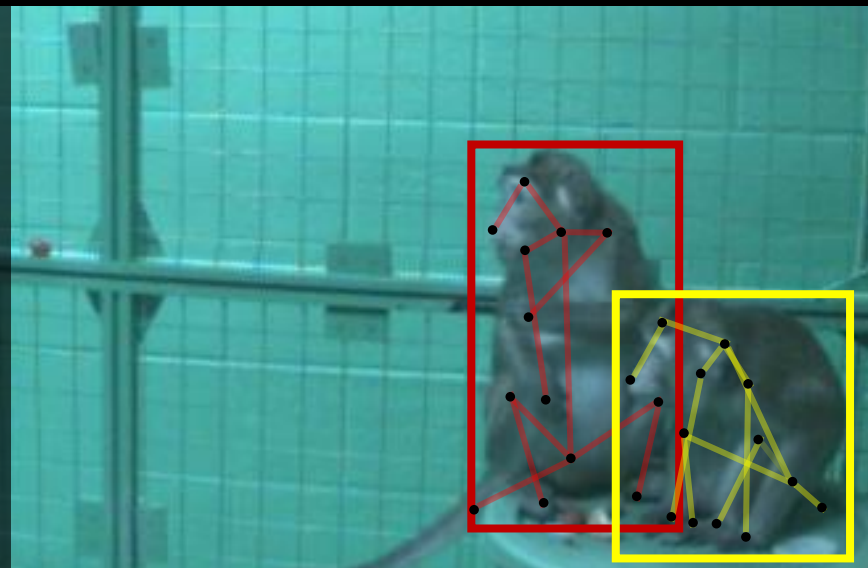
METHODS FOR LEARNING THE LANGUAGE OF BEHAVIOR

OBSERVING, REPRESENTING, AND PREDICTING ACTION

Praneet C Bala

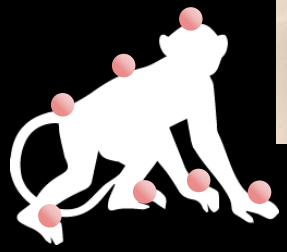
Committee Members:

Volkan Isler (Chair)
Stephen J Guy
Hyun Soo Park
Jan Zimmermann
Nikolaos Papanikolopoulos





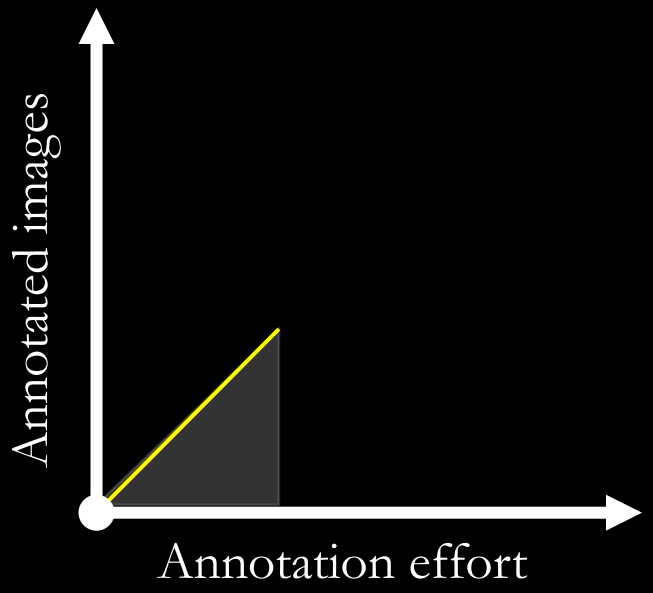




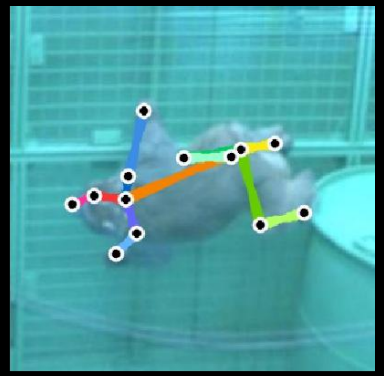


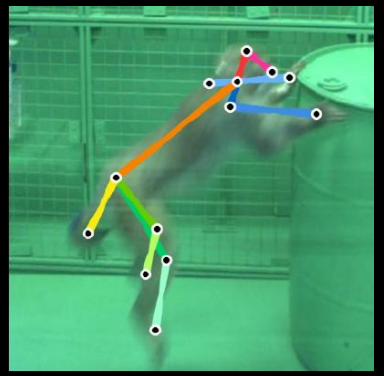
Markers

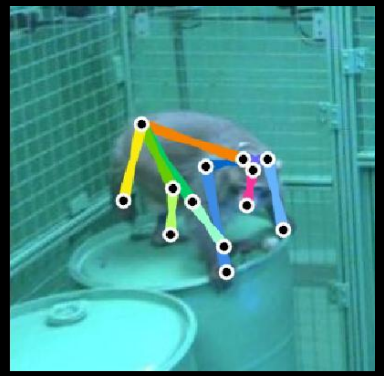


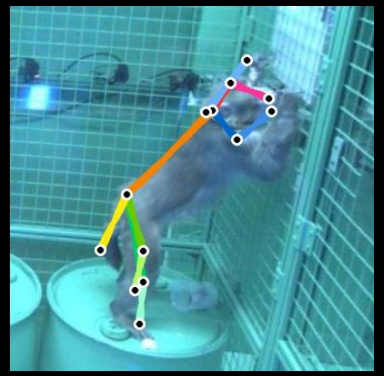


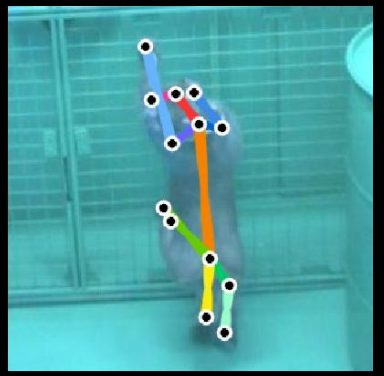












No. of images	Time per image	Hourly rate	Annotation cost
2.5M	3min	$45/hr	$5M

Across viewpoints

Across environments



Across granularity

Across time

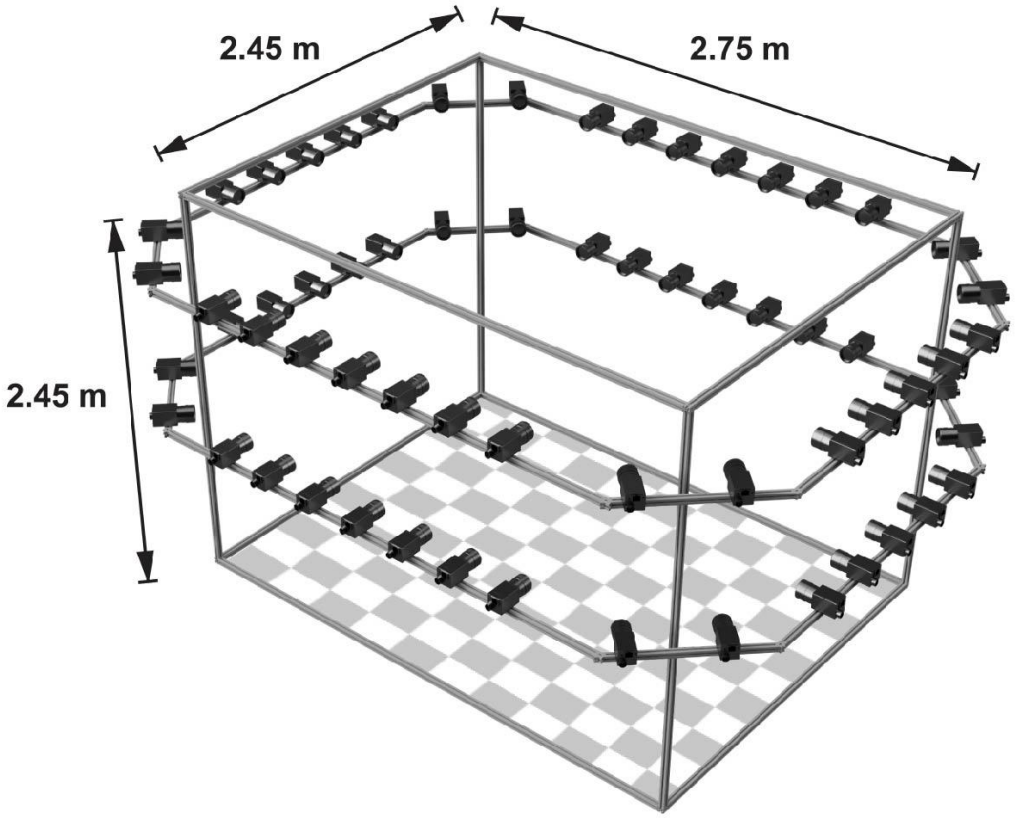
The behavioral record

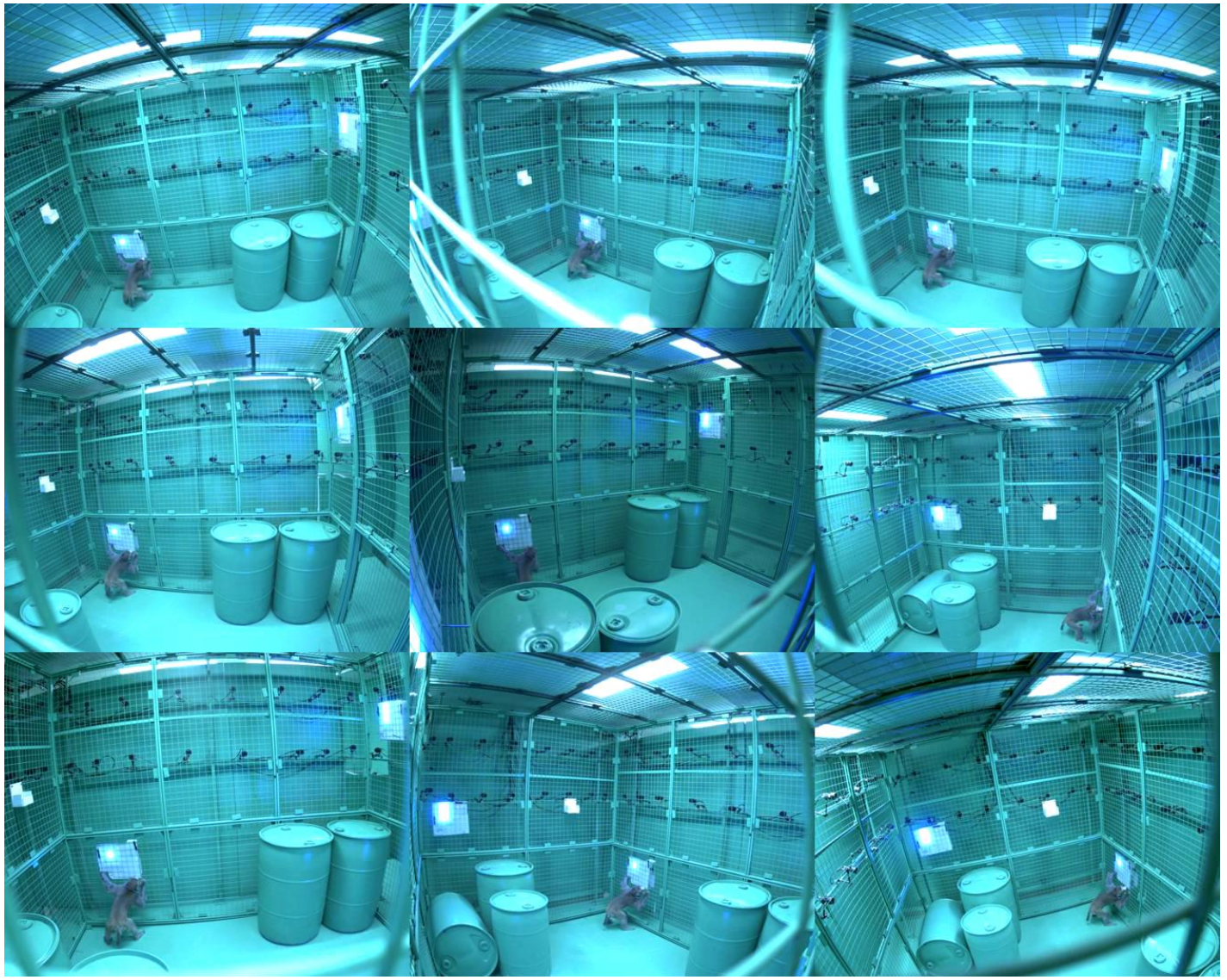
"Can we build the record at all?"



OpenMonkeyStudio

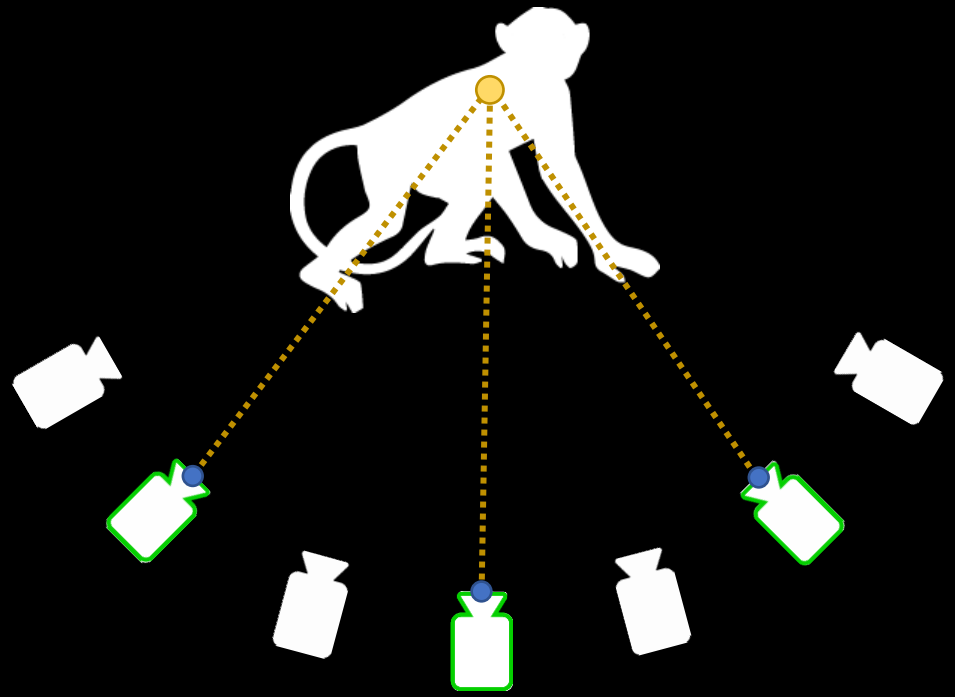
Multi-camera system

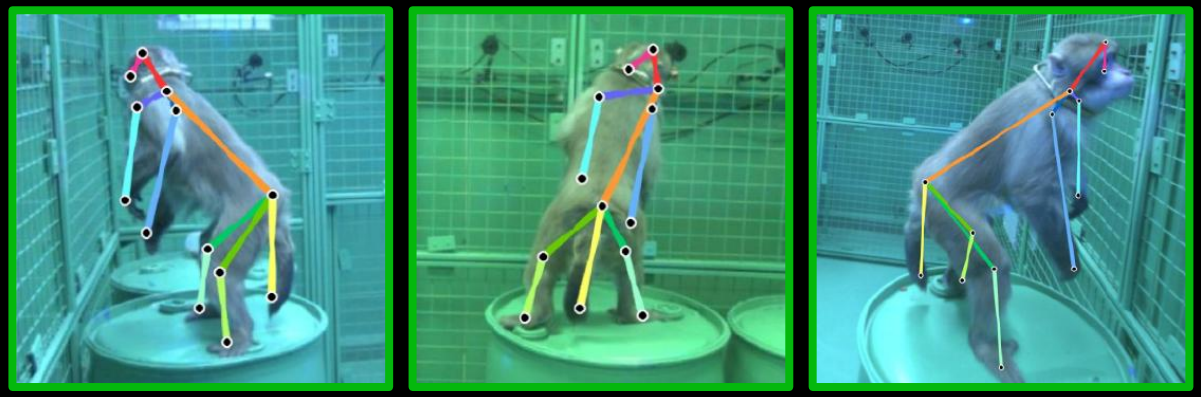




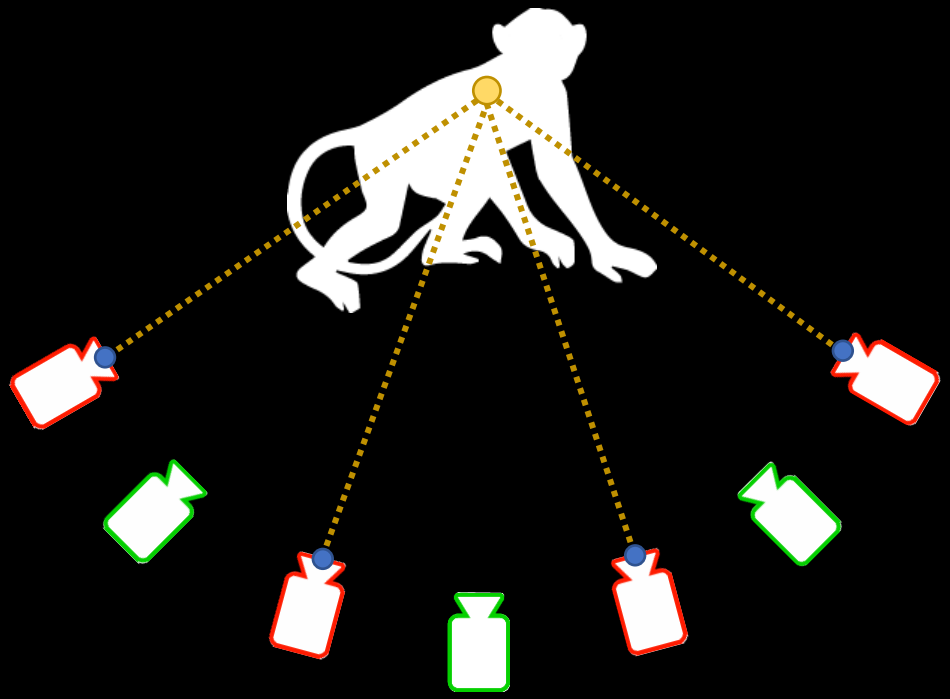
Multiview setup:

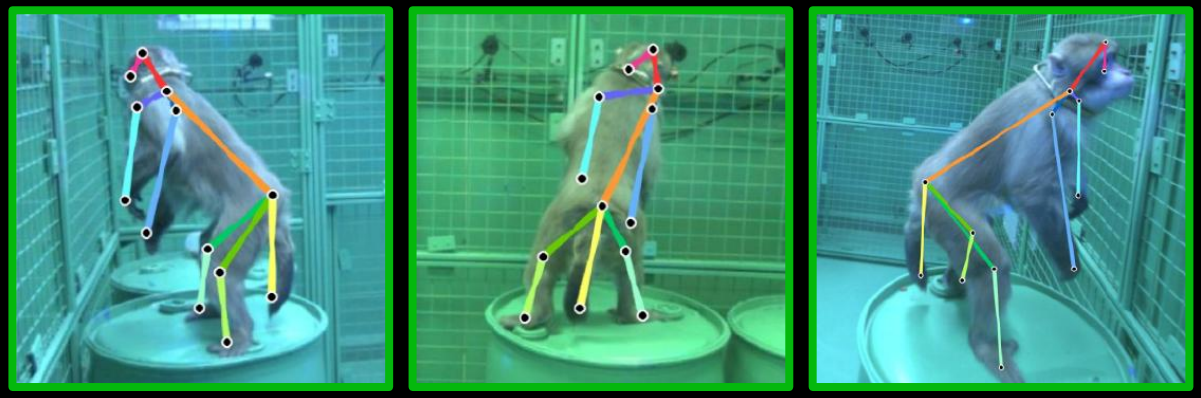
- 62 synchronously triggered cameras
- 30 frames per second



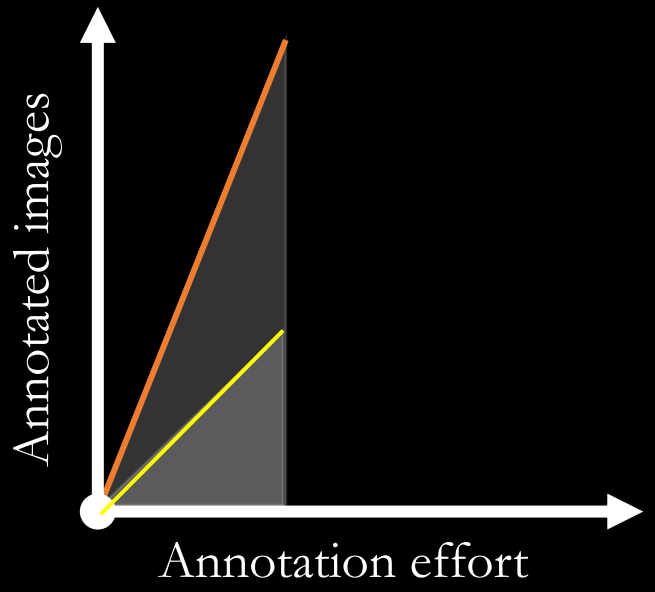


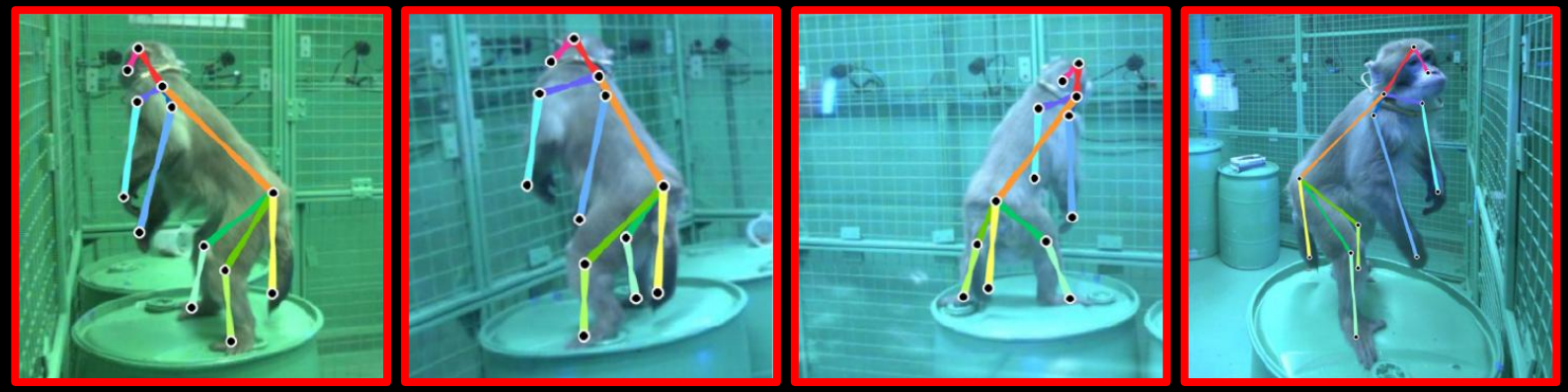
Annotated Data



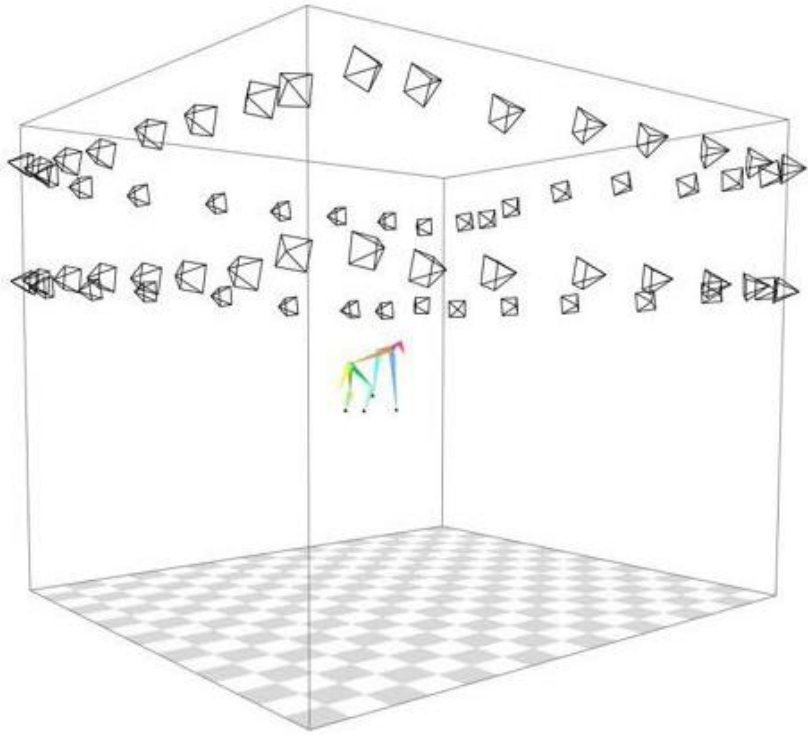


Annotated Data

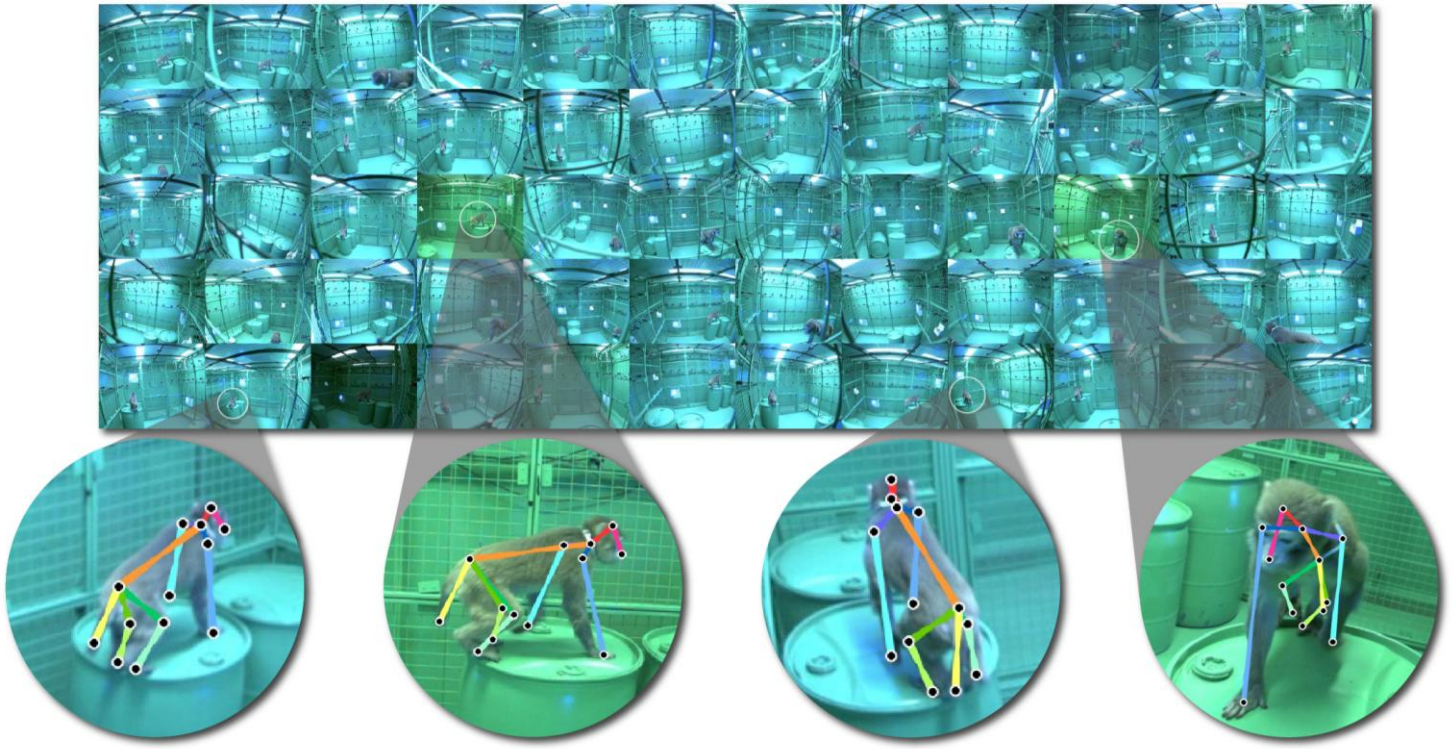




Data augmentation via Geometric reasoning



Triangulation



OpenMonkeyStudio Dataset

- 6-7x scale up from triangulation
- 195k images from 33k labeled images (Hive AI)
- 13 annotated landmarks

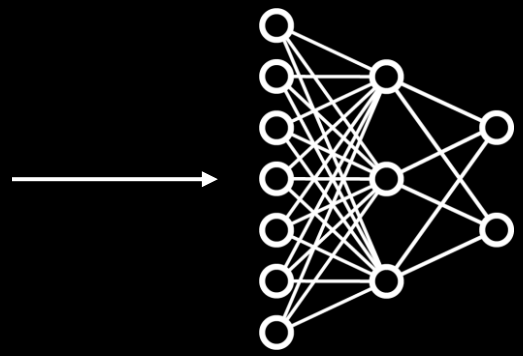
A significant contribution to the macaque pose estimation community





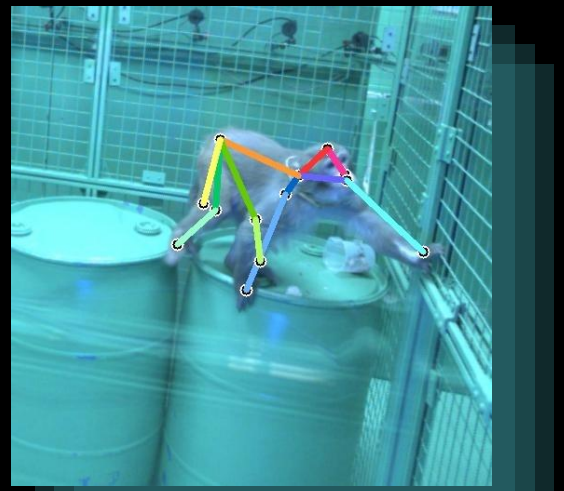
OpenMonkeyStudio Dataset





Conv. Pose Machine

$\Psi(I_i)$

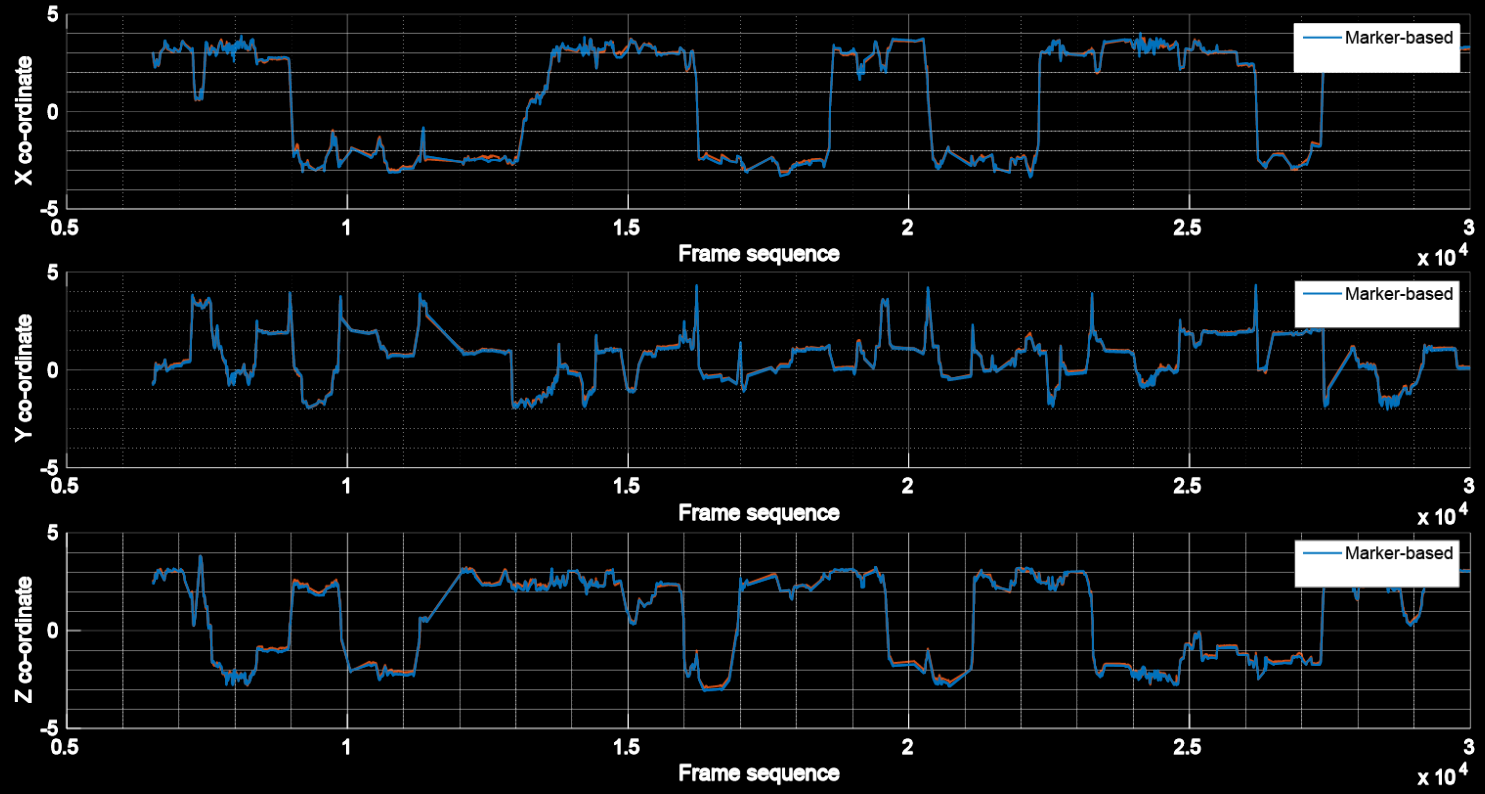


Markerless vs Marker-based



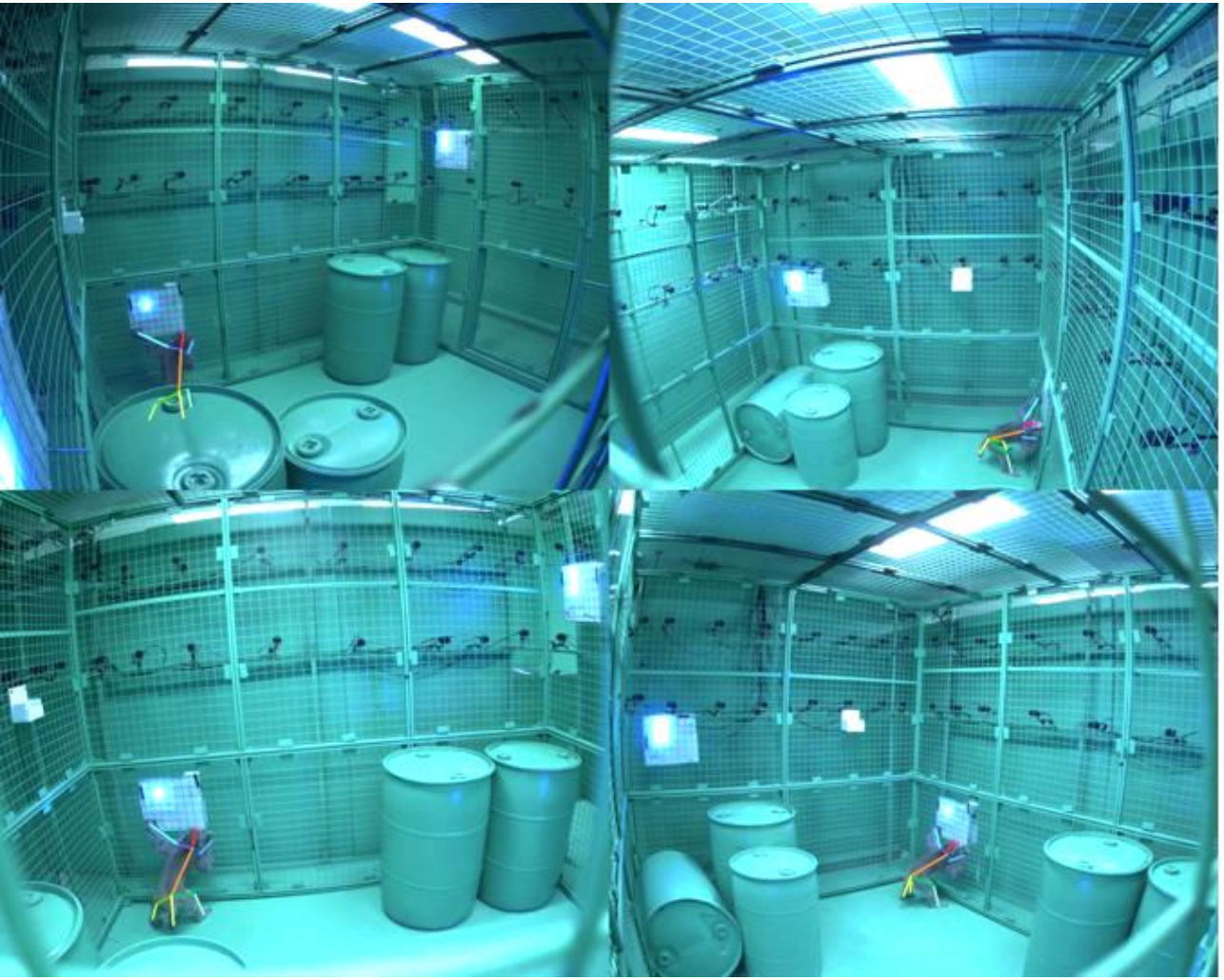
Markerless vs Marker-based

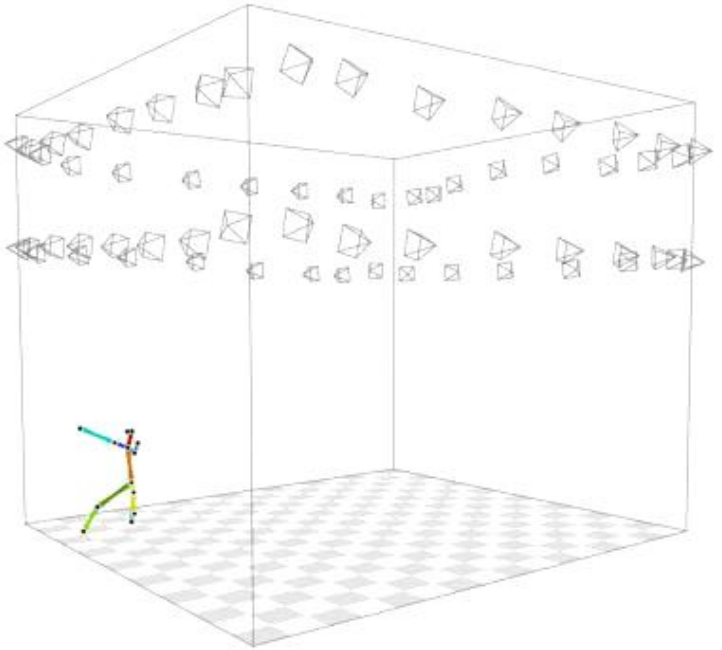
Co-ordinate analysis: markerless vs marker-based



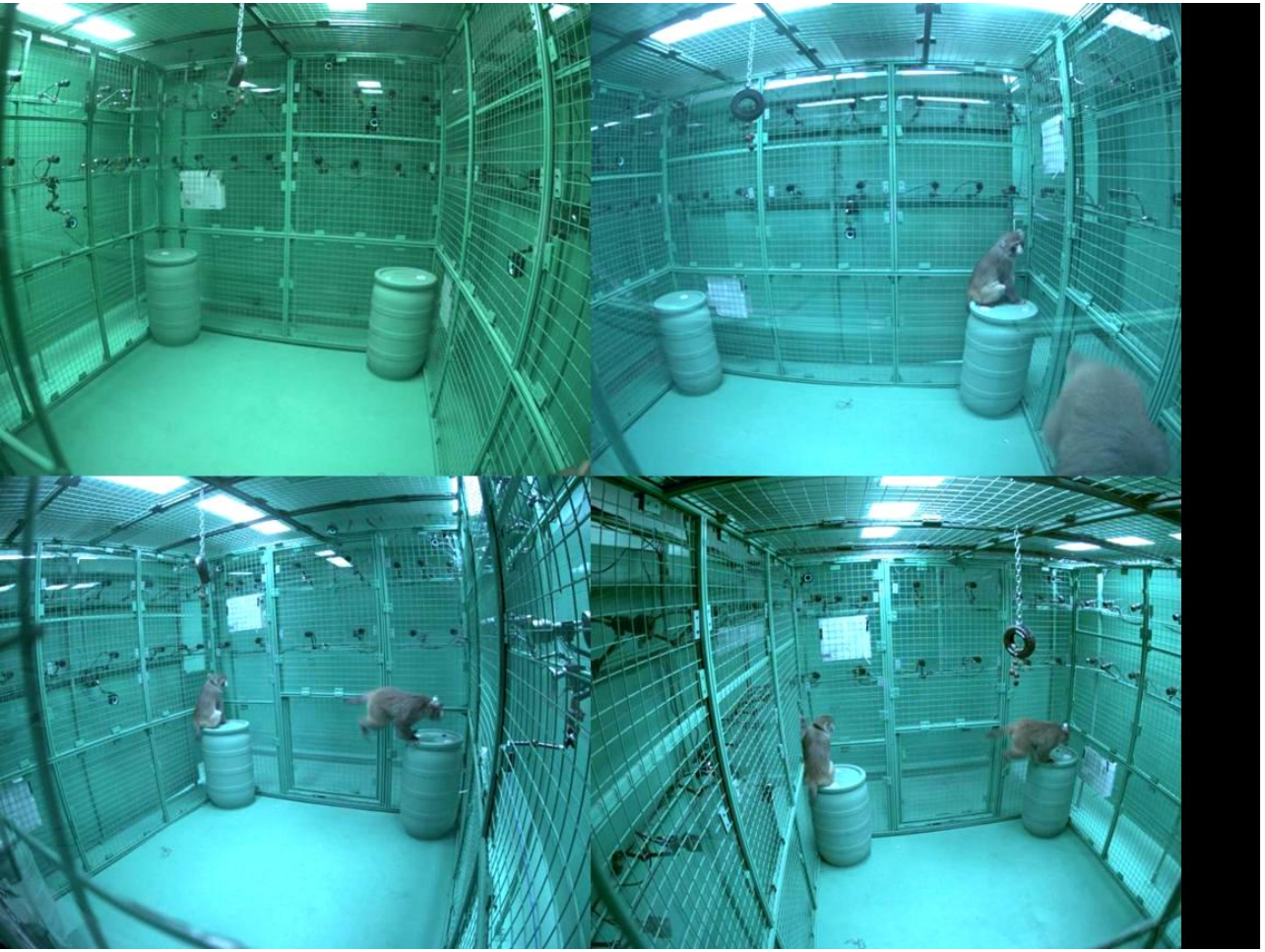


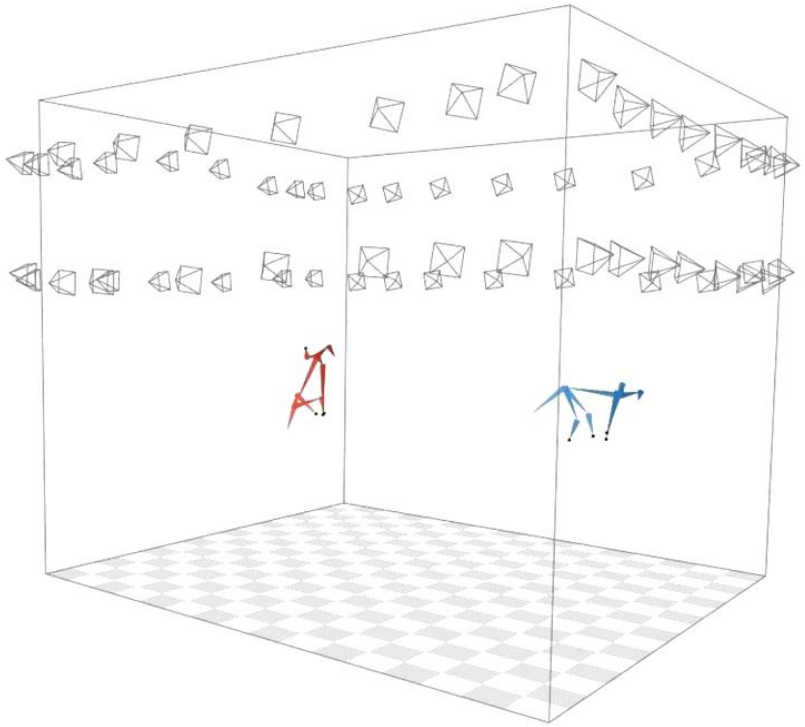
Pose Estimation with 3D Output





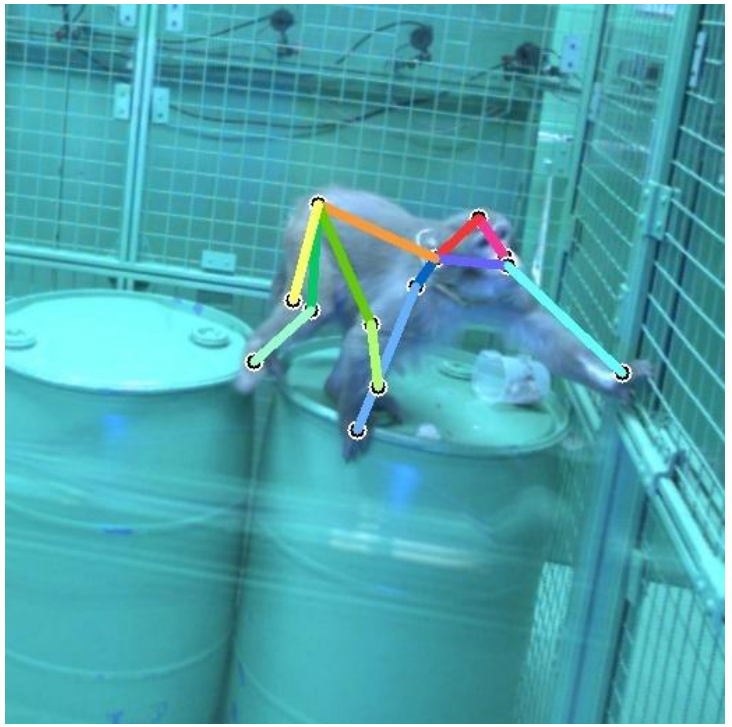
Social Monkey Data





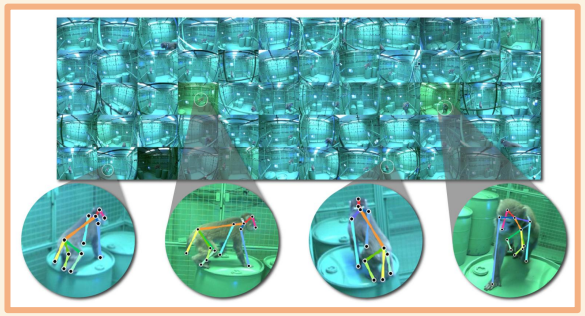
OMS

- ✓ 195k images
- ✓ 13 landmarks
- ✓ 6-7x augmentation
- ✓ Marker level accuracy
- ✓ Robust performance across viewpoints



Across viewpoints

- 3D geometry between cameras
- 195k images from 33k annotations
- *OpenMonkeyStudio*



Across environments



The behavioral record

Across granularity

Across time

"Does it hold outside the lab?"

OpenMonkeyChallenge: Bridging the Lab-to-Wild Gap

- OpenMonkeyStudio: 195,228 images, lab-only, controlled
- Existing wild datasets: Too small
- Gaps: Models trained in labs fail in natural habitats

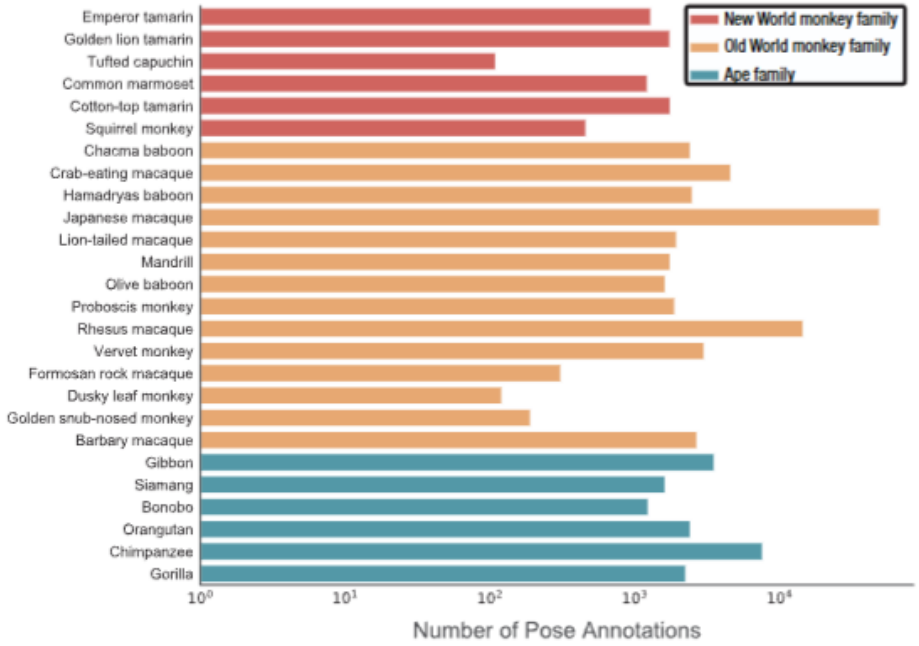


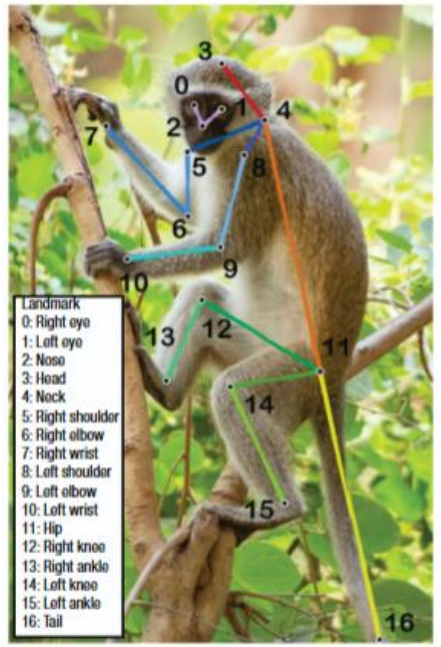
Large-scale, diverse dataset for generalizable primate pose estimation

OpenMonkeyChallenge Dataset

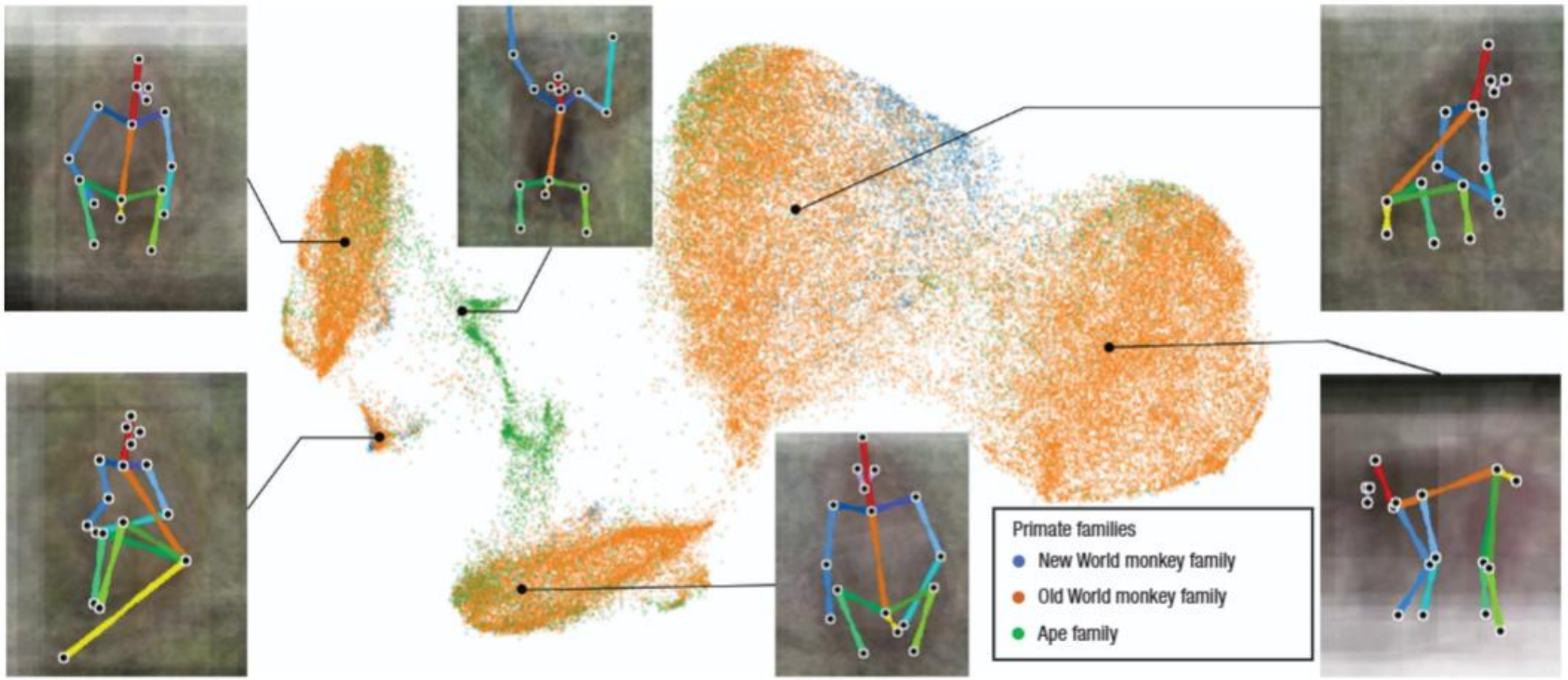
- **111,529** images
- **26** species (***6*** *New World,* ***14*** *Old World,* ***6*** *apes*)
- **17** landmarks per image







Dataset Covers Diverse Poses Across Species

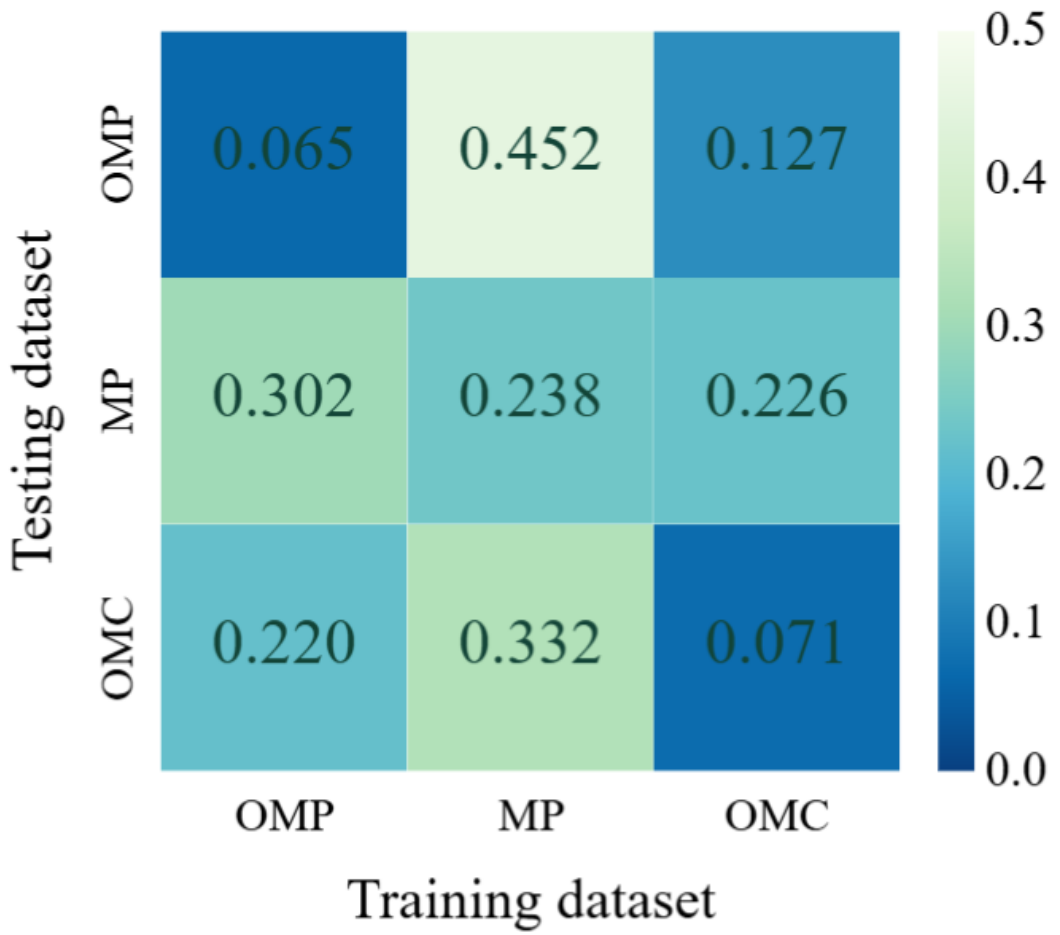


UMAP visualization shows OMC covers comprehensive pose space across primate families

Evaluation Results

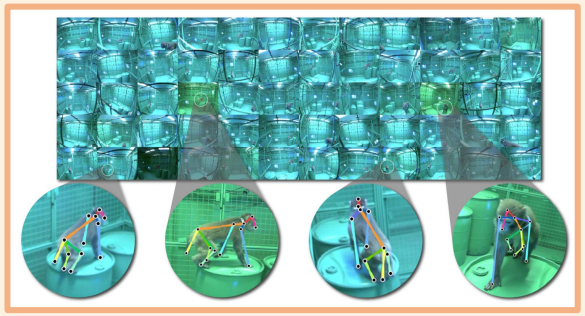
Cross-Dataset Generalization

- Models trained on OMC generalize best across datasets
- Key insight: ***Diversity improves generalization***



Across viewpoints

- 3D geometry between cameras
- 195k images from 33k annotations
- *OpenMonkeyStudio*





The behavioral record

Across environments

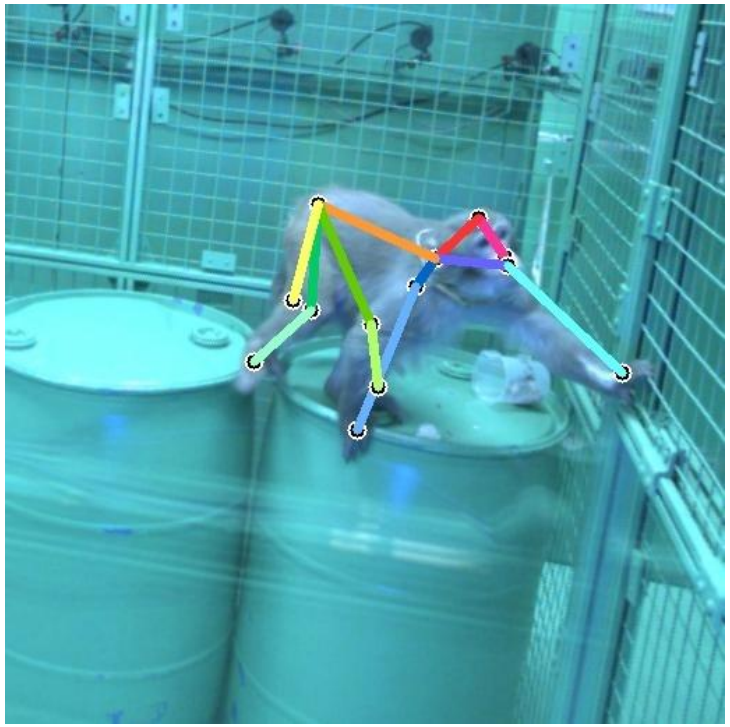
- 2D geometry of the skeleton
- 111,529 images, 26 species
- OpenMonkeyChallenge



Across granularity

Across time

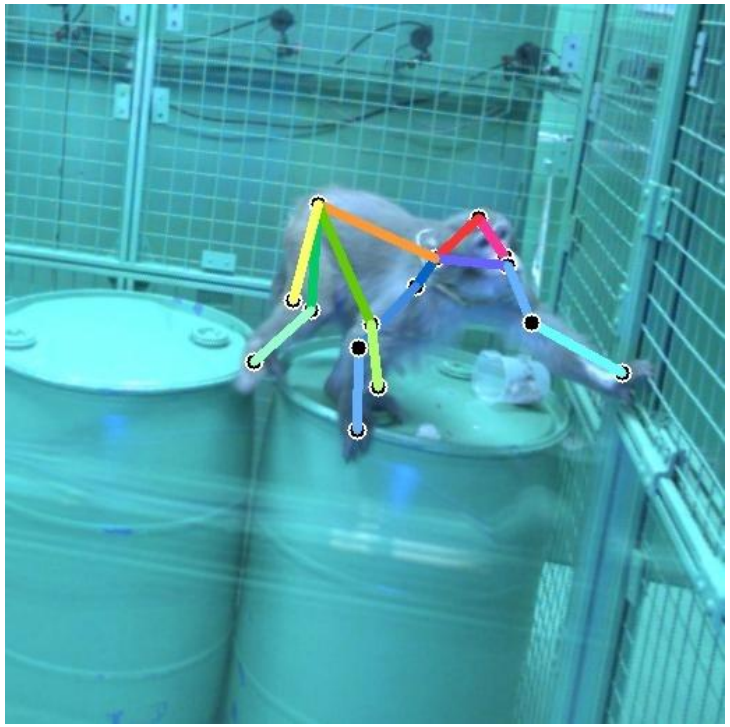
"Can we get denser without annotating more?"

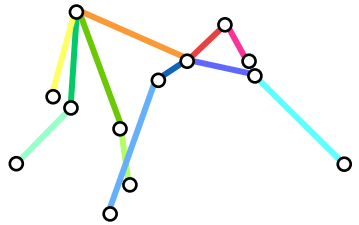


Annotated Landmarks:

- Nose
- Head
- Neck
- Right Shoulder
- Right Hand
- Left Shoulder
- Left Hand
- Hip
- Right Knee
- Right Foot
- Left Knee
- Left Foot
- Tail

13 Landmarks!

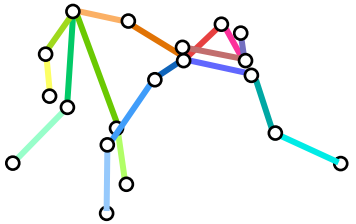




Primary Landmarks:

Animated Landmarks:

- Nose
- Head
- Neck
- Right Shoulder
- Right Hand
- Left Shoulder
- Left Hand
- Hip
- Right Knee
- Right Foot
- Left Knee
- Left Foot
- Tail



Secondary Landmarks:

Additional Landmarks:

- Right Elbow
- Left Elbow
- Eyes
- Ears
- Spine

We could label the data manually!



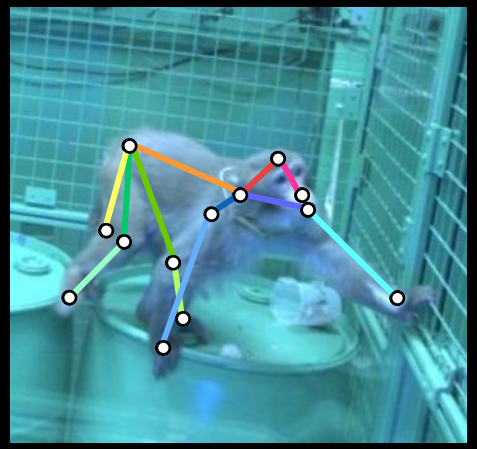
OpenMonkeyStudio Dataset

OpenMonkeyStudio+ ?

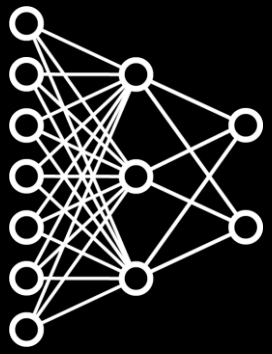


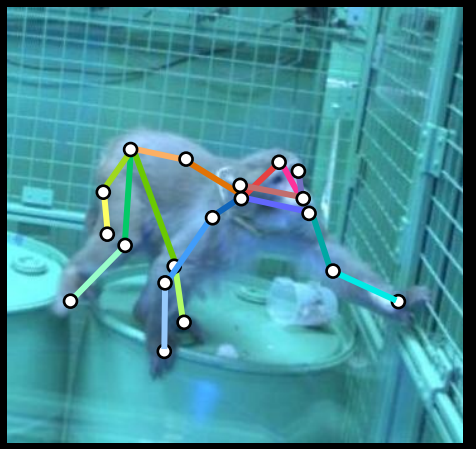
Challenges

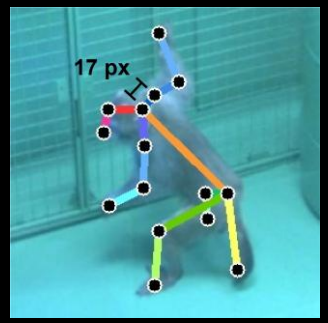
Leverage the existing data?

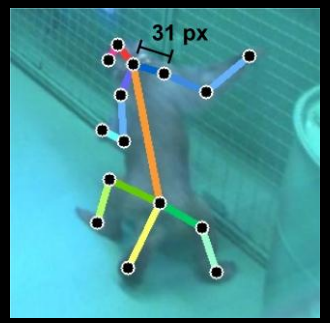


Prediction

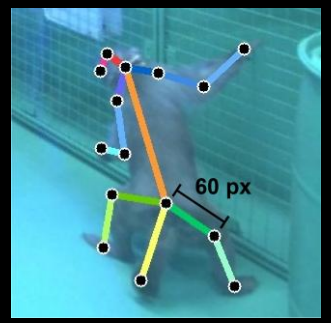


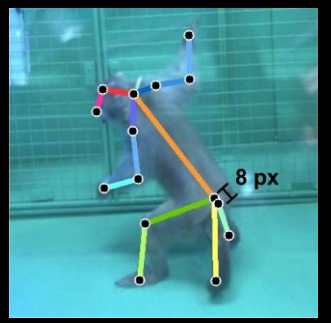






Neck to Right Shoulder

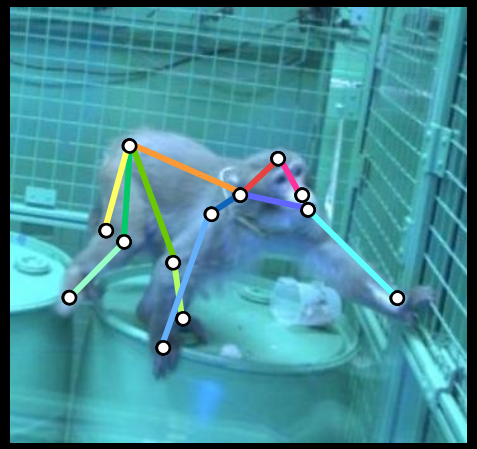




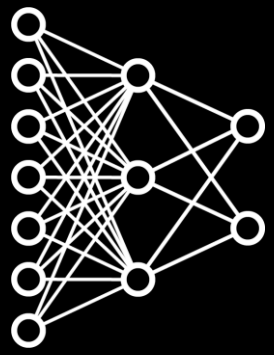
Hip to Knee

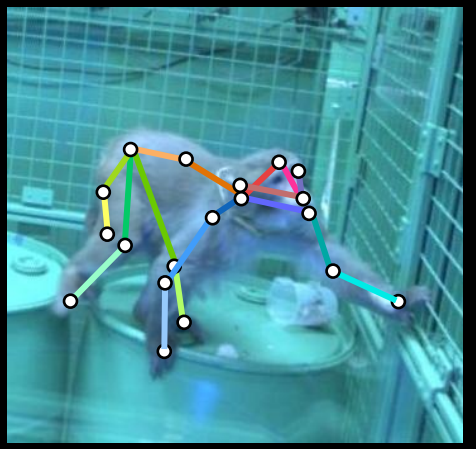
2D images ambiguous for limb references

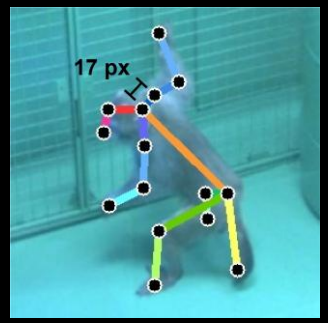
Leverage the existing data?

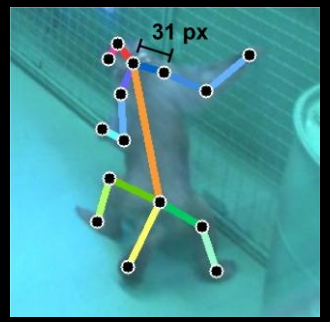


Prediction

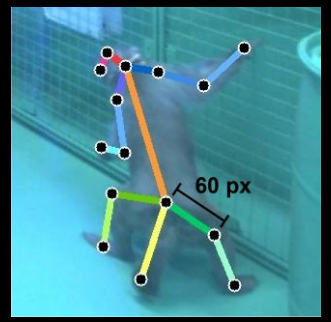


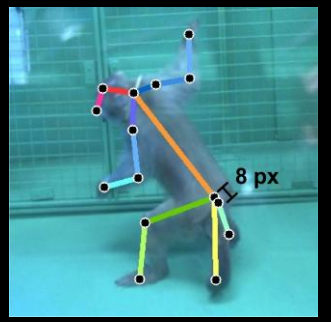






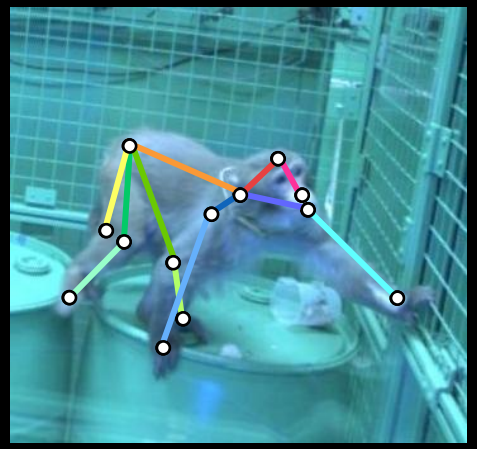
Neck to Right Shoulder



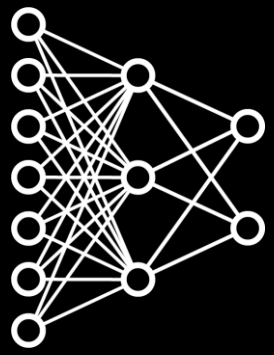


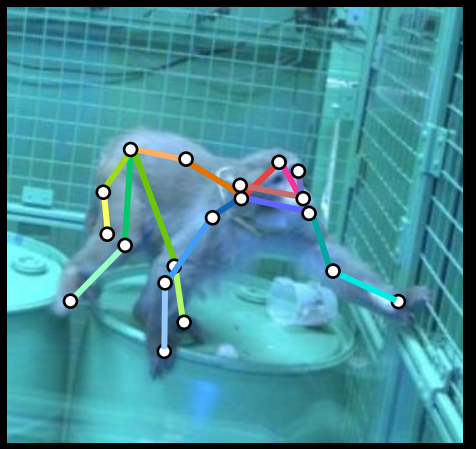
Hip to Knee

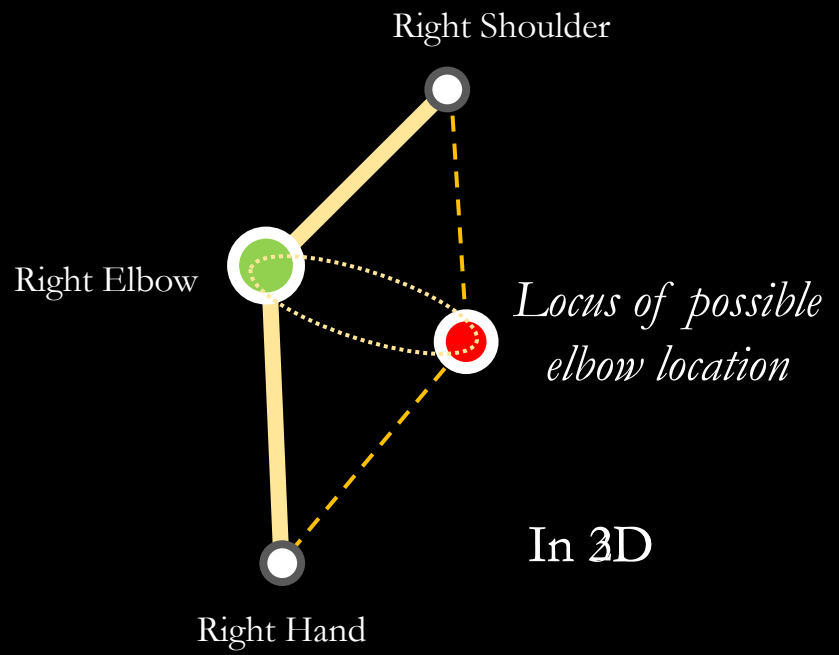
2D images ambiguous for limb references

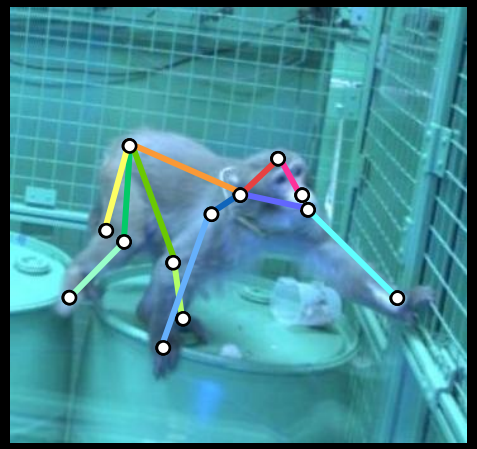


Prediction

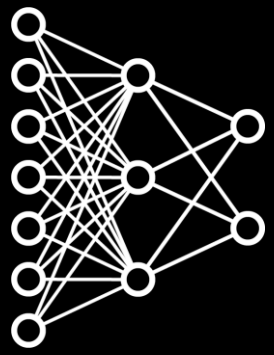


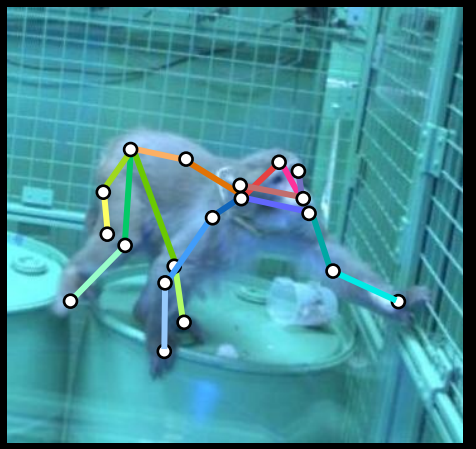


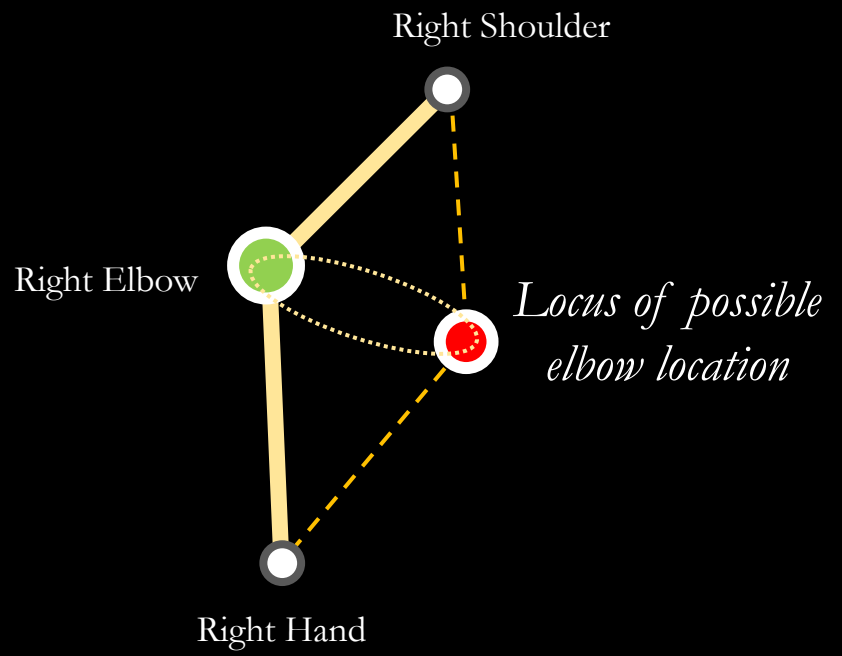

Right Shoulder
Right Elbow
Locus of possible elbow location
In 3D
Right Hand



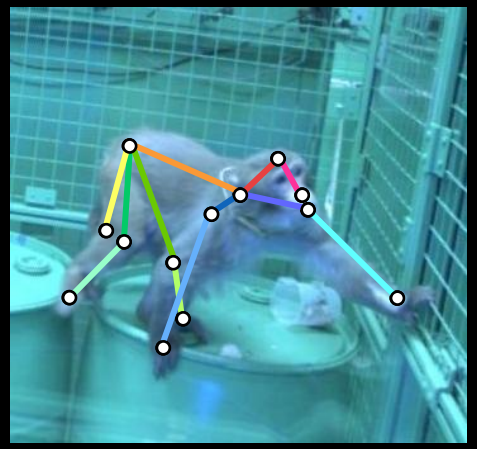
Prediction





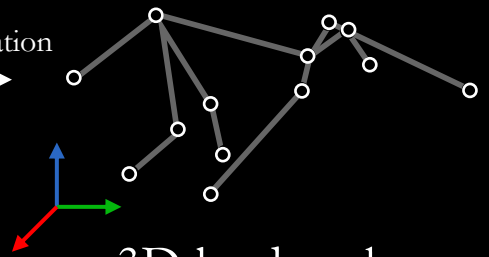

Right Shoulder
Right Elbow
Locus of possible elbow location
Right Hand

The Idea

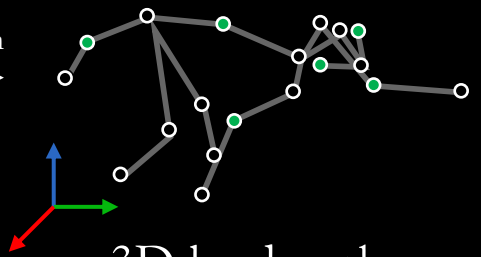




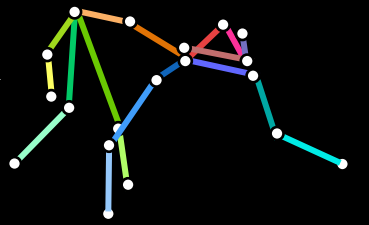
Lifting operation



Prediction



Reprojection



2D landmarks
(Primary)

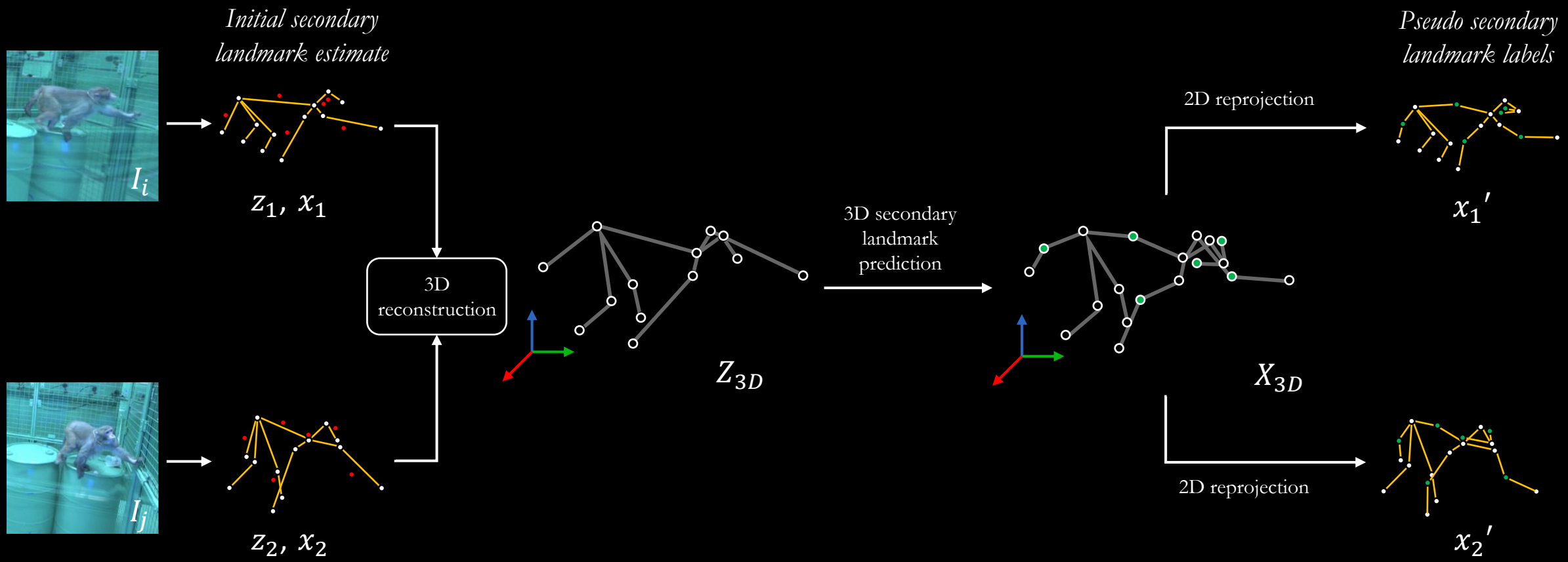
3D landmarks
(Primary)

3D landmarks
(Primary + Secondary)

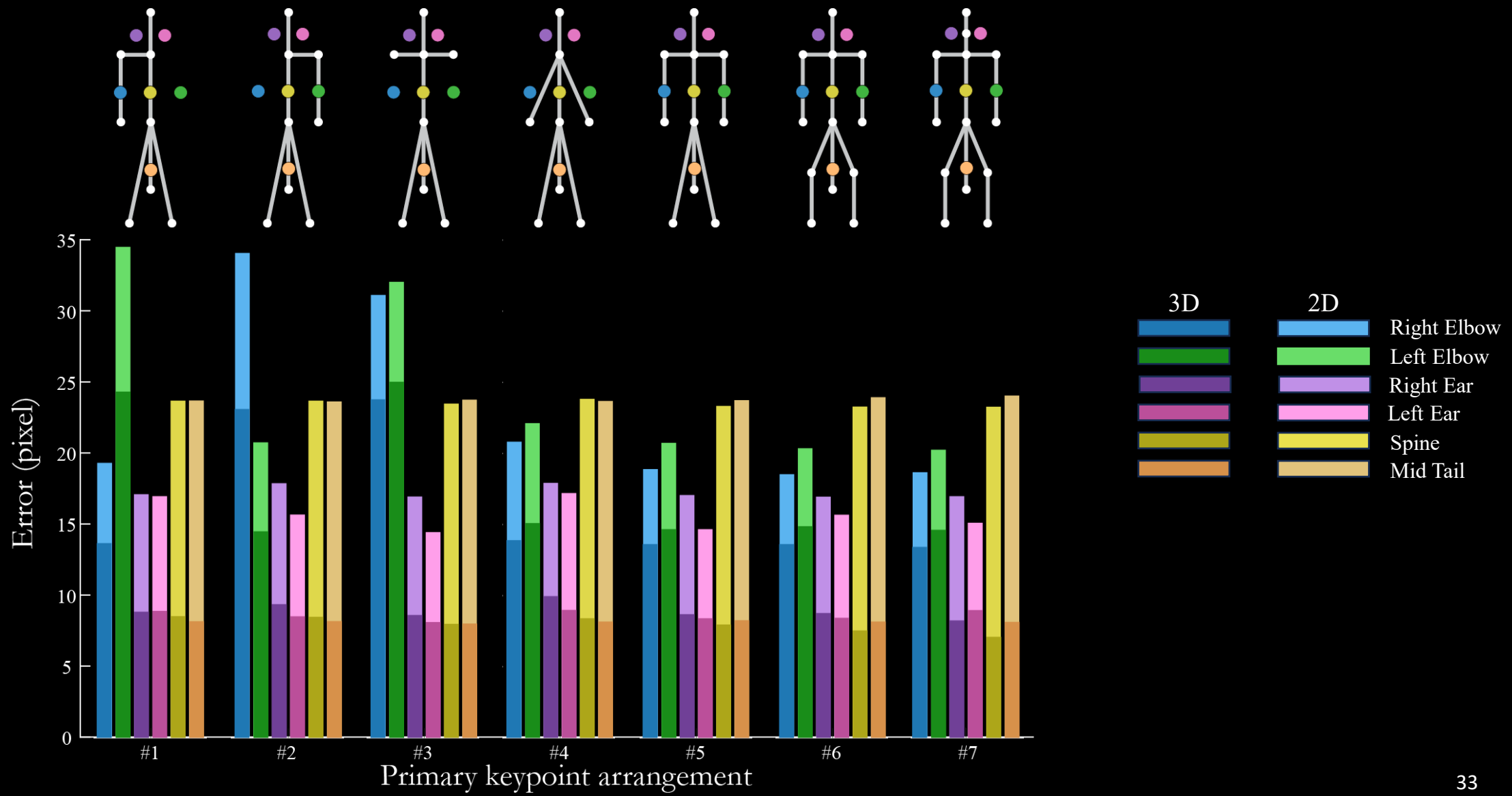
2D landmarks
(Primary + Secondary)

Method

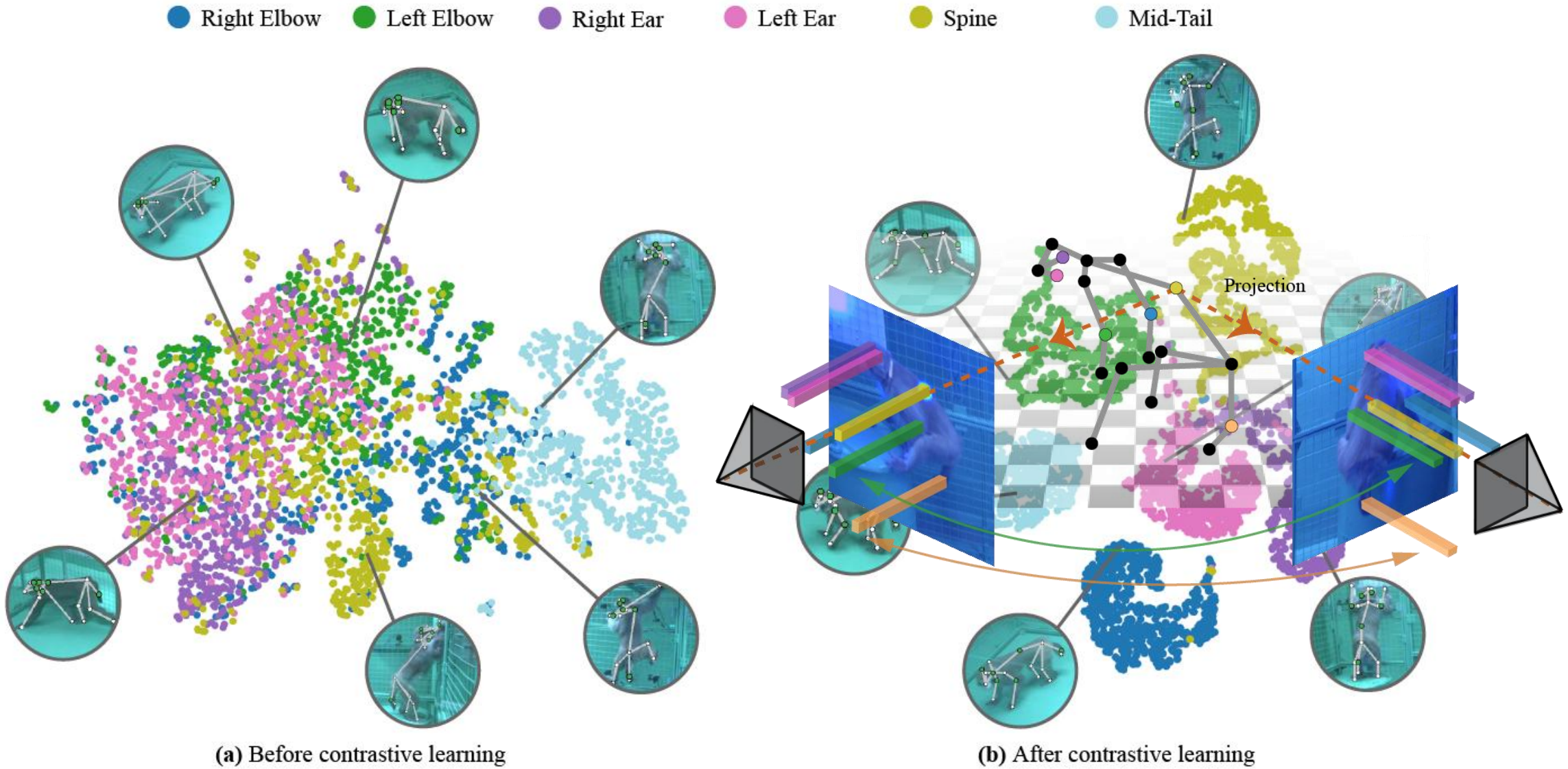
1. Abide by the 3D property of the body
2. 2D discriminative



Shared Representation Analysis – Why 3D and not 2D?



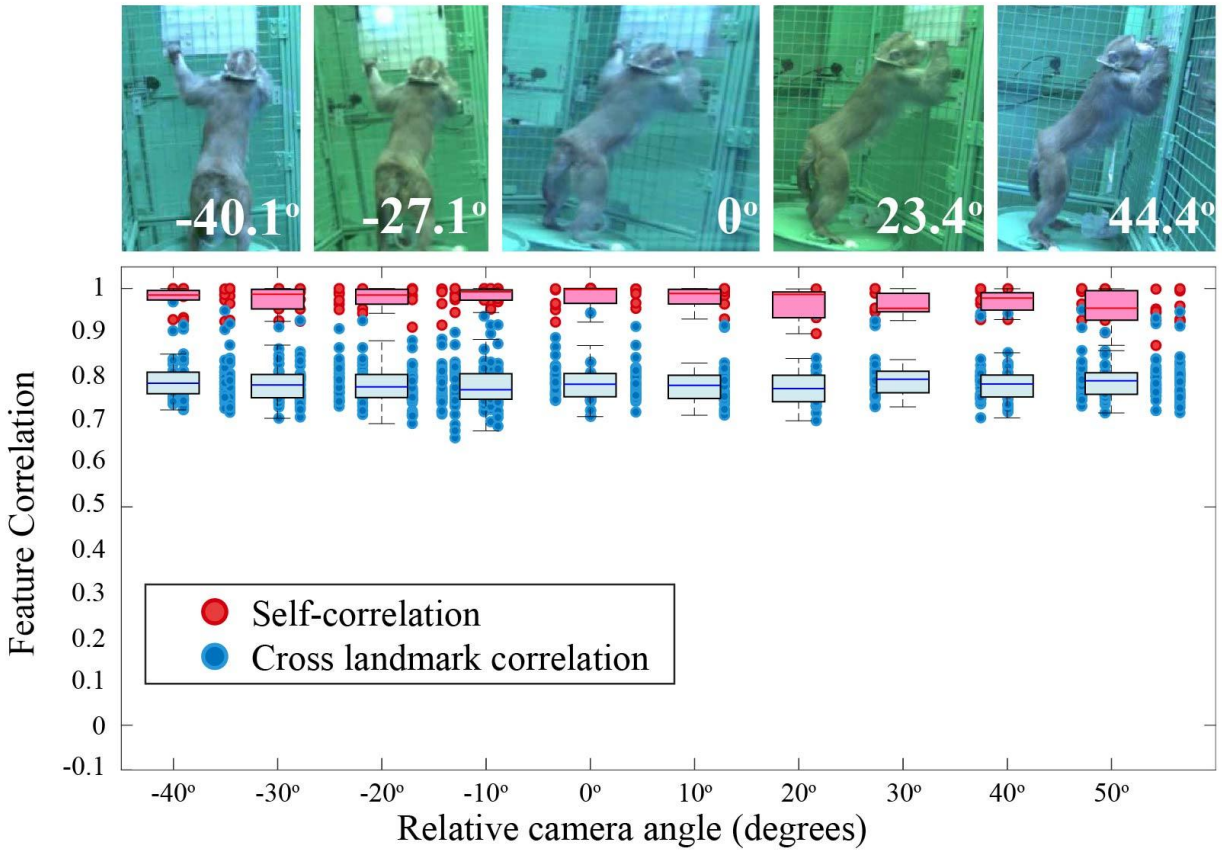
Impact of Contrastive Learning



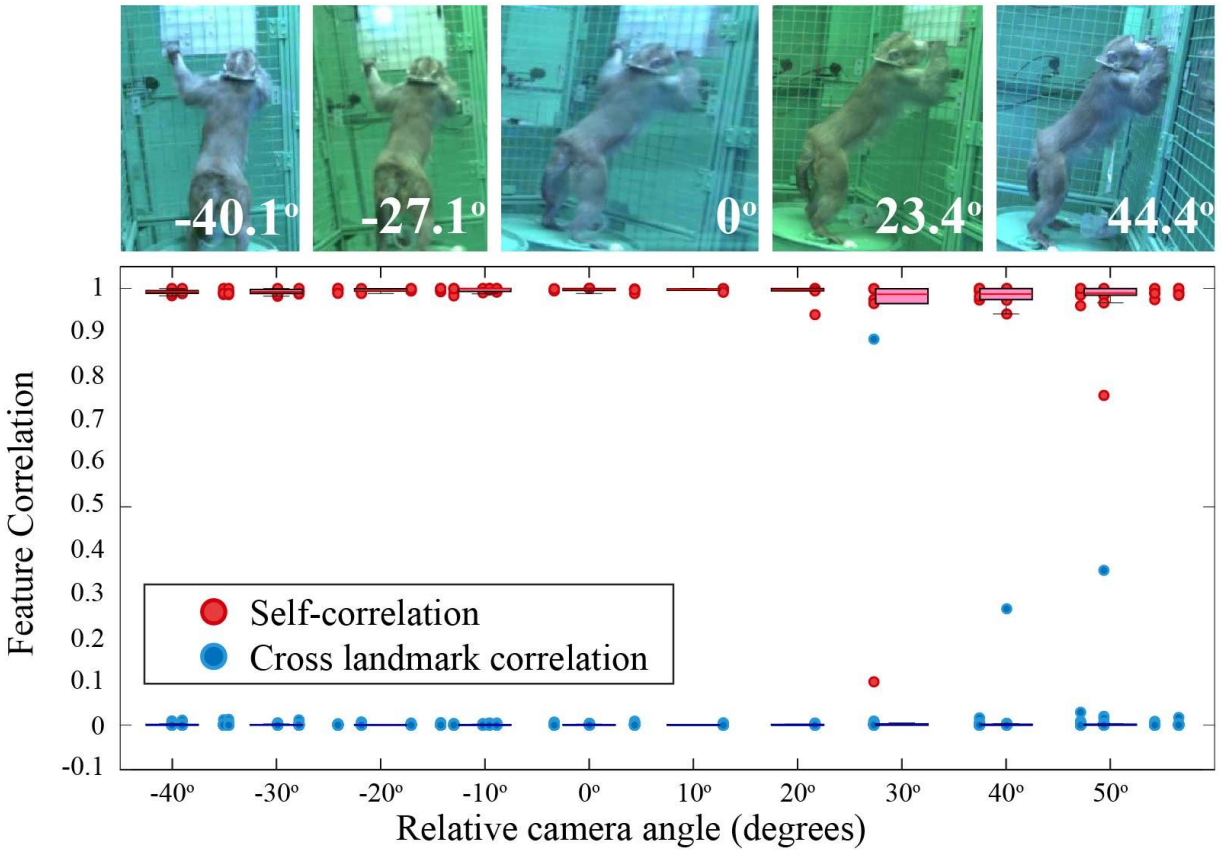
(a) Before contrastive learning

(b) After contrastive learning

Contrastive Learning: Quantitative



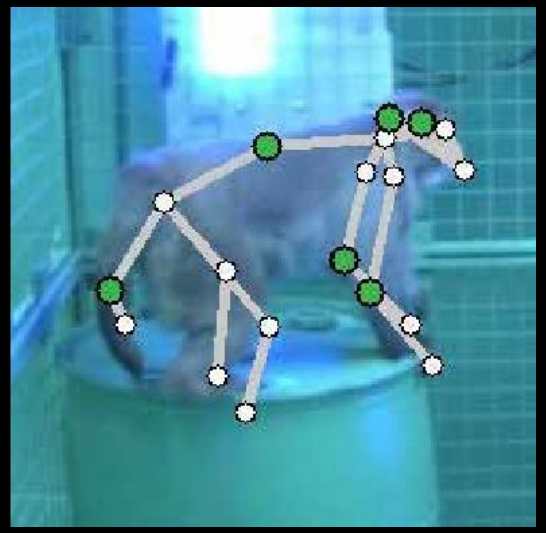
(a) Before contrastive learning



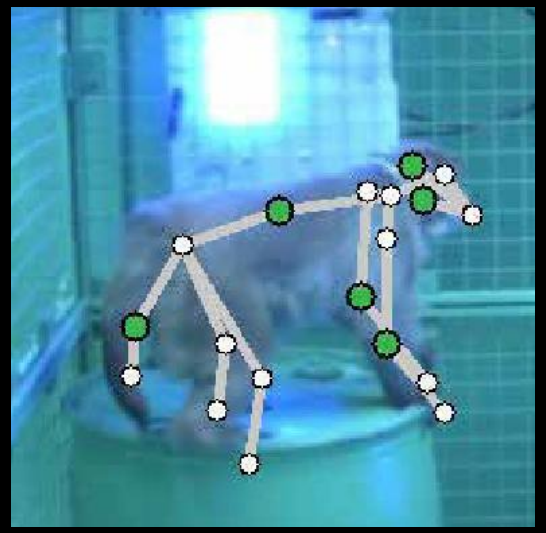
(b) After contrastive learning

Qualitative comparison

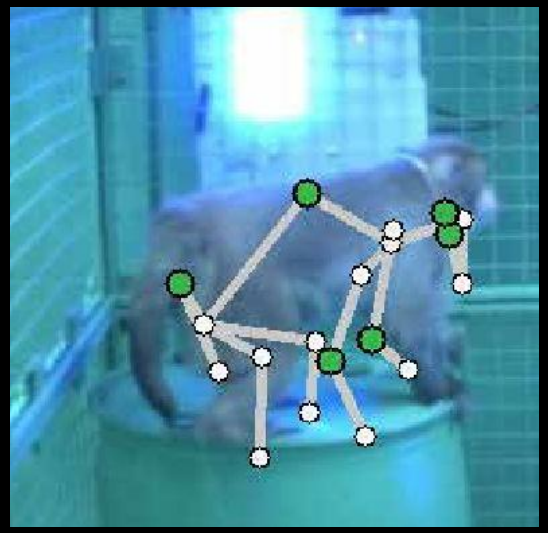
GT



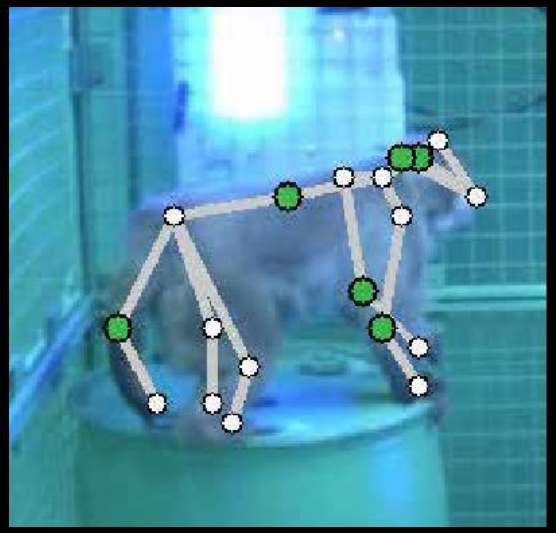
VAE(2D)



VAE(3D)

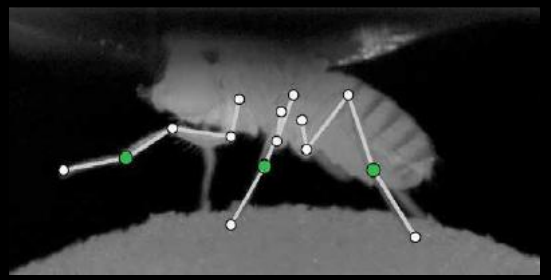


Ours

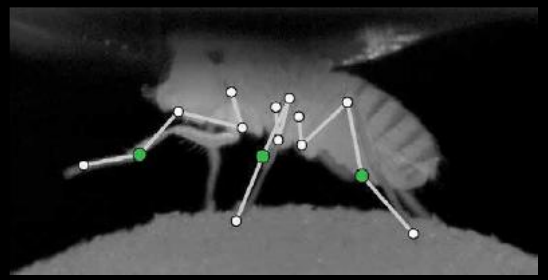


Qualitative comparison

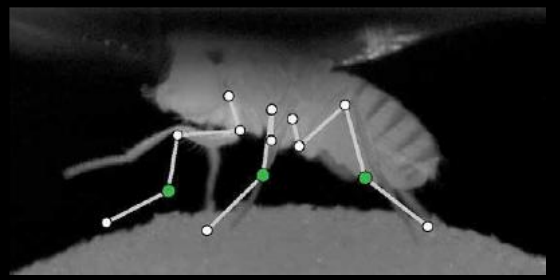
GT



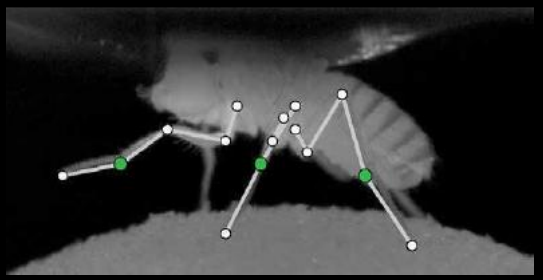
VAE (2D)



VAE (3D)

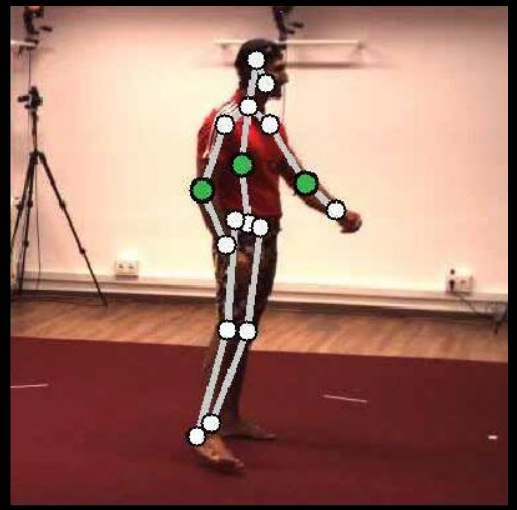


Ours

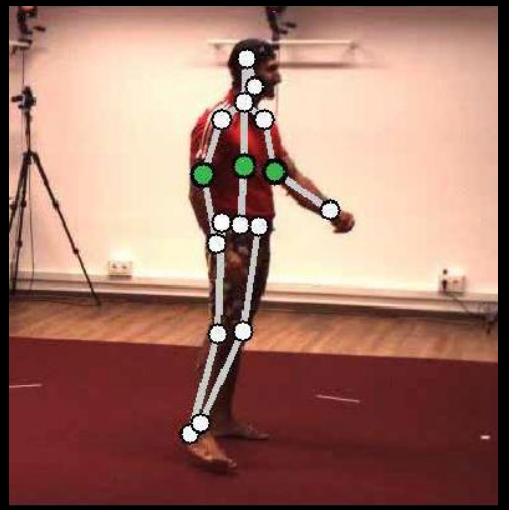


Qualitative comparison

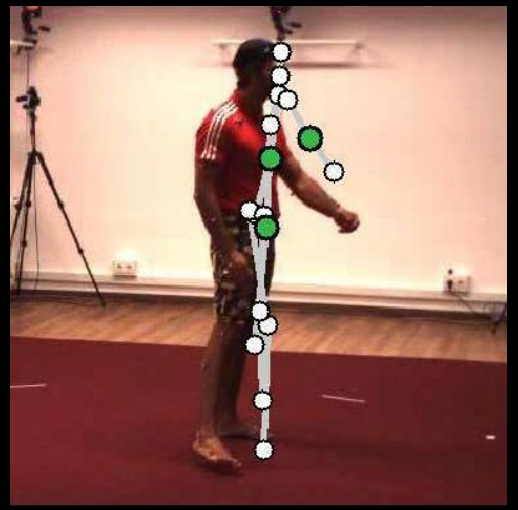
GT



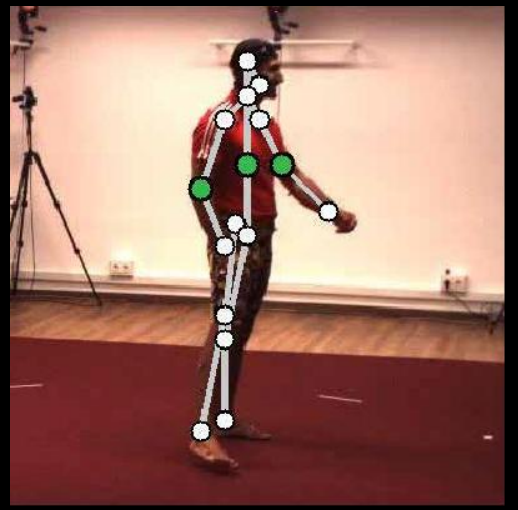
VAE(2D)



VAE(3D)



Ours

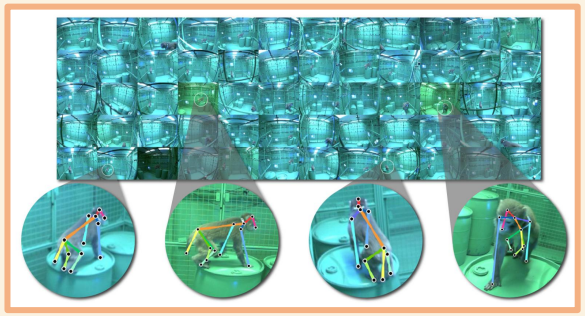


Secondary Landmark Detection

- ✓ 5 additional landmarks detected
- ✓ Works on humans, macaques, and flies: no dependence on body plan
- ✓ The relationship: 3D geometry of the skeleton

Across viewpoints

- 3D geometry between cameras
- 195k images from 33k annotations
- *OpenMonkeyStudio*





The behavioral record

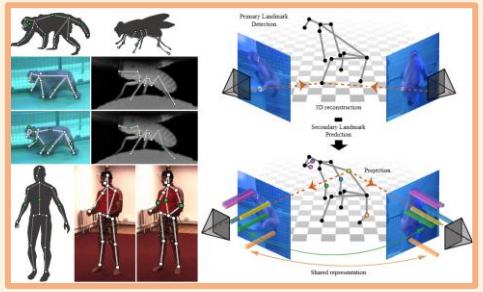
Across environments

- 2D geometry of the skeleton
- 111,529 images, 26 species
- OpenMonkeyChallenge



Across granularity

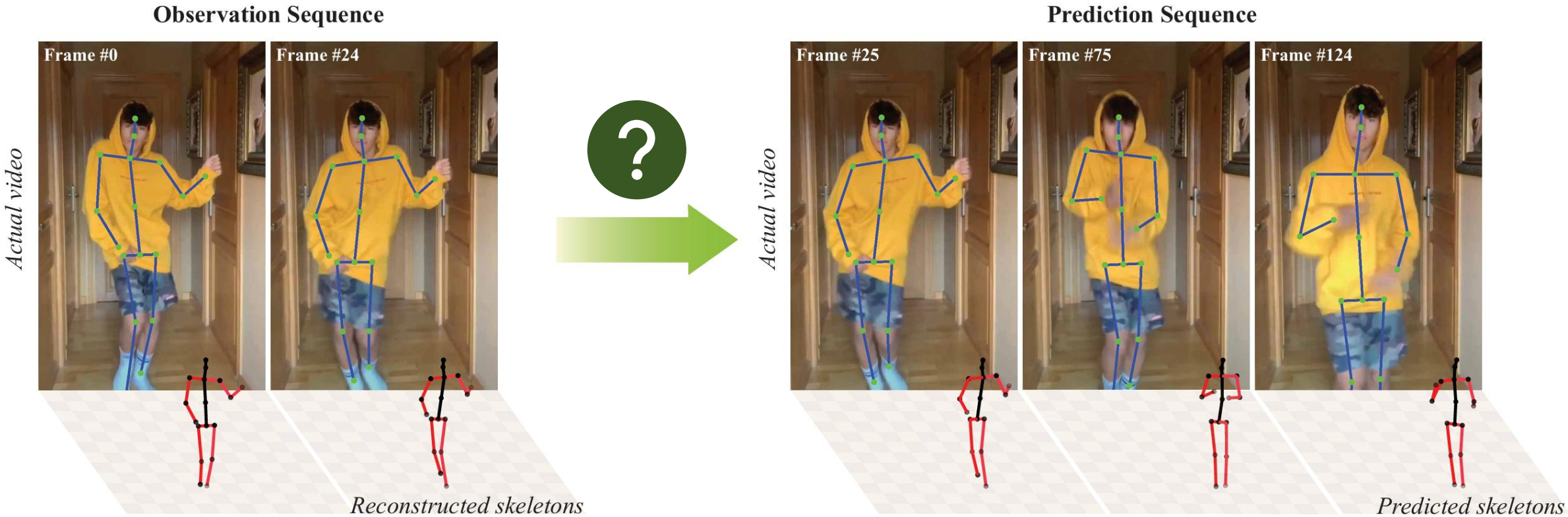
- 3D geometry of the skeleton
- 13 → 19 landmarks, no new annotation campaign
- Secondary Landmark Detection



Across time

"Can we predict what comes next?"

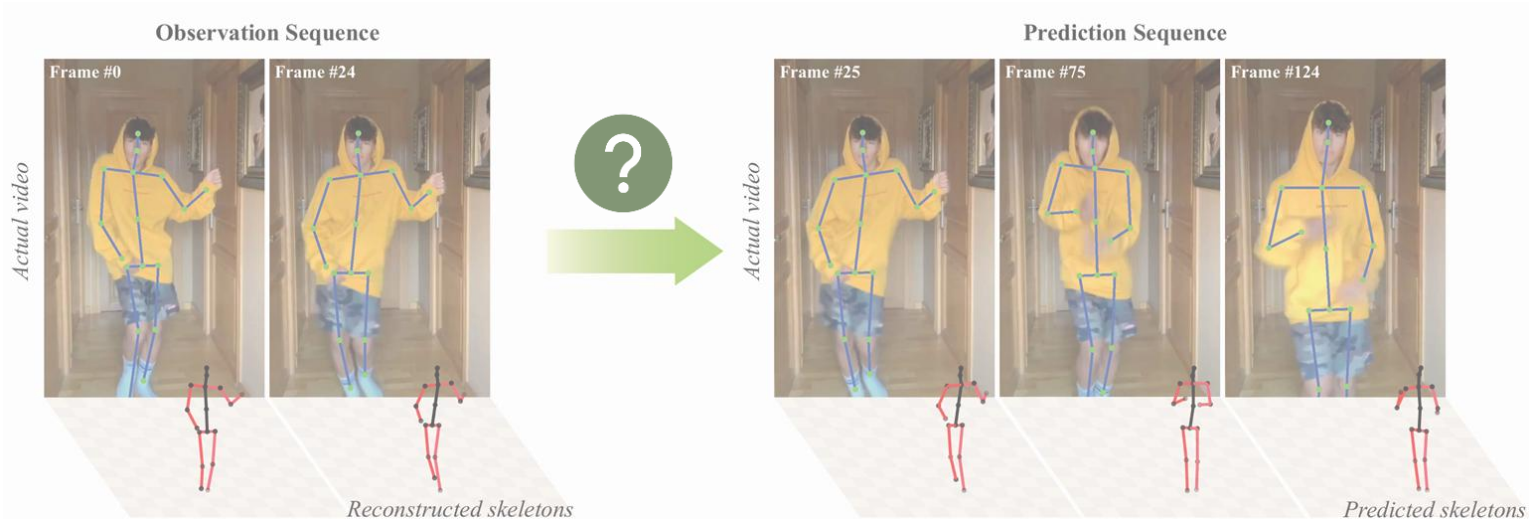
Motion Forecasting: From 2D Observations to Future 3D Motion



Motion Forecasting: From 2D Observations to Future 3D Motion

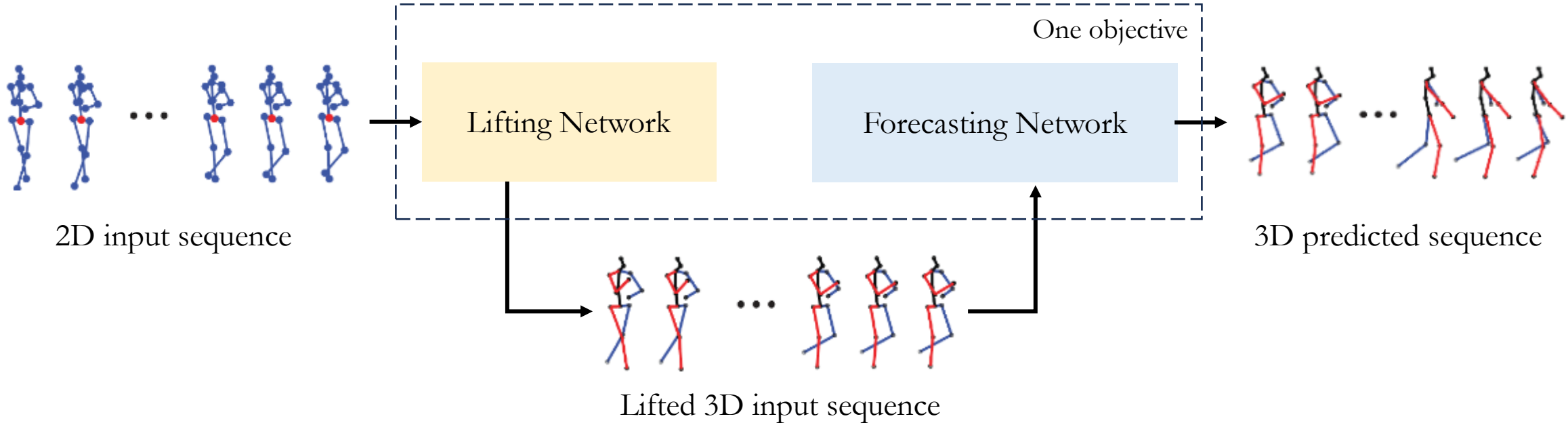
Challenges

- Hard long-term prediction
- Domain shift
- Limited 3D supervision



Motion Forecasting: From 2D Observations to Future 3D Motion

Should we separate 2D → 3D lifting and forecasting, or unify them?



Coupled objective, decoupled parameters

Motion Forecasting: From 2D Observations to Future 3D Motion

Adapting to a new domain

Pre-training (Human 3.6M)

- 2D sequences + 3D ground truth
- Lifting network and forecaster coupled through one objective



Adaptation (new domain)

- 2D keypoints only
- Lifting network produces pseudo-3D labels
- No 3D ground truth required

Motion Forecasting: From 2D Observations to Future 3D Motion

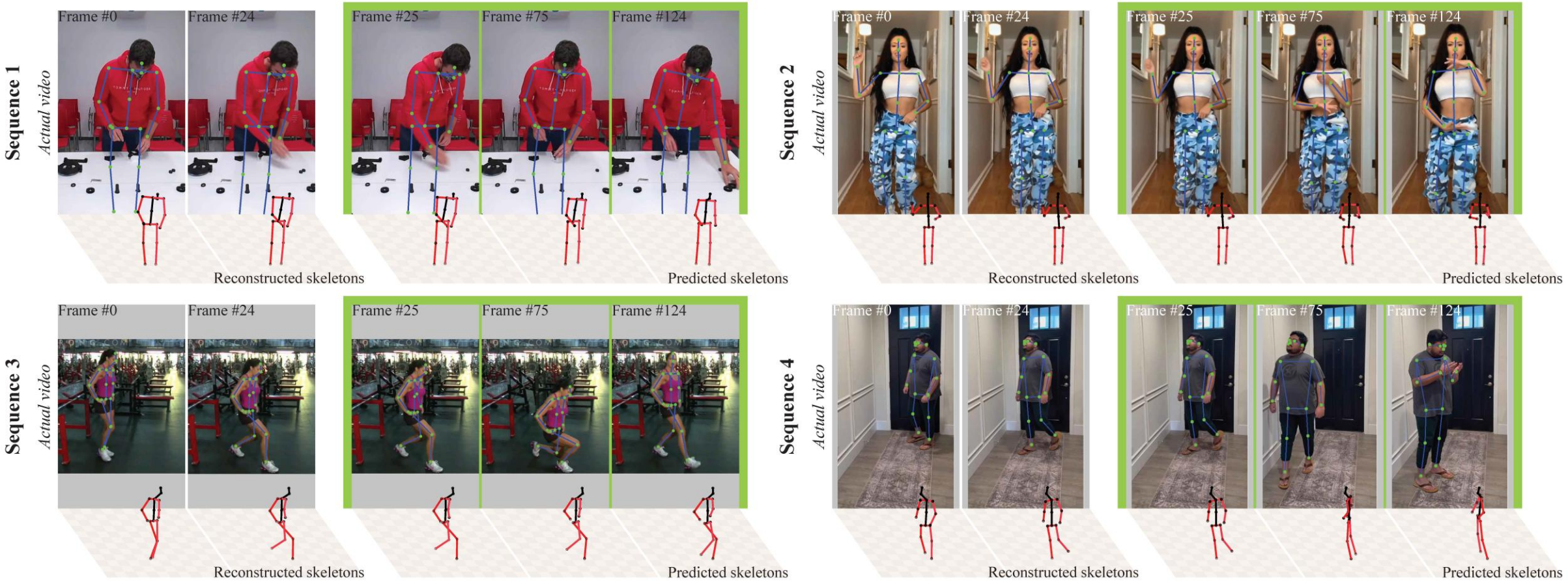
Why human benchmarks?

- Adaptation needs **continuous single-view 2D**: ~4 s, one camera, observation and future.
- OMS gives continuous 3D. It does not give continuous single-view 2D.
- Evaluated on Human3.6M and MPI-INF-3DHP.

Motion Forecasting: From 2D Observations to Future 3D Motion

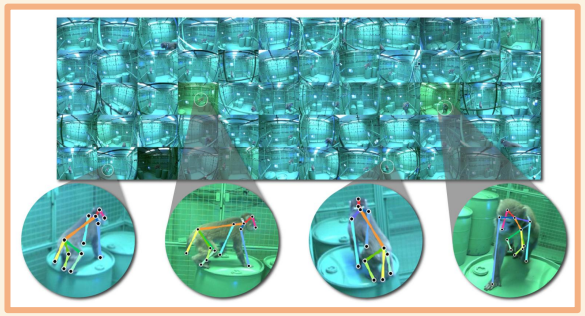
Qualitative results

In-the-wild video (YouTube, TikTok): no ground truth



Across viewpoints

- 3D geometry between cameras
- 195k images from 33k annotations
- *OpenMonkeyStudio*





The behavioral record

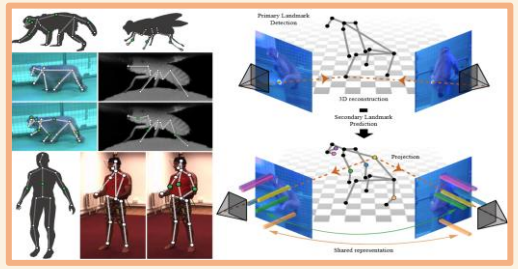
Across environments

- 2D geometry of the skeleton
- 111,529 images, 26 species
- OpenMonkeyChallenge



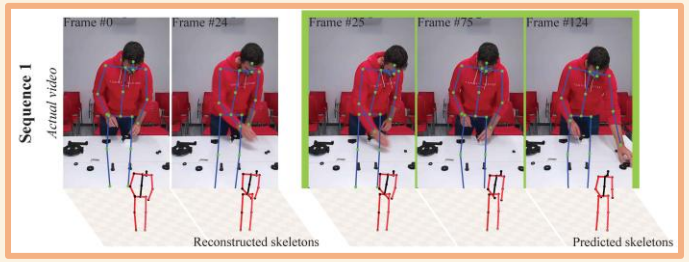
Across granularity

- 3D geometry of the skeleton
- 13 → 19 landmarks, no new annotation campaign
- Secondary Landmark Detection

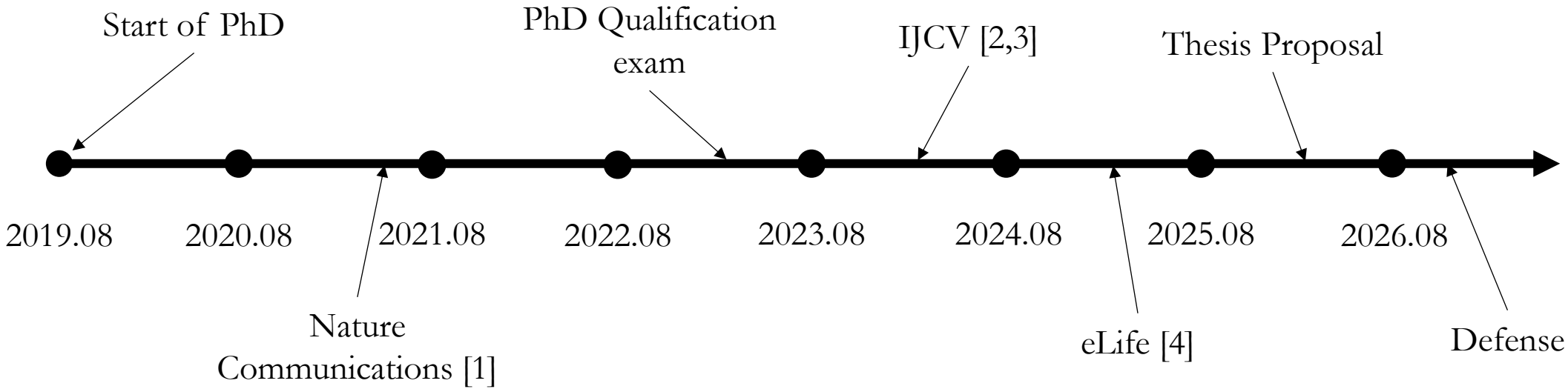


Across time

- How the skeleton carries forward
- 100 frames from 2D; adapts with no 3D ground truth
- Motion Forecasting



Timeline



[1] PC Bala, BR Eisenreich, SBM Yoo, BY Hayden, HS Park, J Zimmermann. "Automated Markerless Pose Estimation in Freely Moving Macaques with OpenMonkeyStudio", Nature Communications 2020.
[2] Y Yao, P Bala, A Mohan, E Bliss-Moreau, K Coleman, SM Freeman, CJ Machado, J Raper, J Zimmermann, BY Hayden , HS Park. "OpenMonkeyChallenge: Dataset and Benchmark Challenges for Pose Estimation of Non-Human Primates", IJCV 2023.
[3] P Bala, J Zimmermann, HS Park, BY Hayden. "Self-supervised Secondary Landmark Detection via 3D Representation Learning", IJCV 2023.
[4] N Desai, P Bala, R Richardson, J Raper, J Zimmermann, BY Hayden. "OpenApePose: A Dataset of Annotated Ape Photographs for Pose Estimation", eLife 2024

Where this goes next?

- **Beyond geometry**
  Pairing metric ground truth with vision-language models to attach meaning to trajectories
- **Beyond forecasting**
  Discovering behavioral units from motion instead of defining landmarks in advance
- **Beyond one animal**
  The relational structure between animals: distance, orientation, approach, coordination
- **Beyond cameras**
  Audio, depth, accelerometry, neural recordings, aligned on calibrated 3D pose

METHODS FOR LEARNING THE LANGUAGE OF BEHAVIOR

OBSERVING, REPRESENTING, AND PREDICTING ACTION

Thank You!

