3D Multimodal Dataset and Token-based Pose Optimization

Mahir Patel, Yiwen Gu, Lucas Carstensen, Dr. Michael E. Hasselmo, Dr. Margrit Betke

BOSTON UNIVERSITY

Project Website

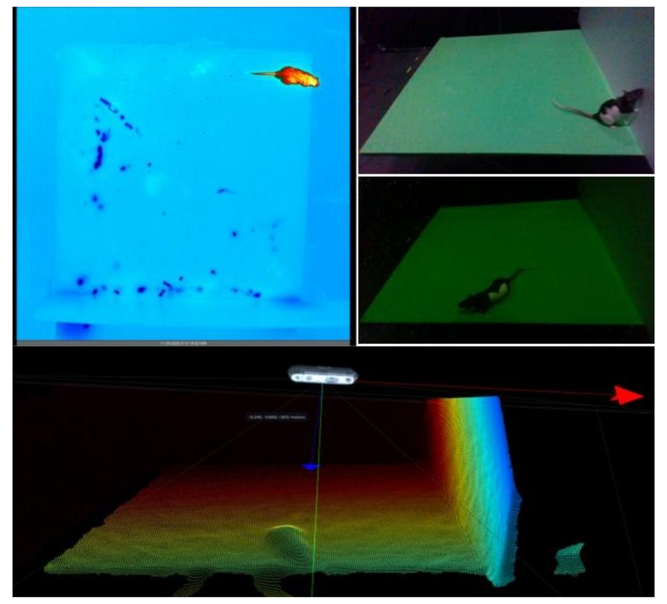
CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

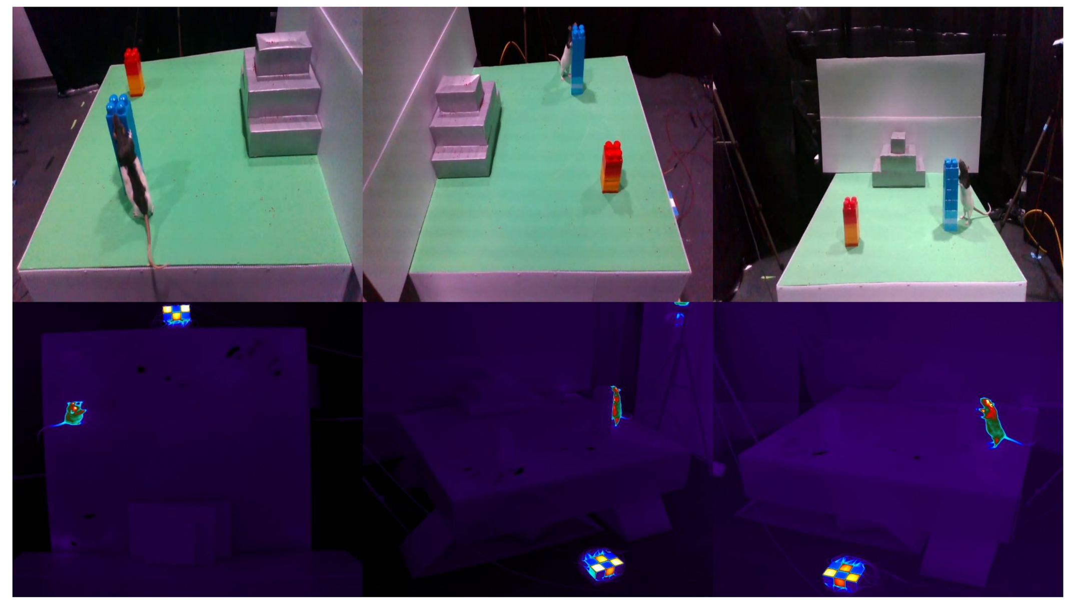
Supported by Office of Naval Research MURI N00014-19-1-2571

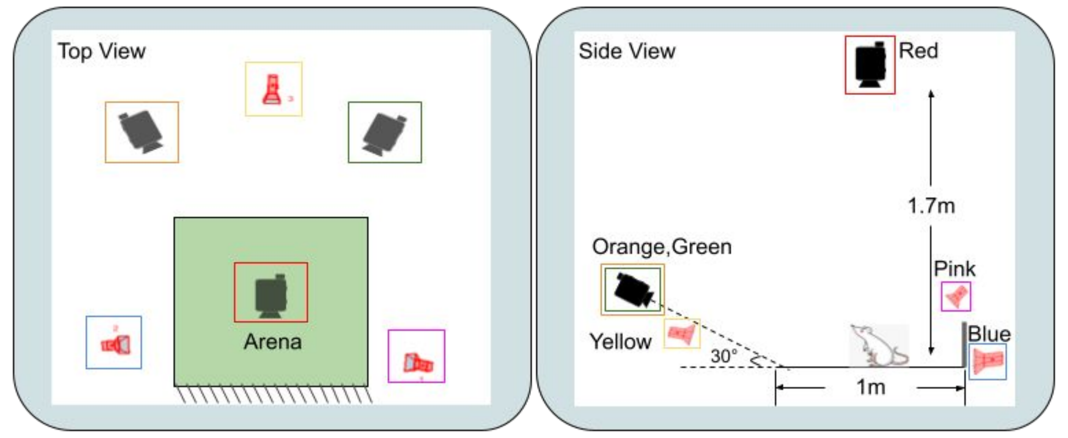
Rodent3D Dataset

We introduce the **Rodent3D** dataset that records animals exploring their environment with multiple cameras and modalities (**RGB, depth, thermal infrared**).

- 200 minutes of multimodal video recordings from up to three thermal and three RGB-D synchronized cameras (approximately **4 million frames**).
- Thermal cameras with 1024x1024 spatial resolution recorded at 60 or 120 Hz through a FLIR High Speed Data Recorder.
- RGB-D Cameras with 848x480 spatial resolution recorded at 30 or 60 Hz. Dropped frames from the RGB-D cameras can be inferred through hardware timestamps.
- All cameras hardware synchronized to an external TTL signal.
- 2D Markers generated by two DeepLabCut [Mathis'18] models for the thermal and RGB modalities respectively.
- Depth data aligned with RGB, stored as pickled numpy arrays of dimension 480x848.



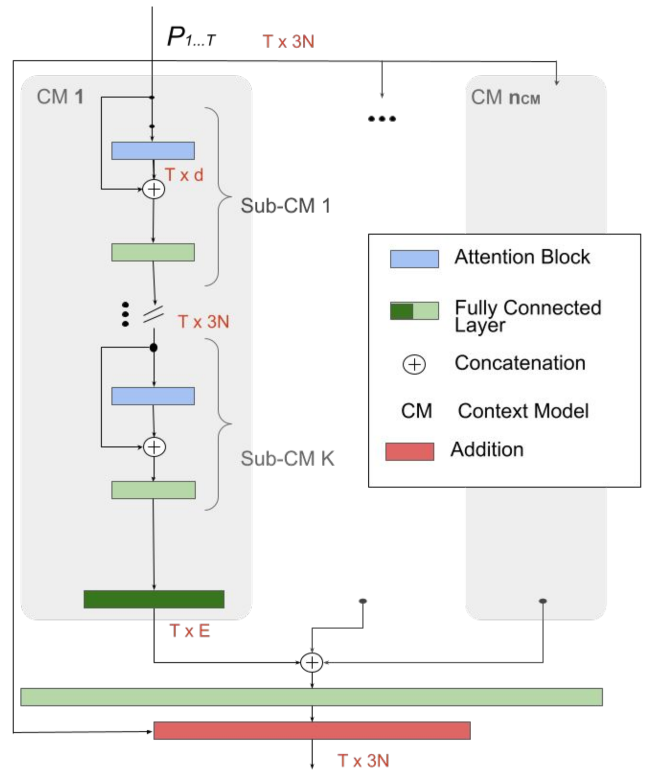


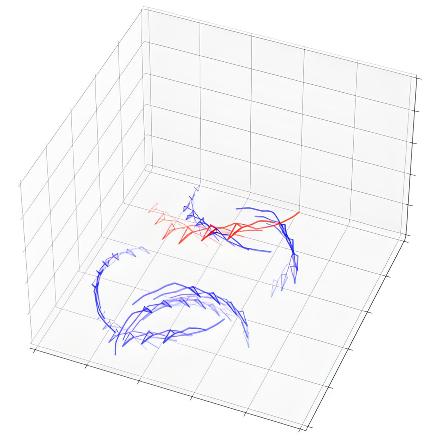


OptiPose Model

For the task of optimizing estimates of pose sequences provided by existing pose estimation methods, we provide a baseline model called **OptiPose**. While deep-learned attention mechanisms have been used for pose estimation in the past, **with OptiPose, we propose a different way by representing 3D poses as tokens for which deep-learned context models pay attention to both spatial and temporal keypoint patterns**.

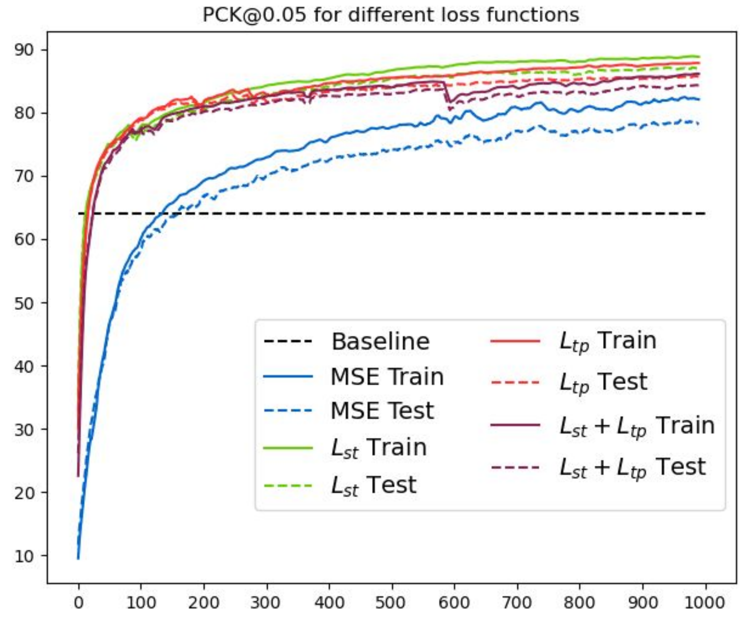
- OptiPose treats each pose with N keypoints as a token. The flattened 3N vector is considered as the input embedding.
- OptiPose uses Parallel Context Models (PCMs) that contribute towards different learned combinations of keypoints.
- Each PCM has a set of sub-Context Models which detects patterns from the specific combination of keypoints, targeted by their respective PCM.
- OptiPose uses random masking, similar to the Masked Language Modelling, to learn how to optimize keypoints.
- Since OptiPose operates on 3D data directly, data augmentation involves synthesizing 3D pose sequences through rotation and translation.
- Structural and Temporal Loss functions promote accelerated learning.



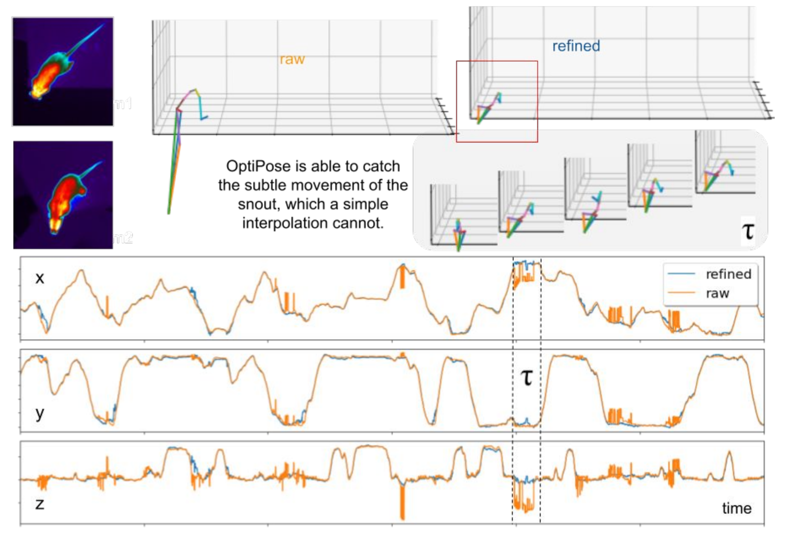


$$\mathcal{L}_{st} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\|x_i - x_j\|_2 - \|\hat{x}_i - \hat{x}_j\|_2\right)^2$$

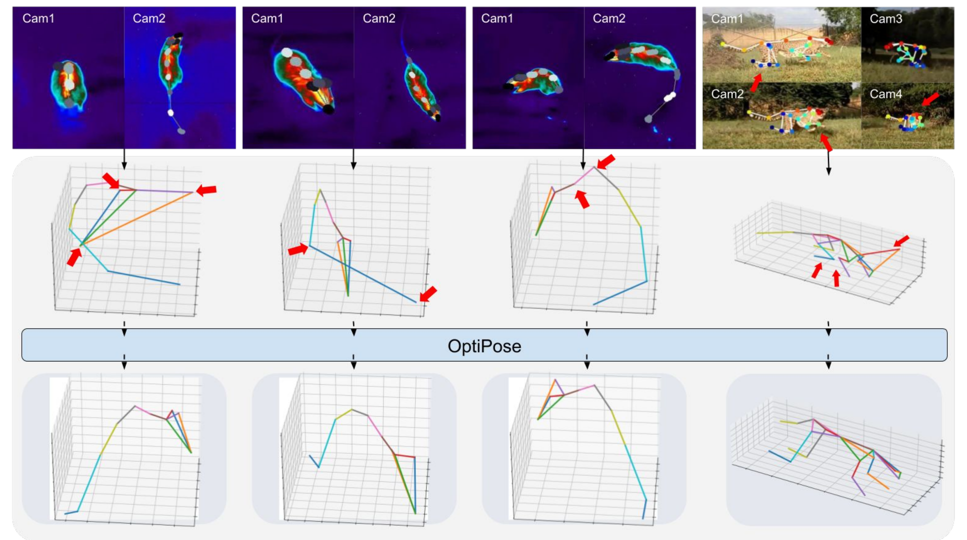
$$\mathcal{L}_{tp} = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|(x_i^{(t)} - x_i^{(t-1)}) - (\hat{x}_i^{(t)} - \hat{x}_i^{(t-1)})\right\|$$



Results and Analysis



Non-linear tracking of Snout



OptiPose on Rodent3D and AcinoSet [Joska'21]

Table 5.2: Average PCK accuracy of OptiPose per keypoint over 6,500 sets of $T \leq F$ consecutive poses on the Rodent3D dataset (Top: color module, $F = 30$. Bottom: thermal module $F = 60$).

Rodent3D:	Snout	RightEar	LeftEar	HeadBase	Mid	TailBase	TailMid	TailTip	Avg
Baseline $H(P)$	64.50	64.52	64.73	64.86	65.00	65.68	64.97	64.90	64.90
PCK@0.05	74.08	83.61	85.04	85.82	78.51	80.59	84.16	80.66	81.56
PCK@0.10	**85.37**	**92.54**	**92.90**	**93.38**	**90.25**	**90.62**	**91.41**	**87.34**	**90.48**
Rodent3D:	Snout	RightEar	LeftEar	HeadBase	Mid	TailBase	TailMid	TailTip	Avg
Baseline $H(P)$	65.32	65.77	65.21	65.76	66.14	65.24	65.85	65.37	65.56
PCK@0.05	78.66	82.11	82.19	78.11	83.76	82.24	80.33	74.34	80.91
PCK@0.10	**89.42**	**92.69**	**92.95**	**94.22**	**93.14**	**92.05**	**90.4**	**86.35**	**91.89**

Videos

OptiPose Tracking with Noisy 2D Markers

Side-By-Side Recording Session

Rodent's View Represented as Binocular Vision