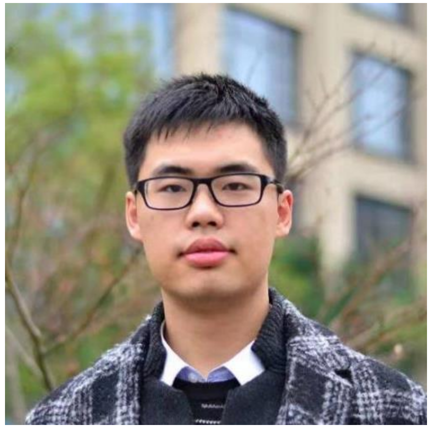
Unsupervised Learning of Visual 3D Keypoints for Control

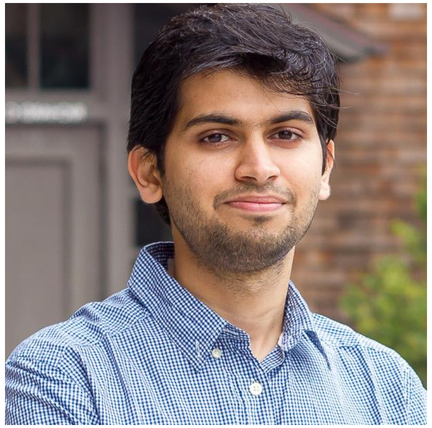


Boyuan Chen

UC Berkeley

Pieter Abbeel

UC Berkeley



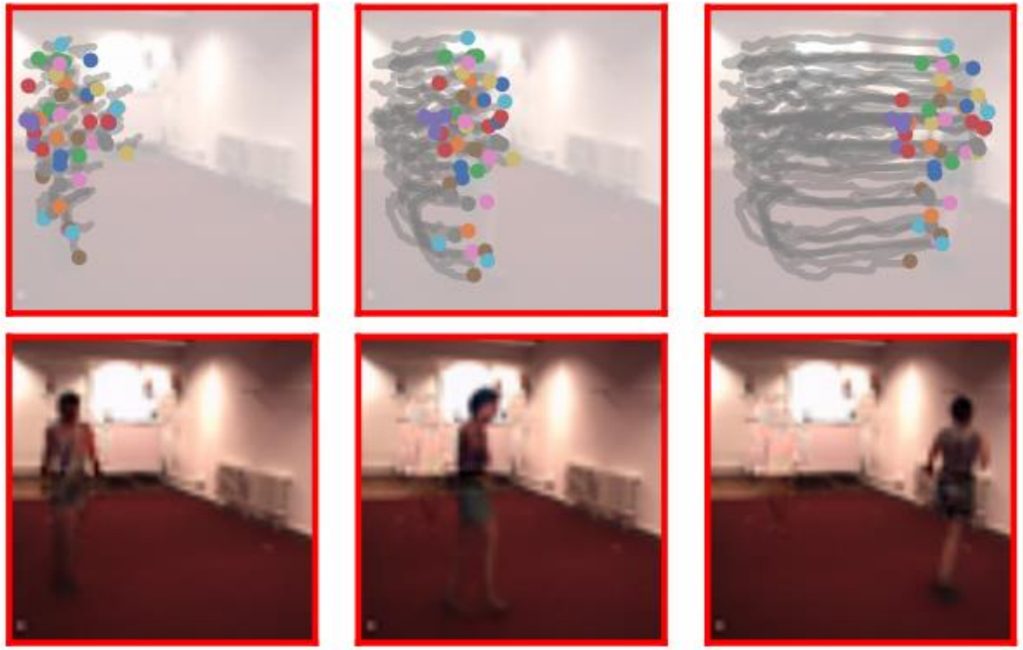
Deepak Pathak

CMU

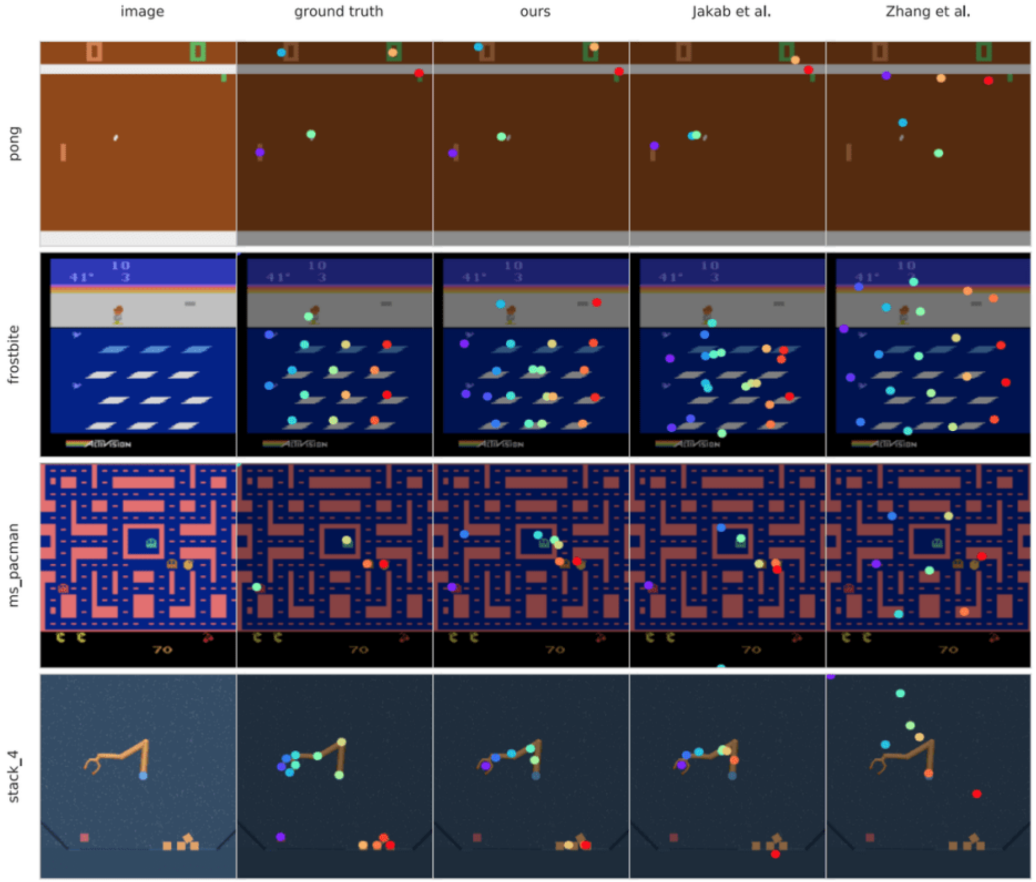
Learning keypoints from pixels

OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Cao et al., 2018

Unsupervised 2D Keypoints learning



Unsupervised Learning of Object Structure and Dynamics from Videos, Minderer et al., 2019

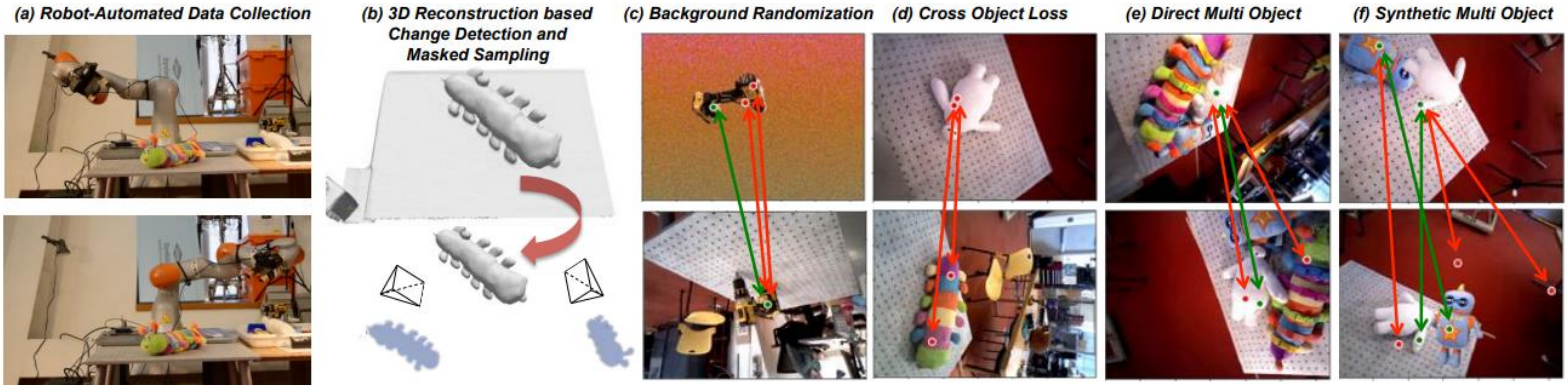


Unsupervised Learning of Object Keypoints for Perception and Control, Kulkarni, Gupta et al., 2019

Unsupervised 3D Keypoints

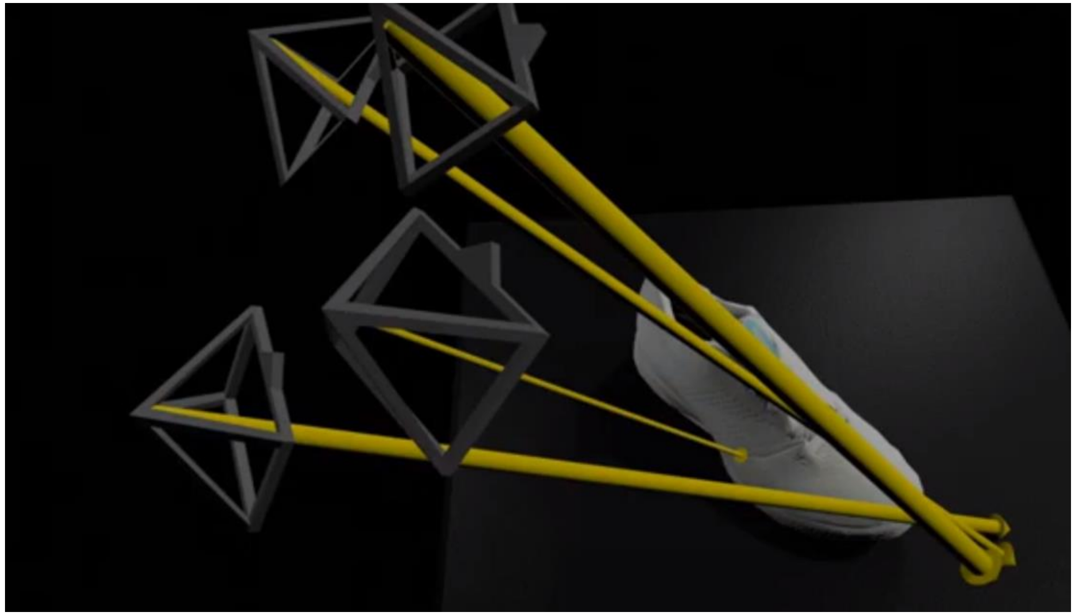
Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning, Suwajanakorn et al., 2018

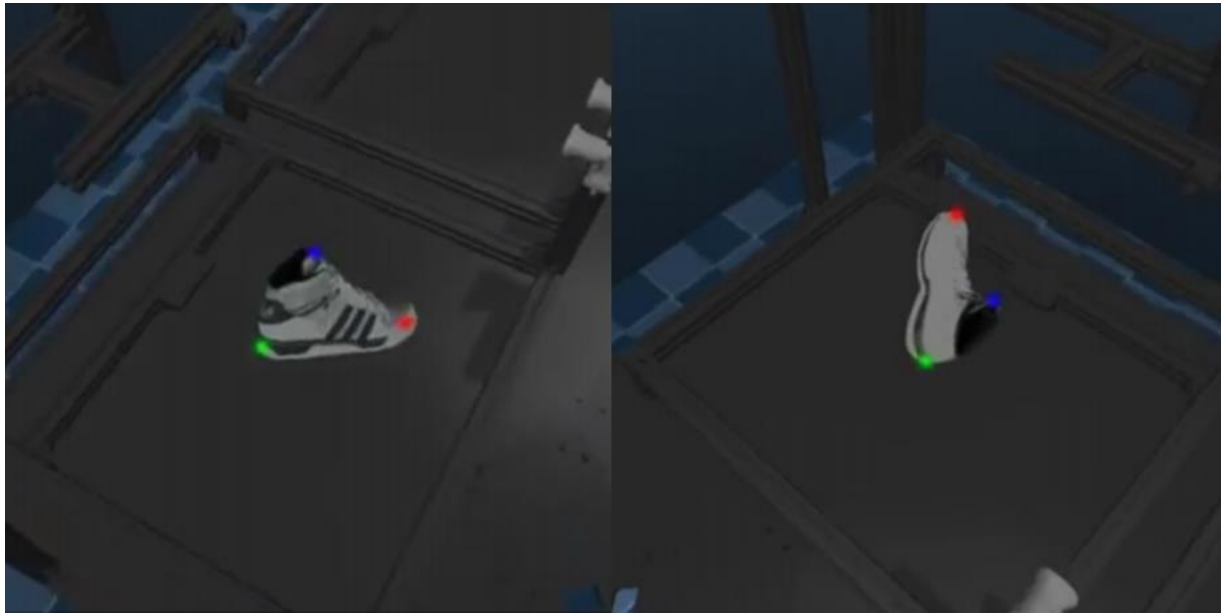
Self-supervised 3D structure learning



Dense Object Nets: Learning Dense Visual Object Descriptors By and For Robotic Manipulation, Florence et al., 2018

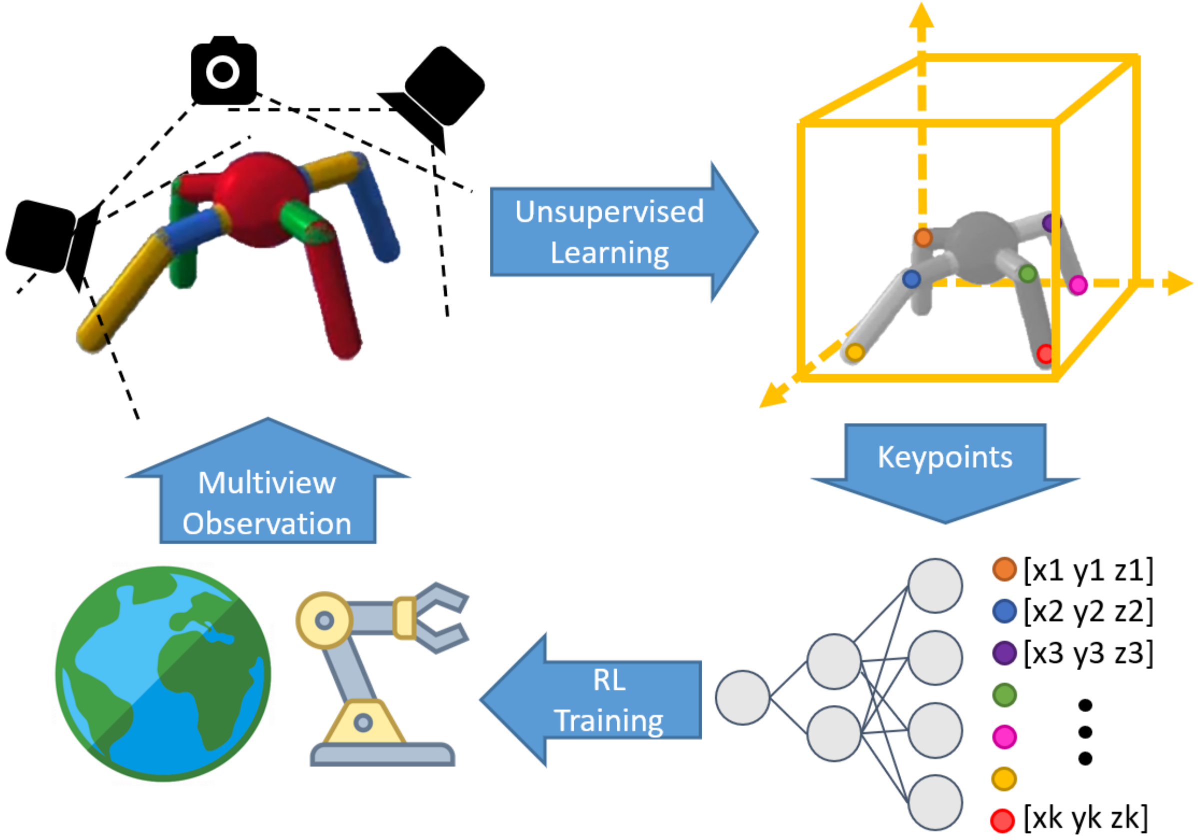
Semi-supervised 3D Keypoints





S3K: Self-Supervised Semantic Keypoints for Robotic Manipulation via Multi-View Consistency, Sushkov et al., 2020

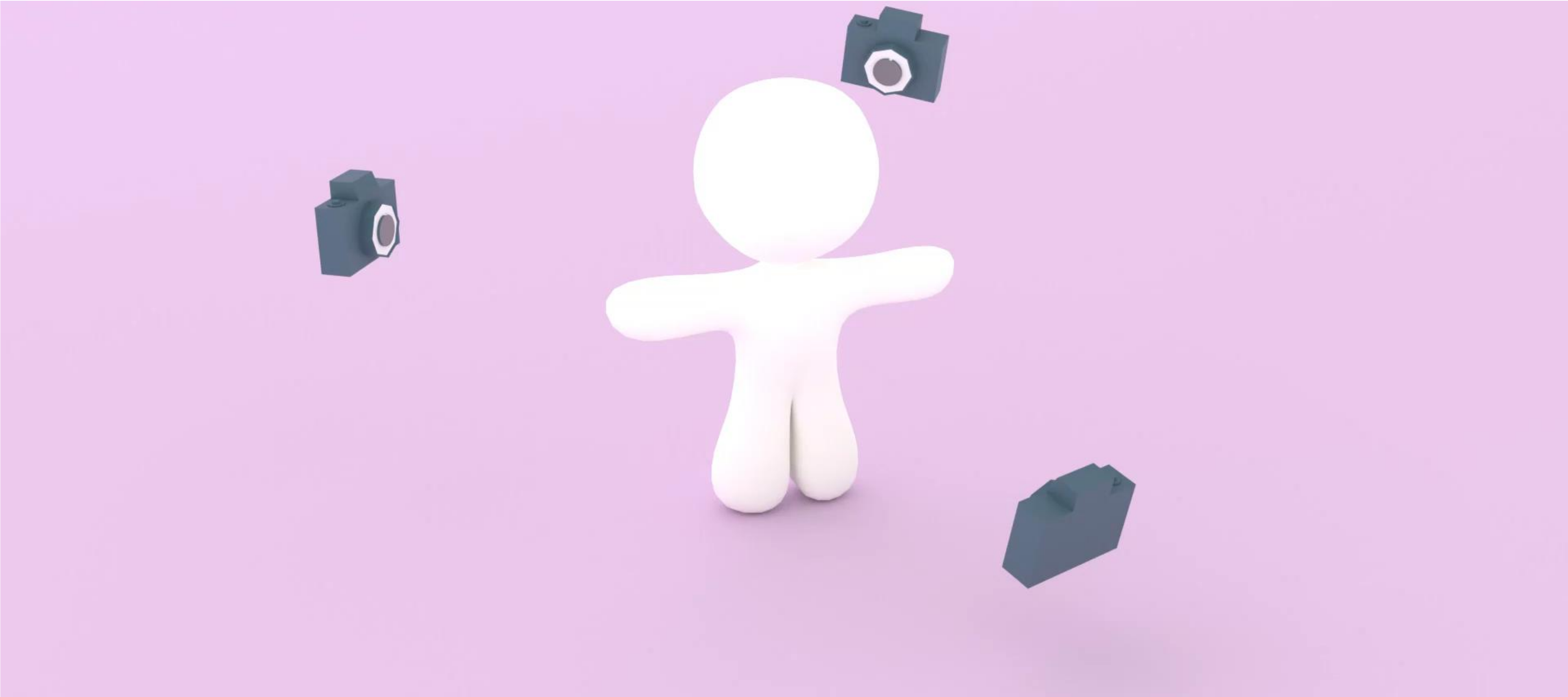
Our work: Keypoint 3D



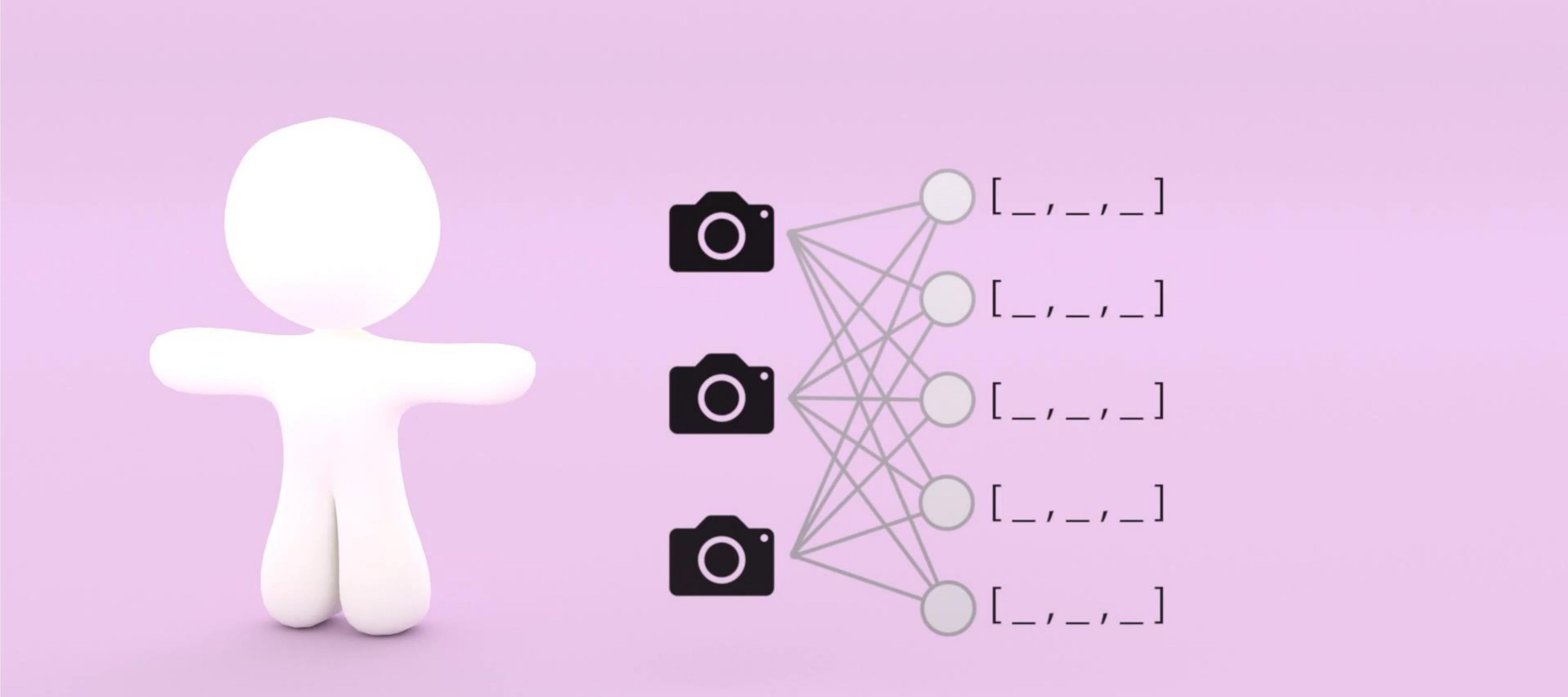
We hope to learn keypoints:

- in 3D world coordinates
- without supervision
- are good representation for control

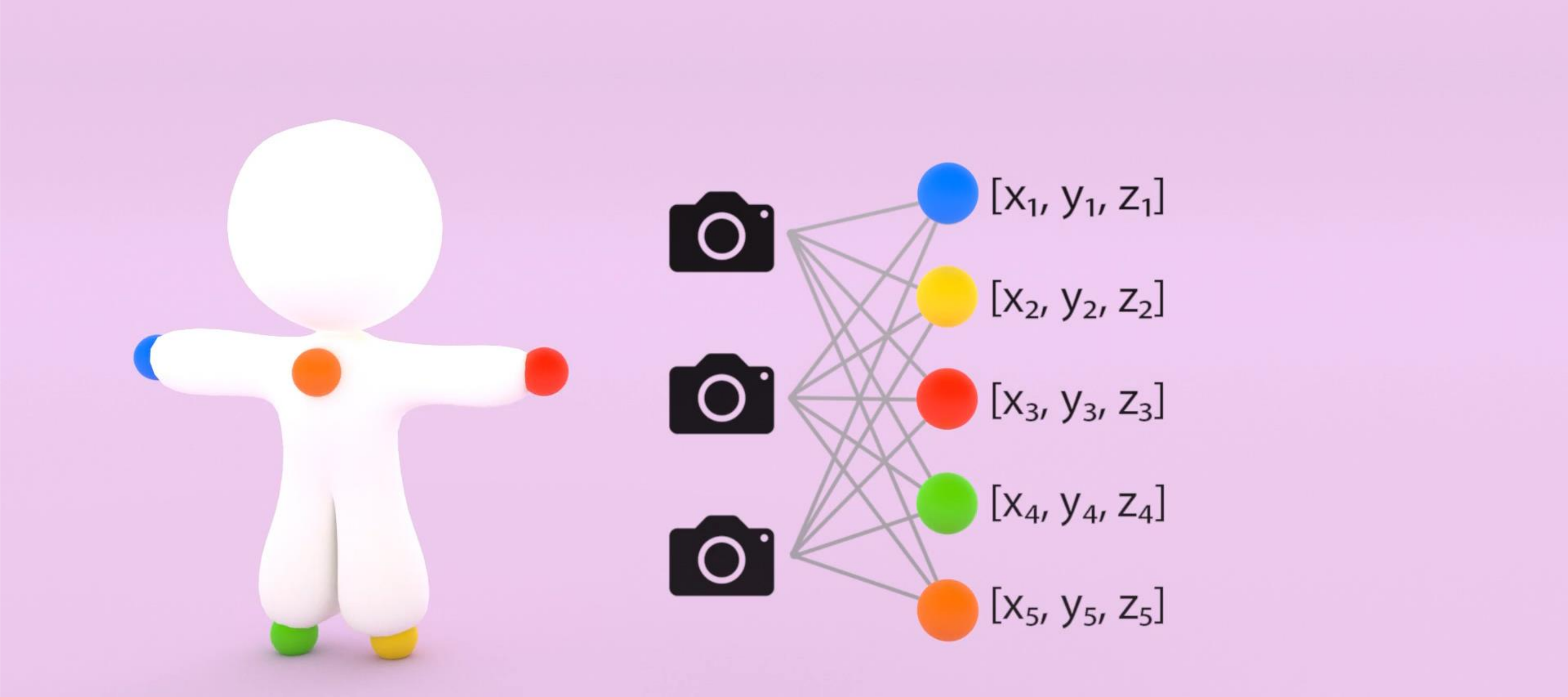
Our 3D Keypoint: Setup



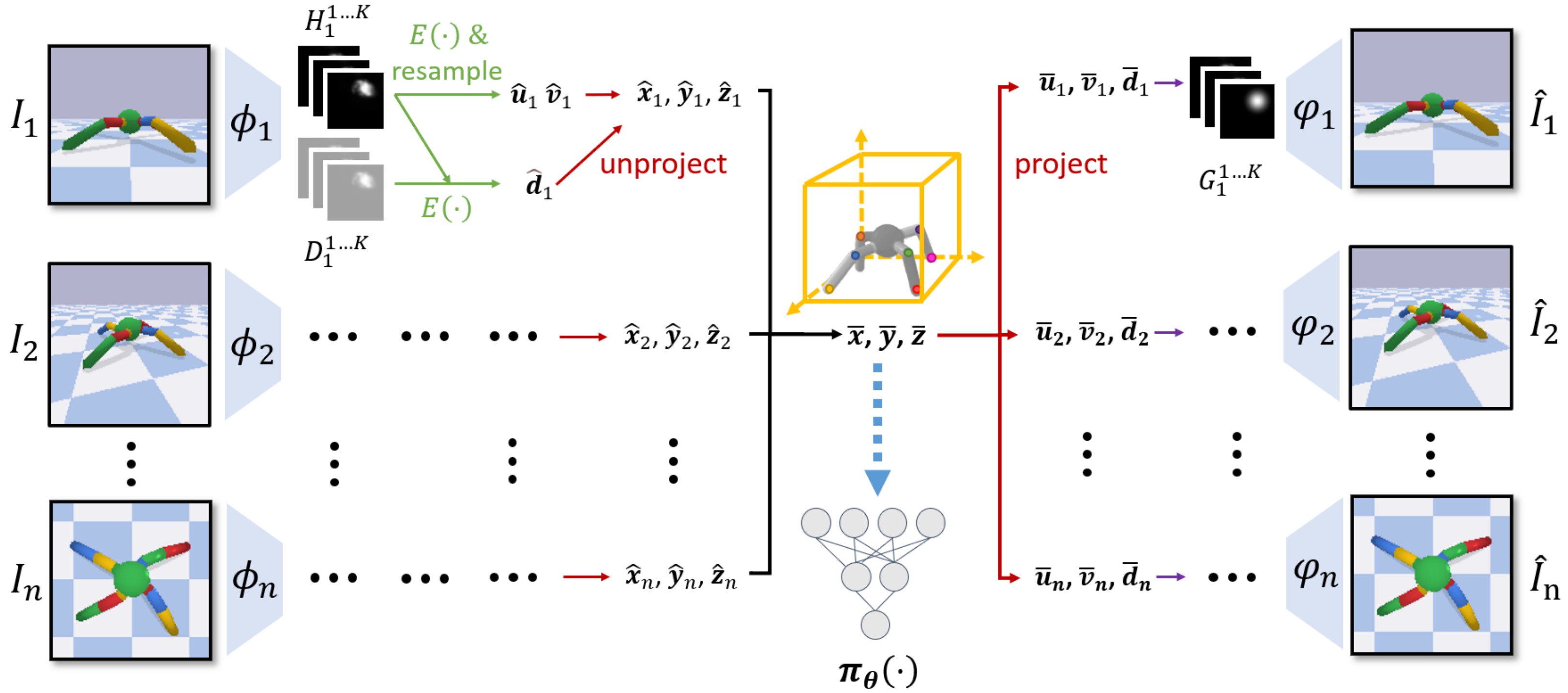
Our 3D Keypoint: Keypoint learning



Our 3D Keypoint: policy

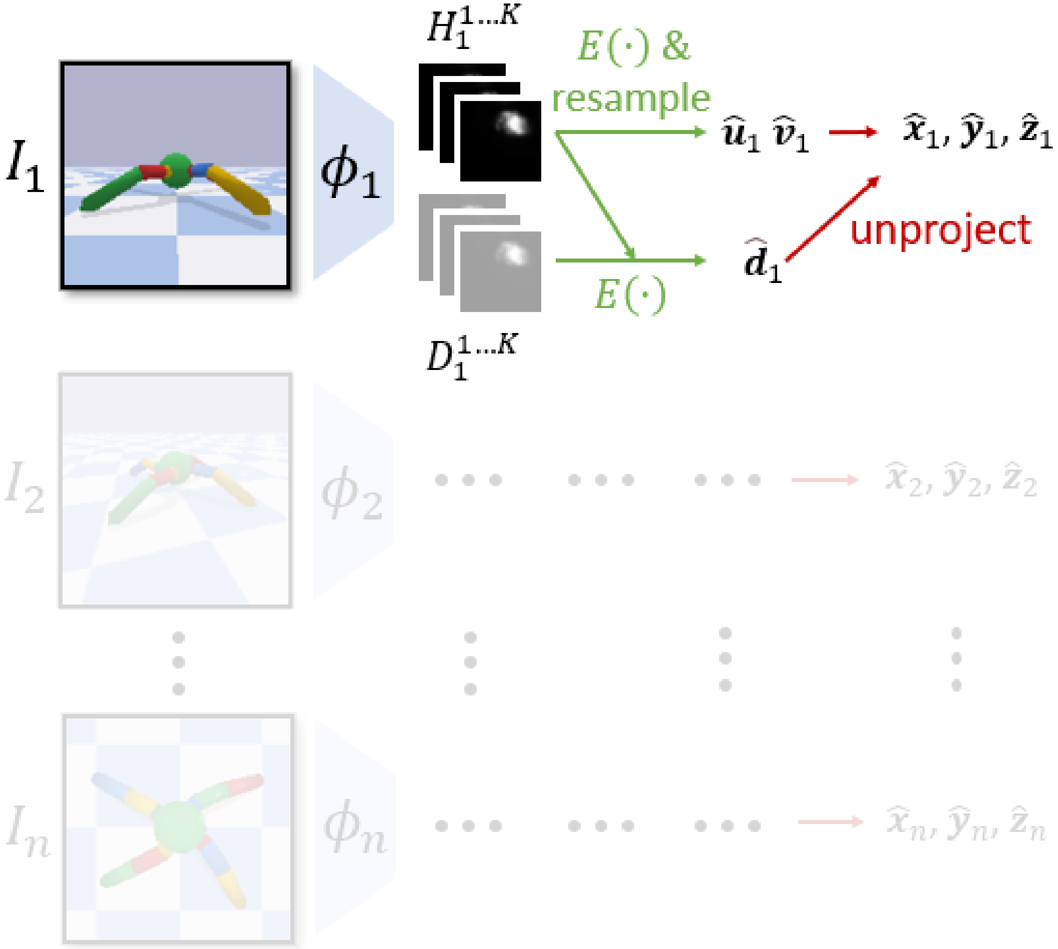


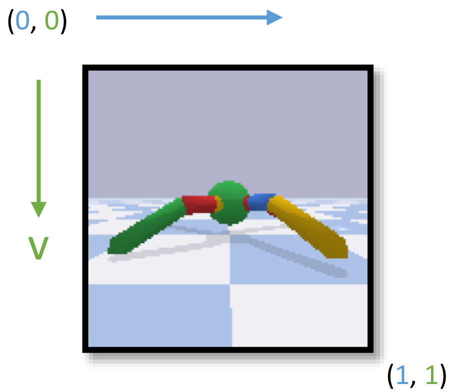
Method

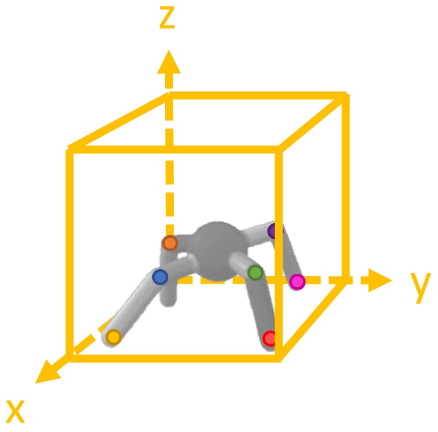


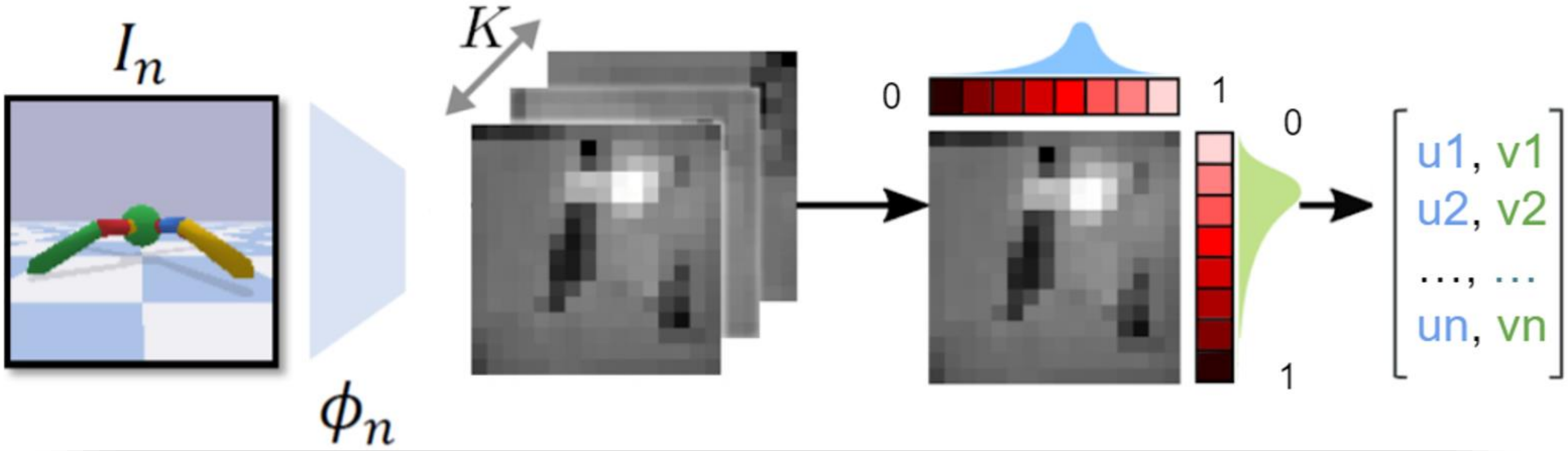
I_1
I_2
I_n
ϕ_1
ϕ_2
ϕ_n
$H_1^{1...K}$
$D_1^{1...K}$
$E(\cdot)$ & resample
$E(\cdot)$
$\hat{u}_1\ \hat{v}_1$
$\hat{d}_1$
unproject
$\hat{x}_1, \hat{y}_1, \hat{z}_1$
$\hat{x}_2, \hat{y}_2, \hat{z}_2$
$\hat{x}_n, \hat{y}_n, \hat{z}_n$
$\bar{x}, \bar{y}, \bar{z}$
$\pi_\theta(\cdot)$
project
$\bar{u}_1, \bar{v}_1, \bar{d}_1$
$\bar{u}_2, \bar{v}_2, \bar{d}_2$
$\bar{u}_n, \bar{v}_n, \bar{d}_n$
$G_1^{1...K}$
φ_1
φ_2
φ_n
$\hat{I}_1$
$\hat{I}_2$
$\hat{I}_n$

Method: encoder



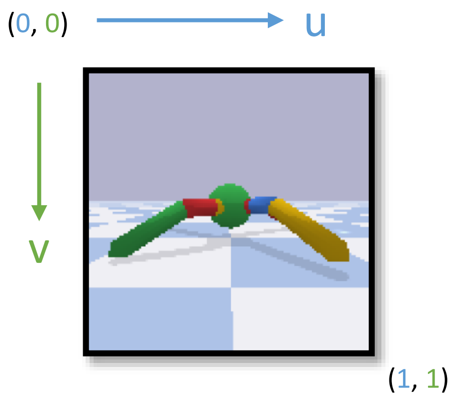


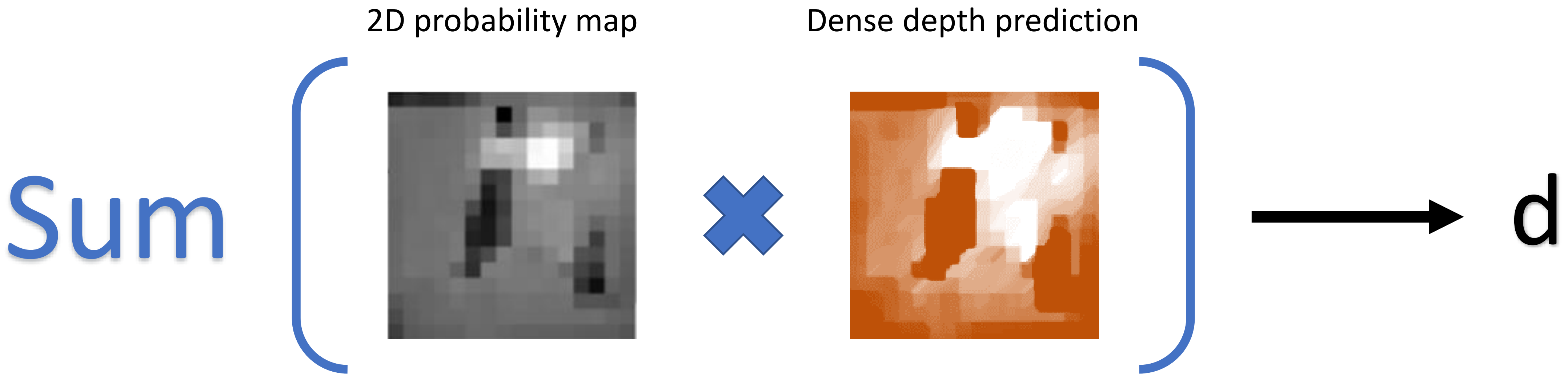




• Unsupervised learning of object landmarks through conditional image generation, Jakab et al., 2018

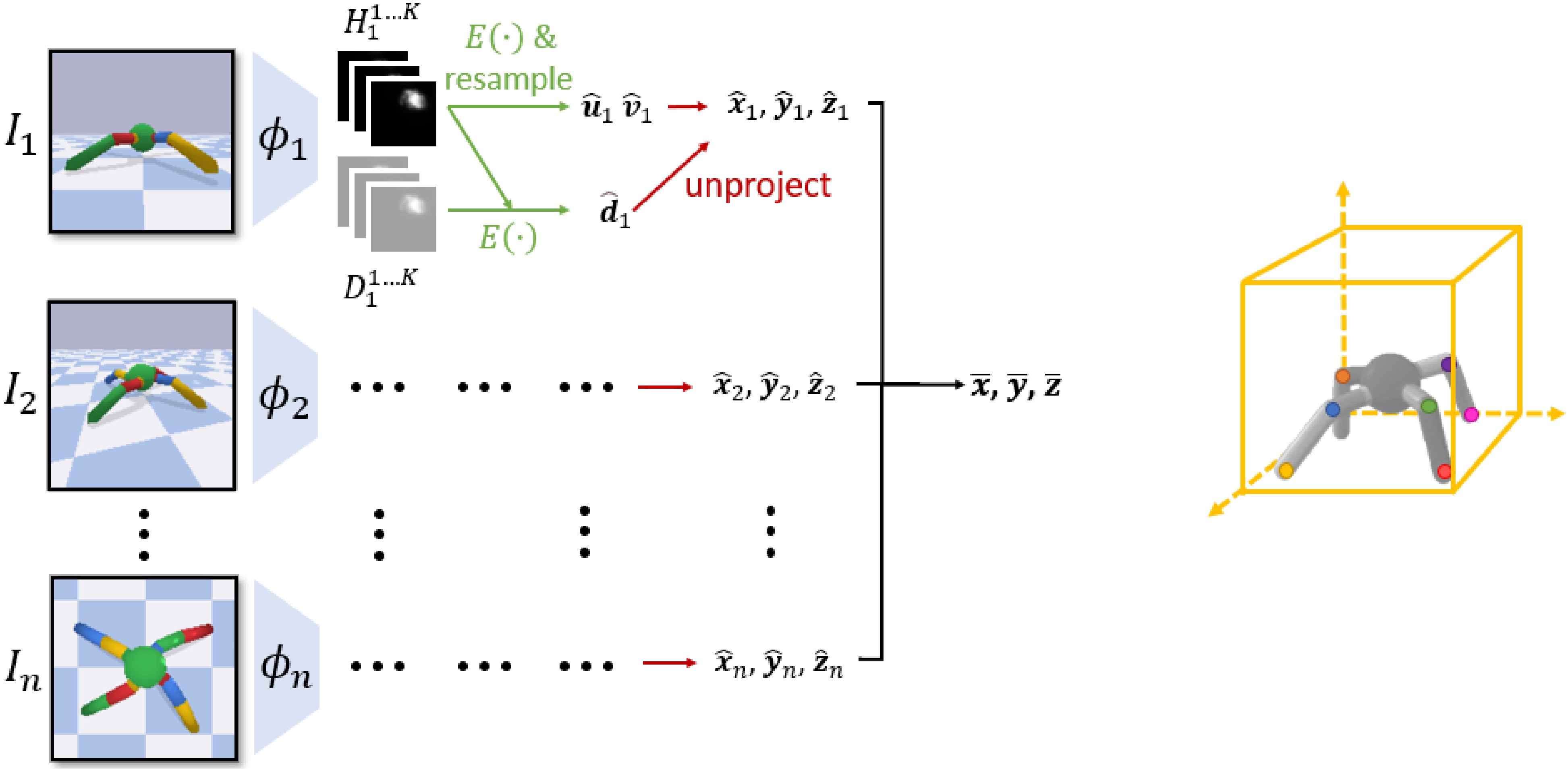
Fully differentiable keypoint bottleneck





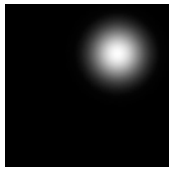
Depth parameterization

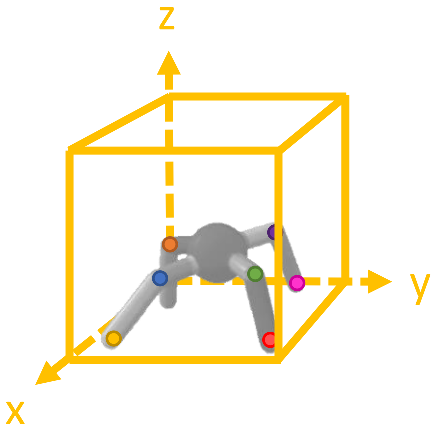
Method: encoder

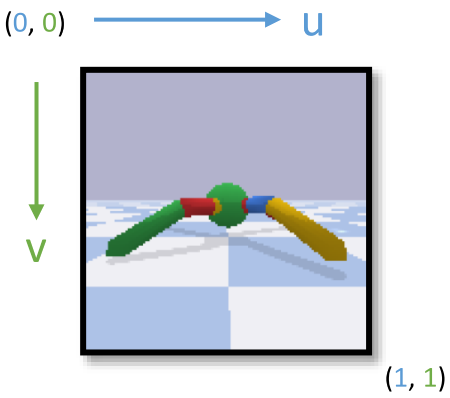


Method: decoder

$\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}, \bar{\boldsymbol{z}} \longrightarrow \bar{\boldsymbol{u}}, \bar{\boldsymbol{v}}, \bar{\boldsymbol{d}} \longrightarrow$



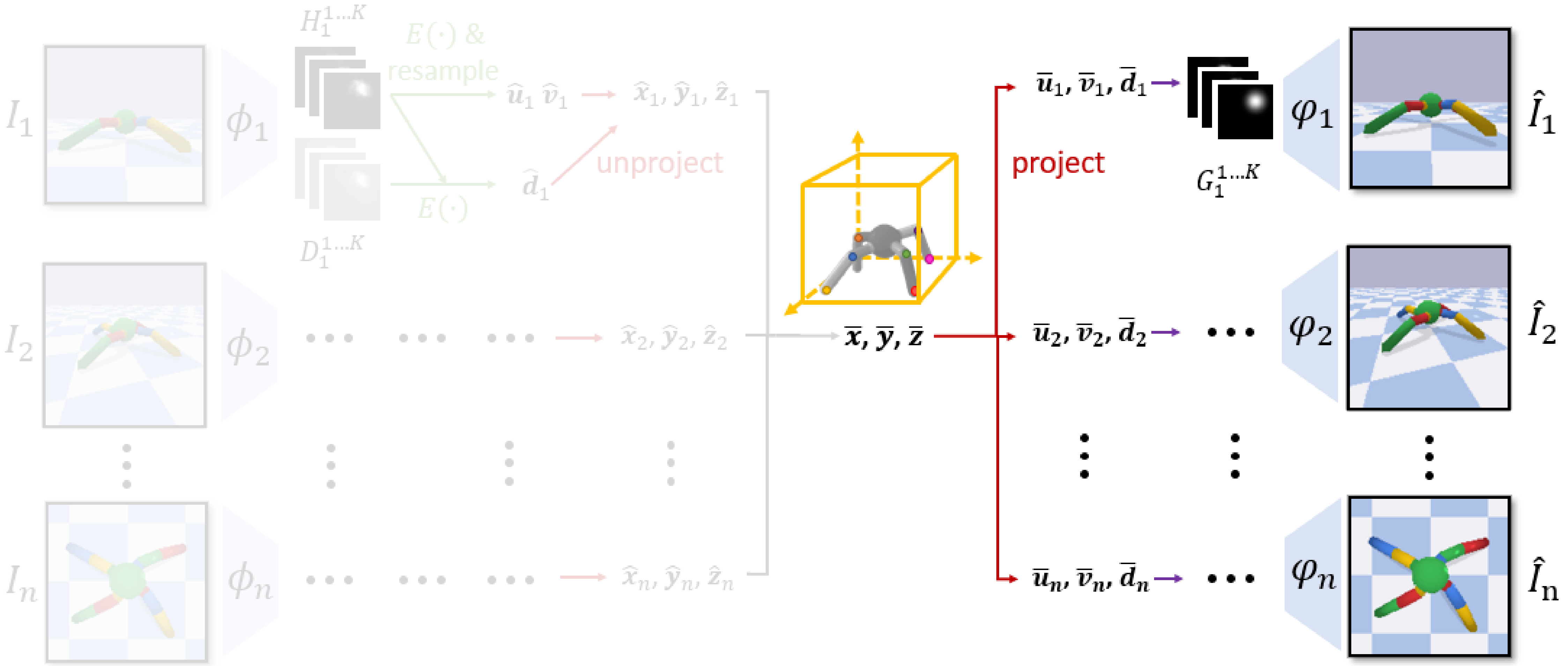




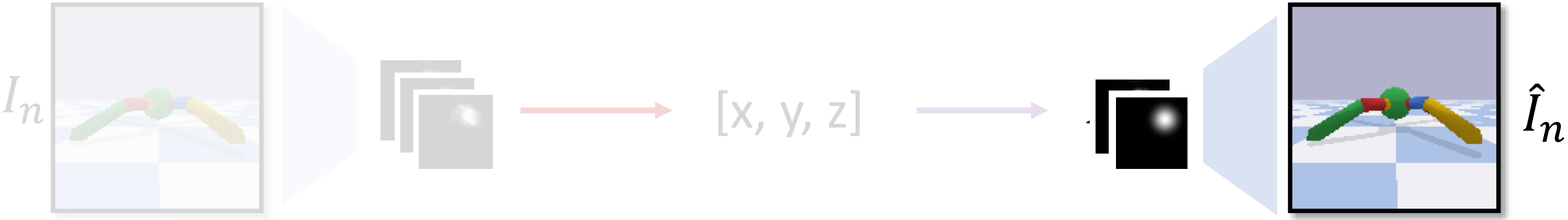
2D Gaussian with
Mean at $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$
Std $\propto 1 / \bar{\boldsymbol{d}}$

xyz coordinate regains 2D structure in a fully differentiable way!

Method: decoder

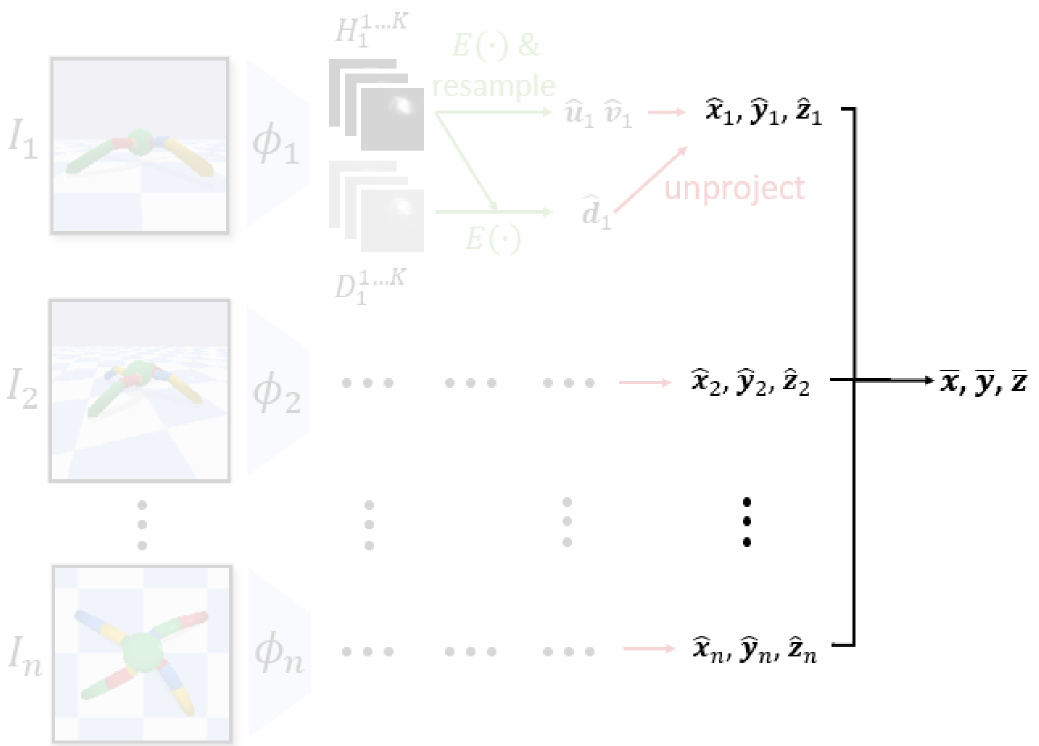


Method: auto-encoding loss



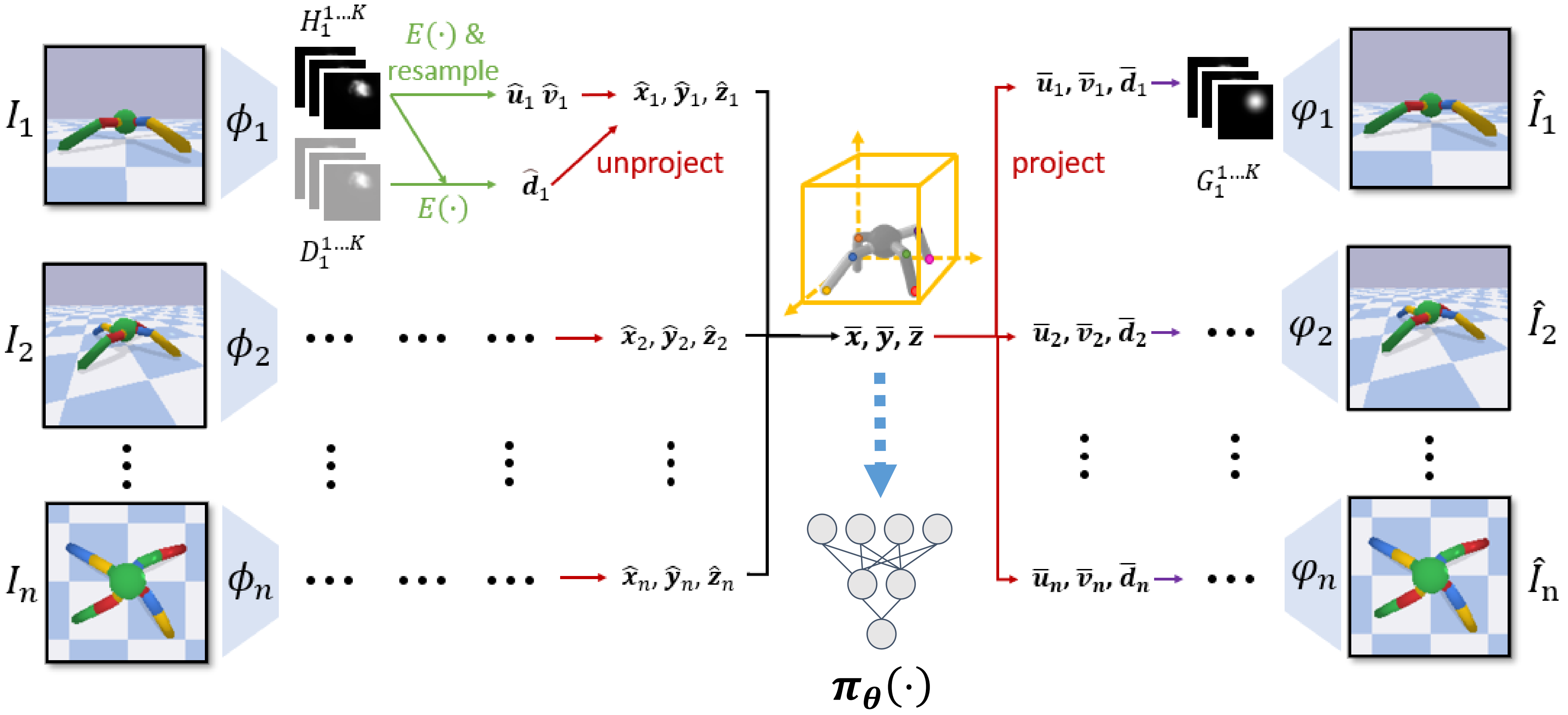
Core intuition:
To best decode to original image, the 2D gaussians have centers aligned with meaningful points

Method: multi-view consistency loss

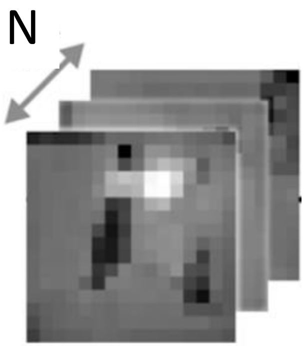


Core intuition:
Some point movements are visible from camera A but not camera B,
B must learn to "hallucinate" these points to minimize disagreement

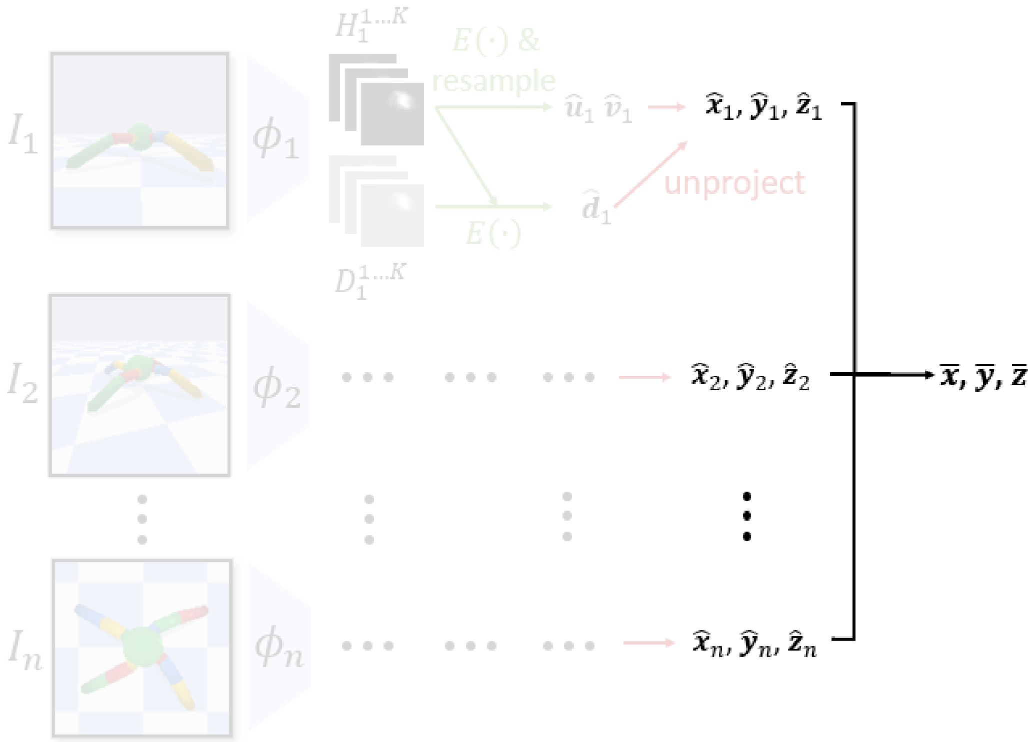
Method: policy



Method: attention

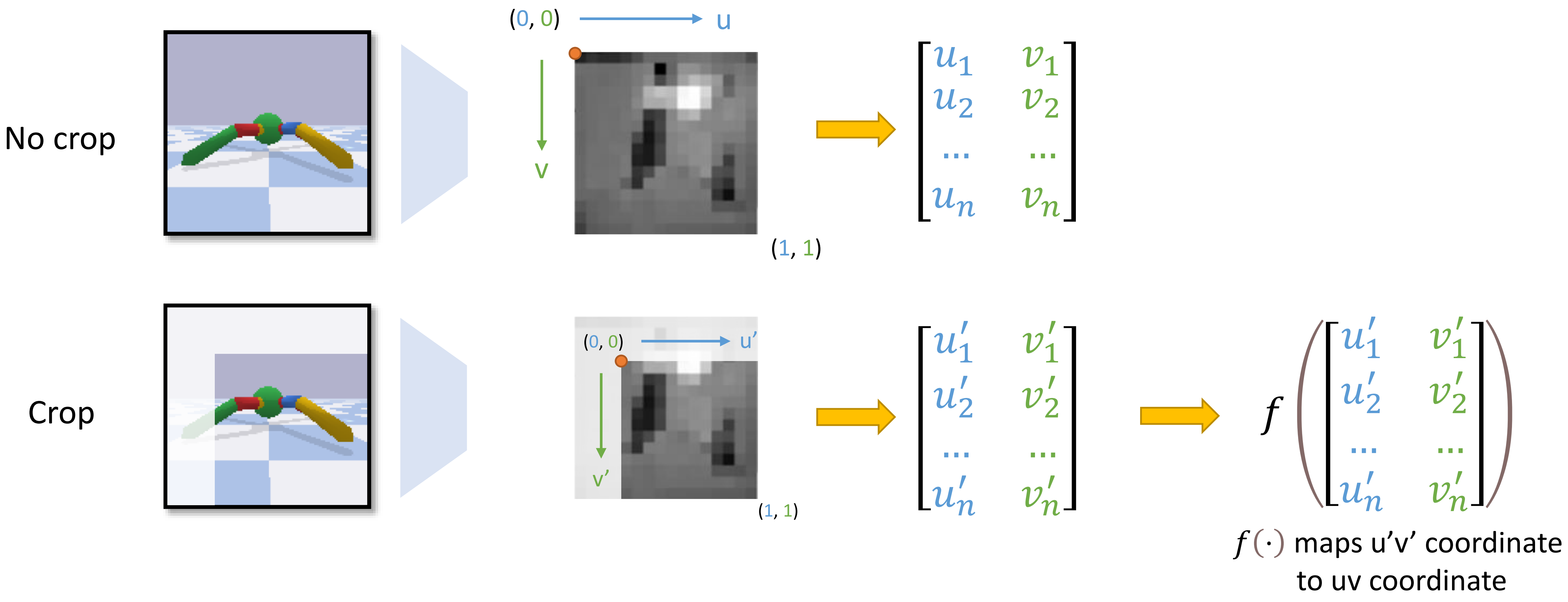


Softmax along # camera dimension
Use mean logits of each map as attention logit



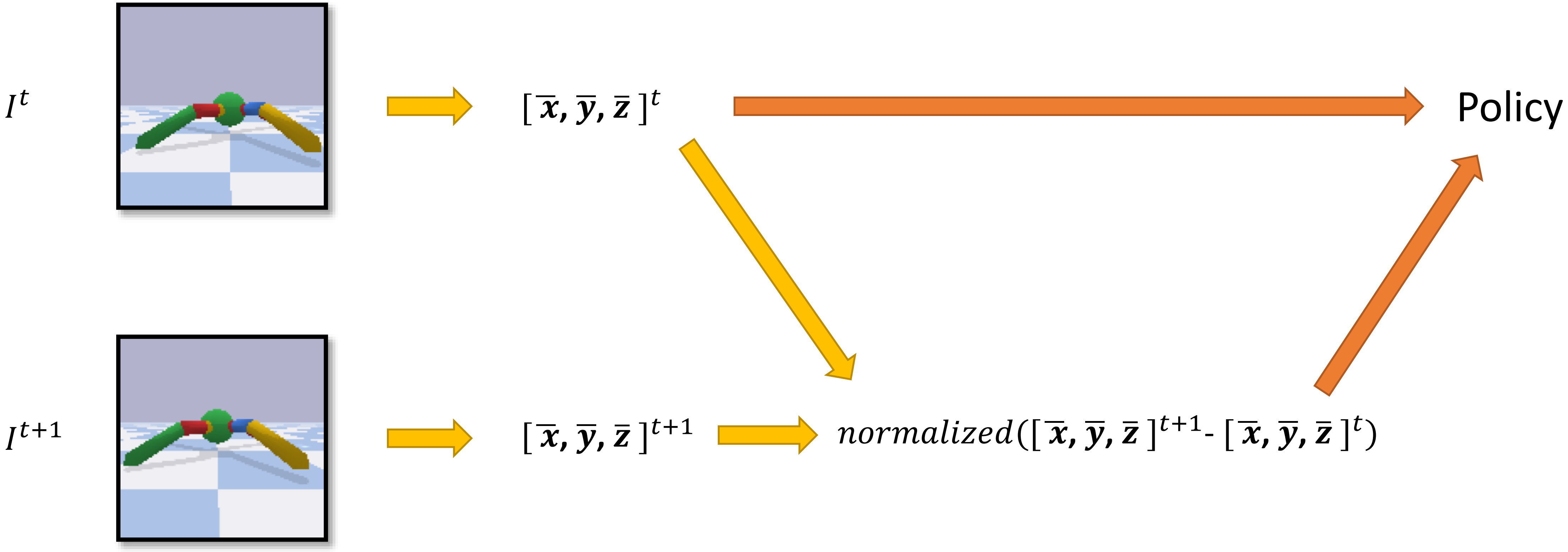
Allow model to ignore unconfident estimations!

Random crop as self-supervision



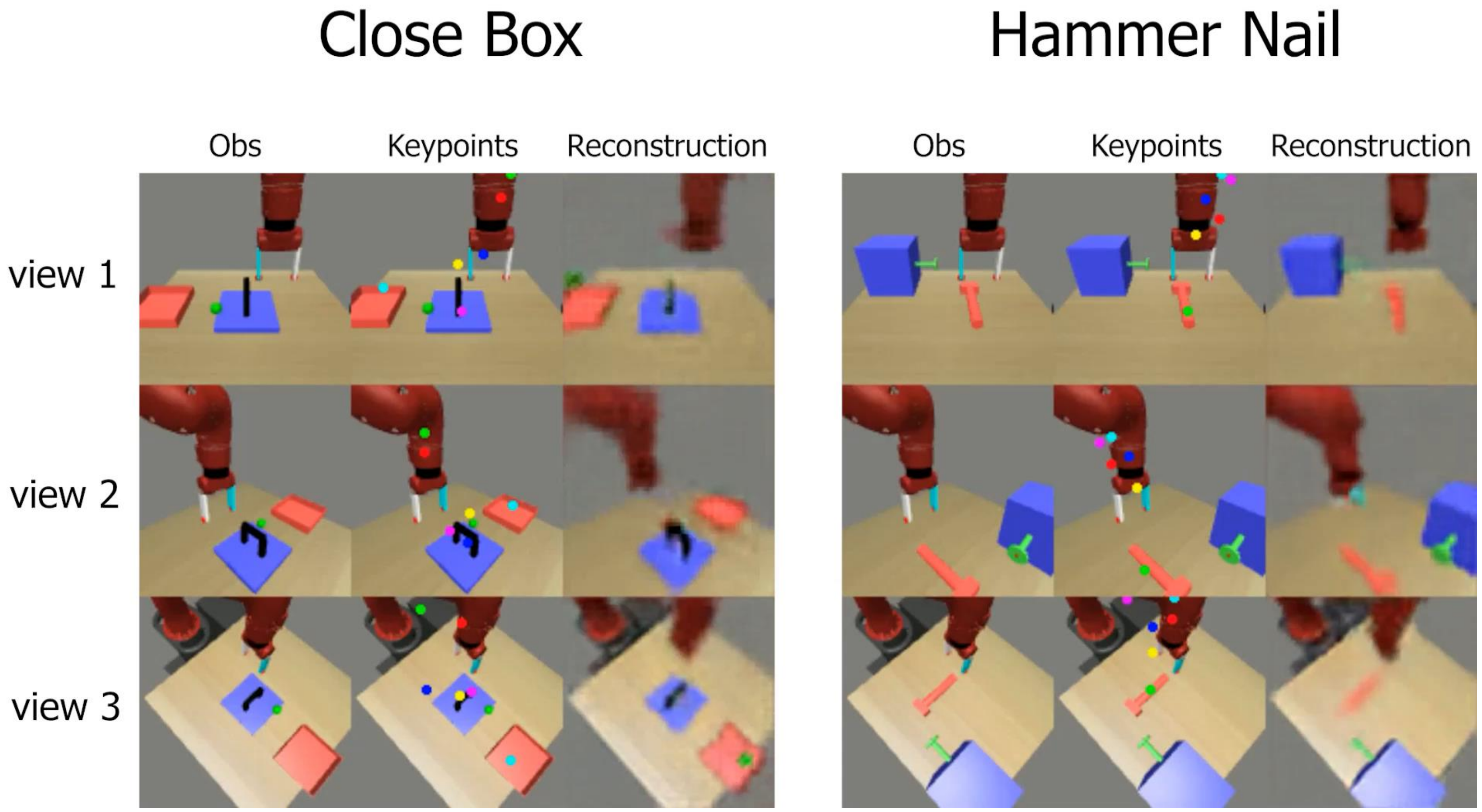
Coordinates must align with the random cropping to predict well

Temporal variant



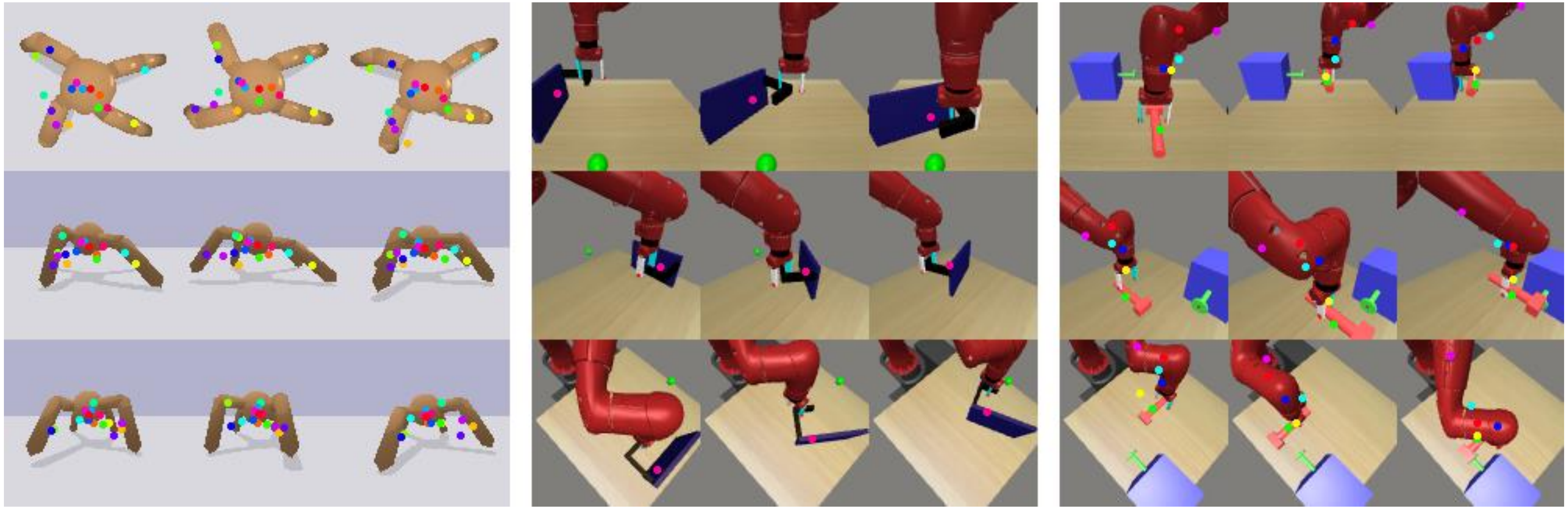
Differences between keypoint prediction is velocity vector! Explicitly normalize as movement feature

Visualization of Learned 3D Keypoints

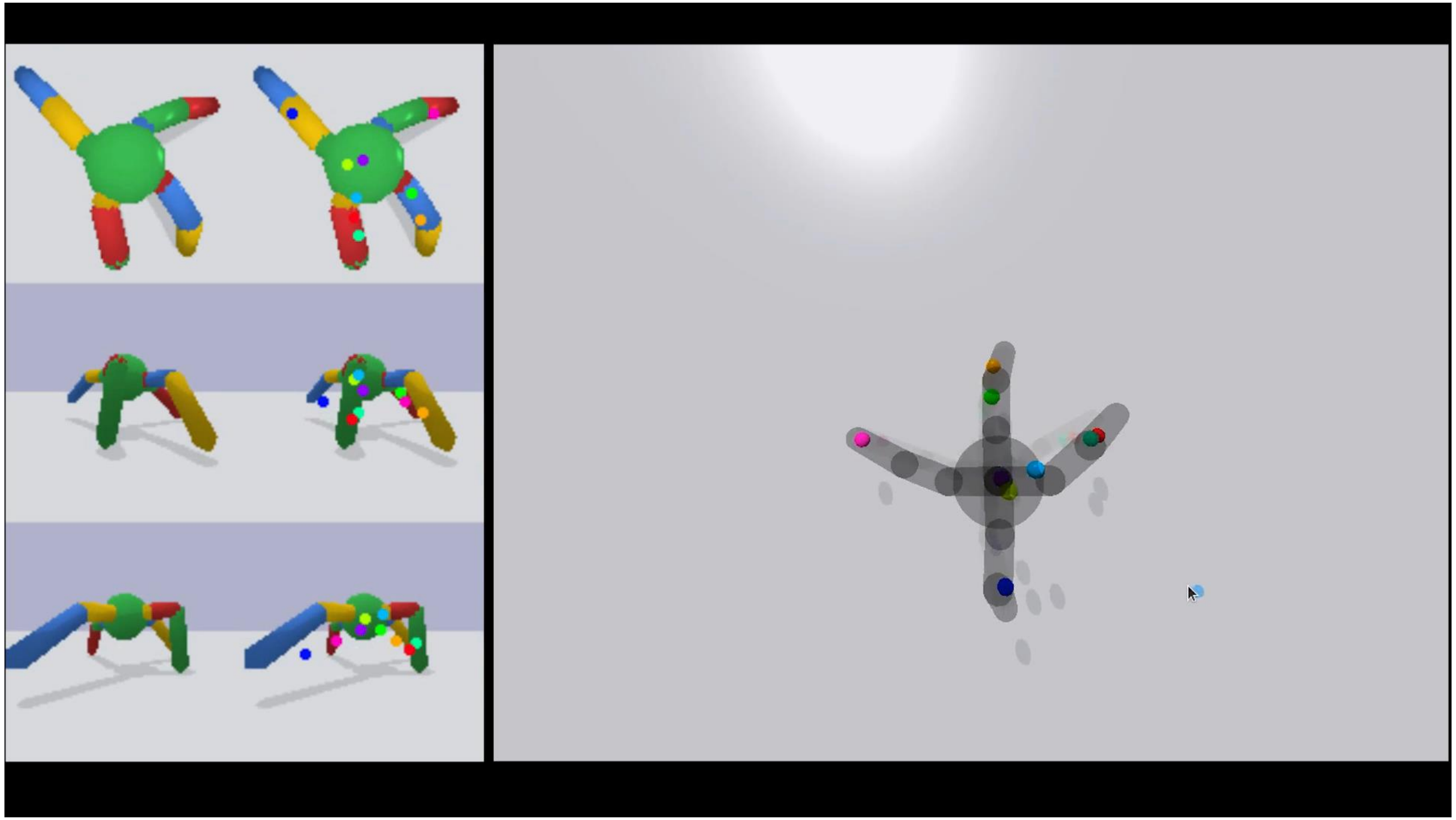


Visualization of Learned 3D Keypoints

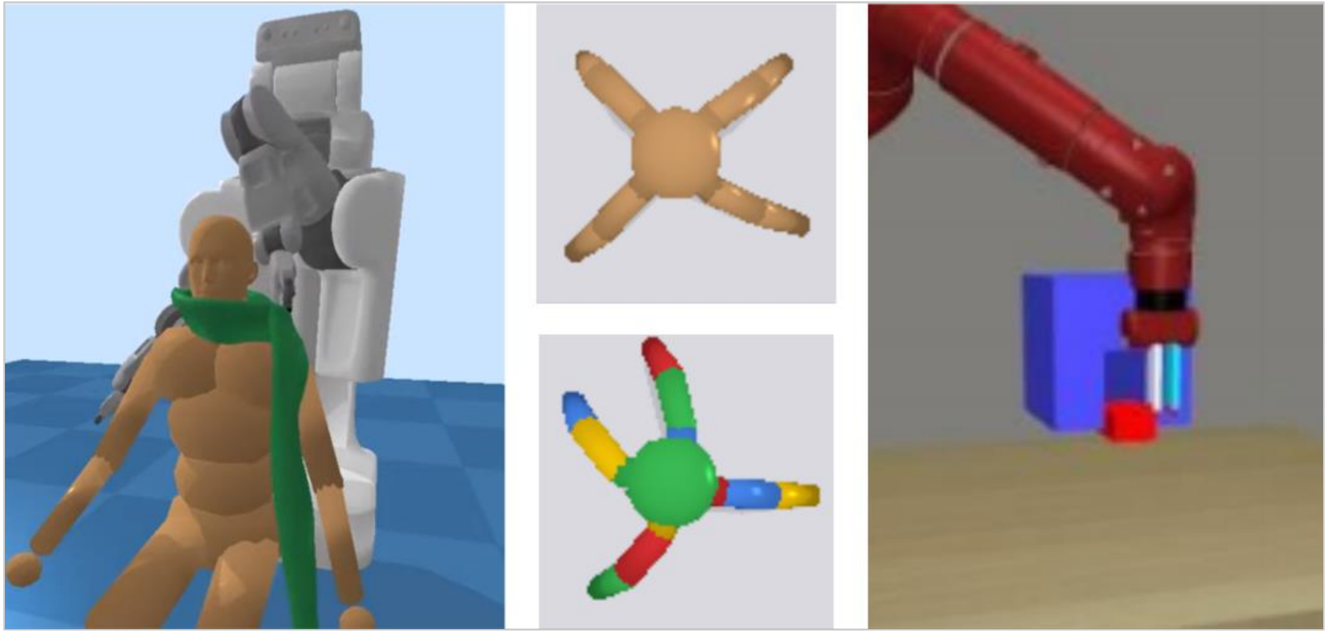
*attention can be used to filter out unconfident predictions with a threshold!



Visualization of Learned 3D Keypoints



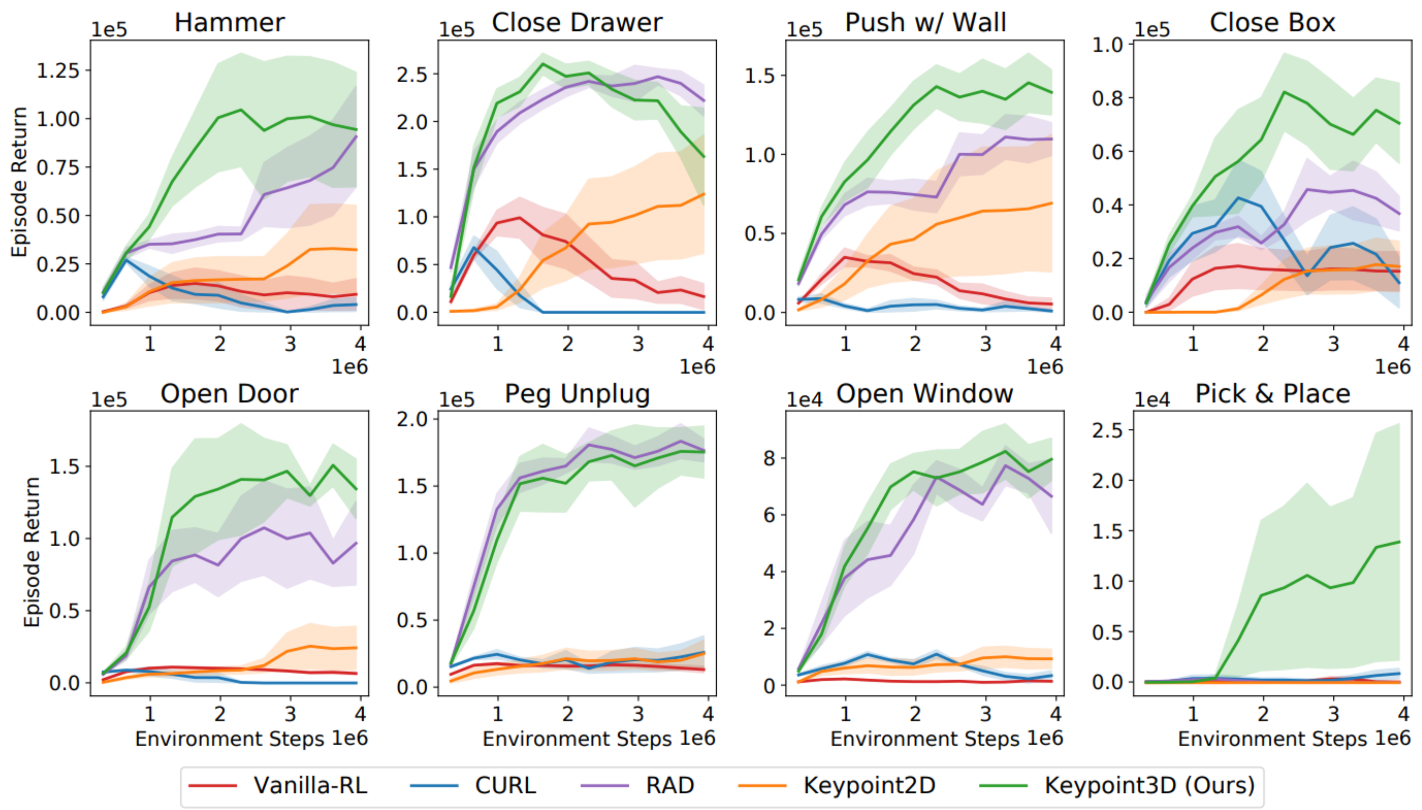
Experiments Overview



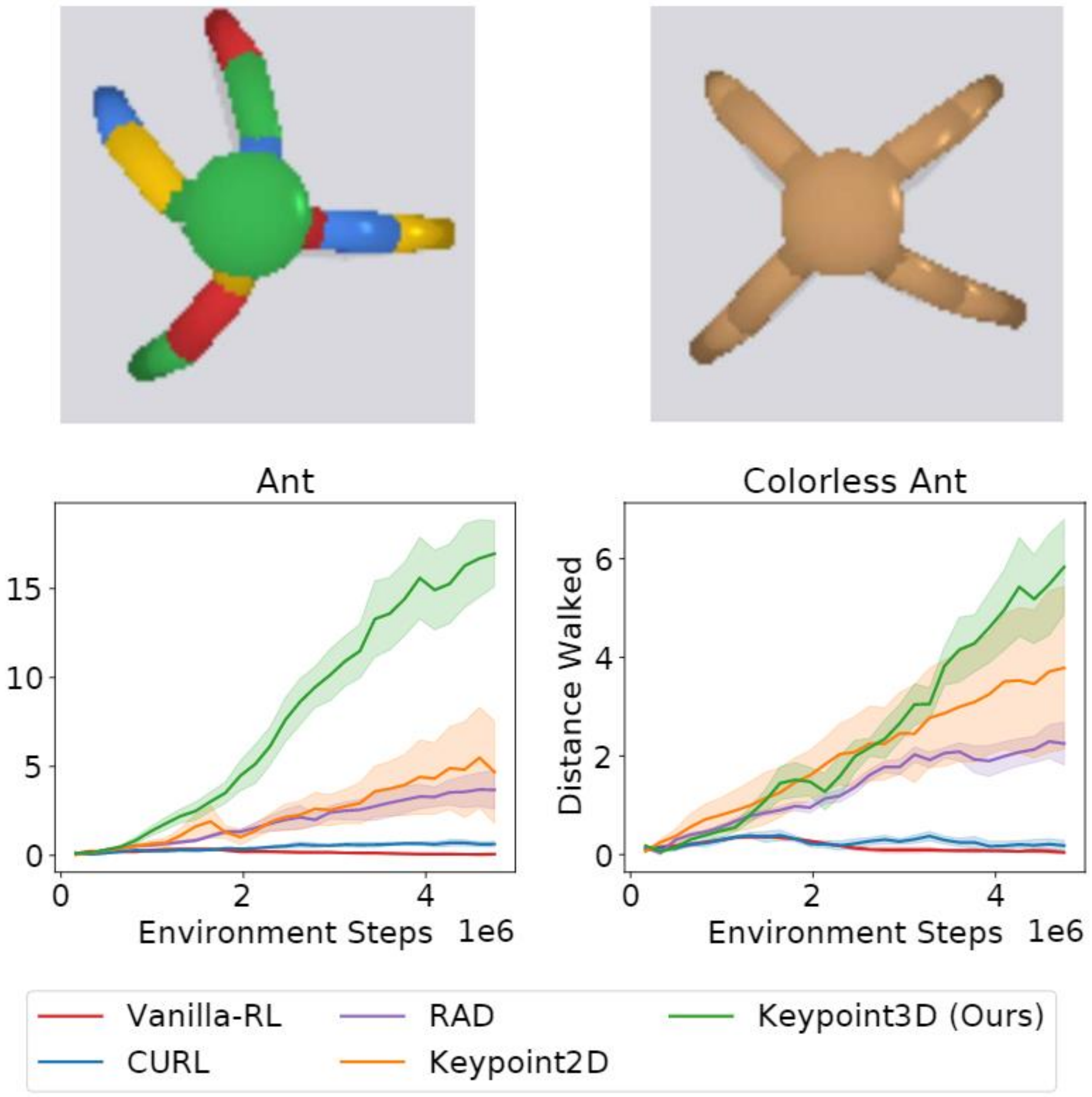
Effectiveness of 3d keypoints for control

- Sample efficiency compared to other representation
- Scalability to higher dimensional control problems(pybullet ant)
- Effectiveness on low-textured objects
- Ability to adapt to deformable objects (scarf manipulation)

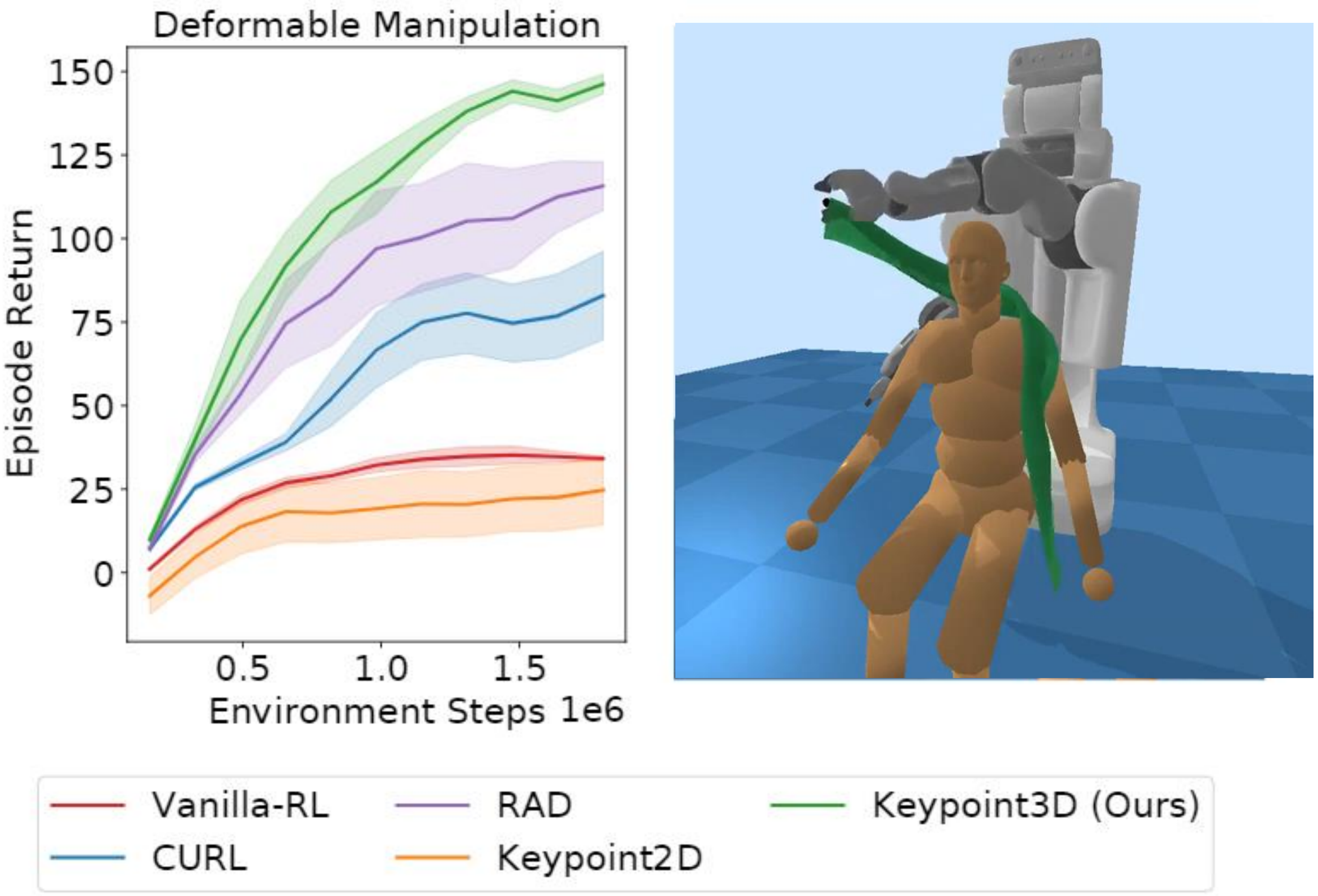
Sample efficiency in manipulation



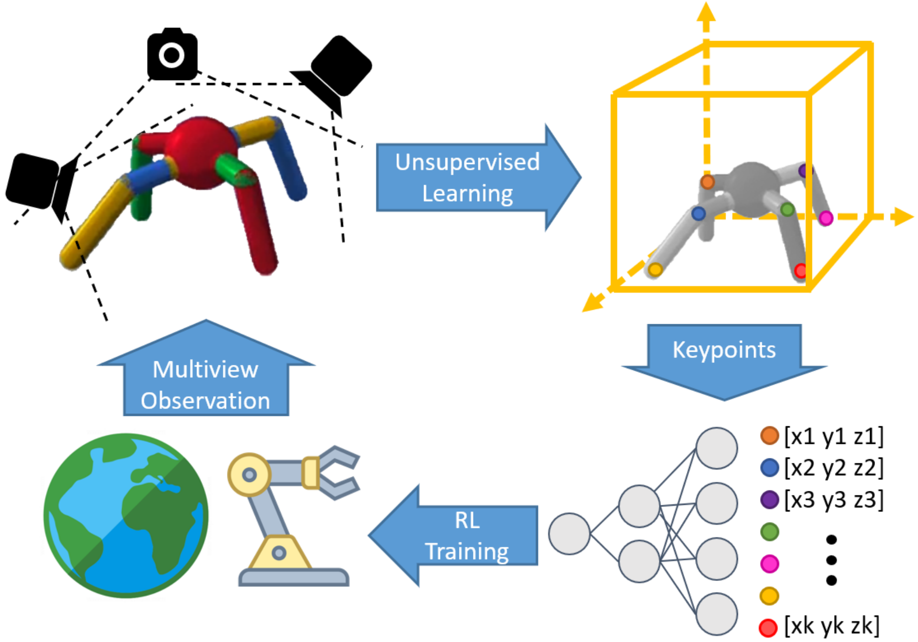
High dimensional control and low textured variant



Deformable manipulation



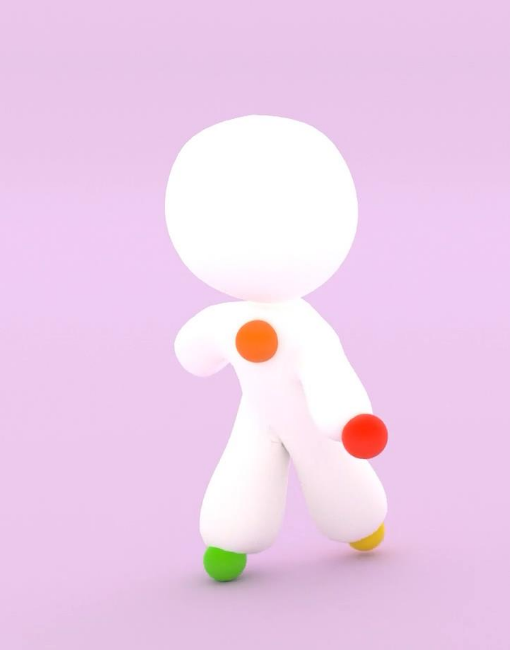
Summary



- We propose a framework to learn 3D keypoints without supervision for continuous control
- We leverage multi-view auto-encoding with a 3D keypoint bottleneck to learn meaningful 3d keypoints; We jointly train policy learning in conjunction with keypoint learning
- Our method achieves significant sample efficiency improvement in a variety of 3D environments.
- The 3D keypoints learned by our algorithm are consistent across space and time.

We hope our method serves as a bridge between pixel domain and 3D control tasks.

More Details:



- [Website] https://buoyancy99.github.io/unsup-3d-keypoints
- [Code] https://github.com/buoyancy99/unsup-3d-keypoints
- [Paper] https://arxiv.org/pdf/2106.07643.pdf