Hierarchical Representation Learning of Dog Behavior via Single-View 3D Pose Estimation

Nanako Miyai
Nara Institute of Science and Technology
miyai.nanako.mr4@naist.ac.jp

Takatomi Kubo
Nara Institute of Science and Technology
takatomi-k@is.naist.jp

Maaya Saito
Azabu University

Kazunori Ohno
Tohoku University

Takefumi Kikusui
Azabu University

Miho Nagasawa
Azabu University

Kazushi Ikeda
Nara Institute of Science and Technology

Abstract

Dogs exhibit diverse behaviors that function as important signals in human–dog communication. Automatic analysis of such behaviors is increasingly needed in both scientific and applied contexts. However, conventional methods for behavior analysis face two major challenges: (i) 3D pose estimation typically requires multi-camera setups or prior training with complex calibration, and (ii) behavior classification relies heavily on predefined labels, limiting the ability to detect previously unseen behaviors. To address these limitations, we combine D-Pose, a model that estimates 3D dog poses from a single camera by learning pose representations, with h/BehaveMAE, a self-supervised framework that learns hierarchical behavior representations from pose sequences without predefined labels. Using a dataset of annotated dog behaviors, we perform preliminary evaluation by applying linear probing on the learned embeddings. Our results suggest that this approach provides a flexible and generalizable pipeline for behavior analysis, enabling promising representation learning from videos. While this study focuses on dog behavior, the proposed framework may serve as a step toward uncovering the mechanisms of animal communication in the future.

1 Introduction

Dogs and humans have coexisted for thousands of years, forming a unique interspecies relationship. Dog–human interactions have been shown to promote human welfare and health, for example, by reducing stress [1] or lowering the risk of cardiovascular disease [2]. At the same time, dogs themselves have developed social cognitive abilities to respond to human communicative cues such as pointing gestures and gaze [3]. These dog behaviors function as important signals in communication and are increasingly studied in both scientific and applied contexts, including training and veterinary care.

Despite this demand, existing methods face two major limitations. First, 3D pose estimation typically requires multi-camera setups and/or extensive pretraining [4], which restricts its accessibility. Second, behavior analysis depends on predefined labels, making it difficult to detect novel or complex behaviors that do not fit existing ethograms. These constraints hinder the applicability and generalizability of current approaches.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for non-human animal communication.

To address these challenges, we explore a new pipeline that combines D-Pose [5], which enables 3D dog pose estimation from a single camera, with h/BehaveMAE [6], a self-supervised framework that learns hierarchical behavior representations directly from pose sequences [6].

This approach offers three main advantages: it reduces dependence on multi-camera systems, allowing analysis from simple video recordings; it avoids reliance on predefined behavior labels, enabling more flexible detection of diverse or previously unseen behaviors; and it provides hierarchical representations that can describe complex behaviors, thereby lowering annotation costs and facilitating future studies of animal communication.

In this paper, we report preliminary results on a dataset of annotated dog behaviors and discuss how this approach may contribute to future research on animal communication.

2 Methodology

Our proposed pipeline consists of three main stages: dog detection, 3D pose estimation, and hierarchical representation learning (Fig. 1).

Dog detection From single-camera video recordings, individual dogs are localized before pose estimation (see Supplementary Information A for details). This step provides cropped image sequences of dogs for subsequent analysis.

3D pose estimation with D-Pose Each cropped frame is processed with D-Pose [5], which estimates 3D dog poses from a single camera. D-Pose employs a vision-transformer backbone and predicts joint locations across multiple projection planes. In this study, we use the DigiDogs pretrained checkpoint, capable of detecting 26 joints, including the head, spine, pelvis, tail, and limbs. This enables reliable single-camera 3D pose estimation without requiring multi-camera setups.

Hierarchical behavior representation with h/BehaveMAE Pose trajectories produced by D-Pose are then input to h/BehaveMAE [6], a self-supervised autoencoder that learns hierarchical embeddings of animal behavior. The model was trained from scratch using dog 3D pose trajectory data obtained above, and embeddings were generated on a per-frame basis. The model divides pose trajectories into spatiotemporal patches, applies random masking, and reconstructs the missing parts, thereby capturing multi-level representations of behavior. The learned embeddings are subsequently evaluated with linear probing, providing a simple yet effective way to assess whether the representations capture meaningful behavioral distinctions.

3 Preliminary results

3.1 Experiment

Dataset

We evaluated our approach on a dataset provided by Azabu University, consisting of approximately one hour of video per dog from 34 individuals. Each recording was annotated with thirteen behavioral categories (tail-high, tail-low, tail-wag, neck-high, neck-low, stand, sit, down, up, focus, sniff, whine, licking nose) with five experimental phases (rest, box, owner, control, condition)[1]. The dataset is annotated in an event-based manner, where each behavior is labeled with its onset time, offset time, and category. Although annotations were derived from multi-view observations, only side-view videos were used in this study.

Preprocessing

Videos were segmented according to experimental phase and cropped to dog regions (see Supplementary Information A). 3D poses were estimated for each frame using D-Pose, and sequences were constructed in segments of 1800 frames. Sequences with excessive missing joints (>20% of frames) were discarded, and shorter sequences were padded when possible. Behavioral annotations were aligned at the frame level as binary labels.

[1] These labels are annotated mainly for another study on dogs' behavior analysis.

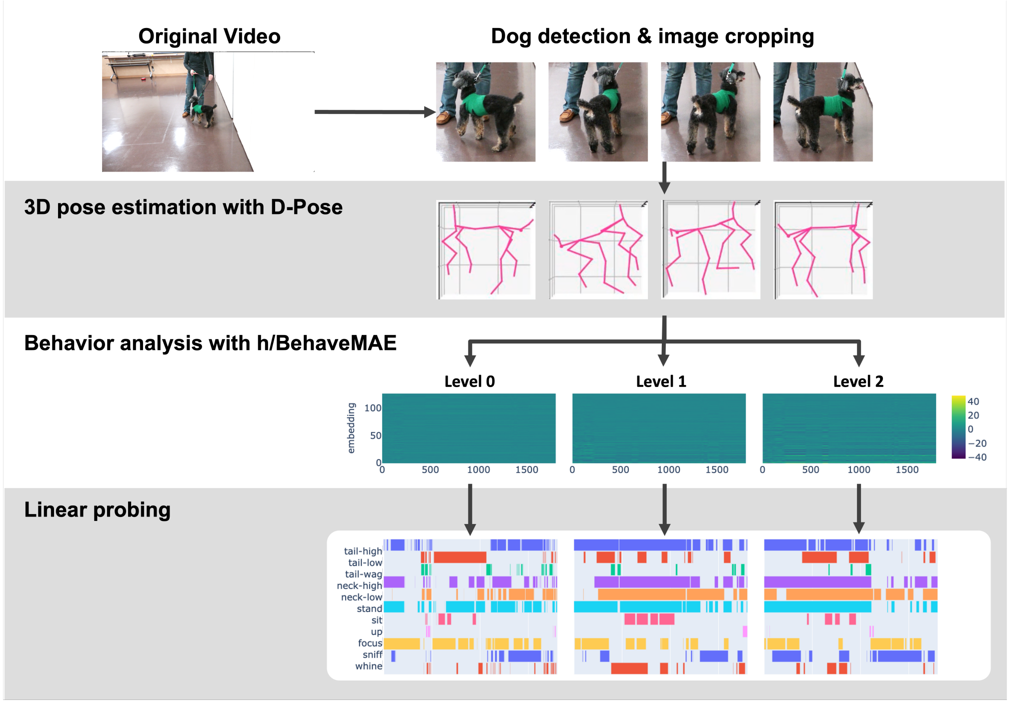

Figure 1: Conceptual illustration of the proposed method.

Evaluation protocol

The representation learning was performed with h/BehaveMAE, and the resulting frame-wise embeddings were assessed using linear probing. The dataset was split at a 4:4:2 ratio into training data for h/BehaveMAE, training data for linear probing, and test data for linear probing, ensuring that all dogs appeared in each split. For linear probing, we employed the RidgeClassifier implementation from scikit-learn, treating each behavioral task as a binary classification problem. Performance was evaluated using the F1 score from scikit-learn, following the evaluation protocol of the previous study [6]. As a baseline, we used 2D pose estimation from DeepLabCut (DLC) [7, 8] with 2D h/BehaveMAE.

3.2 Quantitative results

Table 1 compares F1 scores between 2D and 3D representations across behaviors. While tail- and neck-related behaviors performed better with 2D poses, several postural behaviors (stand, sit, down) and low-level actions (sniff, whine) benefited substantially from 3D poses. Notably, all behaviors with 3D input exceeded chance level, and the mean F1 score was higher for 3D than 2D inputs (improvements in mean F1 score: 0.044 [18.62%] at level 0, 0.022 [9.47%] at level 1, and 0.014 [5.91%] at level 2). As the hierarchy progresses from levels 0 to 2, temporal receptive field expands, enabling the representation of longer temporal patterns. Because most behaviors can be discriminated primarily based on static posture information in our study, classification was possible even within the short spatiotemporal window of level 0. In the future, we will validate our approach with a new video dataset that includes more dynamic behaviors representing complex temporal patterns.

3.3 Qualitative results

Fig. 2 shows examples of estimated poses, ground-truth annotations, and predicted intervals at different hierarchical levels. D-Pose produced stable estimates overall, and h/BehaveMAE embeddings captured temporal structure. However, sit and down remained difficult to classify due to their low frequency in the dataset, and mutually exclusive categories (e.g., neck-high vs. neck-low) were sometimes predicted simultaneously, reflecting the limitations of linear probing without using correlation among categories.

Table 1: Comparison of F1 scores between 2D and 3D approaches with hierarchical levels.

Task ID	2D			3D			Chance level
	lv 0	lv 1	lv 2	lv 0	lv 1	lv 2	
tail-high	**0.228**	0.194	0.251	0.227	0.178	0.219	0.108
tail-low	**0.300**	0.238	0.237	0.262	0.225	0.192	0.085
tail-wag	0.023	0.064	**0.076**	0.039	0.054	0.044	0.012
neck-high	**0.222**	0.161	0.205	0.221	0.208	0.221	0.131
neck-low	**0.186**	0.139	0.134	0.176	0.168	0.126	0.049
stand	0.242	0.203	0.238	0.244	0.258	**0.266**	0.173
sit	0.185	0.191	0.202	0.345	0.331	**0.407**	0.029
down	0.451	0.588	0.471	**0.742**	0.572	0.497	0.012
focus	0.532	0.568	**0.575**	0.557	0.549	0.564	0.376
sniff	0.226	0.240	0.258	0.268	**0.274**	0.256	0.073
whine	0.021	0.006	0.009	**0.024**	0.023	0.023	0.009
mean	0.238	0.236	0.242	**0.282**	0.258	0.256	0.096

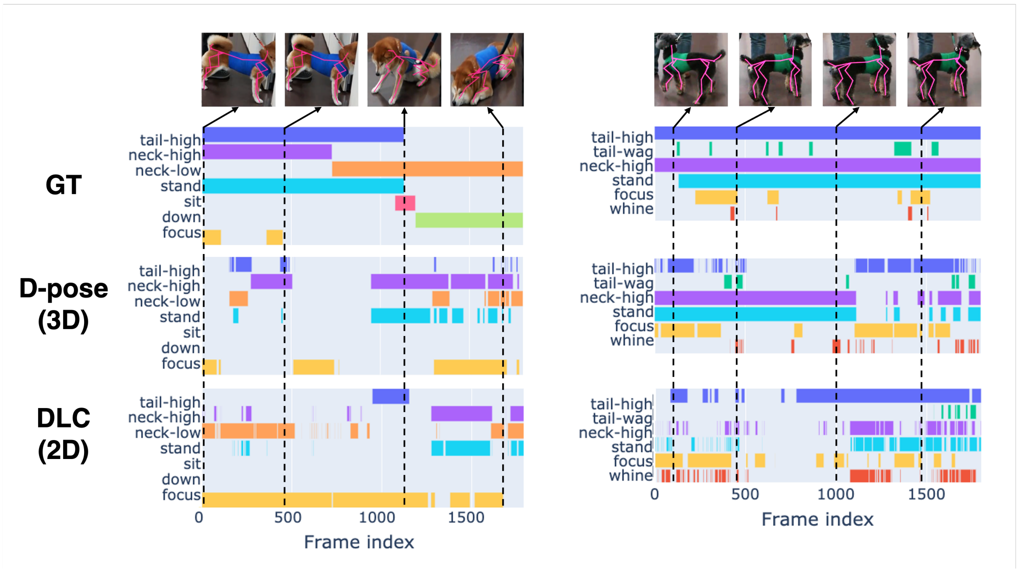

Figure 2: Frame-wise visualization of D-Pose (3D) and DLC (2D) poses, ground-truth behaviors, and predicted behaviors by the best-performing linear probe among hierarchical levels.

3.4 Summary

These results demonstrate that combining D-Pose and h/BehaveMAE enables learning of meaningful hierarchical representations of dog behavior from single-camera videos. At the same time, they highlight areas for improvement, including handling rare behaviors, reducing pose estimation errors, and moving beyond simple linear probes for classification. While our results are preliminary, they highlight that rare behaviors remain bottlenecks. We anticipate that increasing data diversity will enhance downstream classification.

4 Discussions

In this work, we presented a framework for dog behavior analysis that combines single-camera 3D pose estimation with D-Pose and hierarchical representation learning with h/BehaveMAE. Our findings demonstrate that meaningful behavioral representations can be learned without relying on predefined labels, suggesting the feasibility of a label-free approach to animal behavior analysis.

Training an end-to-end 2D video model requires dense frame-level annotations and significant computational resources. In contrast, our goal was to isolate the benefit of geometric priors via explicit 3D pose representation obtained from the pretrained D-Pose, which would be obscured in usual end-to-end 2D settings. The explicit 3D intermediate representation improves interpretability and may reduce overfitting via geometric constraints. By avoiding reliance on annotated ethograms, this approach could extend to tasks such as detecting atypical or previously unseen behaviors, with potential applications in anomaly detection for welfare monitoring and clinical contexts. Moreover, because hierarchical representations can capture complex behaviors, even simple linear probing achieved reasonable classification, indicating the potential to reduce annotation costs in future behavior studies. Overall, this study shows that integrating single-camera 3D pose estimation with self-supervised representation learning offers a promising pipeline for analyzing animal behavior. While our current focus is on behavior recognition, such methods may ultimately help uncover the principles of animal communication.

5 Acknowledgments

This work was supported by JST, CREST Grant Number JPMJCR22P2, Japan.

References

[1] Kerstin Meints, Victoria L Brelsford, Mirena Dimolareva, Laëtitia Maréchal, Kyla Pennington, Elise Rowan, and Nancy R Gee. Can dogs reduce stress levels in school children? effects of dog-assisted interventions on salivary cortisol in children with and without special educational needs using randomized controlled trials. *Plos one*, 17(6):e0269333, 2022.

[2] Mwenya Mubanga, Liisa Byberg, Christoph Nowak, Agneta Egenvall, Patrik K Magnusson, Erik Ingelsson, and Tove Fall. Dog ownership and the risk of cardiovascular disease and death–a nationwide cohort study. *Scientific reports*, 7(1):15821, 2017.

[3] Miho Nagasawa, Shouhei Mitsui, Shiori En, Nobuyo Ohtani, Mitsuaki Ohta, Yasuo Sakuma, Tatsushi Onaka, Kazutaka Mogi, and Takefumi Kikusui. Oxytocin-gaze positive loop and the coevolution of human-dog bonds. *Science*, 348(6232):333–336, 2015.

[4] Rim Yu and Yongsoon Choi. Okeydoggy3d: a mobile application for recognizing stress-related behaviors in companion dogs based on three-dimensional pose estimation through deep learning. *Applied Sciences*, 12(16):8057, 2022.

[5] Moira Shooter, Charles Malleson, and Adrian Hilton. Digidogs: Single-view 3d pose estimation of dogs using synthetic training data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 80–89, 2024.

[6] Lucas Stoffl, Andy Bonnetto, Stéphane d'Ascoli, and Alexander Mathis. Elucidating the hierarchical nature of behavior with masked autoencoders. In *European conference on computer vision*, pages 106–125. Springer, 2024.

[7] Alexander Mathis, Pranav Mamidanna, Taiga Abe, Kevin M Cury, Venkatesh N Murthy, Mackenzie W Mathis, and Matthias Bethge. Markerless tracking of user-defined features with deep learning. *arXiv preprint arXiv:1804.03142*, 2018.

[8] Shaokai Ye, Anastasiia Filippova, Jessy Lauer, Steffen Schneider, Maxime Vidal, Tian Qiu, Alexander Mathis, and Mackenzie Weygandt Mathis. Superanimal pretrained pose estimation models for behavioral analysis. *Nature communications*, 15(1):5165, 2024.

[9] Glenn Jocher and Jing Qiu. Ultralytics YOLO11, 2024.

Supplementary information

A Training YOLOv11 for Dog Detection

To accurately extract dog regions from the Azabu University dataset, we trained a YOLOv11 detector [9]. The dataset was constructed from the training videos used for h/BehaveMAE. For each video, two types of images were sampled:

- Positive samples: five frames uniformly sampled across each video where dogs were present.
- Negative samples: frames extracted at 0.1-s intervals from segments where dogs were not detected during pre-processing with YOLO tracking.

The resulting dataset contained 1029 positive and 669 negative images, split into training, validation, and test sets in a 70:15:15 ratio. We ensured that each split contained images from all individual dogs.

Training was performed by fine-tuning the pretrained weights yolo11n.pt with 50 epochs and early stopping (patience=10). During inference, we selected the model checkpoint that achieved the best validation performance. If multiple bounding boxes were detected in a frame, the box with the highest confidence score was used to generate the cropped image.