Available Online: 29 August 2025 DOI: 10.54254/2977-3903/2025.26435

Animal group behavior analysis and pose estimation based on deep learning

Jiayi Zhou

SWJTU-LEEDS JOINTSCHOOL, Southwest Jiaotong University, Chengdu, China

ryann0712@outlook.com

Abstract. Animal behavior analysis plays a pivotal role in neuroscience, behavioral ecology, animal welfare, and precision agriculture. However, traditional manual observation methods are often subjective, labor-intensive, and insufficient for large-scale quantification. The advent of deep learning has revolutionized this field, enabling automated, high-throughput and accurate analysis particularly in complex group settings. This review provides a comprehensive overview of recent advances in deep learning-based animal group pose estimation and behavior analysis. It systematically outlines the key stages from data acquisition to behavior interpretation, including object detection, multi-animal tracking, pose estimation, and individual identification. Representative models and tools are critically evaluated, along with their applications across various species and experimental contexts. While notable advancements have been attained—including refined occlusion handling via part affinity fields and augmented temporal behavior recognition through video transformers—several core challenges persist. These include robustness in wild environments, rare behavior detection and long-term identity preservation. Future research should focus on end-to-end joint modeling, data-efficient learning paradigms and multimodal data integration for advancing robust and intelligent systems. This review aims to provide researchers with a panoramic view of the field, highlighting key methodologies and directions for future development.

Keywords: group behavior, deep learning, pose estimation, multi-animal tracking, behavior recognition

1. Introduction

For social animals, group behaviors such as the processes of collective decision-making, and sophisticated social interactions reveal system dynamics that far exceed the complexity of individual behavior. Therefore, accurate and efficient analysis of group behavior is essential for understanding the biological foundation, social organization and operational mechanisms of whole ecosystems [1, 2]. Nevertheless, conventional methodologies in animal behavior research face substantial limitations, such as time-consuming and labor-intensive observational procedures, which hinder the implementation of large-scale quantitative analyses. Conventional approaches based on sensors can record certain aspects of individual activity, but fall short in capturing critical spatial relationships among group members, which is particularly limited when analyzing complex behaviors such as the coordinated movement of fish schools [3].

Presently, computer vision technologies with deep learning at their core have yielded transformative breakthroughs in addressing this dilemma. By enabling automated and high-throughout analysis of video, researchers could objectively quantify animal behavior with unprecedented precision. Most importantly, the advent of markerless pose estimation techniques based on deep neural networks is seen as a significant milestone. These technologies not only enable precise tracking of individuals but also grasp the dynamic postures of most members within a group, providing a robust data foundation for embedded analysis of collective social interactions.

Despite rapid technological advancements, substantial challenges persist in applying these methodologies to complex group scenarios — particularly in contexts involving severe inter-individual occlusion, high visual similarity, and analogous complexities. Accordingly, this study seeks to provide researchers with an overview of the technical landscape and future research directions by synthesizing key methodologies, representative applications, and core challenges in group-level pose estimation and behavior analysis. Meanwhile, it highlights pressing real-world problems for scholars in the computer vision community, thus promoting interdisciplinary collaboration and fostering scientific discovery in the field of animal group behavior.

2. Key technologies and core models

2.1. Data acquisition and preprocessing

Deep learning analysis of group animal behavior is a multi-stage technical pipeline. Moreover, the initiation of this process lies in the acquisition of image data, whose quality directly dictates the efficacy of all subsequent analyses. Modern studies extensively employ both 2D and 3D imaging systems, ranging from high-resolution fixed cameras with laboratory settings to large-scale aerial imaging using drones in outdoor environments [2, 3]. In contrast to the relatively controllable backgrounds in laboratory settings, data acquisition in farm and wild environments confronts substantial challenges, including fluctuating lighting conditions, intricate backgrounds, and highly variable animal postures. These factors also impose stringent requirements on the robustness of computer vision models [1, 3].

2.2. Animal detection and multi-target tracking

When datasets are acquired, the primary task is to accurately detect and constantly track each individual within the group across complex scenes. The process commonly follows the tracking-by-detection paradigm. Specifically, object detection algorithms are first utilized to localize all animals within each frame, followed by the employment of data association methodologies to link detections across frames into continuous individual trajectories [1].

Nevertheless, the high degree of phenotypic similarity among individuals, frequent physical interactions, and occlusions render the long-term preservation of identity a pivotal challenge. To address this, advanced tracking frameworks have begun to incorporate re-identification modules. For instance, Lauer et al. Proposed a multitask learning architecture that simultaneously performs pose estimation and predicts individual identity. This enables more precise re-identification and re-linking of trajectories subsequent to brief occlusions, a capability that is imperative for analyzing long-term social interaction patterns [4].

2.3. Key point pose estimation

Pose estimation lies at the core of the entire analytical pipline, aiming to lo locate the predefined key points on each animal's body. For individual subjects, the framework exemplified by LEAP [5] transforms input images into a sequence of keypoint "heat maps" via deep convolutional neural networks—specifically encoder-decoder architectures—where the brightness of each pixel in the map denotes the probability of keypoint occurrence at that location, thereby enabling precise localization. Upon expansion to group scenarios, task complexity escalates drastically. Current mainstream solutions obey two primary paradigms. The top-down approach first detects each individual and subsequently estimates poses independently, whereas the bottom-up approach identifies all keypoints within the image and assembles them into individual skeletons using spatial cues such as Part Affinity Fields (PAFs) [1]. The bottom-up method is particularly effective in crowded scenes with heavy overlap, which does not depend on accurate individual detection. This method has made significant progress in handling occlusion issues in dense scenes through PAFs and performs well in laboratory environments. However, its robustness still needs to be further improved in extremely dense groups or complex backgrounds such as outdoor or underwater scenes.

2.4. Behavior recognition and analysis

Animal behavior analysis which represent the ultimate goal of pose estimation seeks to convert visual data into meaningful behavioral labels. After obtaining individual detection, two main paradigms are adopted by researchers.

The first approach is based on keypoint sequence analysis. Pose estimation tools such as DeepLabCut are used to extract spatiotemporal trajectories of body keypoints, resulting in high-dimensional time-series data [4]. These sequences are then input into temporal models such as recurrent neural networks, long short-term memory networks or Transformers to perform supervised behavior classification, identifying predefined behaviors such as walking or standing [3]. Furthermore, this paradigm also supports unsupervised behavior discovery, where dimensionality reduction and clustering are applied to pose data in the absence of predefined labels, enabling automatic identification and segmentation of species-specific stereotyped behaviors [5]. This method has shown remarkable success in controlled laboratory settings with model organisms. The second approach is an end-to-end video-based method, which bypasses explicit keypoint extraction by feeding raw video clips directly into spatiotemporal learning models—such as 3D convolutional networks or video Transformers like SlowFast and its improved variant ViTAM-SlowFast to recognize complex behavior patterns [6]. This method excels at capturing subtle visual cues that may be difficult to represent with keypoints alone, such as interactions between individuals and objects or fine-grained texture changes.

As tasks evolve, analysis extends to temporal action localization, spatio-temporal detection and dynamic scene graph generation. The techniques will support detailed behavior understanding across both controlled environments and challenging

wildlife monitoring.

3. Major application areas and public datasets

3.1. Neuroscience and basic biology

In neuroscience and basic biology, these technologies are widely employed in controlled laboratory settings to enable accurate and high-throughput quantification of behavior in model organisms such as mice and fruit flies. Techniques such as DeepLabCut [4] and LEAP [5] empower researchers to track three-dimensional movement trajectories of multiple keypoints on an animal's body with unprecedented precision, thereby enabling the quantification of a broad spectrum of behaviors—spanning from gait dynamics and grooming patterns to complex social interactions. These high-resolution behavioral datasets provide a solid foundation for establishing direct associations between neural activities such as those recorded through optical fibers or obtained through two-photon imaging and specific behavioral outputs, greatly enhancing our understanding of the neural circuit mechanisms behind behaviors [7].

3.2. Smart agriculture and aquaculture

Building on their success in neuroscience, these technologies have also found significant applications in smart agriculture and aquaculture, where the goal shifts from basic research to real-world monitoring and management. In the field of smart agriculture, these technologies offer powerful solutions for non-invasive, automated monitoring of animal health and welfare. In livestock farming, continuous monitoring of group behavior such as that of pigs enables the analysis of changes in feeding, drinking, lying, and activity patterns. Such behavioral insights facilitate the early detection of disease symptoms or stress responses, supporting precision management and timely intervention [6]. In aquaculture, computer vision models are employed to address a range of unique challenges, including the automatic identification of underwater fish species, biomass estimation and the analysis of key behaviors such as feeding and stress. These applications are helpful for optimizing feeding strategies, monitoring water quality and detecting abnormal behaviors of fish populations, thereby enhancing both production efficiency and animal welfare [3].

3.3. Ecology and wildlife conservation

Extending further into natural environments, these technologies are increasingly applied in ecology and wildlife conservation to observe animal behavior in the wild. Utilizing drones, infrared cameras, and camera traps, researchers can conduct long-term, non-invasive observations of wild animals such as primates and ungulates [1, 2]. This enables the study of social structures, foraging strategies, migration patterns, and responses to human activity and climate change without disrupting natural behavior. For example, facial recognition and tracking of wild chimpanzees have allowed for in-depth analysis of their complex social networks and kinship relationships [8].

3.4. Public datasets and benchmarks

The development and deployment of these applications across diverse domains have been robustly underpinned by the growing accessibility of public datasets and benchmark resources. Standardized, high-quality open datasets are essential for driving algorithmic innovation, ensuring research reproducibility, and enabling fair comparisons across different models. In recent years, several representative datasets have emerged in the field. For instance, Lauer et al. released benchmark datasets featuring multiple species—including mice, fish, and marmosets—in complex social environments [4]; Vogg et al. focused on behavioral analysis of wild primates [1]; additionally, specialized datasets such as Fish4Knowledge, which target specific application scenarios, have garnered widespread recognition [1]. Additionally, researchers often leverage large-scale and general-purpose vision datasets like MS COCO [9] for model pretraining, thereby enhancing generalization to animal-specific tasks. The open sharing of these datasets has significantly accelerated progress across the domain.

4. Challenges and solutions

4.1. Open problems

Despite the remarkable progress in deep learning-based analysis of group animal behavior, several key challenges still need to be addressed including the high cost of data annotation, severe occlusion issues and problems related to the accuracy of cross-domain generalization capabilities in complex real-world scenarios.

At the data level, the exorbitant annotation cost constitutes the primary bottleneck impeding the widespread adoption of this technology. Whether it is identifying primates in wildlife footage or precisely marking dozens of body keypoints in laboratory mice, the process requires a large amount of human effort and expert domain knowledge. This presents a considerable obstacle for large-scale, long-term studies [1,6]. Compounding this issue are problems of small sample size and long-tail distribution. Many rare but crucial behaviors occur extremely infrequently in datasets, making it difficult for the model to learn meaningfulrepresentations. Additionally, the cross-domain generalization remains a huge challenge. Models trained in a controlled laboratory setting tend to exhibit a significant performance degradation when directly deployed to real-world environments featuring fluctuating lighting, intricate backgrounds, and frequent occlusions [3,6], or to underwater scenarios characterized by low contrast and high noise levels [3]. Extending further into the model level, two central challenges are severe occlusion in dense group and high visual similarity among individuals. When animals aggregate closely, frequent body occlusion makes the localization of key points and the continuous tracking of individual identities extremely difficult [6,7]. Bottom-up methodologies such as those reorganizing body components via Part Affinity Fields offer a viable strategy for addressing occlusion challenges [9], yet their robustness remains to be further enhanced. Although some affinity field methods have made progress in handling group occlusion, in the wild environment, the occlusion problem becomes more complex due to changes in light, complex backgrounds, and high-density groups. Moreover, accurately reconstructing 3D poses from 2D videos remains a technical challenge, requiring more advanced depth estimation and multi-view fusion techniques [6]. Finally, computational efficiency remains a critical concern. While tools like LEAP offer fast inference [5], many high-accuracy models are too resource-intensive for real-time applications, limiting their use in closed-loop experiments or field-based monitoring systems.

4.2. Future research directions

To address the above-mentioned challenges, future research should focus on developing integrated and intelligent analysis frameworks that directly tackle data sparsity, model robustness and scene complexity. A critical direction is end-to-end joint modeling which combines detection, tracking, pose estimation and behavior recognition within a single network. This unified method can mitigate error accumulation in conventional pipelines, especially in dense and occluded group settings [1,4,6].

Given the reliance on large annotated datasets, data-efficient learning is crucial. Techniques such as self-supervised learning on unlabeled field videos can help bridge the lab-to-field domain gap [2,6], while active learning enables efficient annotation by prioritizing the most informative samples [1, 5]. In visually adversarial environments such as turbid water [2] or high occlusion scenarios [4,6], multimodal fusion—specifically the integration of video, audio, depth, and thermal data—enhances perceptual capabilities and analytical precision.

Ultimately, the objective is to advance from mere behavioral description toward deeper behavioral comprehension.This requires developing models with higher interpretability and causal inference capabilities. It is imperative not only to discern behavioral patterns but also to explore the underlying drivers, decision-making mechanisms, and social dynamics that underpin them, thereby truly attaining a rigorous scientific comprehension of the “why” behind animal group behavior [10].

5. Conclusion

In conclusion, this review has systematically charted the landscape of deep learning-based animal group pose estimation and behavior analysis, underscoring its transformative impact across diverse scientific domains. From high-throughput quantification of social interactions in neuroscience to non-invasive welfare monitoring in precision agriculture and large-scale behavioral studies in wildlife conservation , deep learning has provided a powerful paradigm for overcoming the limitations of traditional methods. We have detailed the key technological pipeline, from data acquisition and multi-target tracking to sophisticated pose estimation and behavior recognition, critically assessing cornerstone frameworks like DeepLabCut, LEAP and SlowFast.

Despite these significant advancements, the field is not without its challenges. The successful implementation of these technologies—particularly within complex group scenarios and unstructured natural environments—remains impeded by challenges such as severe occlusion, exorbitant annotation costs, and the imperative for enhanced model generalization. These challenges, however, illuminate clear directions for future innovation. As we have discussed, the development of end-to-end joint modeling frameworks, the adoption of data-efficient learning strategies, and the integration of multimodal data hold immense promise for building more robust, scalable, and accurate systems.

Ultimately, the goal of this field extends beyond mere quantification to a deeper, mechanistic understanding of animal behavior. The journey from pose estimation to behavioral analysis is a journey from "what" to "why". By fostering closer interdisciplinary collaboration between computer vision experts and biologists, we can continue to refine these powerful tools, pushing the boundaries of what is possible in the automated analysis of animal behavior. This synergy will not only accelerate scientific discovery but also provide critical insights for addressing pressing global challenges, from biodiversity conservation to sustainable food production.

References

[1] 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. (2024). Computer Vision for Primate Behavior Analysis in the Wild. arXiv preprint arXiv: 2401.16424.

[2] Koger B, Deshpande A, Kerby JT, Graving JM, Costelloe BR, Couzin ID. (2023) Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*, 92: 1357–1371.

[3] Yang L, Liu Y, Yu H, Fang X, Song L, Li D, Chen Y. (2021) Computer vision models in intelligent aquaculture with emphasis on fish detection and behavior analysis: A review. *Archives of Computational Methods in Engineering*, 28: 2785–2816.

[4] Lauer J, Zhou M, Ye S, Menegas W, Schneider S, Nath T, Rahman MM, Di Santo V, Soberanes D, Feng G, Murthy VN, Lauder G, Dulac C, Mathis MW, Mathis A. (2022) Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, 19: 496–504.

[5] Pereira TD, Aldarondo DE, Willmore L, Kislin M, Wang SSH, Murthy M, Shaevitz JW. (2019) Fast animal pose estimation using deep neural networks. *Nature Methods*, 16: 117–125.

[6] WuJ, Yang Q, Xiao D, Wu M, Chen Z, Hong Q. (2025) Quantifying behavioural patterns for group-housed pigs based on deep learning and statistical analysis. *Computers and Electronics in Agriculture*, 237: 110521.

[7] Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., & Couzin, I. D. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife,* 8, e47994.

[8] Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, 5(9), eaaw0736.

[9] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In European Conference on Computer Vision (pp. 740-755). Springer.

[10] Erekalo, K. T., Gemtou, M., Kornelis, M., Pedersen, S. M., Christensen, T., & Denver, S. (2025). Understanding the behavioral factors influencing farmers' future adoption of climate-smart agriculture: A multi-group analysis. *Journal of Cleaner Production*, 510, 145632.