TOWARDS REPRESENTATION LEARNING FOR PHENOTYPING BEYOND ANIMAL POSE ESTIMATION

Takatomi Kubo, Nina Nakajima, Nanako Miyai, Midori Osaki & Suzuka Higashitsutsumi
Graduate School of Science and Technology
Nara Institute of Science and Technology
Ikoma, Nara, Japan
takatomi-k@is.naist.jp

1 INTRODUCTION

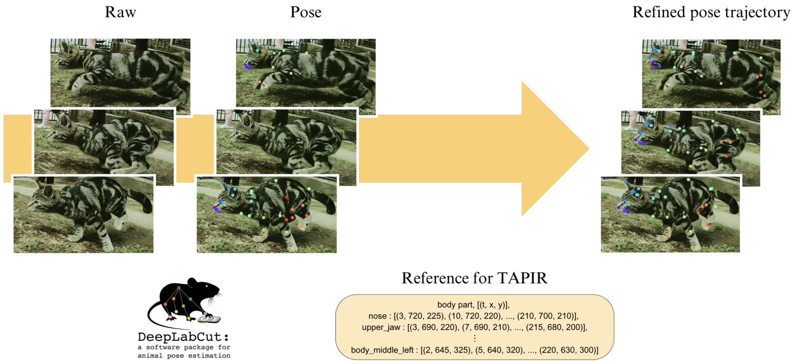
Understanding biological phenomena requires capturing phenotypic manifestations, among which behavior is crucial. Pose estimation methods such as DeepLabCut (DLC) Mathis et al. (2018) and dimensionality reduction techniques like CEBRA Schneider et al. (2023) have been applied to analyze behaviors and their neural representations. However, characterizing the underlying manifolds as behavioral representations remains an open challenge. In this study, we propose introducing TAPIR to complement DLC-based pose estimation. By using TAPIR (Tracking Any Point with per-frame Initialization and temporal Refinement) Doersch et al. (2023) to enhance long-term pose tracking and applying CEBRA to the processed data, we aim to explore the latent manifold of behavior and facilitate its representation learning. Furthermore, by comparing the extracted behavioral manifold with other biological data, we seek to assess its potential for providing deeper insights into phenotypes. Ultimately, our approach aims to contribute to the ongoing development of a foundation for comprehensive representation learning of biological phenomena.

2 PROPOSED METHOD: DLC-TAPIR INTEGRATION

We integrate TAPIR with DLC to improve tracking of highly flexible animal postures (Figure 1). DLC, using the pre-trained SuperAnimal-Quadruped model, reliably tracks body parts but struggles with occlusions and rapid movement. To further enhance robustness, we apply DLC with video adapt (a form of pseudolabeling), allowing the model to adjust to the specific characteristics of the video and improve tracking consistency. To ensure high-confidence tracking, we select frames where the likelihood of each predicted body part position exceeds 0.8. From these frames, we extract distributed samples of frame indices to ensure wide temporal coverage. These selected coordinates serve as input to TAPIR, which refines and interpolates missing body parts across frames through per-frame initialization and temporal refinement. This integration enhances tracking robustness, ensuring more consistent pose estimation, particularly for dynamic and complex postures. By combining DLC's structured pose detection with video adapt and TAPIR's ability to handle missing body parts, our approach provides a more reliable and interpretable tracking method for studying animal movement.

3 EXPERIMENT

To validate the effectiveness of our proposed method, we conducted experiments on tracking highly flexible cat postures, particularly under challenging conditions involving extreme flexibility and occlusions. We used a publicly available video (1280 × 720, 29.67 fps) of a running cat from Pexels (https://www.pexels.com/) to evaluate tracking performance. DLC v3 was applied to estimate the coordinates of 39 body parts using the SuperAnimal-Quadruped model pretrained for quadruped animals. To ensure reliable input for TAPIR, we selected frames just before posture transitions where DLC predictions were relatively accurate (i.e., key point confidence > 0.8). These frames provided a stable reference for further refinement. The TAPIR model was applied to the video to refine and interpolate missing or mispredicted body parts, particularly in frames where DLC failed due to occlusions or rapid motion. We compared tracking performance between DLC and the proposed DLC+TAPIR approach, highlighting improvements in continuity and robustness.



```
body part, [(t, x, y)],
nose : [(3, 720, 225), (10, 720, 220), ..., (210, 700, 210)],
upper_jaw : [(3, 690, 220), (7, 690, 210), ..., (215, 680, 200)],
⋮
body_middle_left : [(2, 645, 325), (5, 640, 320), ..., (220, 630, 300)]
```

Figure 1: Overview of the proposed method, integrating DLC and TAPIR.

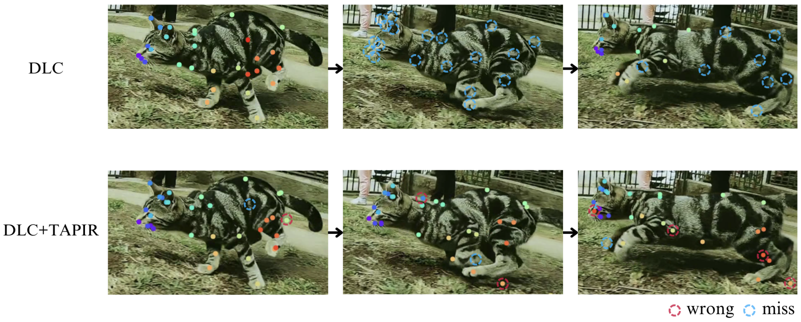


Figure 2: Examples of tracking results for a cat's flexible posture.

4 Results

Figure 2 shows examples where DLC failed to track body parts, while DLC+TAPIR successfully recovered them. DLC worked well in stable postures but failed in rapid motion. TAPIR improved robustness, successfully recovering missing body parts under occlusions. Tracking performance was evaluated by counting detected, undetected, and misdetected (correct/miss/wrong) body parts; DLC: 169/290/43, DLC+TAPIR: 398/13/138, showing significant improvement.

Meaningfulness Statement

Understanding life requires capturing biological phenomena in a structured and interpretable form. Behavior, as a core phenotype, emerges from complex movement patterns and interactions with the environment, yet its reliable quantification remains challenging due to occlusions and missing key points. By integrating TAPIR with DLC, we enhance tracking accuracy, providing more complete input for downstream analysis. Furthermore, applying CEBRA to these refined pose sequences enables learning structured representations of behavior, helping to reveal underlying principles of biological dynamics—essential for advancing our understanding of life.

References

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10061–10072, 2023.

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, 2018.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.