM = SongMeter bioacoustic recording device).

Site ID	Sensor Type	Site Name	Habitat Type	Selection Rationale
TW01	CT	Nomad Ridge Shelter (East)	Elevated structure above lake	High Pere David's deer activity at lake; protected mounting location
TW02	CT	Nomad Ridge Shelter (West)	Elevated structure above lake	Complementary lake coverage; high deer activity observed
TW03	CT	Old Giraffe Feeder (Pasture D)	Feeding structure near lake with salt lick	High deer activity around salt lick; artificial congregation point
TW04	CT	Lake Trail (Northwest facing)	Tree-mounted overlooking lake	High deer activity along lake trail; natural travel corridor
TW05	SM	Lake Trail Gate O	Gate structure with vegetation	Easy maintenance access; high bird activity; sunrise/sunset recording
TW06	SM	Zebra Shelter	Open plains structure	Ungulate activity around shelter; daytime recording schedule
TW07	SM	Zipline Tower	Open field structure	Maintenance accessibility; diverse bird activity; sunrise/sunset recording
TW08	SM	Fence with Dense Vegetation	Pasture edge near lake	Capture different acoustic environment; hourly recording