

WildDepth: A Multimodal Dataset for 3D Wildlife Perception and Depth Estimation

Muhammad Aamir^{1★}, Naoya Muramatsu^{2★}, Sangyun Shin^{1★}, Matthew Wijers³, Jia-Xing Zhong¹, Xinyu Hou¹, Amir Patel^{2,4}, Andrew Loveridge³, and Andrew Markham¹

¹ Department of Computer Science, University of Oxford
{muhammad.aamir,sangyun.shin,xinyu.hou,andrew.markham}@cs.ox.ac.uk

² University of Cape Town

³ Wildlife Conservation Research Unit, Department of Biology, University of Oxford
{matthew.wijers,andrew.loveridge}@biology.ox.ac.uk

⁴ Department of Computer Science, University College London
amir.patel@ucl.ac.uk

Abstract. The ability to remotely measure an animal’s size and shape (i.e. its morphology) in its natural habitat is of key importance for applications such as conservation, re-identification, and biomechanical understanding. Monocular depth estimation (MDE) and 3D reconstruction approaches can be used. This is useful to generate accurate deformable mesh models or to characterize biologically important parameters such as sexual dimorphism or the size distribution of populations. Whilst depth estimation and 3D reconstruction have been extensively studied as core topics in computer vision, ranging from early work on simple, rigid objects, to more recent work on human and animal reconstruction, the majority of existing animal models are trained on video/image based datasets that lack metric scale and ground truth e.g. from camera trap images or from public videos. In addition, they do not fully reflect the challenges of accurately estimating the focal animal’s scale when it is distant from a camera. To address this limitation, we present *WildDepth*, a multimodal dataset and benchmark suite for depth estimation, behavior detection, and 3D reconstruction from diverse categories of animals ranging from domestic to wild environments with synchronized RGB and LiDAR. We provide three focused benchmarks: (1) monocular depth estimation with per-distance and temporal stability analysis, (2) behavior detection, (3) 3D reconstruction and densification. Our results show that large-scale MDE models degrade significantly at long range, while LiDAR-anchored fusion reduces metric error by up to 25% – 30% RMSE in mid-range scenarios. Our aim is to enable camera trap data to be more accurately ‘lifted’ to 3D morphometrics, unlocking a new era of scale-aware animal modelling.

★ M. Aamir, N. Muramatsu, and S. Shin contributed equally and are listed alphabetically by surname.

1 Introduction

Monitoring wildlife behavior and population dynamics has become increasingly crucial as biodiversity faces unprecedented threats from habitat loss, poaching, and climate change [8]. Recent advances in computer vision and remote sensing have opened new possibilities for non-invasive monitoring at scale, enabling the observation of species in their natural environments without human interference [40, 43]. From camera traps [5] to aerial surveys [11, 19], these data have become fundamental in understanding species distribution, morphology, and ecological interactions [10]. However, the development of robust models for wildlife monitoring remains constrained by the scarcity of comprehensive, high-quality, and multimodal field datasets, especially with regards to providing accurate scale. Most existing datasets are limited to RGB images or videos, leaving a significant gap in the training and benchmarking of deep learning methods for ecological applications [17, 35, 42].

To address this gap, we present *WildDepth*, a unified multimodal wildlife dataset comprising three subsets collected in real-world environments that bridge the disciplines of ecology, computer vision, and artificial intelligence. The main innovation is the use of a synchronized RGB-lidar acquisition platform, common in automotive [7, 22, 39] or indoor applications [37, 49], but to the best of our knowledge, not trialled in the challenges of an outdoor, wild environment. The first subset was collected in December 2022 in the Kgalagadi Transfrontier Park, South Africa. A combination of RGB-LiDAR recordings using a zoom-lens camera setup, collected a wide range of wildlife data from apex predators such as lions, cheetahs and leopards to large herbivores such as springbok and gemsbok. The second subset was recorded across 4 days in June 2025 in the Bulyebe Valley Conservancy, Zimbabwe, combining LiDAR and RGB data under various natural illumination and environmental conditions. The third subset was collected in July 2025 at the Longleat Safari Park, United Kingdom, using a suite of multimodal sensors that included RGB cameras, GoPro Hero 11, and LiDAR systems that captured synchronized data. Together, these three subsets form *WildDepth*, one of the first multimodal field collections designed explicitly for wildlife image processing tasks such as pose estimation, 3D reconstruction, morphological analysis and depth prediction.

The objective of assembling these datasets is two-fold. First, they provide rich multimodal data for developing AI systems capable of fine-grained spatial and behavioral understanding of wildlife, particularly by allowing accurate scale to be determined from the paired RGB-LiDAR setup. Second, they serve as benchmark datasets to facilitate reproducible research across various conservation and ecological monitoring tasks. Each modality contributes complementary information, for example, LiDAR aids in accurate depth and morphological measurements [4]. The synchronized collection of RGB-LiDAR supports cross-modal learning, enabling models to infer structure and behavior from incomplete or noisy data. Beyond the focal animal capture, the scenes themselves are challenging for techniques such as MDE as they lack rectilinear shapes as would typically be encountered in urban environments.

In the broader context of wildlife conservation, integration of such datasets into AI workflows aligns with a growing recognition that traditional human observation-based methods are insufficient for large scale ecosystem monitoring. Initiatives like GBIF [20], eBird [38], and the IUCN Red List [24] have demonstrated the importance of open, standardized data sharing, but most rely on 2D or textual information, which limits their applicability to 3D and behavioral modeling. We note that there are excellent classification models available for animal species recognition [5, 15] and even individual recognition [36, 41]. However, the next step to enable true 3D modelling of animals within their wild habitats is lacking in representative datasets. By providing high resolution, multimodal sensor data, the proposed datasets offer new opportunities for advancing computational ecology, data-driven conservation, and automated biodiversity assessment. Our main aim is to enable better training of MDE and 3D reconstruction models that will allow rich, image based datasets (e.g. over 10M labelled images on Lila BC from wildlife camera traps [23]) to be more accurately scaled to 3D. This work draws inspiration from recent efforts in multimodal wildlife behavior datasets, such as MammAlps [14], which demonstrated the potential of integrating video, audio, and segmentation data for multi-view animal behavior analysis. However, our datasets differ in scope and modality, emphasizing 3D perception and cross-domain sensor fusion rather than camera-trap based video analytics. Through experiments on diverse wildlife monitoring tasks including depth and morphology estimation, behaviour analysis and 3D reconstruction, we demonstrate how multimodal sensor fusion enhances model robustness and ecological interpretability. Our contributions can be summarized as follows:

- WildDepth, a unified multimodal wildlife dataset with three subsets collected in ecologically distinct environments using synchronised RGB and LiDAR imaging systems.
- A set of benchmark experiments across depth estimation, behaviour analysis, and 3D reconstruction and densification tasks, establishing baselines for AI-driven conservation research.
- Insights into multimodal learning for ecological data, showcasing how combining heterogeneous sensors can improve accuracy and generalization in field-based wildlife monitoring.

2 Related Work

Recent years have witnessed a convergence between computer vision and ecology, driven by the demand for vision models capable of understanding general objects across domains in wild.

Vision-based Wildlife Monitoring. Traditional ecological monitoring relies heavily on camera traps, aerial imagery, or manual observation. These methods have inspired a series of computer vision applications for animal detection, identification, and behavioral tagging [33]. Datasets such as MammAlps [14], LoTE [25], and BaboonLand [12] have provided essential benchmarks for video-based behavior recognition, yet most are limited to RGB or audio inputs and lack geometric or metric-scale calibration. As a result, these datasets mainly support

classification rather than quantitative 3D reasoning. Furthermore, the majority of existing animal 3D meshing models are derived from internet-based video and image datasets for example, Animal Kingdom [30] and MammalNet [9], which capture the canonical appearance of species but lack metric ground truth.

Monocular Depth Estimation and 3D Understanding. Monocular Depth Estimation (MDE) has become a central topic in computer vision, progressing from classical stereo geometry toward self-supervised and transformer based learning. Models such as MonoDepth [21], DPT [31], DepthPro [6], and DepthAnything [47] have achieved remarkable results in structured indoor or urban environments (e.g., KITTI [22], NYUv2 [37]). However, these benchmarks make assumptions of stable and largely structured geometry, controlled lighting, and static objects - conditions rarely met in wild habitats. In contrast, wildlife scenes are highly unstructured, dominated by irregular terrain, vegetation, and non-rigid subjects that move across large depth ranges. Metric scaled ground truth from such environments is insufficient, which precludes rigorous evaluation of MDEs under realistic ecological conditions. The WildDepth dataset introduces precisely aligned RGB-LiDAR pairs in natural outdoor environments, offering one of the few sources of metric depth for small and distant targets, where monocular methods typically fail.

3D Reconstruction and 4D Deformable Modeling. The use computer vision algorithms in 3D reconstruction of objects including animals in motion extends beyond static depth prediction into the domain of deformable 4D perception, where geometry, texture, and time must be jointly modeled. While human motion capture and synthetic datasets (e.g., AMASS [28], Human3.6M [16]) have enabled rapid progress in human body modeling, equivalent resources for wild animals remain limited. Existing research on animal meshing or skeleton recovery (e.g., SuperAnimal [48], ChimpACT [26]) are trained on video data without metric calibration, preventing accurate validation of predicted poses or surface deformations. In such circumstances, the need for a dataset that provides synchronized 3D data of freely moving, deformable subjects in natural habitats is important and can provide scale awareness for accurate morphometrics.

Multimodal and Domain Robust Perception. The computer vision community has increasingly recognized the value of multimodal learning, combining RGB and depth information (such as from LiDAR, Radar or stereo vision) to enhance perception in various environments [46]. However, most multimodal datasets are still urban or indoor in nature, focusing on autonomous driving or robotics (e.g., KITTI-360 [22], nuScenes [7]). The ecological domain remains an outlier, it combines large dynamic ranges, deformable subjects, and environmental challenges that can degrade each modality differently.

Table 1 presents publicly available datasets and their properties comparative to WildDepth for wildlife monitoring tasks. By integrating RGB-LiDAR sensing across three geographically and ecologically distinct sites, the *WildDepth* dataset enables the study of cross-domain robustness and perception models transfer between savanna grasslands, and semi-controlled environments such as zoos. This

aspect is especially valuable for designing foundation models capable of adapting to unseen environmental conditions.

Table 1: Publicly available datasets for wildlife monitoring

Datasets	Source	# Species	# Frames	RGB	Depth
AnimalKingdom [30]	YouTube	850	30k	✓	✗
MammalNet [9]	YouTube	173	20k	✓	✗
KABR [18]	Drone	3	13k	✓	✗
Meerkats [34]	Zoo	1	848K	✓	✗
ChimpACT [26]	Zoo	1	160k	✓	✗
MammAlps [14]	Camera Traps	5	6k	✓	✗
LoTE [25]	Camera Traps	11	10k	✓	✗
BaboonLand [12]	Drone	1	30k	✓	✗
WildDepth	Camera/Lidar	29	202K	✓	✓

3 WildDepth

3.1 Data Collection

WildDepth comprises three subsets collected in 2022 and 2025 from semi-controlled safari park settings and natural, wild habitats, providing a diverse representation of animal appearances and environmental conditions. Table 2 present the WildDepth dataset statistics. We have carefully selected ‘snippets’ of diverse data to avoid duplication and provide a balanced dataset of approximately 7.5 hours, but also release our full, raw dataset. In total, the dataset required 24 days of field-work across three countries to be collected.

Kgalagadi Transfrontier Park, South Africa The first dataset component was captured in December 2022 in the Kgalagadi Transfrontier Park, South Africa. This data was collected using the WildPose system [29], a long-range 3D motion capture system combining a zoom-lens camera and LiDAR. The system was mounted on a vehicle door for a 13-day survey along fixed routes near the Two Rivers and Nossob camp sites. We recorded synchronized RGB-LiDAR sequences of land animals and large birds, with capture durations ranging from 4 to 345 seconds. This data features a wide range of wildlife, including apex predators like lions and cheetahs, as well as large herbivores such as springbok and gemsbok. To ensure data quality under varying field conditions, we maintained a high shutter speed (1/1000s) and actively managed the lens aperture (e.g., f/11–32) to balance adequate image brightness with sufficient depth of field.

Zimbabwe Wildlife Conservancy The second wild dataset was collected in June 2025 in the Buby Valley Conservancy, a private wildlife area in southern Zimbabwe dominated by mopane and acacia woodland habitats. We built a custom data collection rig comprising a 4K RGB camera linked with LiDAR (Livox Mid-70, DJI). Data were recorded using both a mobile setup, with the rig mounted on a vehicle window, and a stationary setup, where the rig was positioned near animals as they moved and grazed. From the data we collected, we curated 2.35 hours of synchronized RGB-LiDAR recordings of 6 species (giraffe, zebra,

warthog, donkey, goat, and plover birds). This dataset presents natural challenges including changing illumination, occlusions, and long-range motion, making it useful for assessing model robustness in real-world settings.

Longleat Safari Park, United Kingdom To complement the wild datasets, we collected an additional dataset at Longleat Safari Park in Wiltshire, UK, which hosts a wide variety of habituated animal species in spacious, open enclosures with relatively sparse vegetation. This setting allowed us to record a broader range of species under more controlled conditions and with lower scene complexity compared to wild environments. The data collection rig consisted of a RGB camera paired with a LiDAR (Livox Avia, DJI). Similar to our data collection in Zimbabwe, data were recorded using both a mobile setup, with the rig mounted on a vehicle window, and a stationary setup, where the rig was positioned on a tripod on a viewing platform overlooking an enclosure. In total, we recorded 7.5 hours of raw data, where after curation, we have 1.25 hours of RGB-LiDAR data comprising of 70 unique videos of different species (e.g. lion, tiger, cheetah, giraffe, ostrich, zebra, wolf, ox, elephant, rhino).

Across all three environments, the data were collected with precise temporal synchronization (± 10 ms tolerance) between the RGB and LiDAR streams. The combined dataset covers diverse species morphologies, ecological contexts, and terrain structures, producing dense geometric ground truth and a wide range of appearance variations. Together, these three subsets form WildDepth, a comprehensive multimodal wildlife dataset that supports the development and evaluation of perception models for depth estimation, 3D reconstruction, and behavioral understanding. The multi habitat data ensure that models trained on this dataset can generalize across distinct natural domains, from wild savannah and miombo woodlands to semi-controlled enclosures, advancing the goal of scalable and robust computer vision for wildlife conservation.

3.2 Data Processing

We utilize Robot Operating System (ROS) [27] as our framework for collecting data across different environments described above. All the collected data is processed for use in benchmarks, such as depth estimation and 3D reconstruction, which consist of calibrated pairs of RGB images and LiDAR scans. These provide a metric ground truth of depth for each task.

The LiDAR sensors are rated at a 1σ range precision of ≈ 2 cm (Livox TELE-15, Mid-70, and Avia for the Kgalagadi, Buby, and Longleat subsets, respectively). Crucially, for these prism-scanning time-of-flight sensors this precision does not scale as a fraction of range but stays roughly constant across our operating distances: for the TELE-15 it is specified in absolute distance bands (e.g. < 2 cm out to 70 m), while the Mid-70 and Avia are rated at ≤ 2 cm at their 20 m reference distance [1–3]. Prior validation of the WildPose acquisition system (used for the Kgalagadi subset) reports a median absolute error of 45.2 mm even at 160 m [29], orders of magnitude below the metric errors of the evaluated MDE models. We characterise these LiDAR tolerances in more detail in the supplementary material. Since the LiDAR scans at 10 frames per second (FPS), and the camera operates at 30 FPS, the synchronization between the camera

Table 2: WildDepth statistics (total and curation frames). A snippet is a sequence of consecutive frames captured during the same observation period that show the same species (or group of species) in a consistent background setting.

Subset	Species	Snippets	Curated Frames	Range (m)
Kgalagadi	26	131	75,741	12–140
Bubye	6	37	84,700	5–50
Longleat	10	70	42,500	5–50

and LiDAR is based on finding the closest matching frames of the camera for each LiDAR scan, based on their timestamps.

3.3 Data annotation

A semi-automated annotation workflow was developed to efficiently label large scale multimodal data. The process involved two main phases, object detection and tracking, and behavioral labeling. Initial animal detections were obtained using MegaDetector [5] fine-tuned on preliminary RGB samples. For 3D modalities (LiDAR), detections were projected into the 2D camera frame via calibrated extrinsics, and temporal consistency was implemented using ByteTrack [50]. These detections generated continuous tracks representing individual animals across frames and modalities. In the second phase, each frame was manually verified and annotated for species behavior (e.g., walking, grazing, resting). LiDAR point clouds were used to measure body shape and relative dimensions. Please refer to the supplementary material for further details on data annotation.

4 Experiment

This section presents the experimental evaluation conducted on the three subsets of WildDepth: Longleat Safari Park (UK), Kgalagadi Transfrontier Park (South Africa), and Bubye Valley Conservancy (Zimbabwe). Our experiments target the following key computer vision and ecological tasks: distance estimation, pose and behavior detection, and Target Point Densification and 3D reconstruction. Each task was designed to evaluate different aspects of the utility of multimodal data and to establish baselines for future research in wildlife-oriented perception and reconstruction. All experiments were implemented in PyTorch, trained with mixed precision on NVIDIA 3080Ti GPUs. Camera and LiDAR modalities were used for ground truth.

4.1 Depth Estimation

Accurate distance estimation in wildlife imagery is a critical prerequisite for depth-aware behavioral analysis, morphometrics and population abundance estimation (distance sampling). Traditional stereo or structured-light approaches are often impractical in dynamic, open environments where sensor placement and animal motion are uncontrolled. Therefore, monocular depth estimation models, pretrained on large scale synthetic or real-world datasets, were evaluated and adapted for wildlife scenes. Figure 1 represents the pipeline for depth estimation beginning with the acquisition and synchronization/calibration of raw RGB and LiDAR data. The RGB input is processed through different MDE

Models to generate predicted depth maps, while the LiDAR data serves as the high-precision ground truth. These maps are then cross-referenced to finalize the Depth Map Generation, which allows for a direct comparison between predicted and actual distances. We setup and benchmarked five state-of-the-art depth estimation models:

- DepthAnything [47] – a large-scale self-supervised model using unlabeled visual data for universal depth prediction.
- DPT [31] – a robust encoder-decoder architecture that transfers from vision transformers (ViT) for fine-grained depth recovery.
- DepthPro [6] – a diffusion-inspired depth model optimized for geometric consistency.
- MonoDepth2 [21] – a self-supervised photometric model that learns depth from stereo image pairs.
- MoGe2 [45] – a computer vision model that recovers accurate 3D geometry from a 2D image.

In all experiments, we used LiDAR depth maps as ground truth. For evaluation, we computed the RMSE, absolute relative error (AbsRel), and scale-invariant log error (SILog) in the test.

Table 3 and Figure 2 provide a per-distance breakdown of depth estimation performance across the benchmarked monocular models. As expected, error increases monotonically with range, highlighting the difficulty of long-range wildlife perception under unstructured field conditions. No single model consistently achieves the lowest RMSE, AbsRel, and Silog across all distance bins; the strongest model varies with range. However, even the strongest models exhibit substantial degradation beyond 20 m, where $\delta < 1.25$ drops significantly. This confirms that long-range deformable wildlife remains a challenging regime for monocular depth estimation and motivates the use of LiDAR anchoring or multimodal fusion for metric reliability.

Qualitative depth visualisations (Figure 3) confirm these observations: DepthAnything preserved smooth terrain gradients and correctly delineated body contours even under varying illumination, while DepthPro excelled at maintaining fine structural edges near boundaries (e.g., horns, legs or tails). DPT captured the scene layout coherently, but occasionally blurred small-scale features.

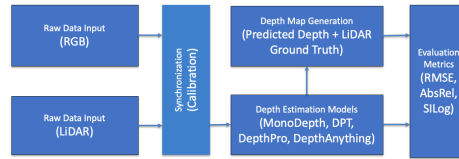


Fig. 1: Depth Estimation Pipeline

4.2 Behavior Recognition

For behavior detection, our aim is to identify naturalistic actions from unconstrained field imagery. Unlike controlled environments, our sequences exhibit

Table 3: Per-distance evaluation on the released WildDepth snippets (11,498 animal detections in the 5–50 m range). The different MDEs (uncalibrated (UC) and LiDAR-calibrated (LC)) camera-to-animal distance error. Lower is better for RMSE, AbsRel, and SIlog; higher is better for $\delta < 1.25$. LC values are bold.

Metric	Model	5–10m		10–20m		20–30m		30–50m	
		UC	LC	UC	LC	UC	LC	UC	LC
RMSE (m)	DepthAnything	6.10	1.61	6.73	3.29	9.93	4.35	15.86	10.44
	DA-v2	3.75	2.06	11.90	5.81	12.84	6.07	21.16	13.88
	DA-v3	8.79	2.01	8.79	3.23	11.21	4.27	16.28	8.98
	DepthPro	14.75	4.09	9.91	4.19	10.66	5.51	14.87	11.96
	DPT	14.27	4.45	13.77	5.45	12.42	5.11	14.21	10.52
	MonoDepth2	18.79	7.36	14.82	6.75	14.40	7.15	18.72	12.33
	MoGe2	6.68	2.07	8.41	4.36	11.78	4.17	23.17	8.08
AbsRel	DepthAnything	0.53	0.16	0.35	0.16	0.30	0.13	0.31	0.19
	DA-v2	0.36	0.19	0.43	0.22	0.41	0.19	0.42	0.29
	DA-v3	0.70	0.17	0.42	0.15	0.35	0.13	0.33	0.18
	DepthPro	1.10	0.38	0.43	0.20	0.31	0.17	0.30	0.24
	DPT	1.25	0.44	0.69	0.27	0.38	0.15	0.28	0.21
	MonoDepth2	1.52	0.68	0.74	0.32	0.44	0.21	0.39	0.26
	MoGe2	0.54	0.17	0.39	0.16	0.34	0.11	0.35	0.16
SIlog	DepthAnything	0.72	0.17	0.67	0.24	0.73	0.20	0.81	0.37
	DA-v2	0.73	0.23	1.02	0.26	1.07	0.20	1.13	0.38
	DA-v3	0.68	0.19	0.77	0.22	0.83	0.19	0.83	0.28
	DepthPro	0.73	0.25	0.68	0.28	0.89	0.25	0.74	0.33
	DPT	0.67	0.26	0.63	0.26	0.65	0.21	0.69	0.30
	MonoDepth2	0.79	0.39	0.96	0.33	0.89	0.29	1.01	0.42
	MoGe2	0.59	0.16	0.78	0.23	0.86	0.18	0.87	0.24
$\delta < 1.25$	DepthAnything	0.40	0.79	0.43	0.72	0.53	0.85	0.47	0.69
	DA-v2	0.40	0.71	0.39	0.57	0.38	0.57	0.43	0.39
	DA-v3	0.41	0.80	0.39	0.78	0.41	0.81	0.47	0.68
	DepthPro	0.31	0.45	0.41	0.74	0.52	0.68	0.48	0.51
	DPT	0.19	0.37	0.30	0.62	0.46	0.76	0.61	0.59
	MonoDepth2	0.16	0.29	0.24	0.54	0.36	0.65	0.39	0.49
	MoGe2	0.45	0.76	0.44	0.78	0.50	0.87	0.51	0.71

large appearance variations, occlusions, and motion blur due to natural lighting and unpredictable animal movements. To address these challenges, we adopt a spatiotemporal modeling approach that builds on recent advances in video-based behavior recognition. We train transformer-based classifiers following the architecture of VideoMAE [44] using RGB to recognize 12 activities (e.g., walking, resting, chasing) and 20 fine-grained actions (e.g., head-turning, grooming, grazing). Each video clip was 5-10 seconds long (≈ 150 frames) and paired with the corresponding Lidar motion signatures. A few examples from our dataset are given in Figure 4. Please refer to our supplementary materials for further details.

4.3 Target Point Densification

We follow a pipeline based on prior work on long-range wildlife 3D reconstruction [29] and subsequent thesis extensions. A state-of-the-art MDE (MoGe2 [45]) predicts the scene’s relative geometry from a single RGB frame, while a segmentation model (SAM2 [32]) provides the animal mask, which we visually inspected to ensure it isolates the target animal. This mask is used to filter the raw LiDAR

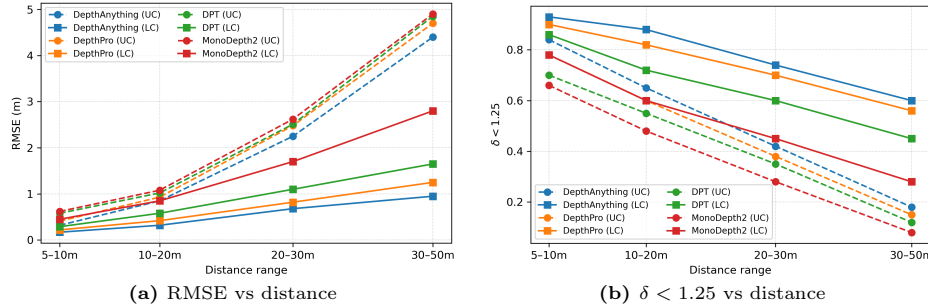


Fig. 2: Per-distance evaluation for MDE models. Solid lines shows the LiDAR-anchored calibration and dashed lines show uncalibrated predictions. LiDAR-anchored calibration of MDEs substantially reduces error and increases threshold accuracy, especially beyond 20 m.

point cloud, isolating a sparse set of metrically accurate “animal points” on the target.

To increase robustness, we include two filtering steps before alignment. First, to mitigate MDE artefacts common in telephoto imagery (e.g., “cliff-edge” transitions from slight defocus), we apply a Sobel-based boundary filter to the MDE output, removing unreliable depth estimates near strong gradients. Second, we employ a RANSAC-based Scale-Shift Alignment (SSA) algorithm that uses the filtered animal points as metric anchors to solve for a global affine transformation (scale and shift) that projects the dense MDE geometry into metric space.

We evaluate this baseline on a challenging real-world sequence of a walking red hartebeest at approximately 97 m, captured with the WildPose system [29]. To quantify performance, we measure the Mean Absolute Error (MAE) and the coefficient of determination (R^2) between the final metric depth map and the ground-truth LiDAR measurements. As a simple but informative baseline, we also consider a Planar Depth model that approximates the animal’s surface with a single 3D plane, fitted via RANSAC to the sparse LiDAR points within the animal’s mask.

Qualitatively, this RGB-LiDAR fusion baseline produces dense and coherent 3D point clouds that capture the overall shape and posture of the animal, preserving key anatomical features only sparsely sampled by LiDAR (Figure 5, bottom). Quantitatively, our analysis highlights the limitations of naive geometric models. As shown in Figure 6 and Table 4, the Planar Depth baseline attains a deceptively low MAE yet an R^2 that is consistently near or below zero. The low MAE is itself an artefact of long-range capture: because the animal’s depth extent is small relative to its ~ 97 m standoff, a single fitted plane lies close to the lowest-residual two-parameter fit and nearly attains the minimal achievable MAE by construction. R^2 , in contrast, normalises the residual by the depth variance induced by body shape, and so exposes what MAE hides. A negative R^2 is not a numerical artefact: by definition it means the model explains *less* variance than a trivial predictor that returns the single mean ground-truth depth of the animal—i.e. approximating an articulated animal by one frontal plane

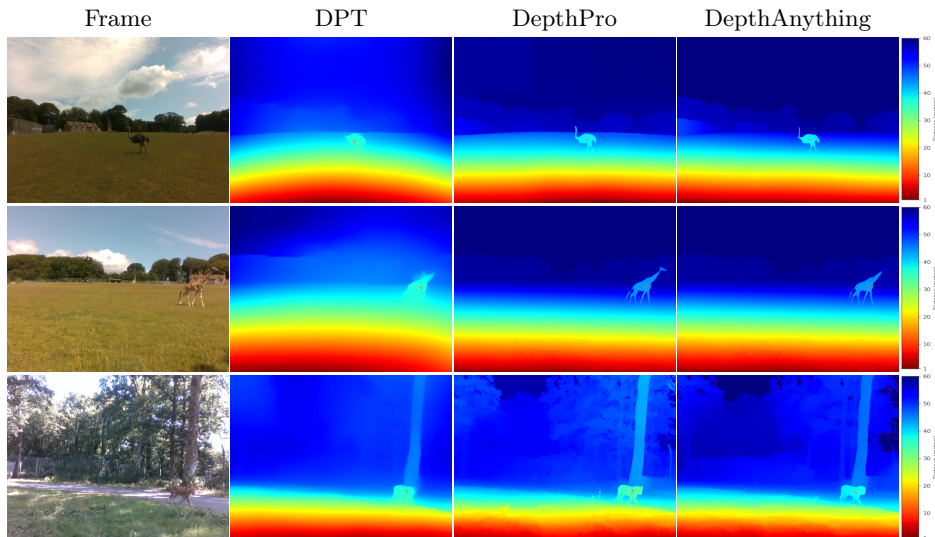


Fig. 3: Camera-to-animal distance by different MDEs for randomly selected frames from WildDepth

is worse than assuming a constant depth. The full RGB–LiDAR fusion (“Full”) reintroduces this non-planar structure from the MDE prior and yields a positive R^2 , showing that a simple global affine alignment anchored by sparse LiDAR recovers geometry that a planar model cannot.

Table 4: Ablation study on the Red Hartebeest sequence (~ 97 m). MAE is comparable across methods, but only the Full pipeline achieves a positive R^2 ; a negative R^2 denotes a fit worse than predicting the constant mean depth. DBF: depth-boundary filter; SDPS: static–dynamic point segmentation; SSA: scale–shift alignment.

Method	MAE ($\downarrow$)	R^2 ($\uparrow$)
Planar Depth	0.13 ± 1.14	-0.10 ± 0.13
Naive SSA	2.40 ± 2.32	-17.19 ± 316.05
+ DBF	1.51 ± 1.53	-0.01 ± 0.22
+ SDPS	0.25 ± 1.28	-4.79 ± 28.65
Full (ours)	0.16 ± 1.12	0.25 ± 1.36

4.4 3D reconstruction

To reconstruct the animals 3D form, for each video, we first isolate the target animal in every frame using SAM2 [32], which provides high-quality segmentation masks. These masks are used to extract and stabilize the foreground region across time, effectively reducing background clutter and motion inconsistencies. The resulting masked monocular clips serve as clean inputs for 4D reconstruction. We then initialize a canonical set of 3D Gaussians in a reference space and employ Gaussian-Splash [13], a Gaussian-splatting-based framework, to optimize a time-dependent deformation field. This field continuously warps

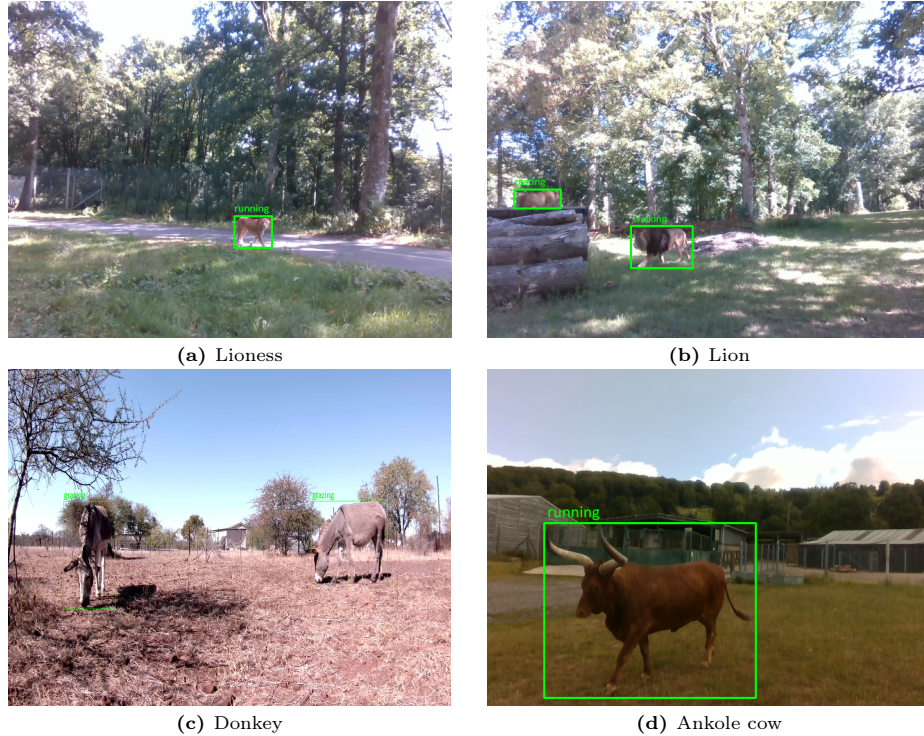


Fig. 4: Behaviour recognition for randomly selected frames from our WildDepth dataset.

the canonical representation to match the observed frames, capturing both fine geometric details and temporal dynamics. Through this process, we obtain temporally consistent 4D reconstructions that preserve the appearance, motion, and structure of the animal across the sequence. The combination of SAM2 based segmentation and Gaussian-Splash reconstruction enables robust modeling even under complex backgrounds and nonrigid motion typical of real-world animal videos. Figure 7 shows random frames and its reconstruction examples.

5 Discussion

WildDepth introduces one of the first multimodal RGB-LiDAR datasets for long-range wildlife perception, providing a testbed for evaluating modern vision models in challenging field conditions. Our experiments reveal several key insights. First, modern monocular depth estimators transfer surprisingly well to wildlife scenes, though no single model dominates across all metrics, with the best performer shifting with range. Second, we show that even extremely sparse LiDAR returns can be fused with dense monocular predictions to recover metrically accurate geometry. Our RGB-LiDAR densification case study on a walking red hartebeest at nearly 100 m confirms that a simple global alignment, anchored by LiDAR, can produce dense depth maps with positive R^2 scores, correctly

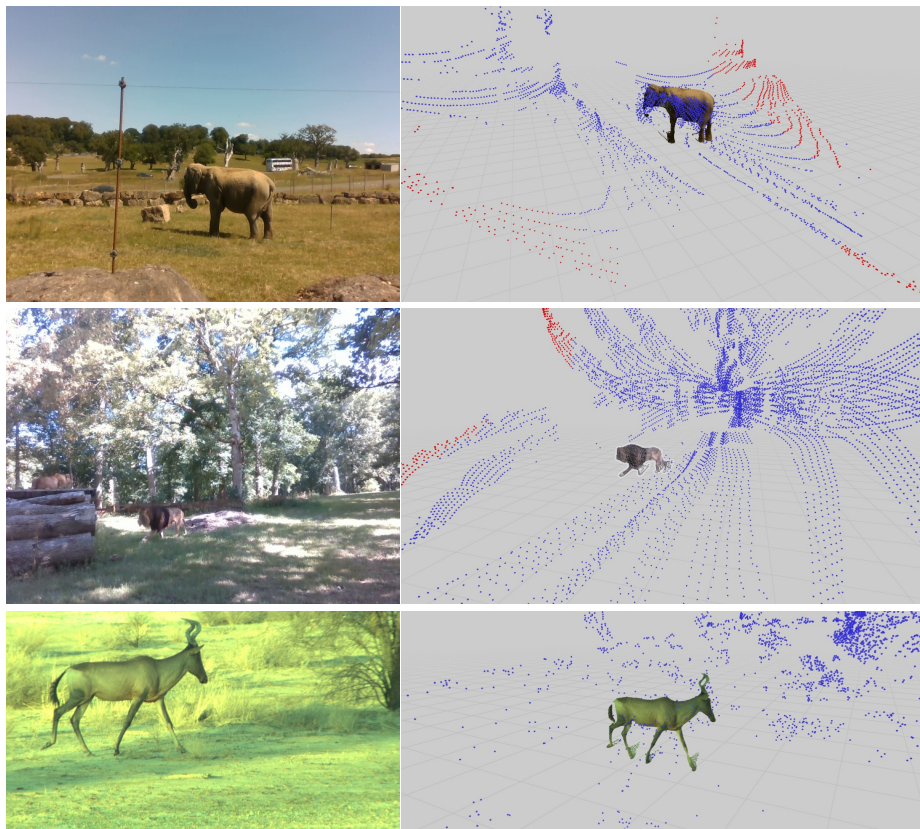


Fig. 5: Qualitative results of target point densification. Each row shows an input RGB frame (left) and the corresponding dense 3D point cloud (right). The Elephant (top) and Lion (middle) are from the Longleat subset, while the Red Hartebeest (bottom) is from the Kalahari subset.

capturing non-planar structure where naive geometric baselines fail. Finally, our qualitative results for 4D reconstruction, which combine SAM2 with a Gaussian-splatting framework, suggest that temporally consistent 3D modeling of animal motion in the wild is becoming attainable.

Our per-distance evaluation (Table 3 and Figure 2) shows that depth error grows steadily with range, and that even the strongest monocular model degrades well before the far end of our captures. We read this as a distribution shift rather than a weakness of any single architecture, since monocular estimators are trained largely on short-range indoor and automotive scenes, whereas WildDepth observes animals from tens of meters through telephoto lenses that suppress texture. This is also the regime where LiDAR anchoring helps most, since calibrating against even sparse LiDAR returns reduces error markedly at longer range and re-anchors the metric scale that monocular depth alone cannot recover at a long standoff.

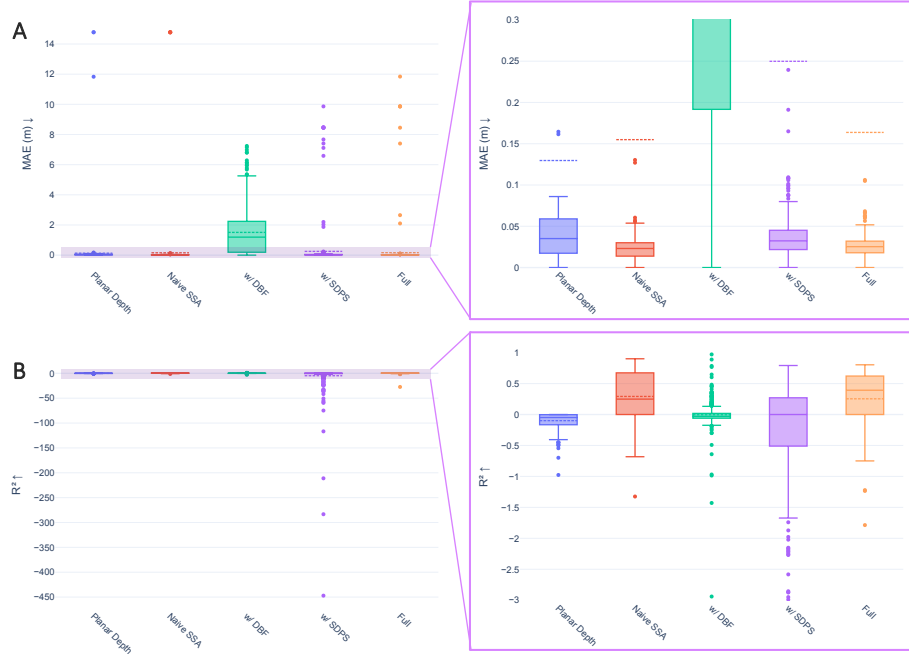


Fig. 6: Per-frame evaluation metrics for the Red Hartebeest sequence across the den-sification ablation. (A) shows the MAE and (B) the R^2 , and the right plot of each row zooms into the shaded near-zero band. Each variant is drawn in a fixed colour. Planar Depth is blue, Naive SSA is red, the variant with depth-boundary filtering (DBF) is green, the variant with static-dynamic point segmentation (SDPS) is purple, and the Full pipeline is orange. Within each plot the box spans the interquartile range, the solid line marks the median, the dashed line marks the mean, and the dots are individual frames. The tall blue and red dots in panel (A) are outlier frames from Planar Depth and Naive SSA. Only the Full method keeps the R^2 consistently positive.

However, we acknowledge several limitations that frame important directions for future work. The current dataset is geographically restricted to Southern Africa and the UK, and does not yet include quantitative benchmarks for higher-level tasks such as pose estimation or behavior analysis, which we leave for future releases. Our experimental validation of RGB-LiDAR fusion is based on a single, albeit challenging, sequence, and its generalization across species and environments requires further investigation. A related limitation is that our metrics aggregate over frames in which the animal’s articulation and orientation vary, since a side-on view exposes more body curvature than a near-frontal pose, and the per-frame R^2 in Figure 6 is correspondingly wide, which is why we report full-sequence distributions rather than a few hand-picked frames that would overstate accuracy. Future work should therefore focus on four key areas: (1) expanding WildDepth’s coverage to new ecosystems and species while enriching annotations for pose and behavior; (2) performing systematic cross-dataset and

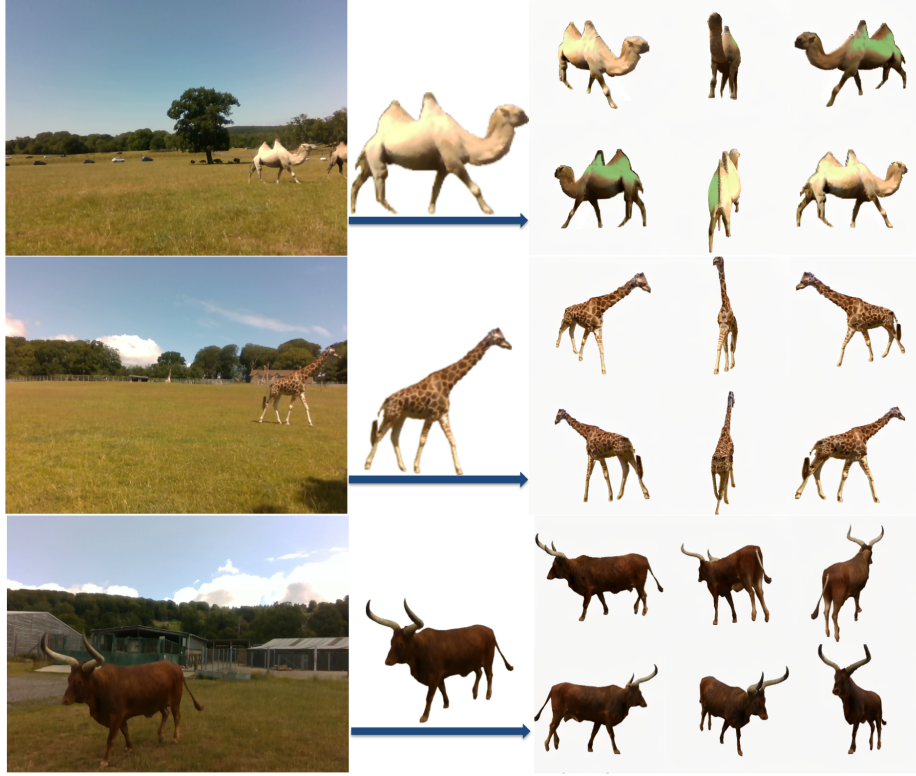


Fig. 7: 3D reconstructions of walking Camel (top), Giraffe (middle), and Ankole cow (bottom).

multimodal ablation studies to quantify domain gaps and the precise utility of LiDAR; (3) connecting geometric outputs to downstream ecological indicators, such as body condition or population density, in collaboration with conservation practitioners; and (4) disentangling shape error from pose, for example by normalizing against tracked keypoints or extending to multi-view capture.

In conclusion, WildDepth provides a new resource and a set of conservative baselines that bridge the gap between standard computer vision tasks and the realities of ecological fieldwork. We hope this work will stimulate further research in robust multimodal perception and 4D understanding of animal behavior in the wild, ultimately contributing to global biodiversity conservation.

6 Acknowledgment

This work was supported by the Natural Environment Research Council [grant number: UKRI037]. The authors extend their sincere gratitude to the management and research staff of the Buby Valley Conservancy, Zimbabwe, South African National Parks, South Africa and Longleat Safari Park, UK for their support and assistance with data collection.

References

1. Livox Avia User Manual
2. Livox TELE-15 User Manual
3. Livox Mid-70 User Manual (Feb 2021)
4. Andualet, T.G., Peters, S., Hewa, G.A., Myers, B.R., Boland, J., Pezzaniti, D.: Channel morphological change monitoring using high-resolution lidar-derived dem and multi-temporal imageries. *Science of The Total Environment* **921**, 171104 (2024)
5. Beery, S., Morris, D., Yang, S.: Efficient pipeline for camera trap image review. *CoRR* **abs/1907.06772** (2019), <http://arxiv.org/abs/1907.06772>
6. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073* (2024)
7. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11621–11631 (2020)
8. Ceballos, G., Ehrlich, P.R., Dirzo, R.: Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *Proceedings of the National Academy of Sciences* **114**(30), E6089–E6096 (Jul 2017). <https://doi.org/10.1073/pnas.1704949114>
9. Chen, J., Hu, M., Coker, D.J., Berumen, M.L., Costelloe, B., Beery, S., Rohrbach, A., Elhoseiny, M.: Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 13052–13061 (2023)
10. Davies, A.B., Asner, G.P.: Advances in animal ecology from 3D-LiDAR ecosystem mapping. *Trends in Ecology & Evolution* **29**(12), 681–691 (Dec 2014). <https://doi.org/10.1016/j.tree.2014.10.005>
11. Delplanque, A., Foucher, S., Théau, J., Bussière, E., Vermeulen, C., Lejeune, P.: From crowd to herd counting: How to precisely detect and count African mammals using aerial imagery and deep learning? *ISPRS Journal of Photogrammetry and Remote Sensing* **197**, 167–180 (Mar 2023). <https://doi.org/10.1016/j.isprsjprs.2023.01.025>
12. Duporge, I., Kholiavchenko, M., Harel, R., Wolf, S., Rubenstein, D.I., Crofoot, M.C., Berger-Wolf, T., Lee, S.J., Barreau, J., Kline, J., et al.: Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos. *International Journal of Computer Vision* pp. 1–12 (2025)
13. Feng, Y., Feng, X., Shang, Y., Jiang, Y., Yu, C., Zong, Z., Shao, T., Wu, H., Zhou, K., Jiang, C., et al.: Gaussian splashing: Unified particles for versatile motion synthesis and rendering. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 518–529 (2025)
14. Gabeff, V., Qi, H., Flaherty, B., Sumbul, G., Mathis, A., Tuia, D.: Mammalps: A multi-view video behavior monitoring dataset of wild mammals in the swiss alps. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 13854–13864 (2025)
15. Gadot, T., Istrate, S., Kim, H., Morris, D., Beery, S., Birch, T., Ahumada, J.: To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images (Oct 2024). <https://doi.org/10.1049/cvi2.12318>
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments.

- IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
17. Jiang, L., Lee, C., Teotia, D., Ostadabbas, S.: Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Computer Vision and Image Understanding* **222**, 103483 (Sep 2022). <https://doi.org/10.1016/j.cviu.2022.103483>
 18. Kholiavchenko, M., Kline, J., Ramirez, M., Stevens, S., Sheets, A., Babu, R., Banerji, N., Campolongo, E., Thompson, M., Van Tiel, N., et al.: Kabr: In-situ dataset for kenyan animal behavior recognition from drone videos. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 31–40 (2024)
 19. Kumar, S., Zhang, B., Gudavalli, C., Levenson, C., Hughey, L., Stabach, J.A., Amoke, I., Ojwang, G., Mukeka, J., Mwiu, S., Ogutu, J., Frederick, H., Manjunath, B.S.: WildlifeMapper: Aerial Image Analysis for Multi-Species Detection and Identification
 20. Lane, M.A., Edwards, J.L.: The global biodiversity information facility (gbif). *Systematics Association special volume* **73**, 1 (2007)
 21. Li, C., Yue, C., Liu, Y., Bie, M., Li, G., Lv, Z., Li, J.: An improved monodepth2 algorithm for vehicle monocular depth estimation. *Optik* **311**, 171936 (2024)
 22. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(3), 3292–3310 (2022)
 23. LILA Sciences: LILA Sciences: Labeled Information Library of Alexandria (2026), <https://lila.science/>, accessed: 2026-03-03
 24. List, I.R.: Iucn red list (2011)
 25. Liu, D., Hou, J., Huang, S., Liu, J., He, Y., Zheng, B., Ning, J., Zhang, J.: Lote-animal: A long time-span dataset for endangered animal behavior understanding. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 20064–20075 (2023)
 26. Ma, X., Kaufhold, S., Su, J., Zhu, W., Terwilliger, J., Meza, A., Zhu, Y., Rossano, F., Wang, Y.: Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. *Advances in Neural Information Processing Systems* **36**, 27501–27531 (2023)
 27. Macenski, S., Foote, T., Gerkey, B., Lalancette, C., Woodall, W.: Robot operating system 2: Design, architecture, and uses in the wild. *Science robotics* **7**(66), eabm6074 (2022)
 28. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5442–5451 (2019)
 29. Muramatsu, N., Shin, S., Deng, Q., Markham, A., Patel, A.: WildPose: A long-range 3D wildlife motion capture system. *Journal of Experimental Biology* **228**(5), JEB249987 (Mar 2025). <https://doi.org/10.1242/jeb.249987>
 30. Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J.: Animal kingdom: A large and diverse dataset for animal behavior understanding. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 19023–19034 (2022)
 31. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 12179–12188 (2021)

32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: SAM 2: Segment Anything in Images and Videos (Aug 2024). <https://doi.org/10.48550/arXiv.2408.00714>
33. Ravoor, P.C., Sudarshan, T.: Deep learning methods for multi-species animal re-identification and tracking—a survey. *Computer Science Review* **38**, 100289 (2020)
34. Rogers, M., Gendron, G., Valdez, D.A.S., Azhar, M., Chen, Y., Heidari, S., Perelini, C., O’Leary, P., Knowles, K., Tait, I., et al.: Meerkat behaviour recognition dataset. arXiv preprint arXiv:2306.11326 (2023)
35. Saad Saoud, L., Sultan, A., Elmezain, M., Heshmat, M., Seneviratne, L., Hussain, I.: Beyond observation: Deep learning for animal behavior and ecological conservation. *Ecological Informatics* **84**, 102893 (Dec 2024). <https://doi.org/10.1016/j.ecoinf.2024.102893>
36. Shinoda, R., Shiohara, K.: Petface: A large-scale dataset and benchmark for animal identification. In: *European Conference on Computer Vision*. pp. 19–36. Springer (2025)
37. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: *European conference on computer vision*. pp. 746–760. Springer (2012)
38. Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S.: ebird: A citizen-based bird observation network in the biological sciences. *Biological conservation* **142**(10), 2282–2292 (2009)
39. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in Perception for Autonomous Driving: Waymo Open Dataset (May 2020). <https://doi.org/10.48550/arXiv.1912.04838>
40. 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., Berger-Wolf, T.: Perspectives in machine learning for wildlife conservation. *Nature Communications* **13**(1), 792 (Feb 2022). <https://doi.org/10.1038/s41467-022-27980-y>
41. do Val, H.G.P., Pezzuti, J.C.B., Carvalho-Jr, E.A.R.: Identifying ocelots by individual coat pattern marks using the Whiskerbook. *Mammalogy Notes* **10**(2), 413–413 (Sep 2024). <https://doi.org/10.47603/mano.v10n2.413>
42. Vally, A., Maswoswere, G., Bowden, N., Paine, S., Amayo, P., Markham, A., Patel, A.: M2S2: A Multi-Modal Sensor System for Remote Animal Motion Capture in the Wild. *IEEE Sensors Letters* pp. 1–4 (2025). <https://doi.org/10.1109/LSENS.2025.3542233>
43. Vogg, R., Lüddecke, T., Henrich, J., Dey, S., Nuske, M., Hassler, V., Murphy, D., Fischer, J., Ostner, J., Schülke, O., Kappeler, P.M., Fichtel, C., Gail, A., Treue, S., Scherberger, H., Wörgötter, F., Ecker, A.S.: Computer vision for primate behavior analysis in the wild. *Nature Methods* **22**(6), 1154–1166 (Jun 2025). <https://doi.org/10.1038/s41592-025-02653-y>
44. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 14549–14560 (2023)

45. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: MoGe-2: Accurate Monocular Geometry with Metric Scale and Sharp Details (Jul 2025). <https://doi.org/10.48550/arXiv.2507.02546>
46. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
47. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
48. Ye, S., Filippova, A., Lauer, J., Schneider, S., Vidal, M., Qiu, T., Mathis, A., Mathis, M.W.: Superanimal pretrained pose estimation models for behavioral analysis. *Nature communications* **15**(1), 5165 (2024)
49. Yeshwanth, C., Liu, Y.C., Niekner, M., Dai, A.: ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes (Aug 2023). <https://doi.org/10.48550/arXiv.2308.11417>
50. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: European conference on computer vision. pp. 1–21. Springer (2022)