Carnegie Mellon

School of Computer Science

Deep Reinforcement Learning and Control

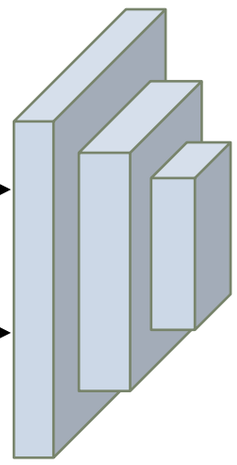
Visual Imitation Learning

Spring 2020, CMU 10-403

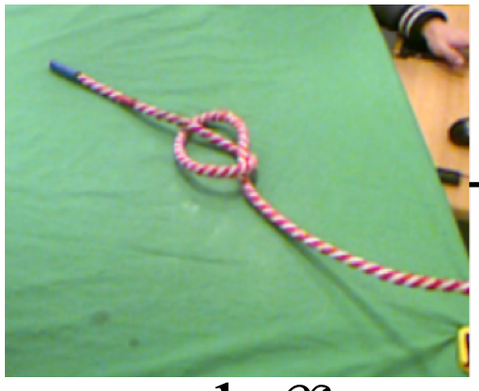
Katerina Fragkiadaki

How likely is to tie a knob only with trial-and-error?

current x_t



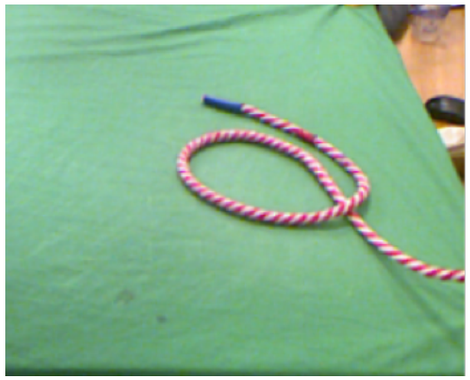
$\longrightarrow$ actions?



goal x_G

Consider a robot with a camera: no access to ground-truth low-dimensional state descriptions of the world.

How likely is to tie a knob only with trial-and-error?



current x_t

$\pi\left(x_t;\theta\right)$
policy

a_t

initially random

goal x_G

r_{t+1}

x_{t+1}

$$\max_{\theta}\left(\sum_{t=1}^{T} r_t\right)$$

Reward depends on how well the resulting state matches embedding-wise the goal state.

Embeddings can be trained with autoencoders, or under a combination of forward and inverse model learning.

Not very likely.

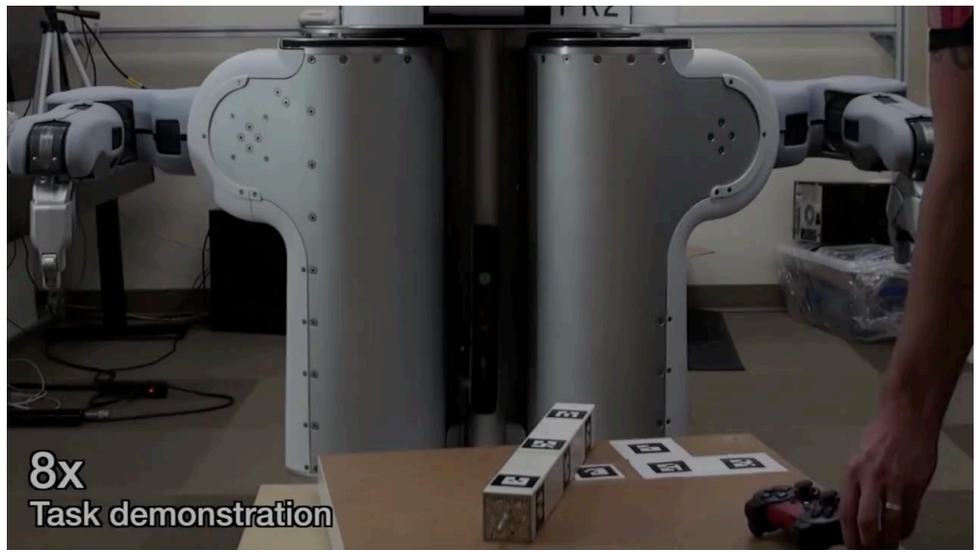
Indeed, we all learnt how to tie knobs with help from our parents

Imitation learning to the rescue

Learning from kinesthetic demonstrations

a.k.a. kinesthetic teaching

Learning skills by having people or other agents performing the skill by taking over the end-effectors of the robot. The expert demos are given in the the action space of the robot agent.

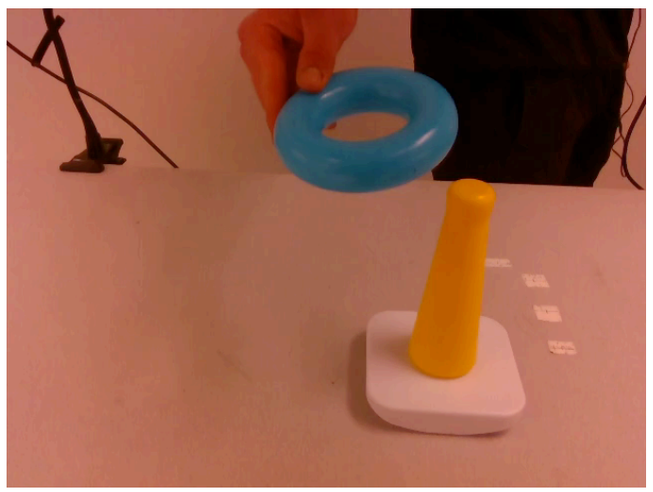


- Behaviour Cloning: simply regressing to the expert's actions
- Adversarial Imitation learning: reinforcement learning with denser reward depending on how well we match the expert's state-action densities
- Adding states from the expert demos in the experience buffer to get a chance to visit them (simple combination of RL and Imitation learning)

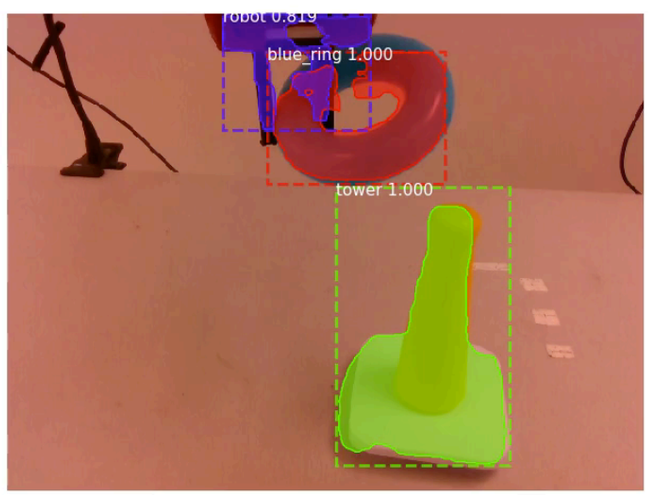
Learning from visual demonstrations

a.k.a. visual imitation or 3rd person imitation
Learning skills by watching people or other agents performing the skill

human demonstration



robot's imitation



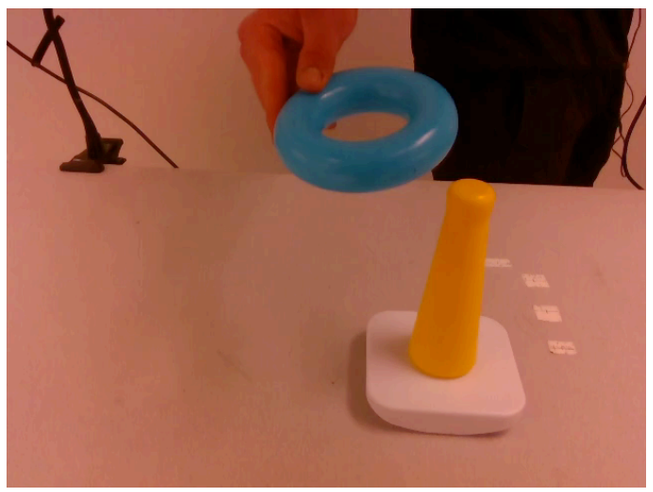
Q: Can we use the imitation methods we have learnt to achieve this?

- Behaviour Cloning: not same action space, so no.
- Adversarial Imitation learning: reinforcement learning with denser reward depending on how well we match the expert's state only densities, so yes.
- Adding states from the expert demos in the experience buffer: no, the state needs to be perceived, and we cannot automatically reset to desired states either.

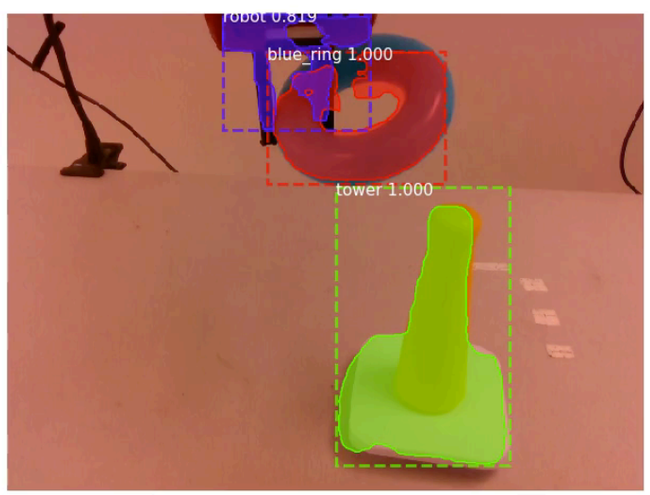
Learning from visual demonstrations

a.k.a. visual imitation or 3rd person imitation
Learning skills by watching people or other agents performing the skill

human demonstration



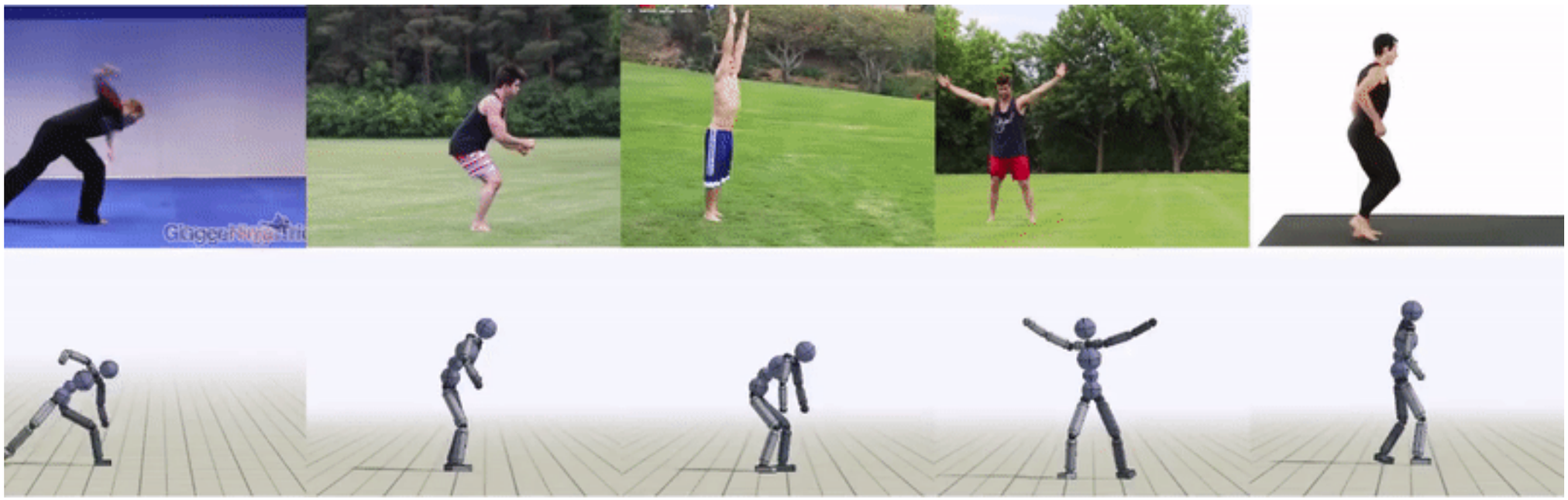
robot's imitation



- Central difficulty in visual imitation is perceiving the world state: where the objects are, in which pose, what velocities, etc.
- World state is available in simulation, e.g., if a human is demonstrating using a glove.
- For visual imitation though demonstration should be given in the real world else we beat the purpose.
- This means that Computer Vision really needs to work.

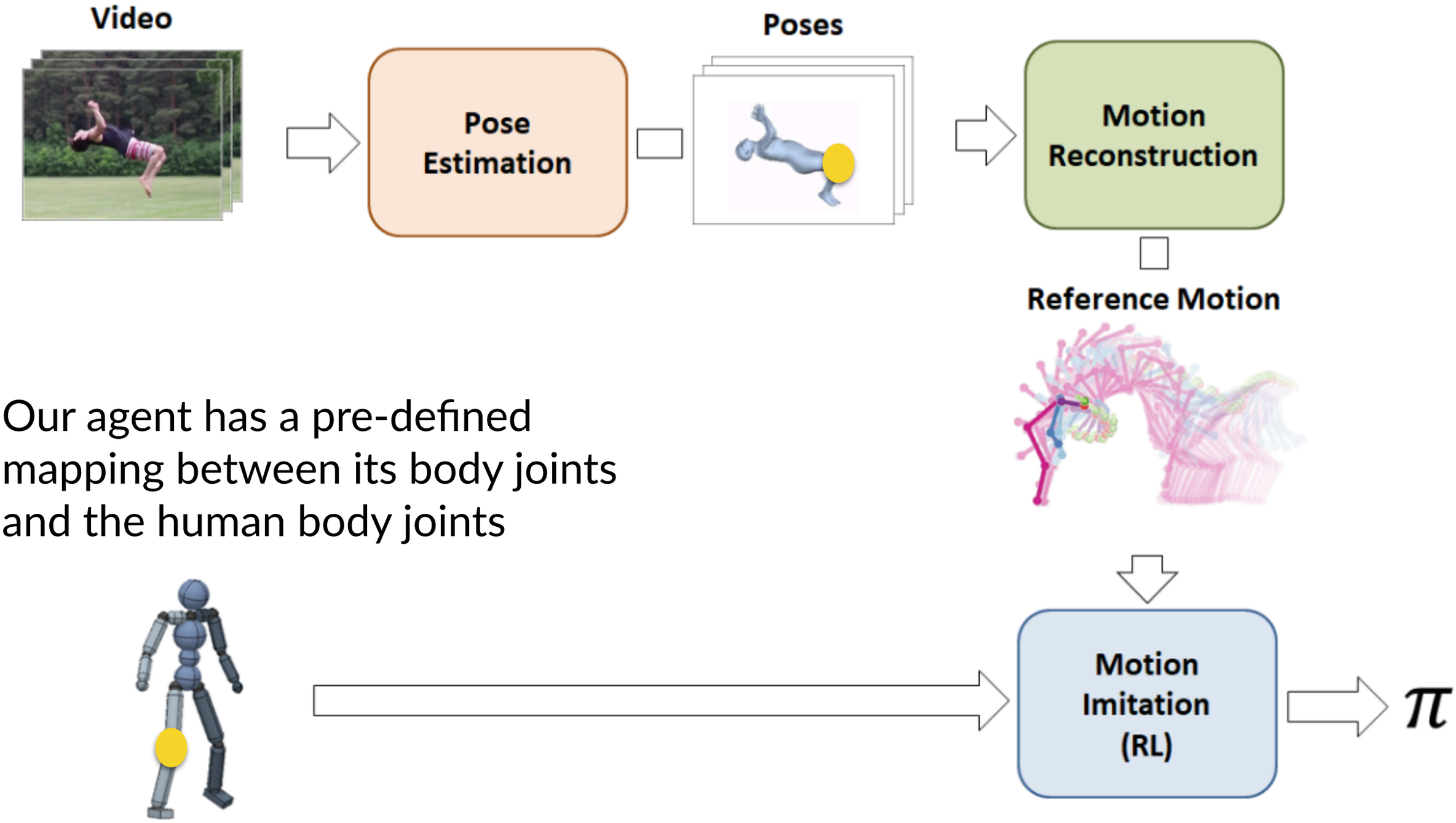
Learning acrobatics from watching Youtube

not by engineering the robot's actions



SFV: Reinforcement Learning of Physical Skills from Videos, Peng et al. 2018

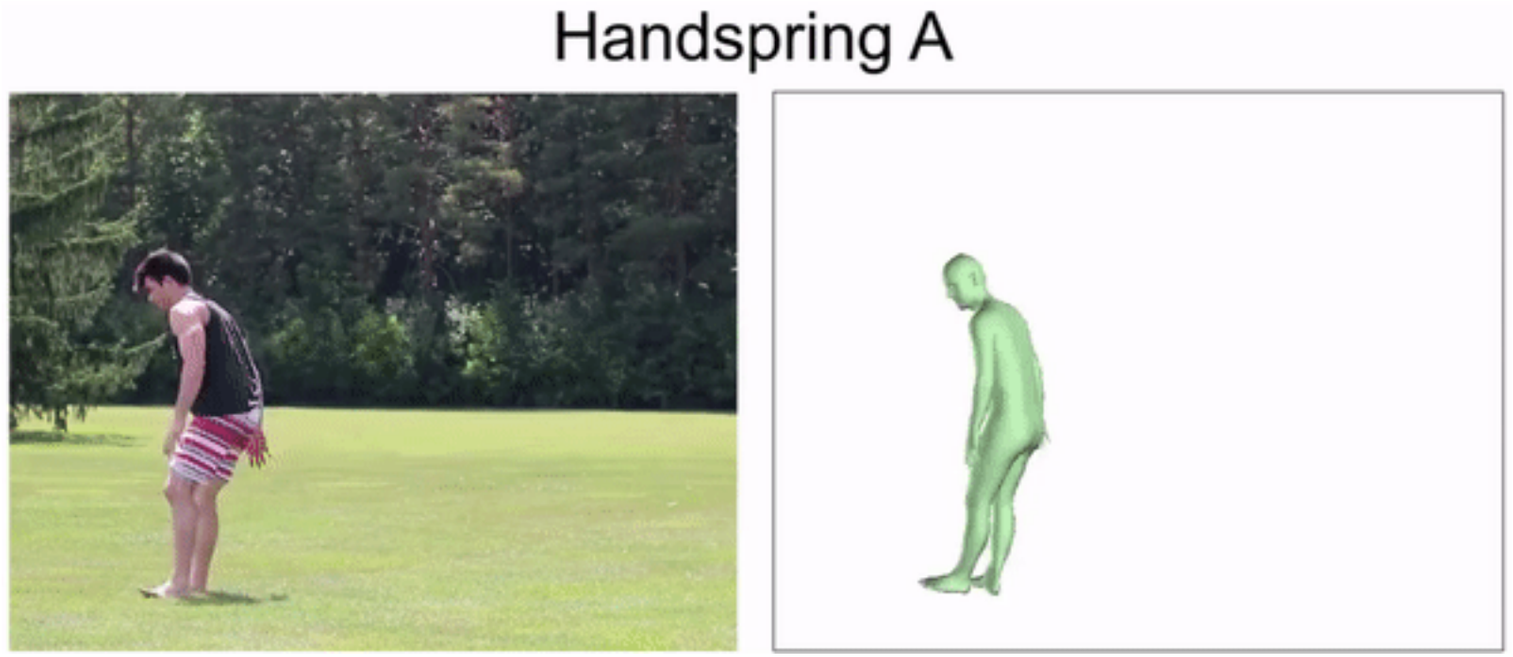
Learning acrobatics from watching Youtube

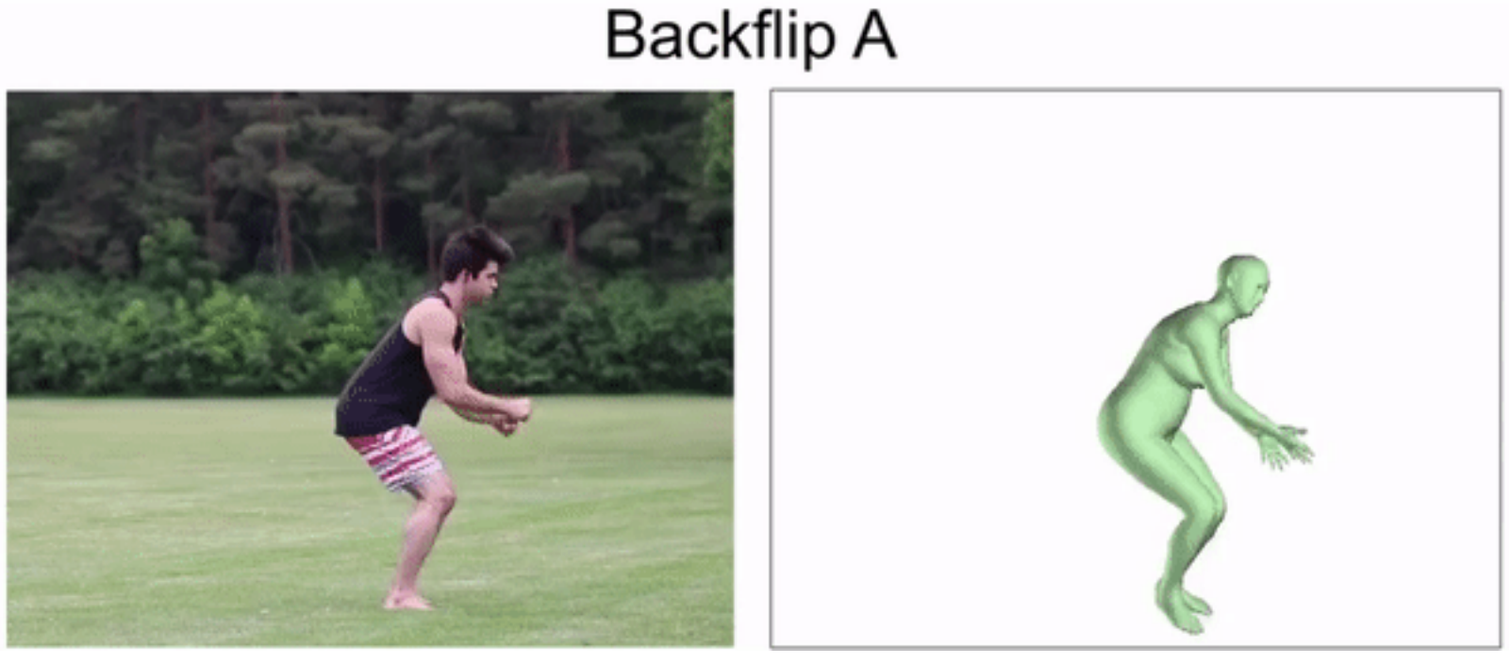


Our agent has a pre-defined mapping between its body joints and the human body joints

Q: Why we need RL? Why we do not just do behaviour cloning to imitate the reference motion sequence?

Imitating Humans by Inferring their 3D Poses





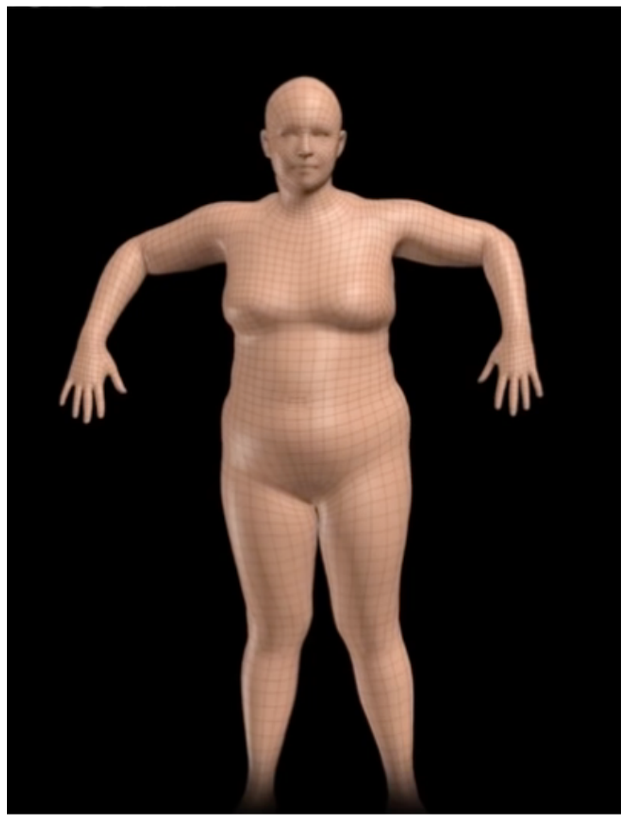
SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β



Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black,
SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

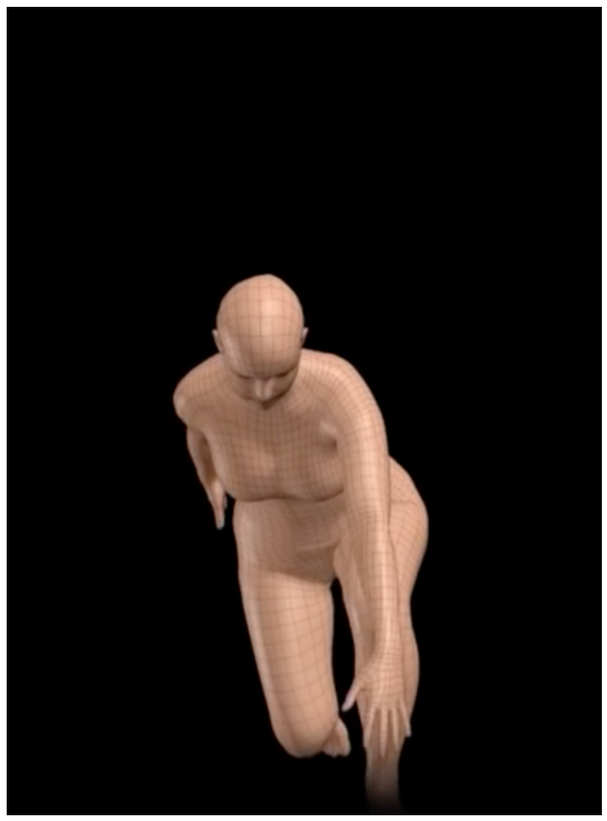
SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β



Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black, SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

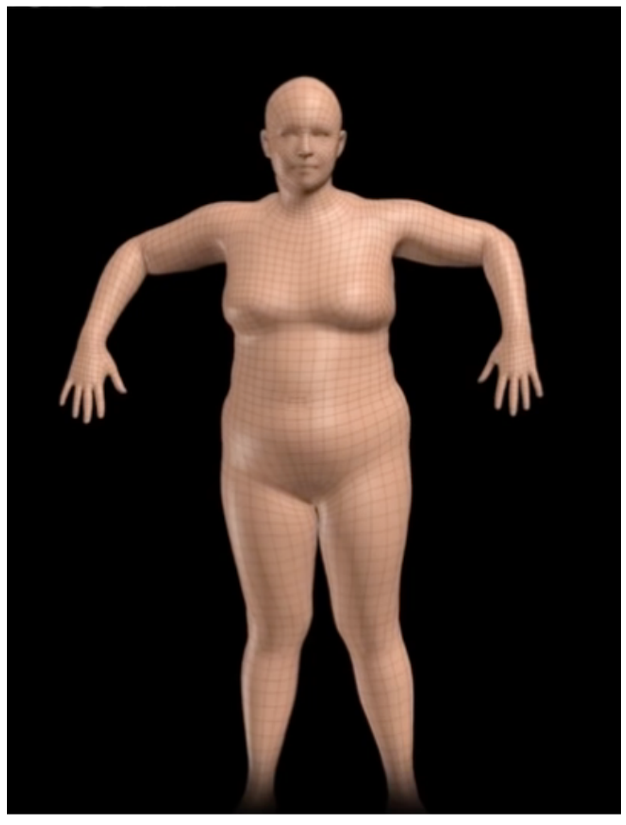
SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β



Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black,
SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

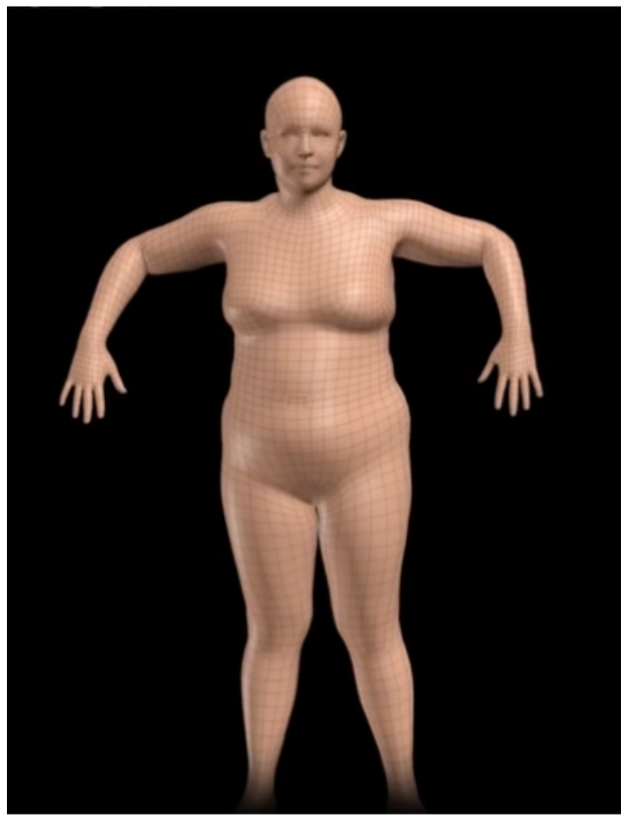
SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β



Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black,
SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black,
SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

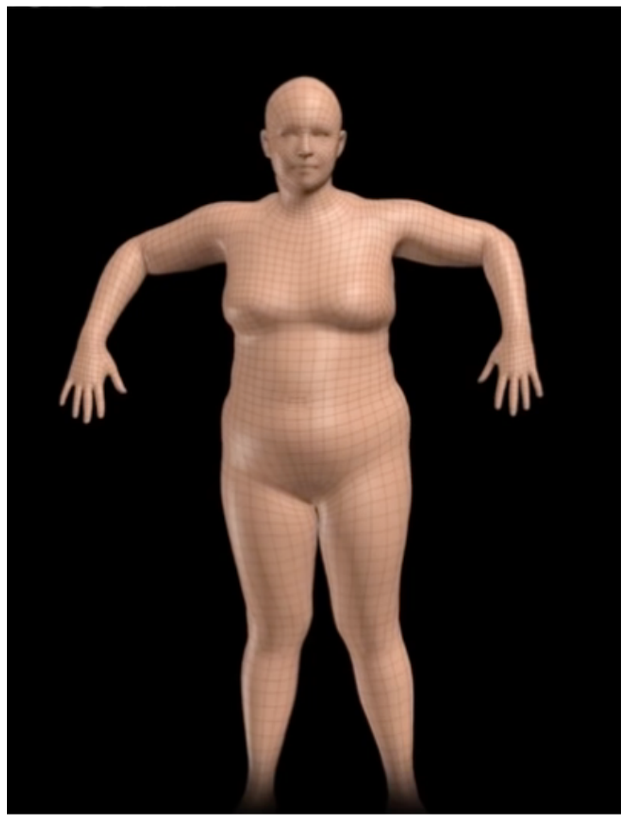
SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

Pose θ

Shape β



Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black,
SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

SMLP: a 3D human shape model

SMPL [M. Loper et al.]: a low-parametric model learned from aligning high-resolution 3D scans.

3D mesh
SMPL(θ , β)

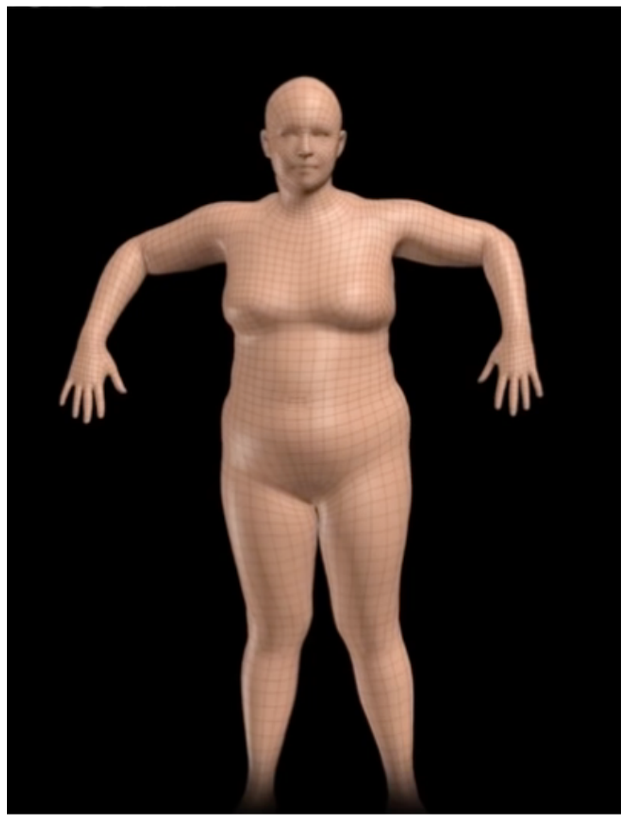
Pose

θ

Shape

β

Differentiable mapping



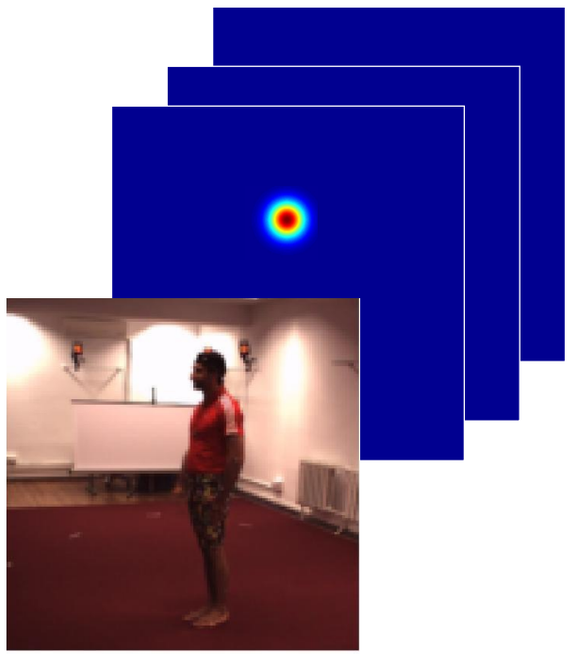
Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, Michael J. Black, SMPL: A Skinned Multi-Person Linear Model (SIGGRAPH Asia 2015)

3D body pose inference

Inputs:

RGB frame

2D keypoint heatmaps



3D body pose inference

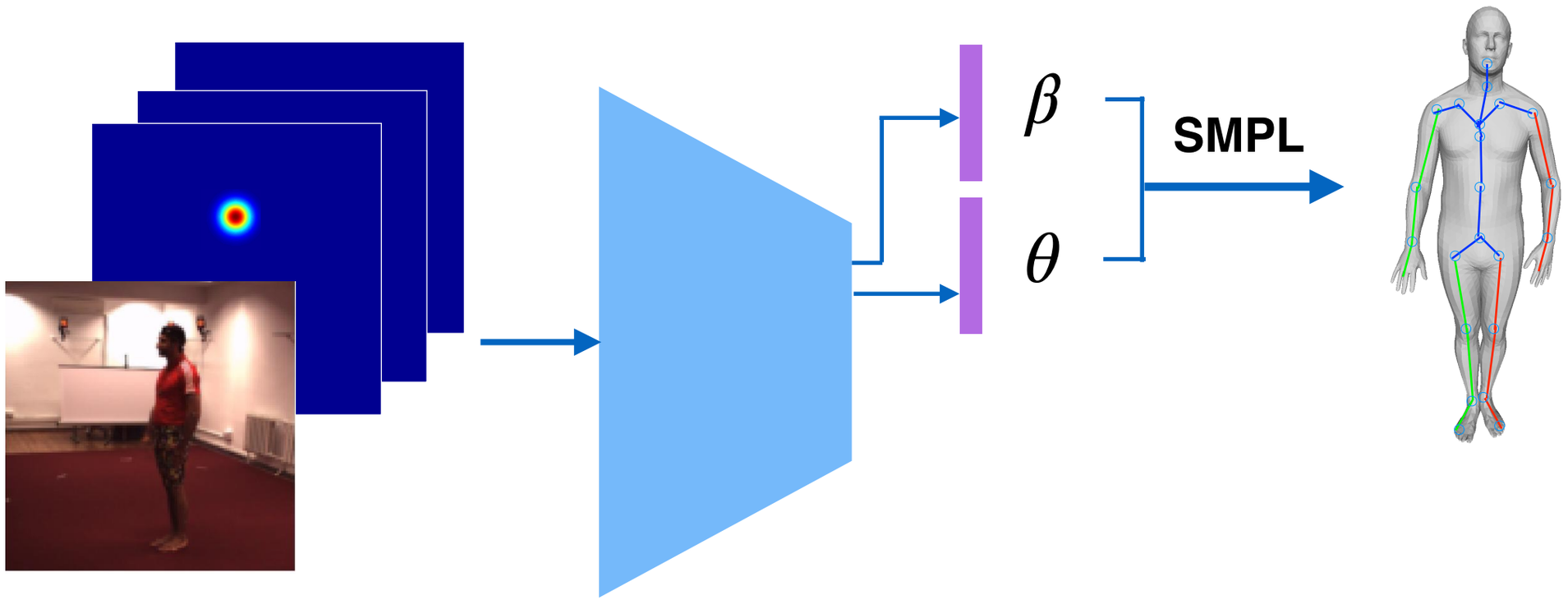
Inputs:

RGB frame

2D keypoint heatmaps

Outputs:

SMPL parameters (θ, β)



3D body pose inference

Inputs:

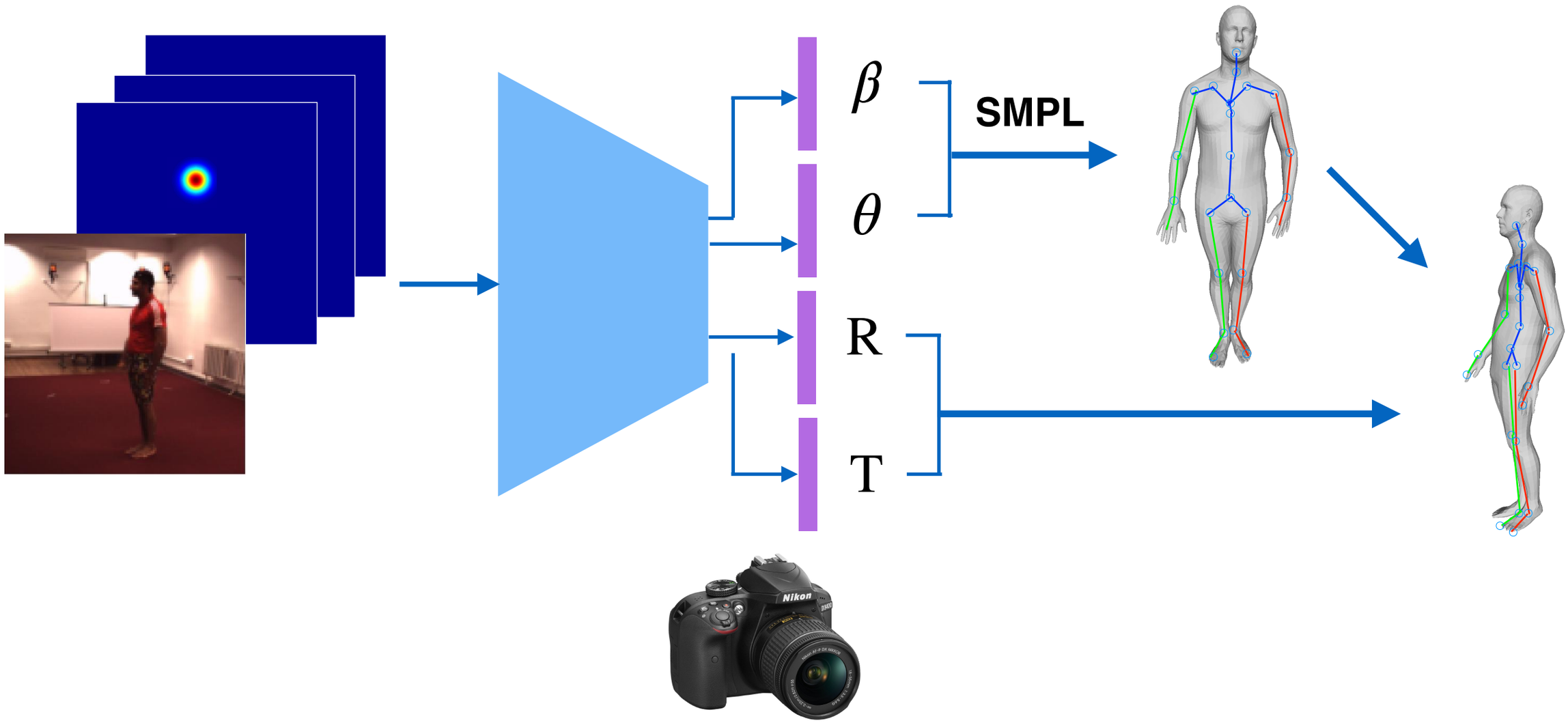
RGB frame

2D keypoint heatmaps

Outputs:

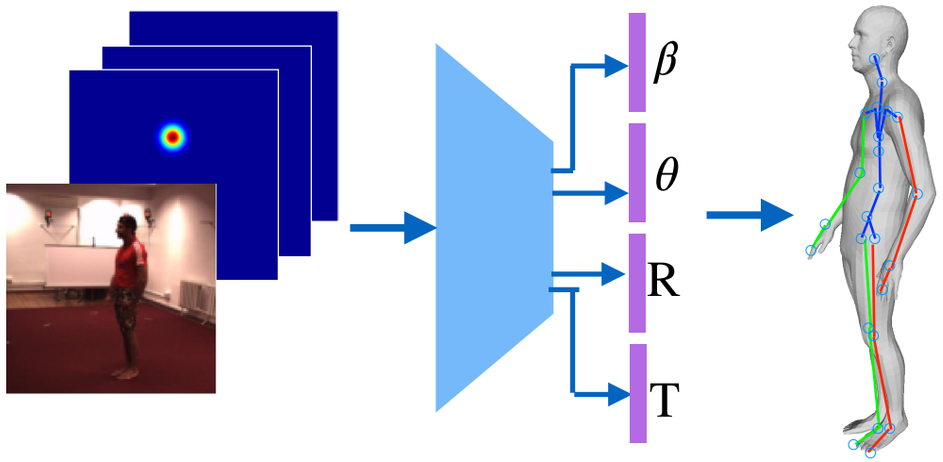
SMPL parameters (θ, β)

camera parameters(R, T)

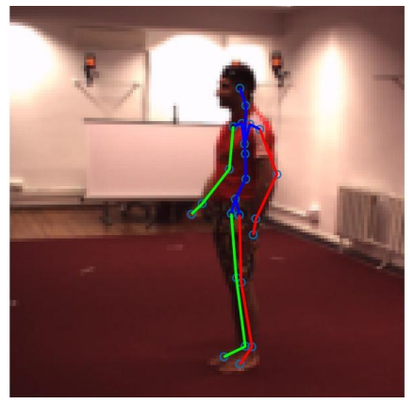


Self-supervised reprojection losses

Frame t

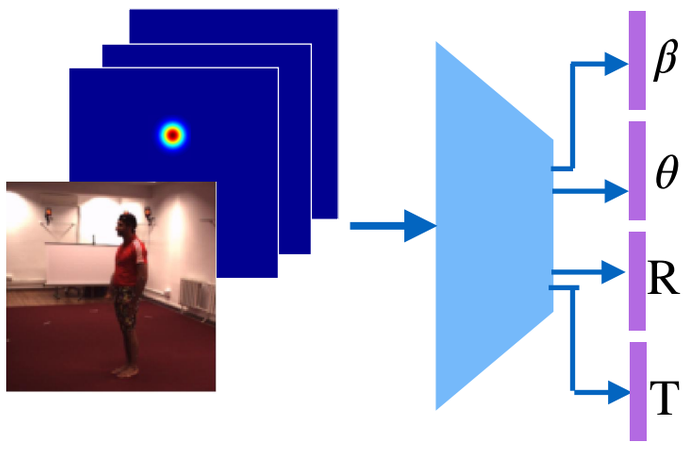


Keypoint
re-projection

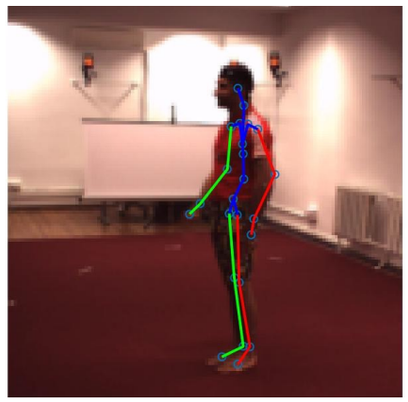


Self-supervised reprojection losses

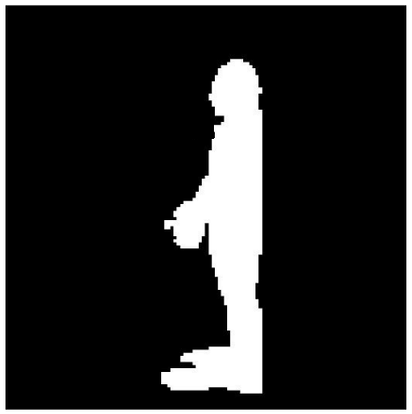
Frame t



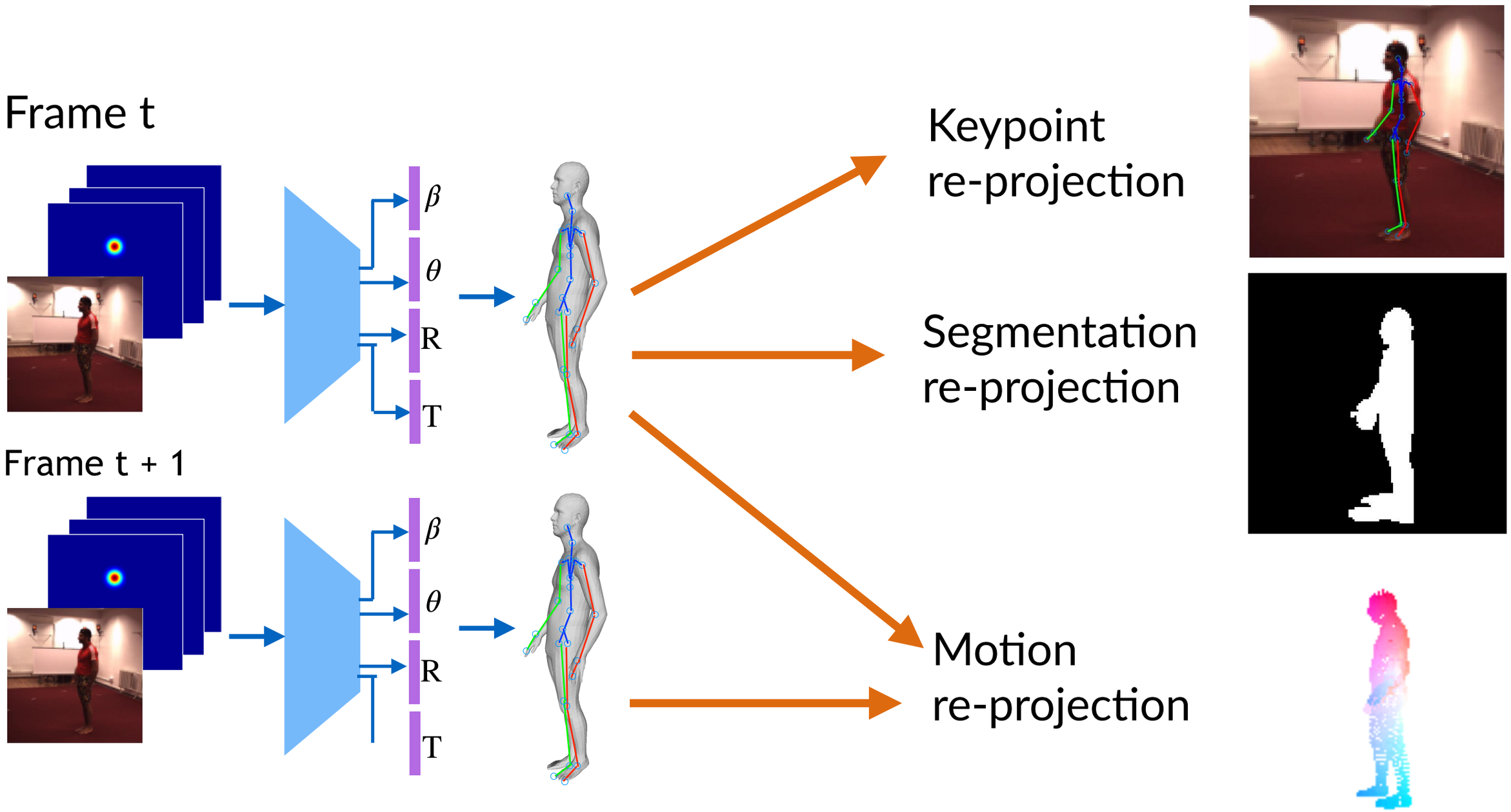
Keypoint re-projection



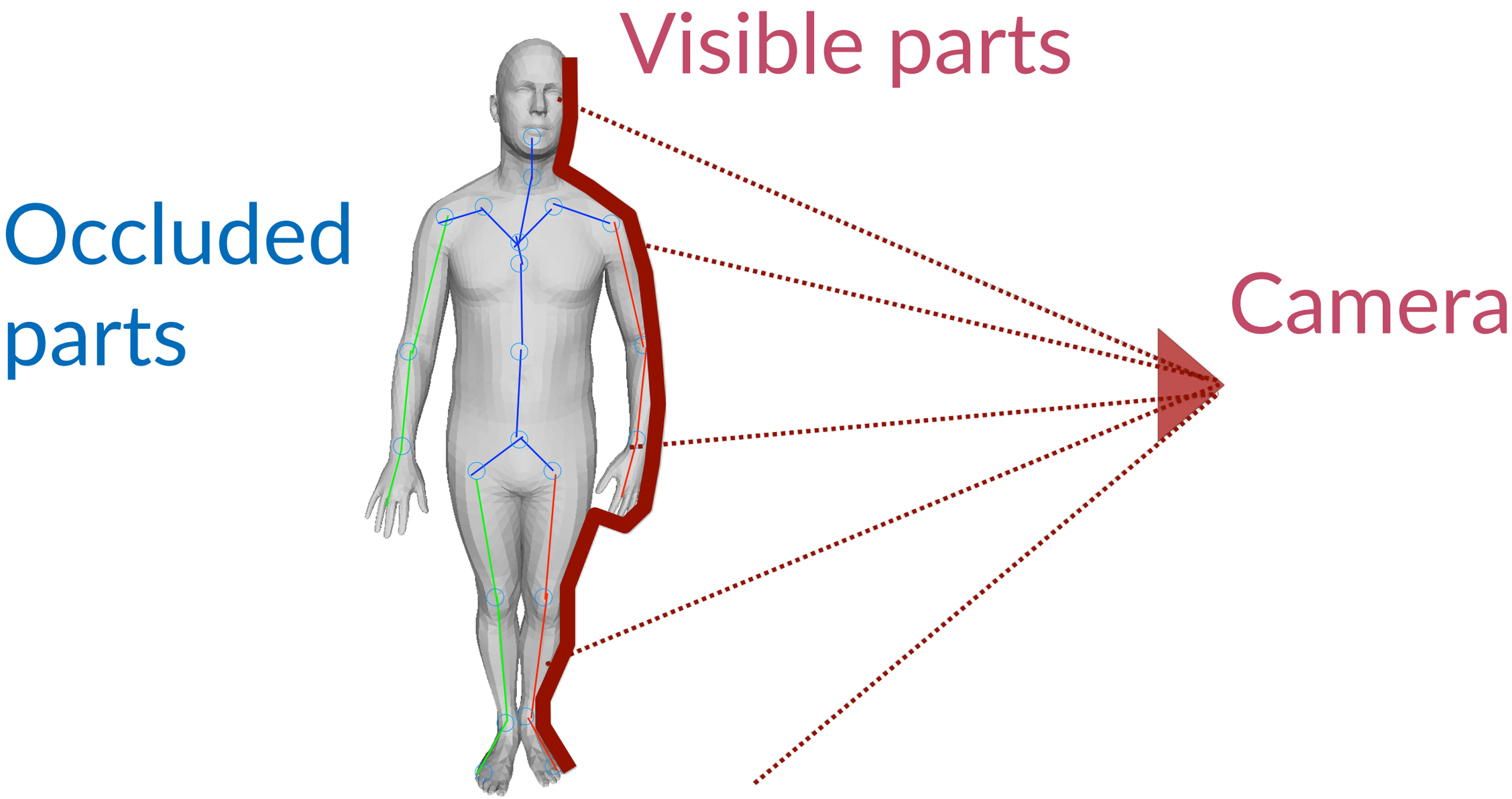
Segmentation re-projection

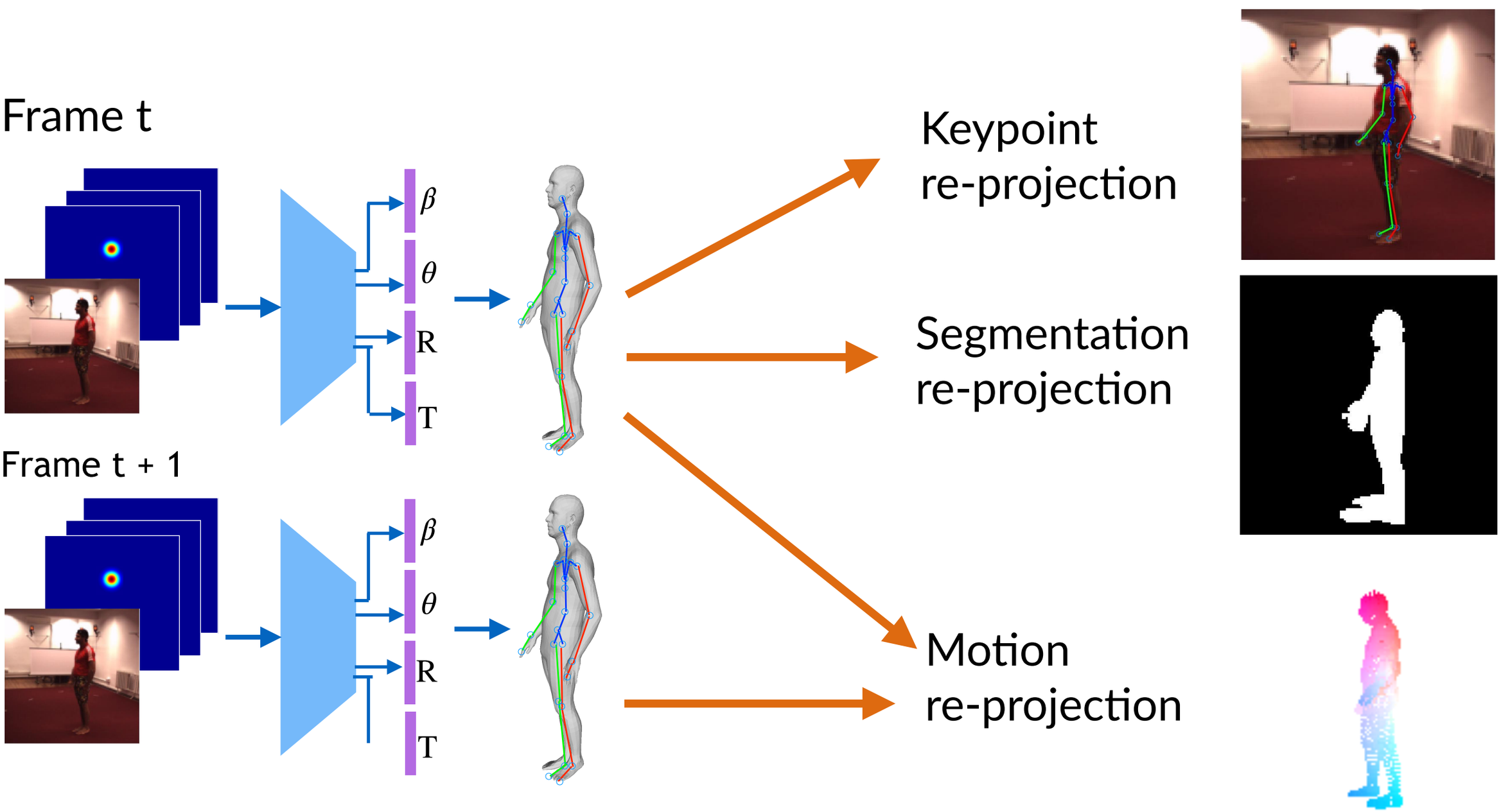


Self-supervised reprojection losses



Visibility-aware reprojection

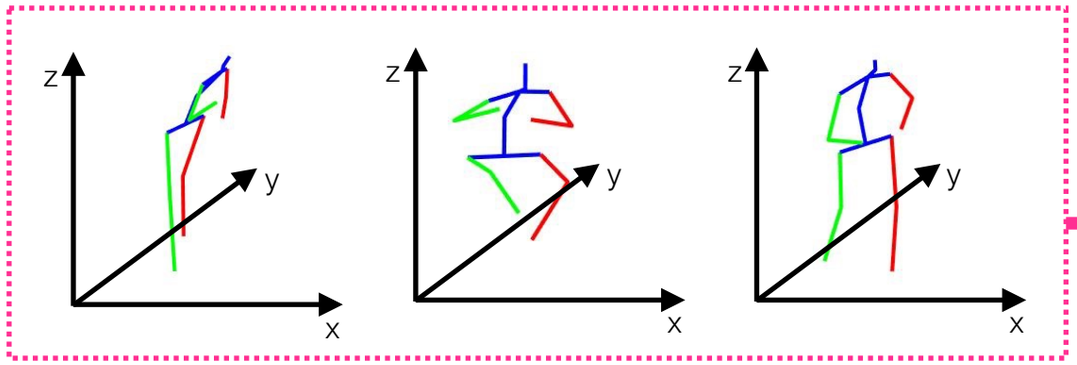




Q: Can such re-projection losses result in a non-anthropomorphic looking 3D human body pose?

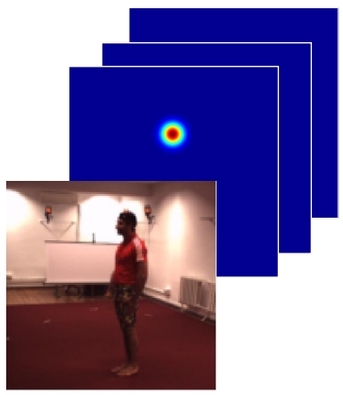
Adversarial matching

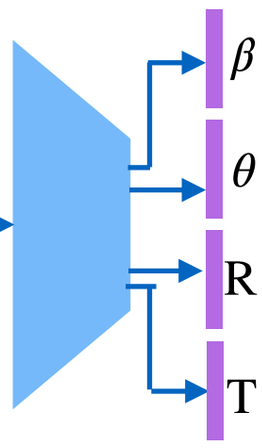
3D human poses

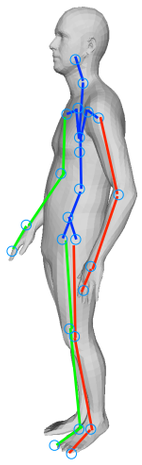


Discriminator

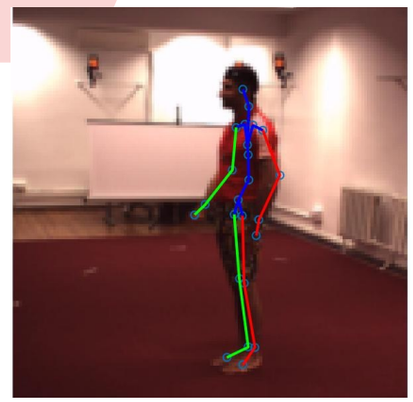
Frame t



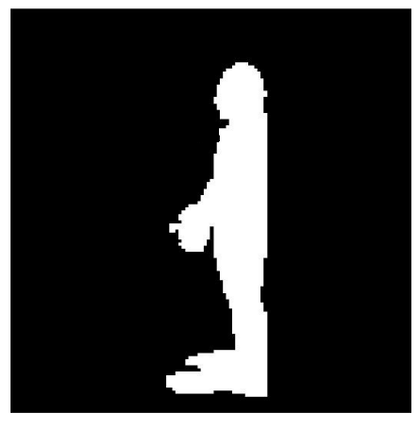




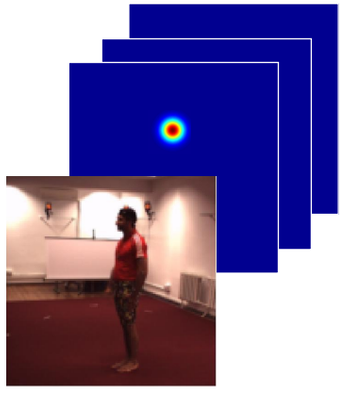
Keypoint
re-projection

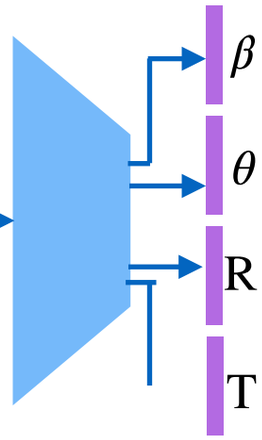


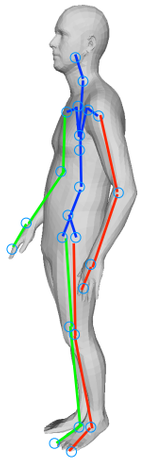
Segmentation
re-projection



Frame t + 1



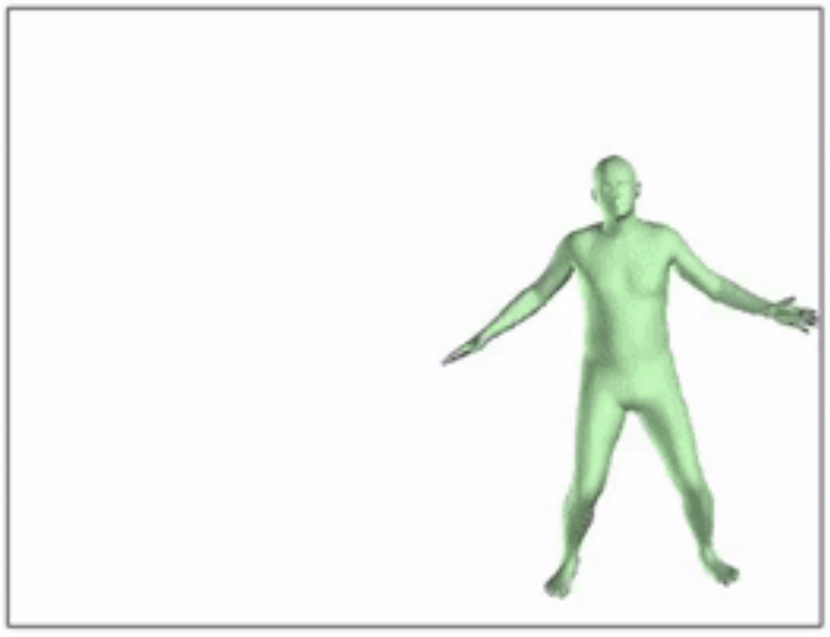




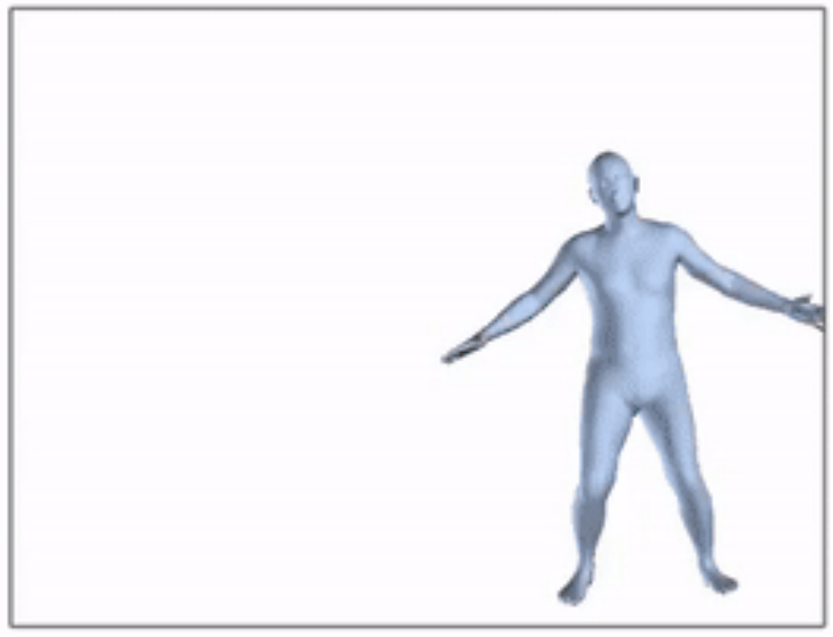
Motion
re-projection

+temporal smoothing

Video: Cartwheel A

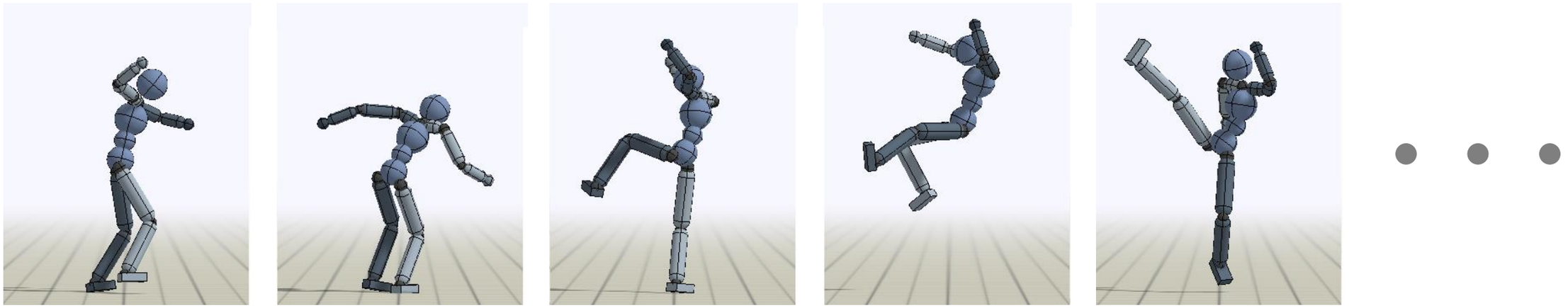


Before Reconstruction



After Reconstruction

Reference Motion

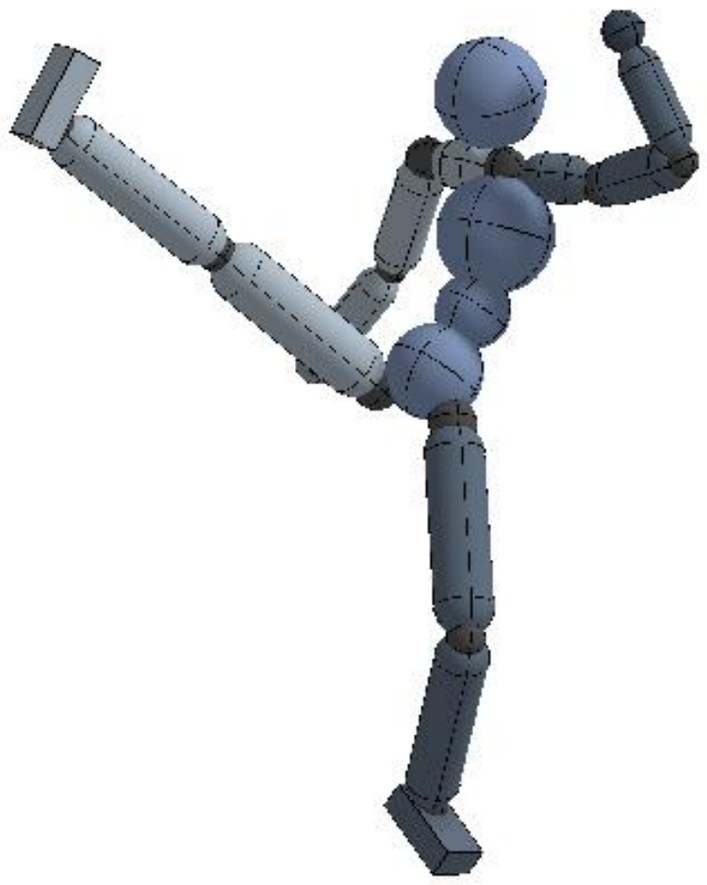


a_0 a_1 a_2 a_3 a_4 • • •

State + Action

State:

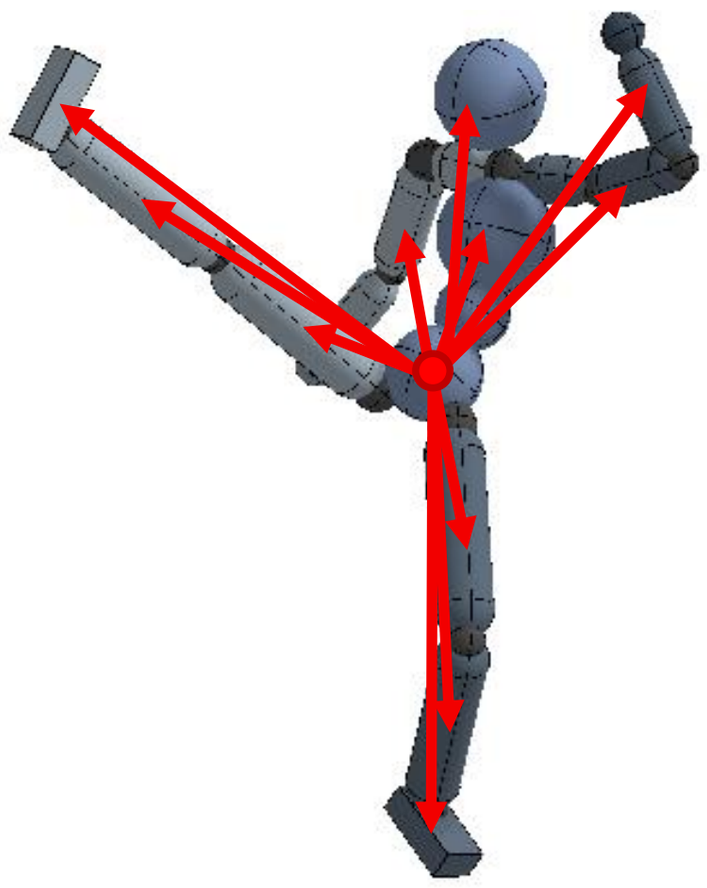
- link positions
- link velocities



State + Action

State:

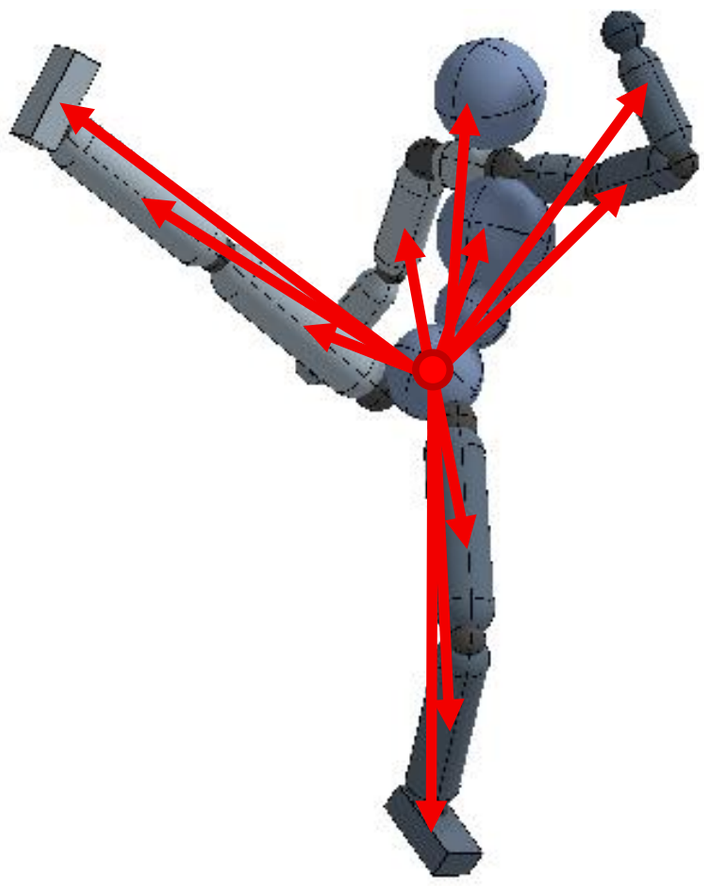
- link positions
- link velocities



State + Action

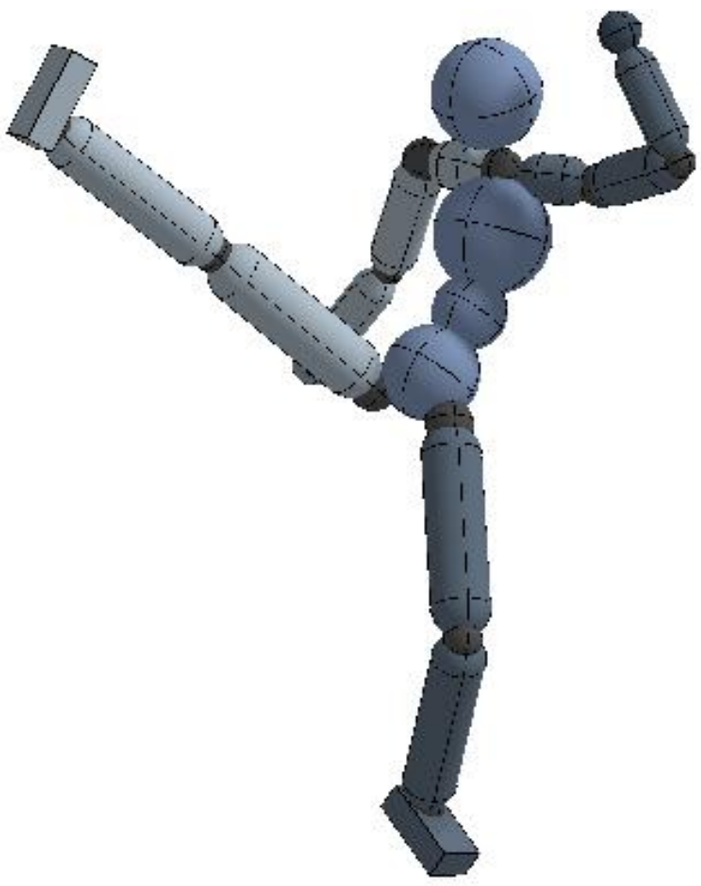
State:

- link positions
- link velocities



Action:

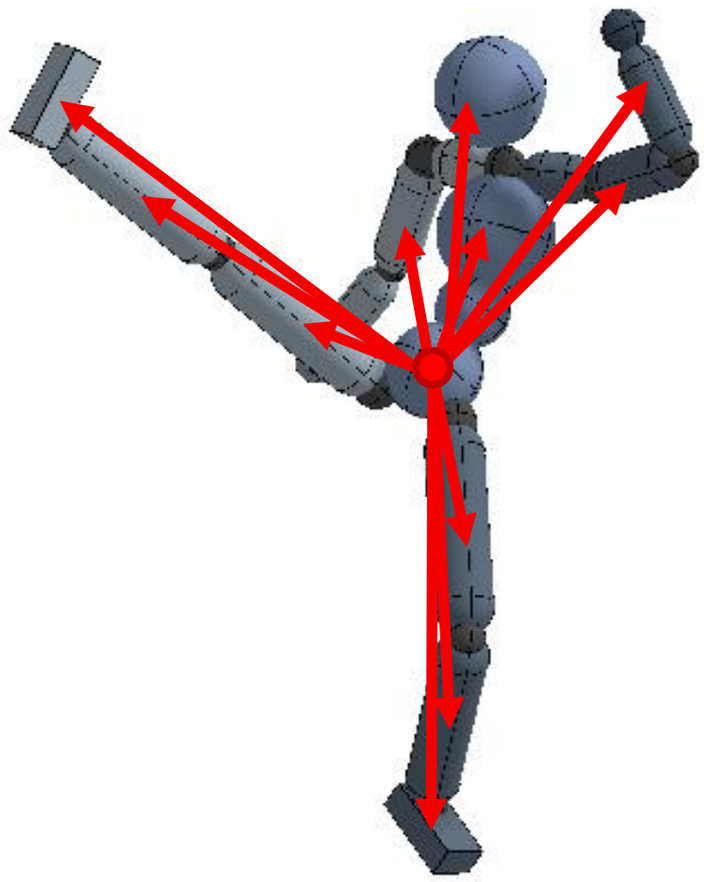
- PD targets



State + Action

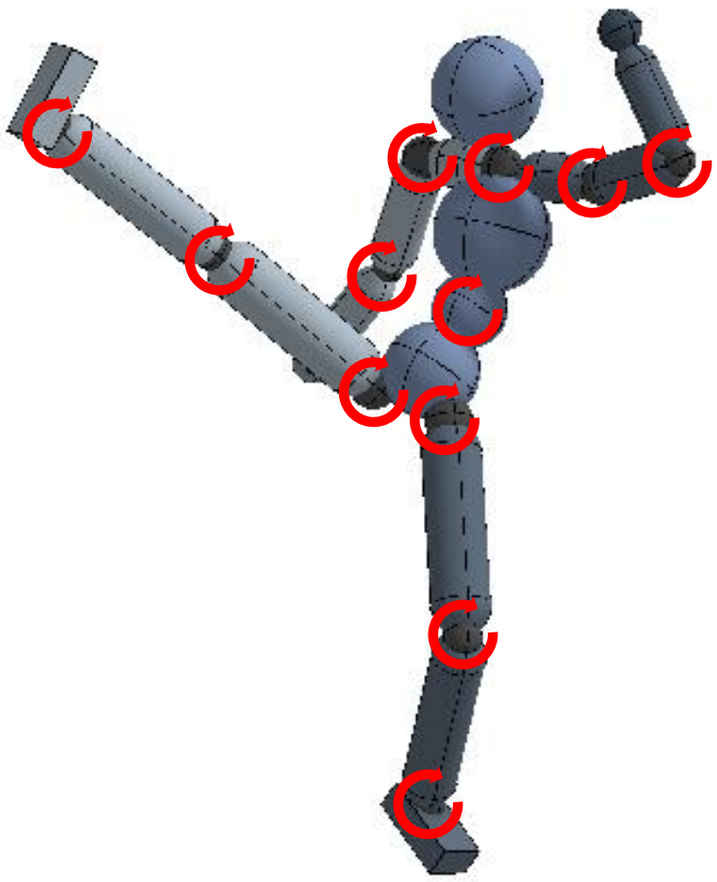
State:

- link positions
- link velocities



Action:

- PD targets

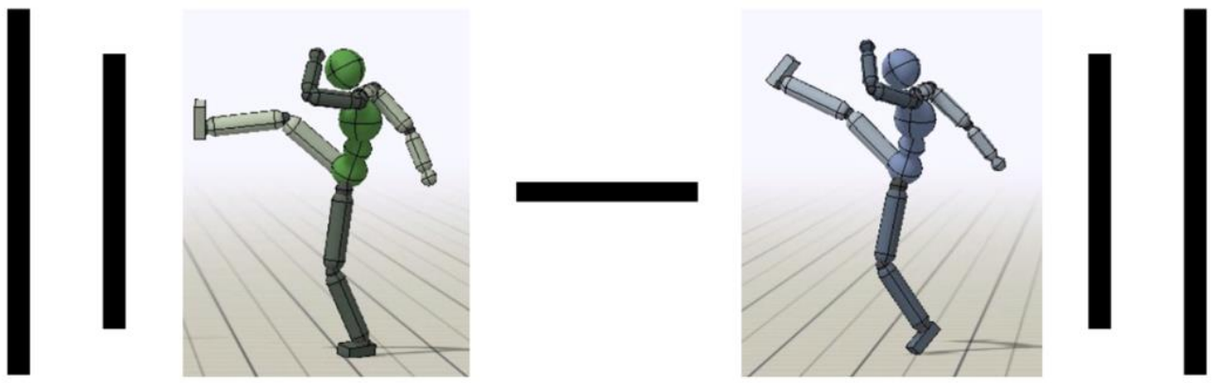


Imitation Objective

the reference trajectory

$$r_t = \exp\left(-2 \left\| \hat{q}_t - q_t \right\|^2\right)$$

Imitation Objective



Proximal Policy Optimization

$$\max_{\theta} \quad J(\theta)$$

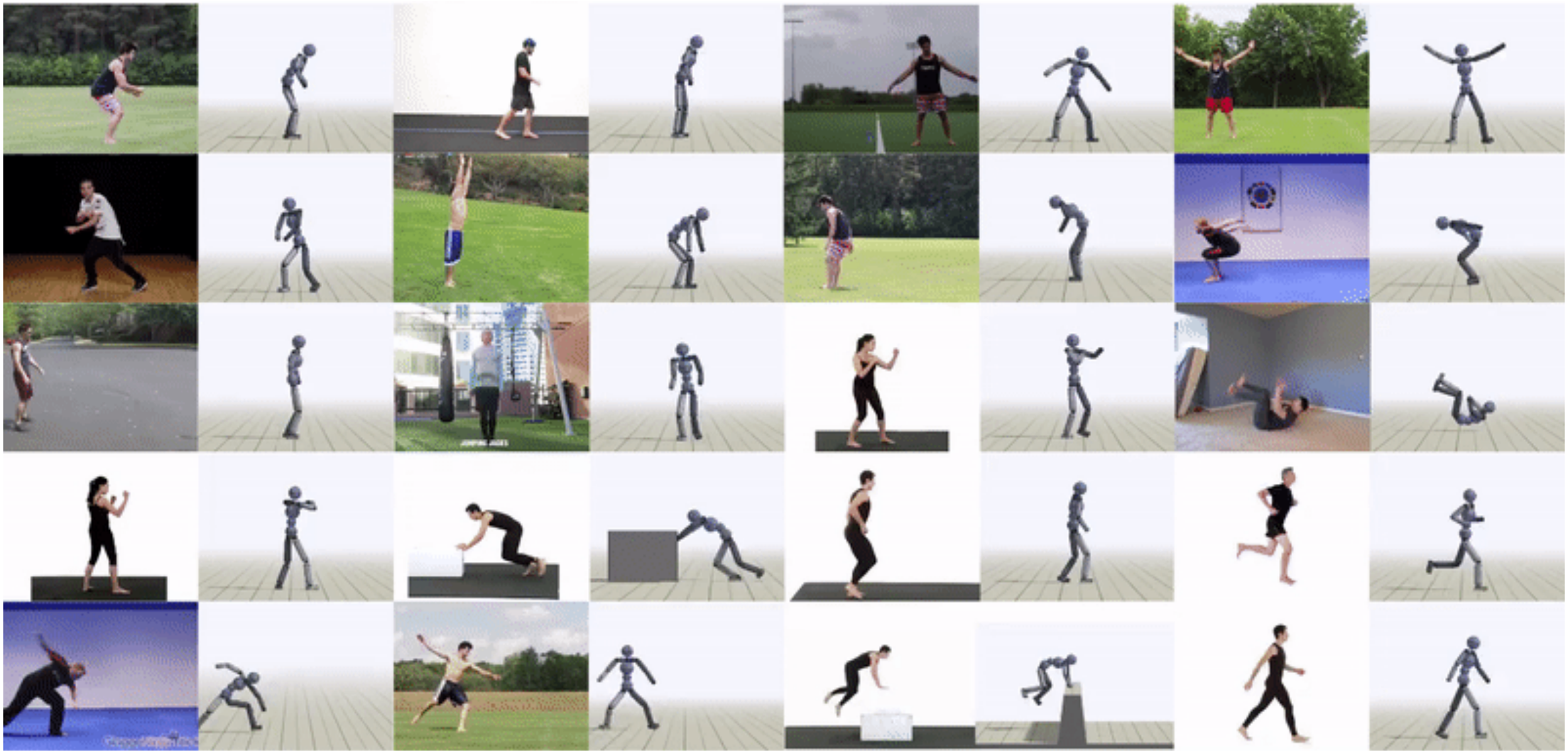
Schulman et al. 2017

Proximal Policy Optimization

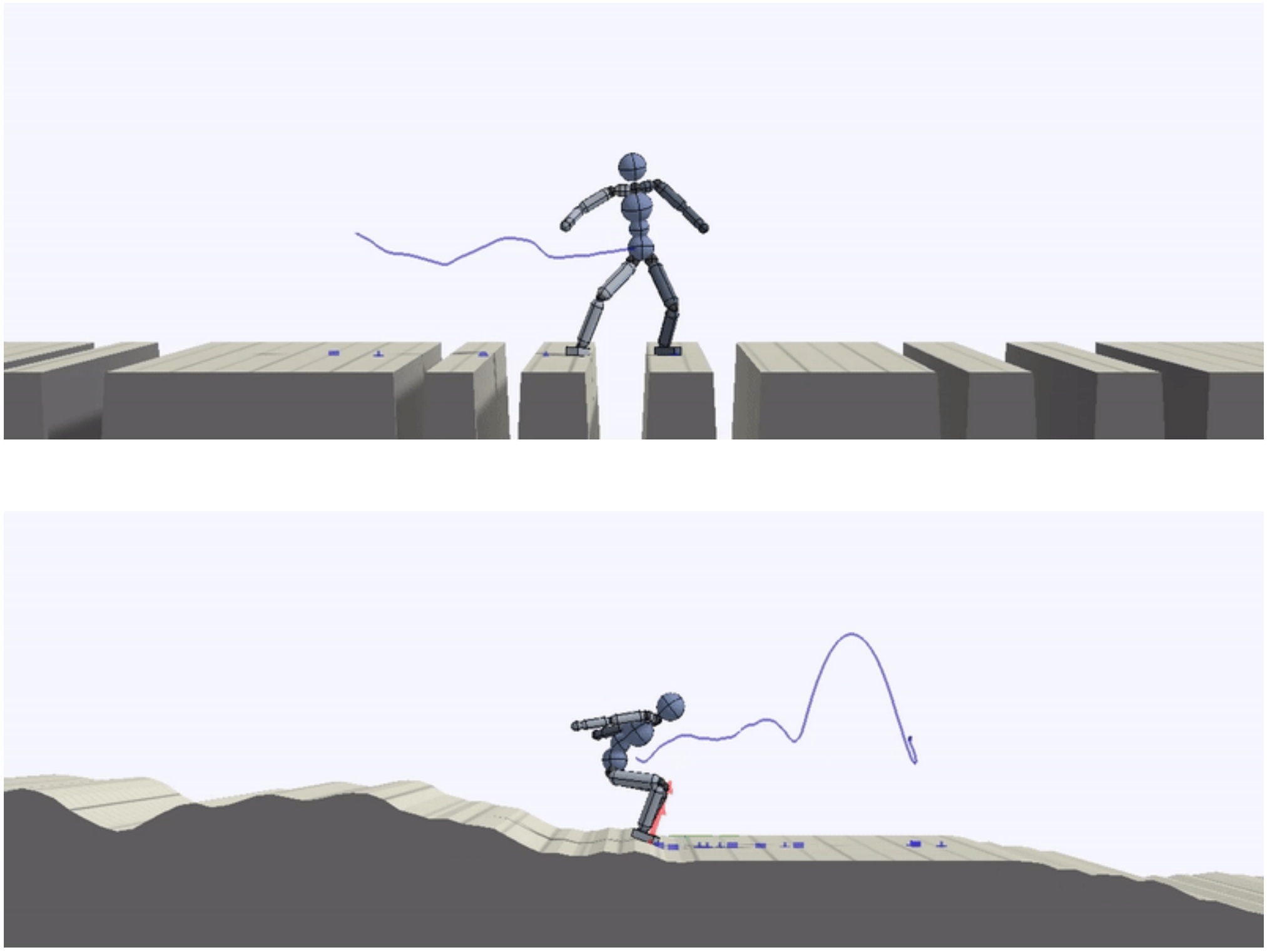
$$\max_{\theta} \quad J(\theta)$$

$$\text{s.t} \quad \mathbb{E}_{S_t \sim d_\theta(s_t)} \left[KL \left(\pi_{\theta_{old}} \left(\cdot \,|\, s_t \right) \,\middle|\, \pi_\theta \left(\cdot \,|\, s_t \right) \right) \right] \leq \delta_{KL}$$

Schulman et al. 2017



Adapting a skill through RL to novel environments



Failure modes

Video: Gangnam Style

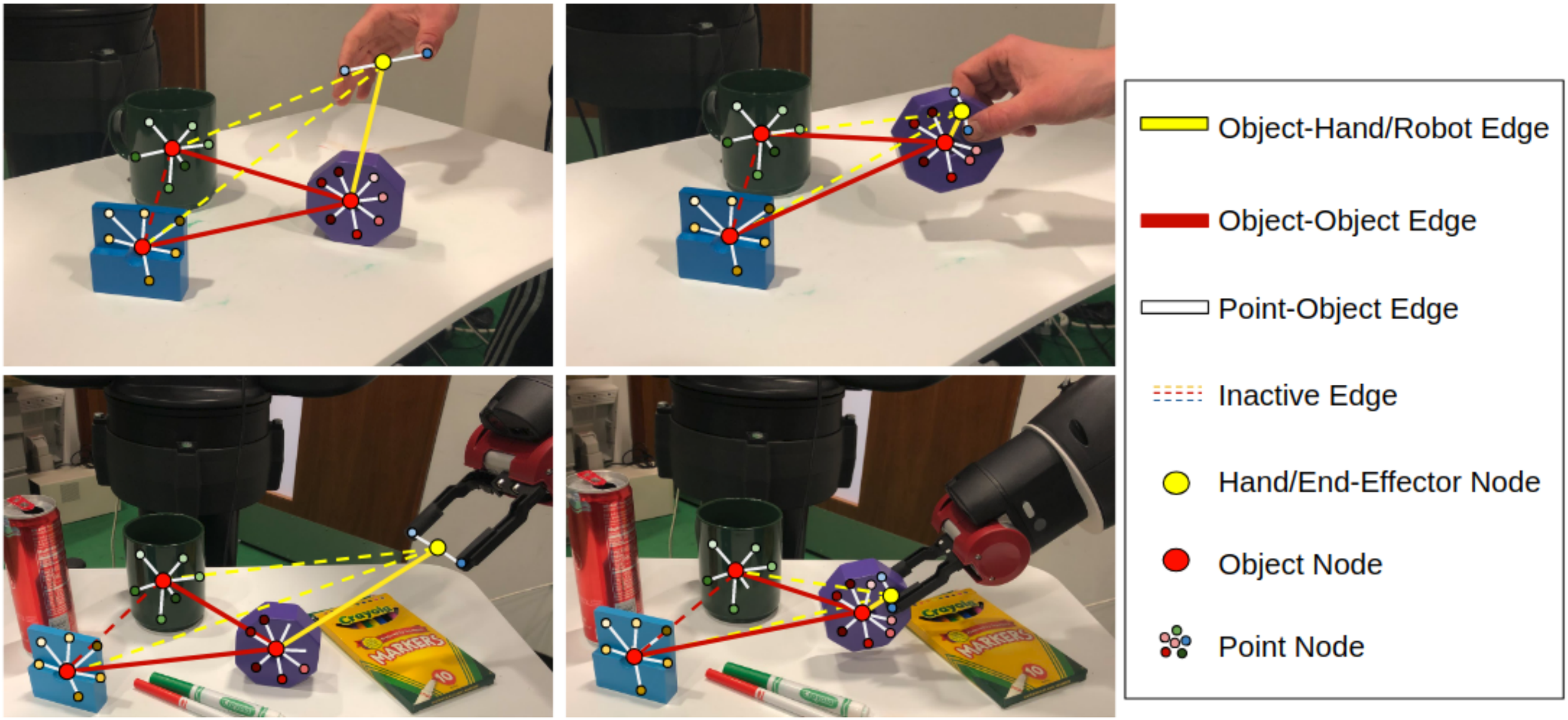
Reference Motion

Simulation

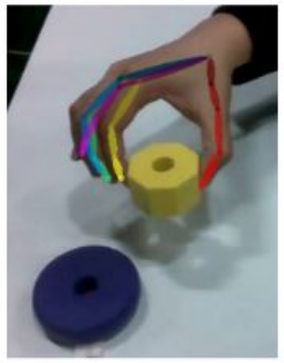
Imitating beyond human body pose

- Imitating human body was possible thanks to progress in Computer Vision that can detect 2D human body keypoints and reconstruct them in 3D very well.
- What about the rest of the objects in the world, that we cannot yet easily 3D reconstruct?
- A: We need wait a bit more for CV to work..

Visual Entity Graphs for Visual Imitation

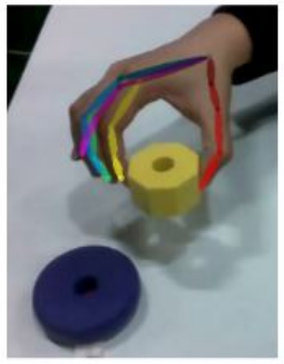


Detecting Visual Entities

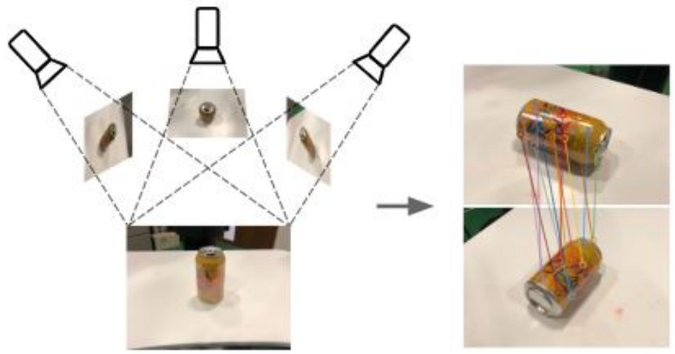


Hand Keypoint Detection

Detecting Visual Entities

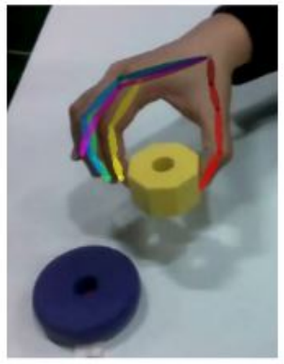


Hand Keypoint Detection

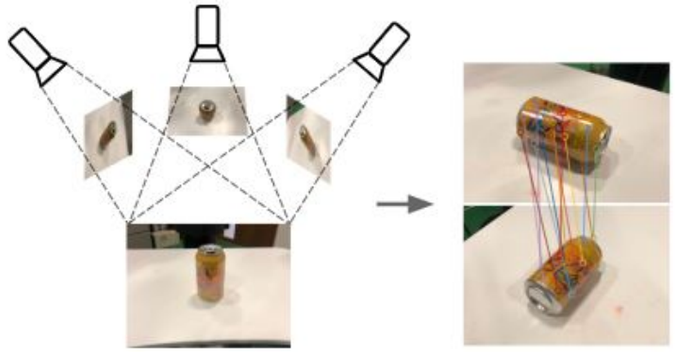


Multi-View Self-Supervised Point-Feature Learning

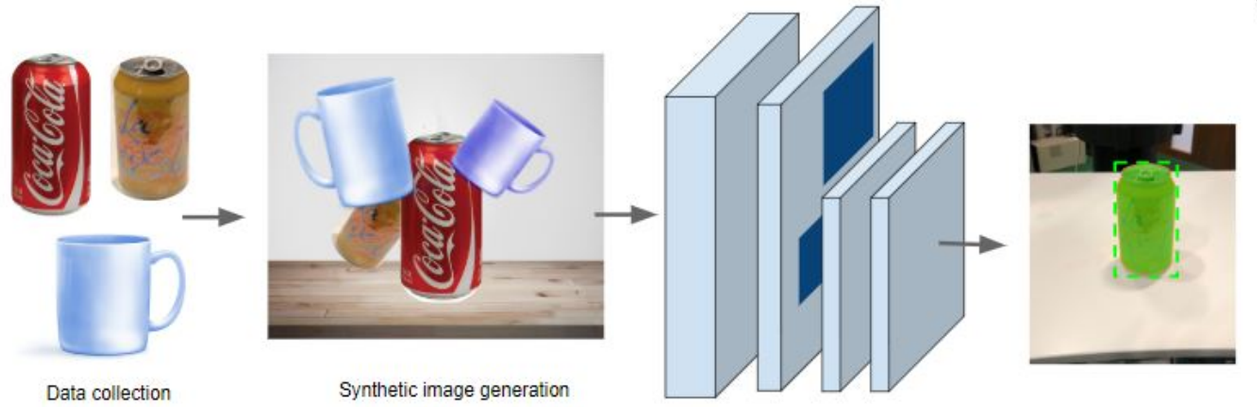
Detecting Visual Entities



Hand Keypoint Detection

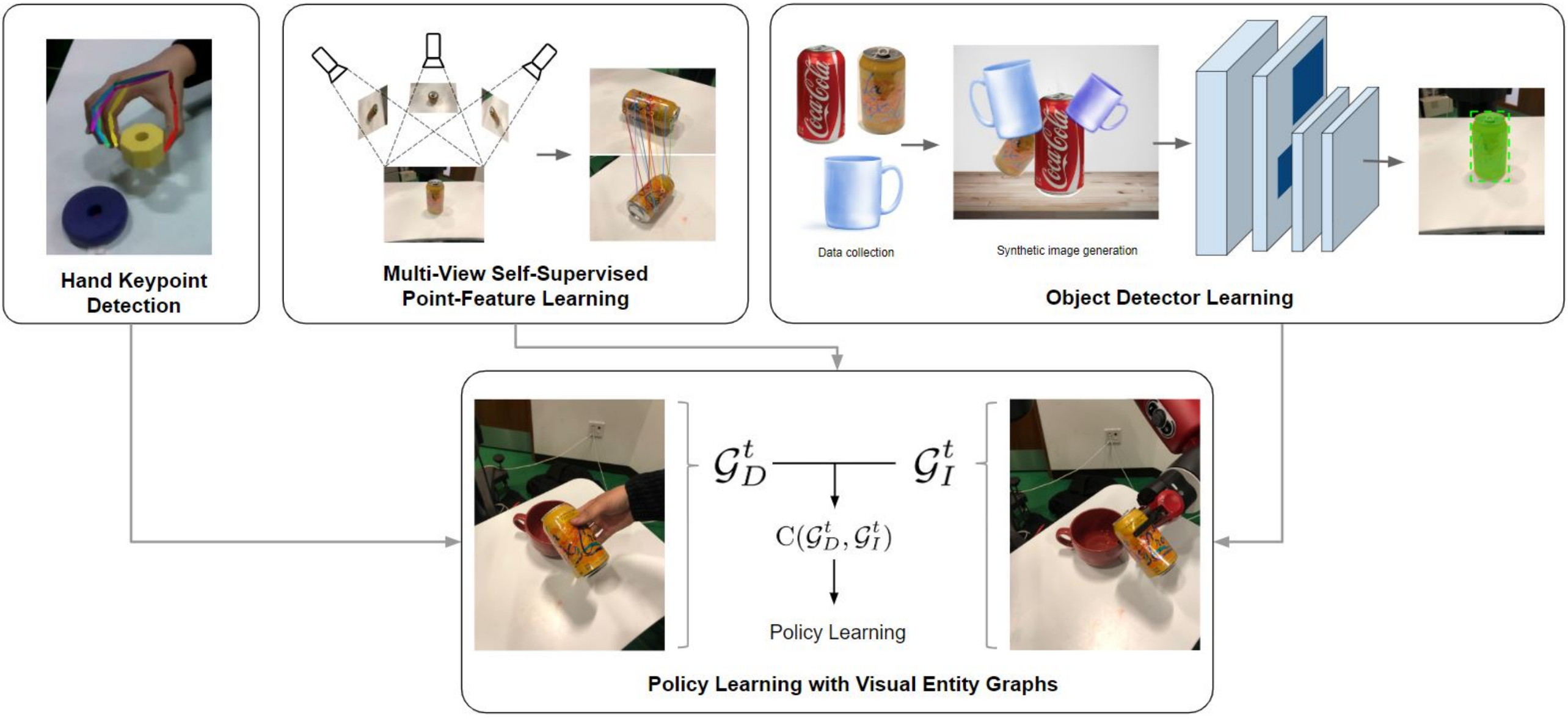


Multi-View Self-Supervised Point-Feature Learning



Object Detector Learning

Detecting Visual Entities



Playing hard exploration games by watching YouTube

Yusuf Aytar*, **Tobias Pfaff***, **David Budden**, **Tom Le Paine**, **Ziyu Wang**, **Nando de Freitas**

DeepMind, London, UK
{yusufaytar,tpfaff,budden,tpaine,ziyu,nandodefreitas}@google.com

- Input: video demonstrations (without rewards).

- Self-supervised visual representation learning to bridge the domain gap between youtube video demonstrations of people playing the game, with the frames the game emits

- Given one video demo, use visual similarity encoded as frame embedding distance as imitation reward, to be added (optionally) to environment rewards.

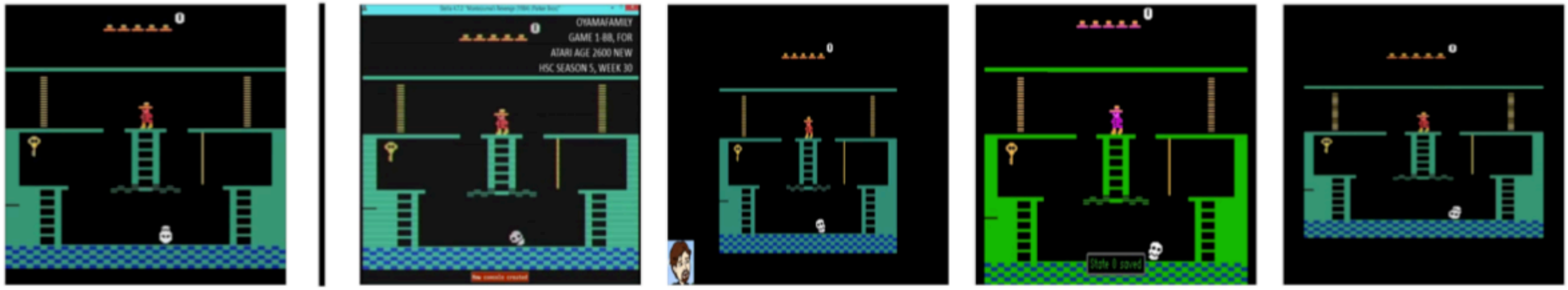
Playing hard exploration games by watching YouTube

Yusuf Aytar*, **Tobias Pfaff***, **David Budden**, **Tom Le Paine**, **Ziyu Wang**, **Nando de Freitas**

DeepMind, London, UK
{yusufaytar,tpfaff,budden,tpaine,ziyu,nandodefreitas}@google.com

- Input: video demonstrations (without rewards).
- Self-supervised visual representation learning to bridge the domain gap between youtube video demonstrations of people playing the game, with the frames the game emits
- Given one video demo, use visual similarity encoded as frame embedding distance as imitation reward, to be added (optionally) to environment rewards.

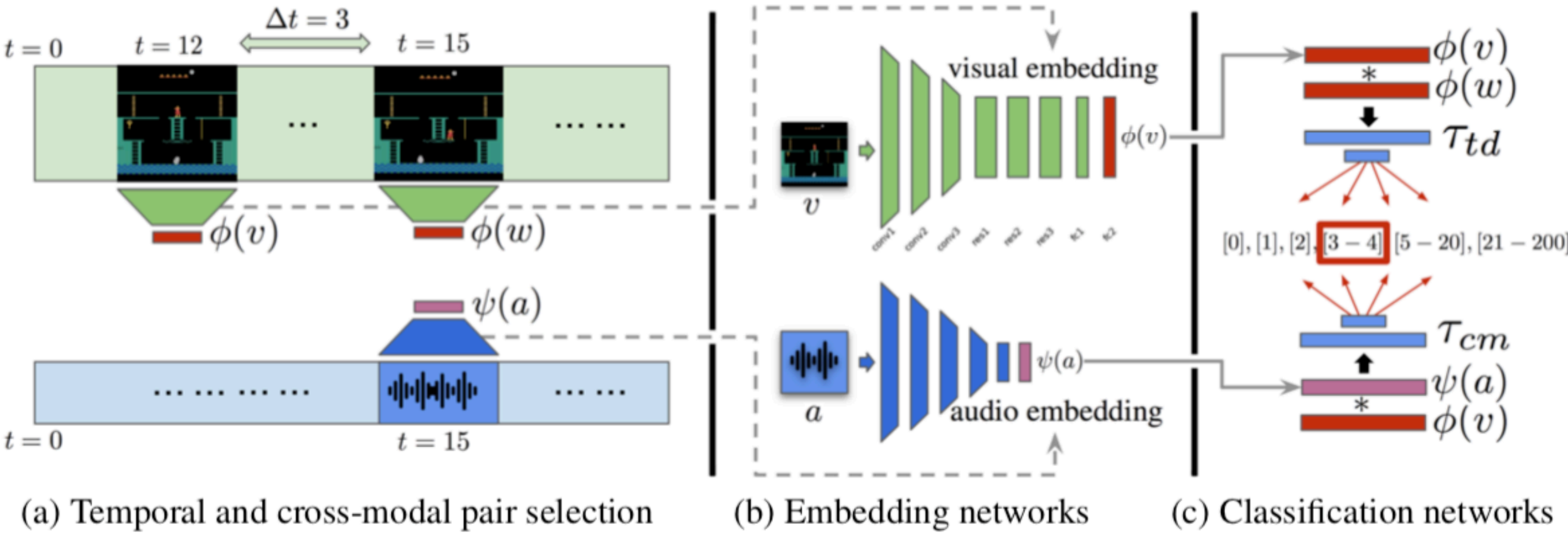
Closing the visual domain gap



(a) ALE frame (b) Frames from different YouTube videos

- Not a huge domain gap, but nonetheless needs to be bridged for comparing frames across the two domains. How?

Closing the visual domain gap



(a) Temporal and cross-modal pair selection (b) Embedding networks (c) Classification networks

- Temporal distance classification: given two frames, clarify their temporal distance into one of k intervals, e.g., {[0],[1],[2],[3−4],[5−20],[21−200]}

- Cross-modal temporal distance classification: given a video frame and an audio snippet, classify their temporal distance into two categories: matching, non-matching.

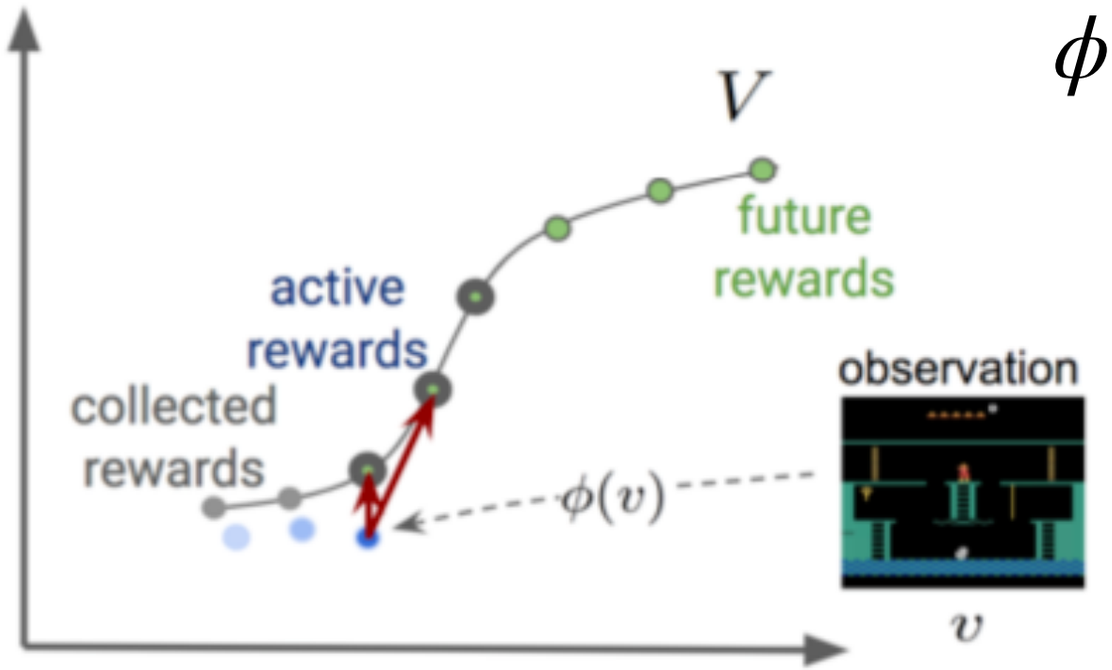
Playing hard exploration games by watching YouTube

Yusuf Aytar*, **Tobias Pfaff***, **David Budden**, **Tom Le Paine**, **Ziyu Wang**, **Nando de Freitas**

DeepMind, London, UK
{yusufaytar,tpfaff,budden,tpaine,ziyu,nandodefreitas}@google.com

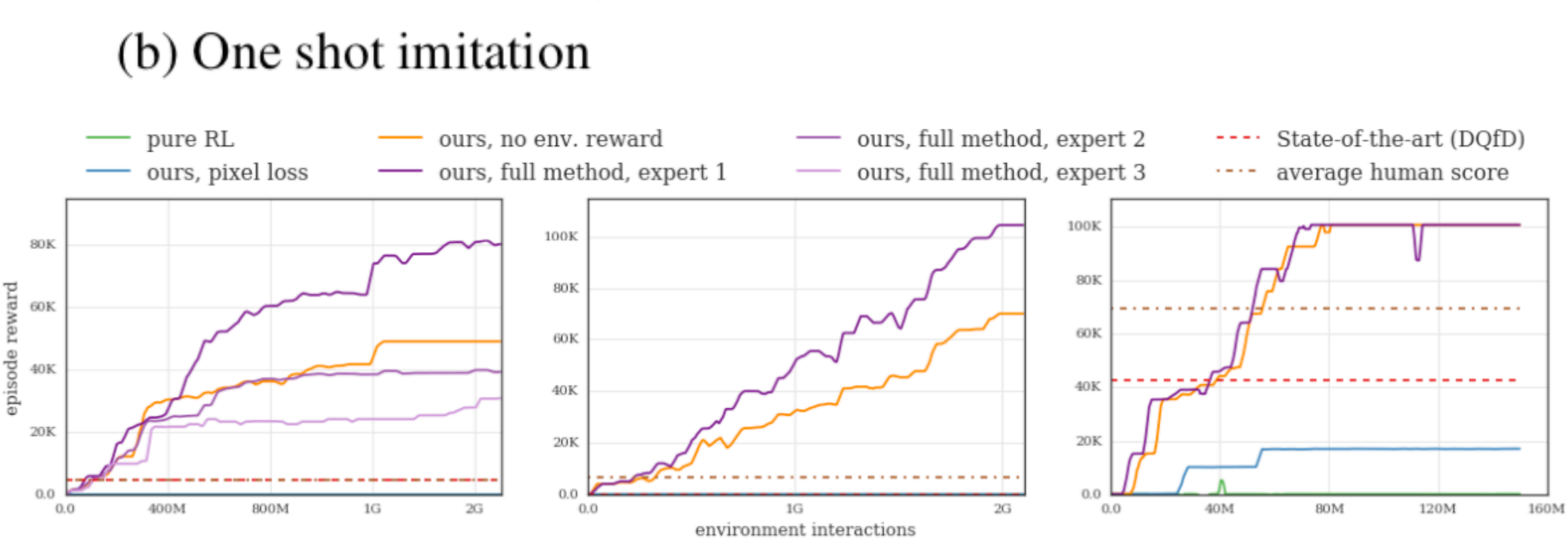
- Input: video demonstrations (without rewards).
- Self-supervised visual representation learning to bridge the domain gap between youtube video demonstrations of people playing the game, with the frames the game emits
- Given one video demo, use visual similarity encoded as frame embedding distance as imitation reward, to be added (optionally) to environment rewards.

Single shot visual imitation



$\phi(v)$: the visual feature encoding

(b) One shot imitation



MONTEZUMA'S REVENGE PITFALL! PRIVATE EYE

Is this how babies imitate?

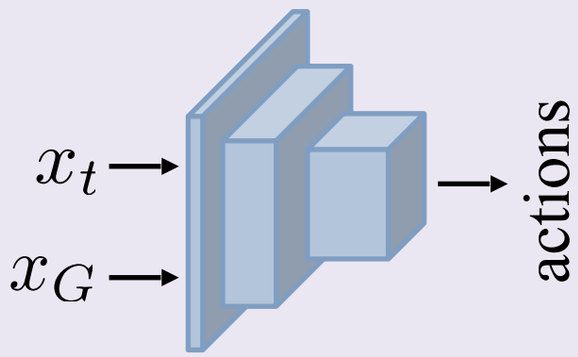
- Doing nothing till someone shows them a visual demonstation, and then they get to work?
- No. They are quite busy even on their own exploring the world and building models for it.
- Then, they make use of those models to accelerate imitation.
- Q: Did the imitation methods we showed used any model knowledge?

Hypothesis of how babies imitate

TRAINING TIME

Explore the Environment

Distill Exploration into skills

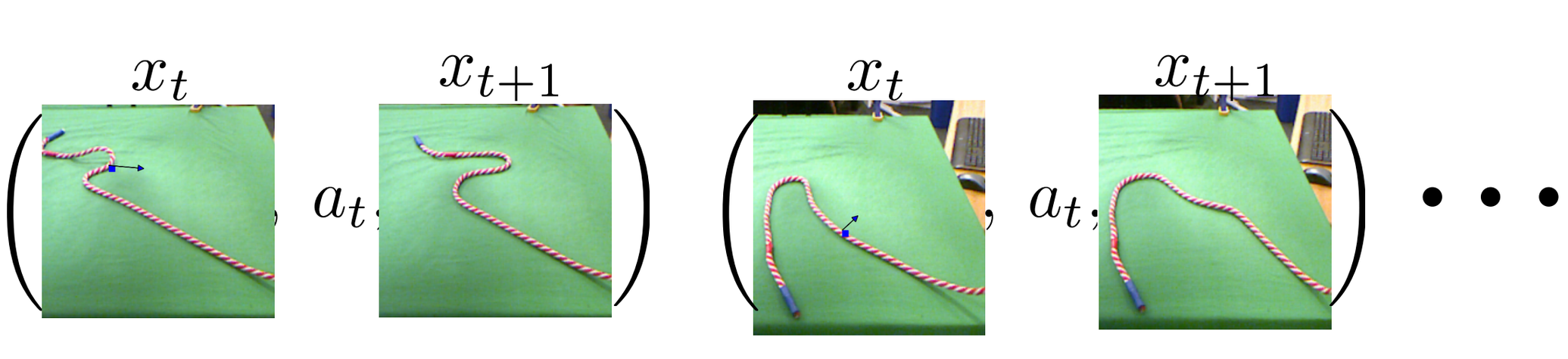


TEST TIME

Perform End-Tasks

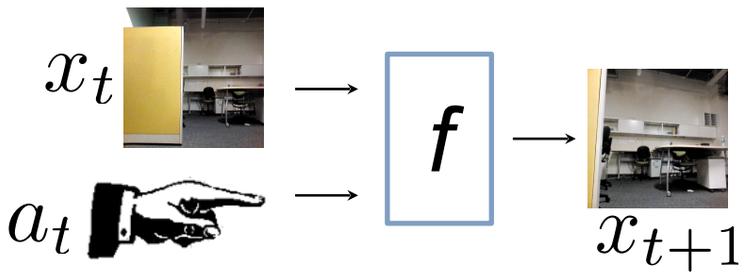
Model learning by random exploration

Rope Manipulation

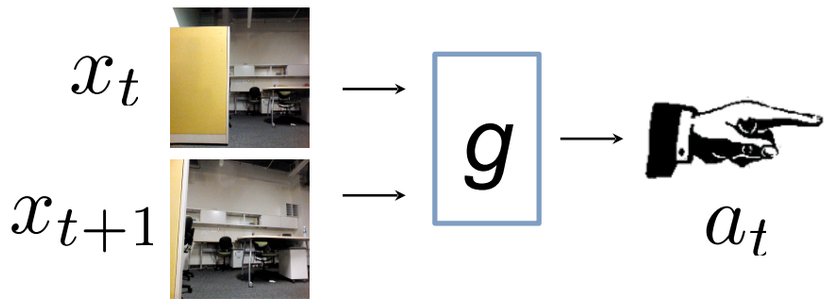


Visual Navigation

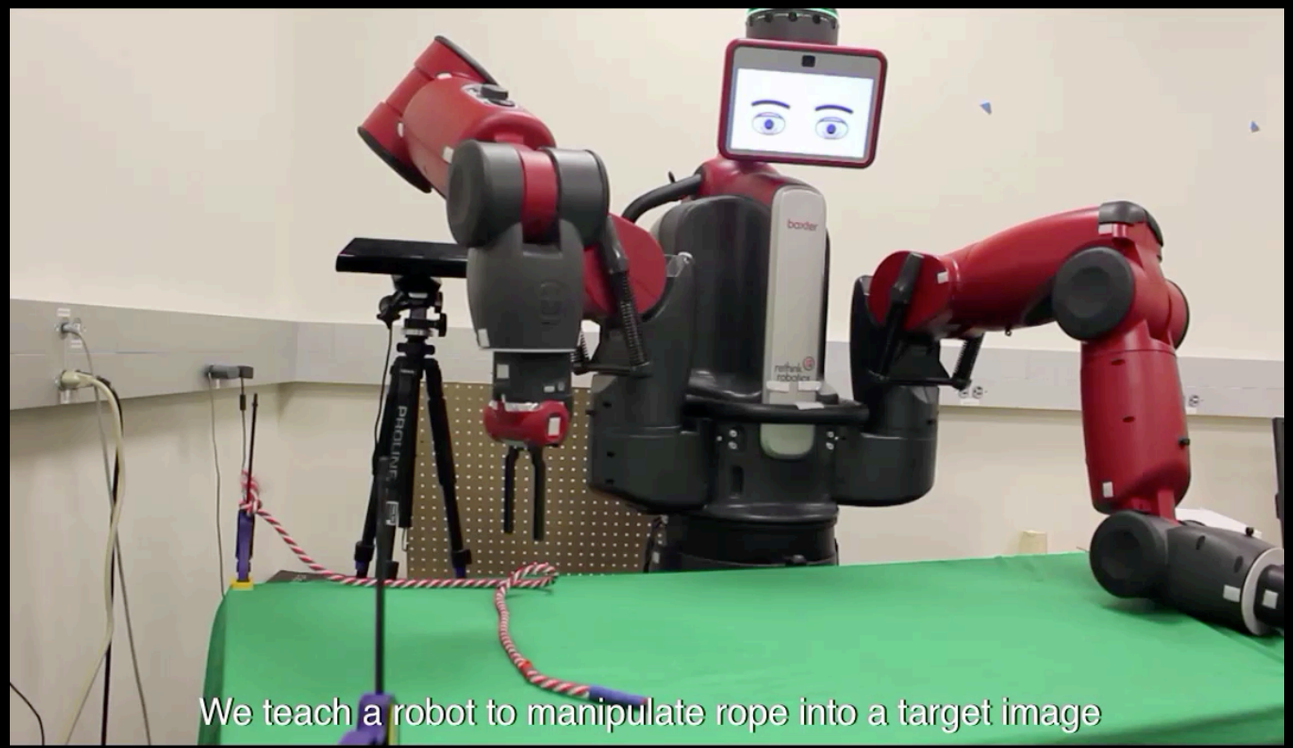
Forward Dynamics



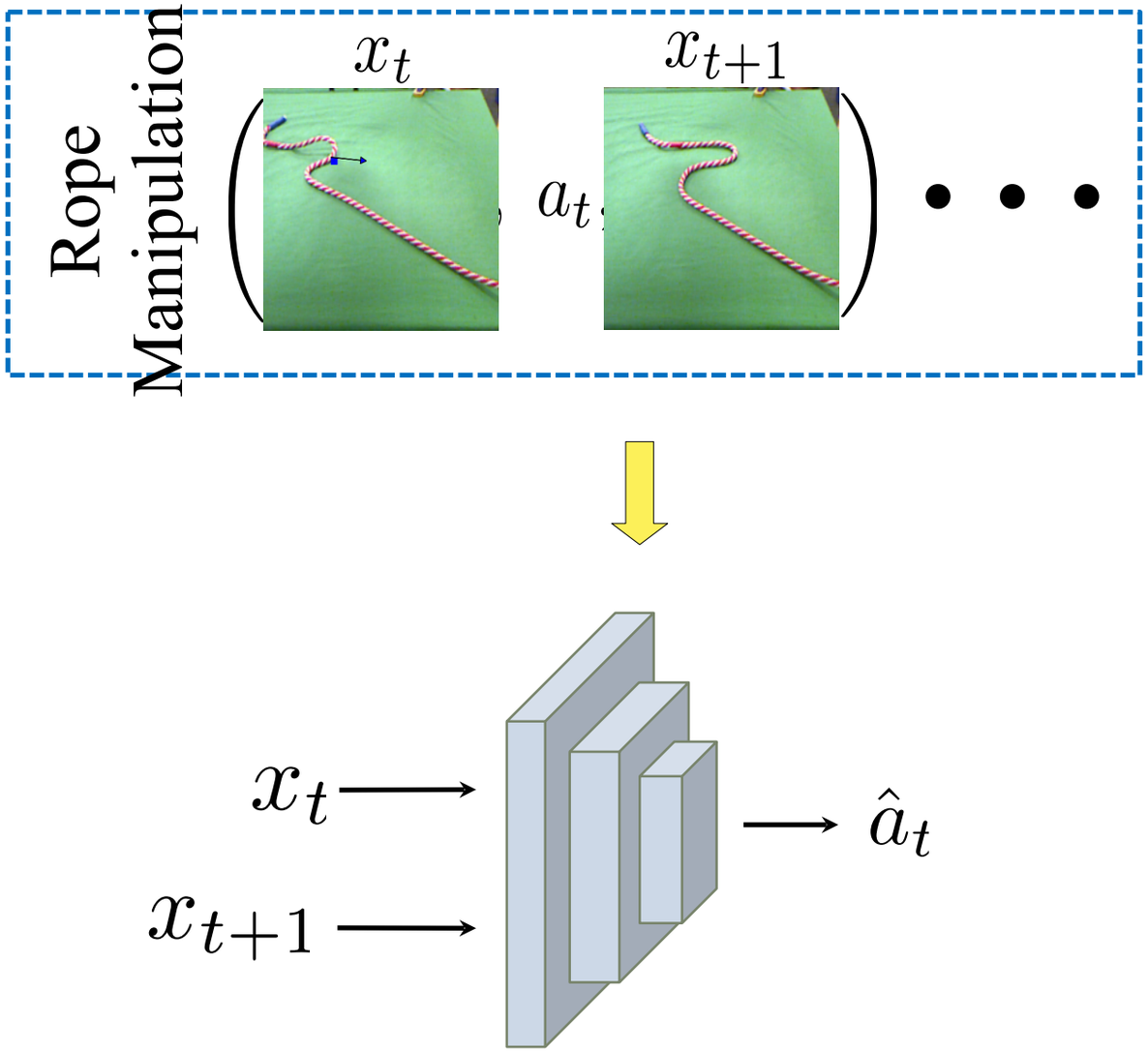
Inverse Dynamics



Exploring the Environment



Model learning by random exploration

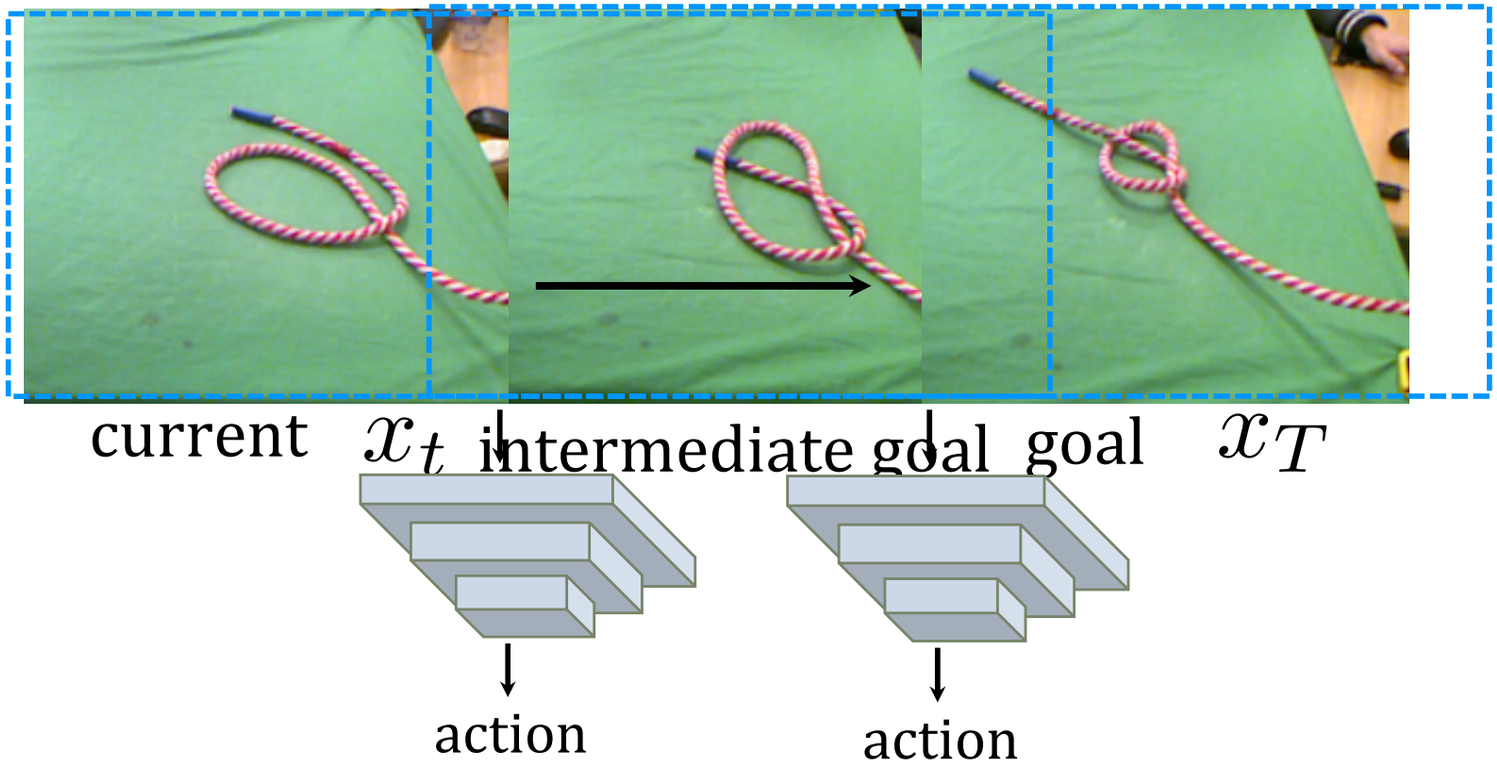


"skill" policy

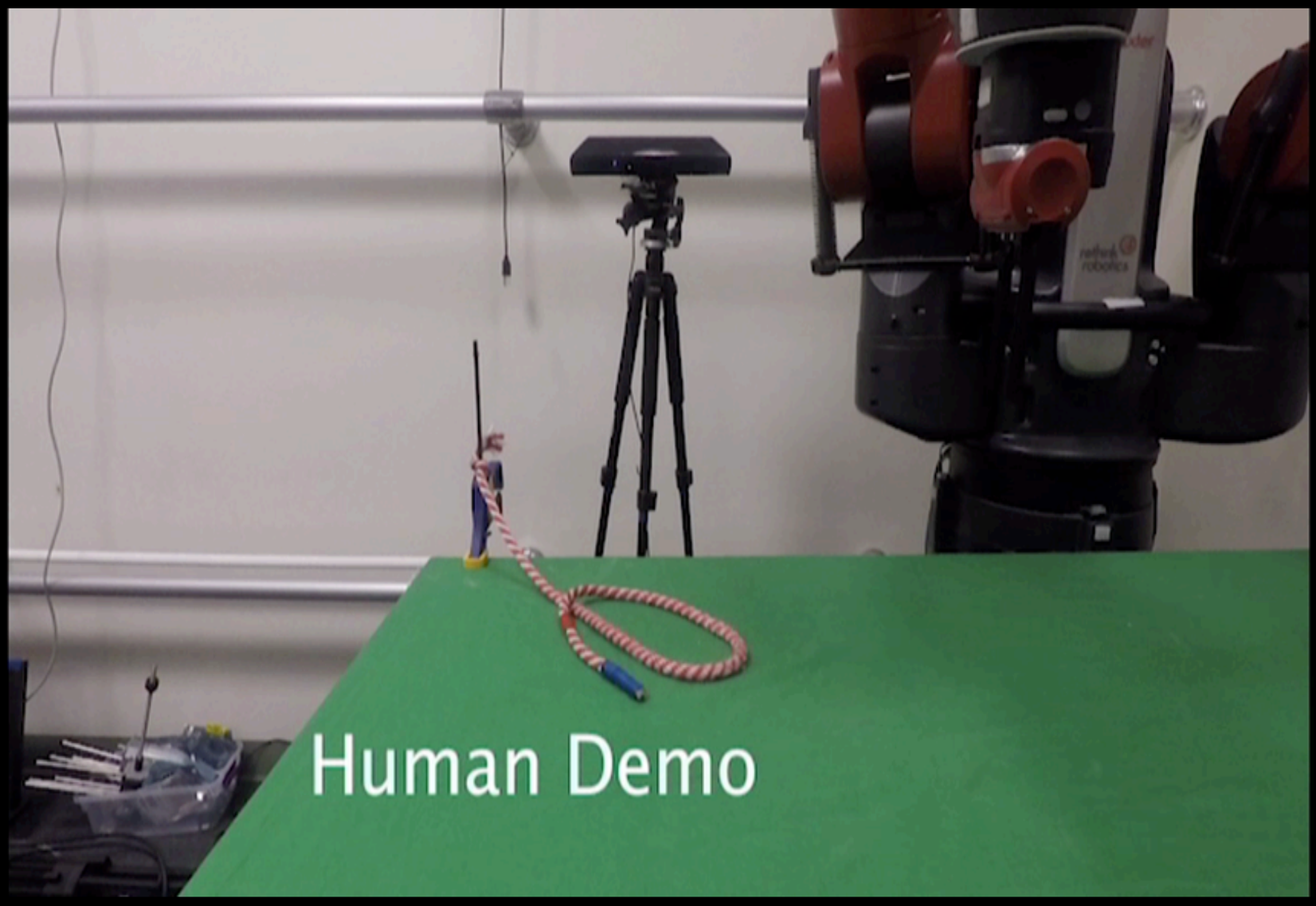
Model-guided Visual Imitation

- Use the visual demonstration to obtain subgoals (waypoints): intermediate states you need to reach

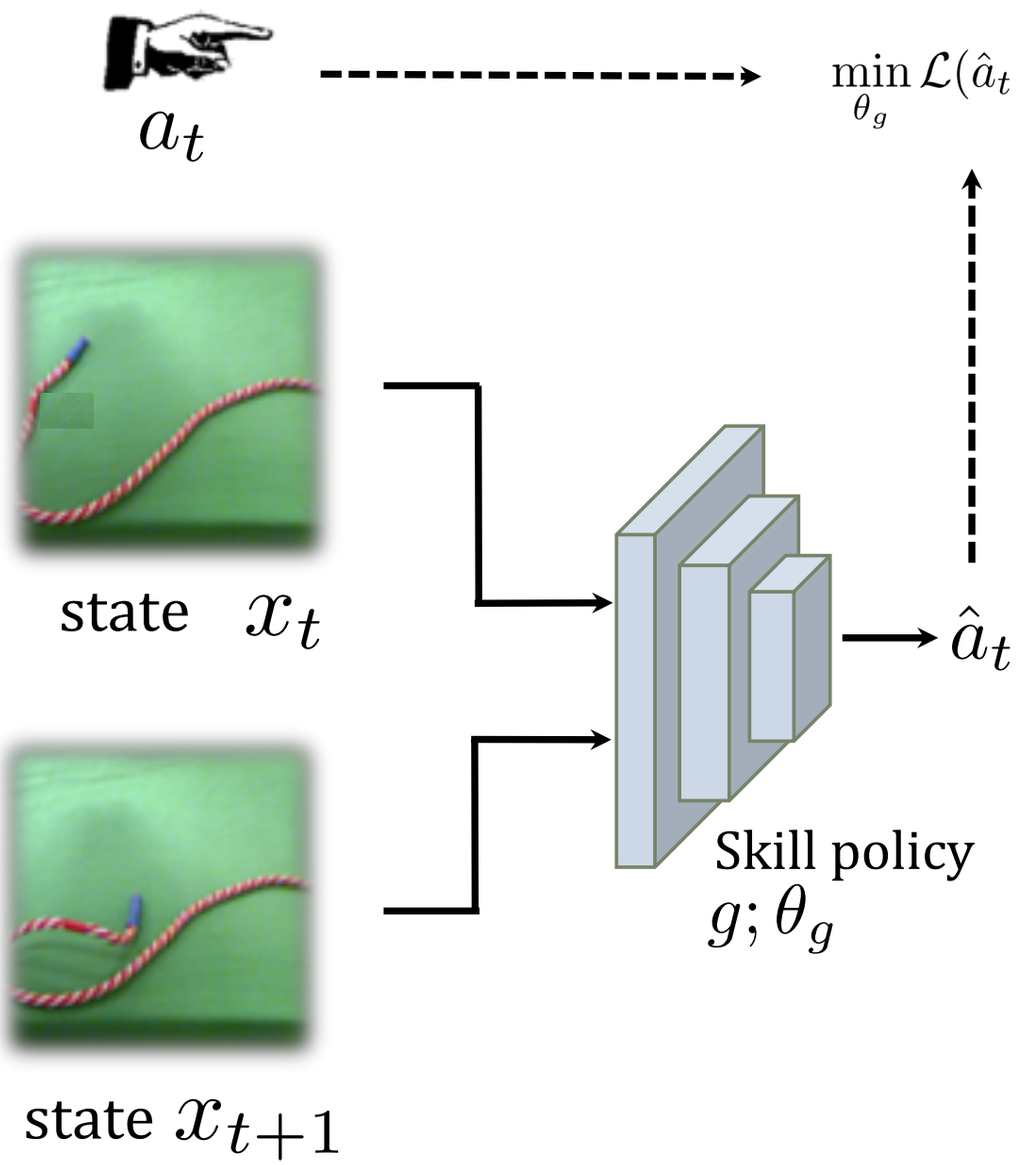
- Use the skills you have learns to reach those subgoals/waypoints



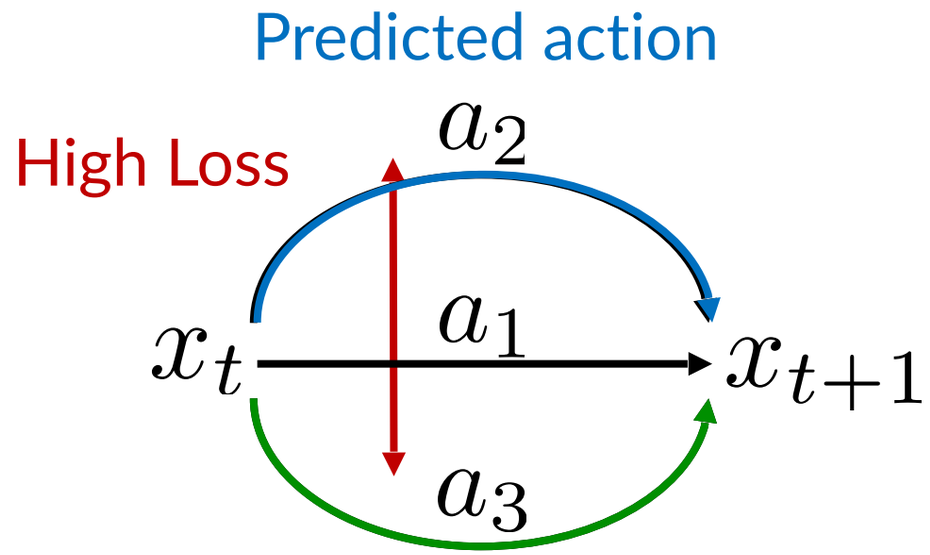
Knot-tying Rope Manipulation



Learning Inverse Models by regressing to actions



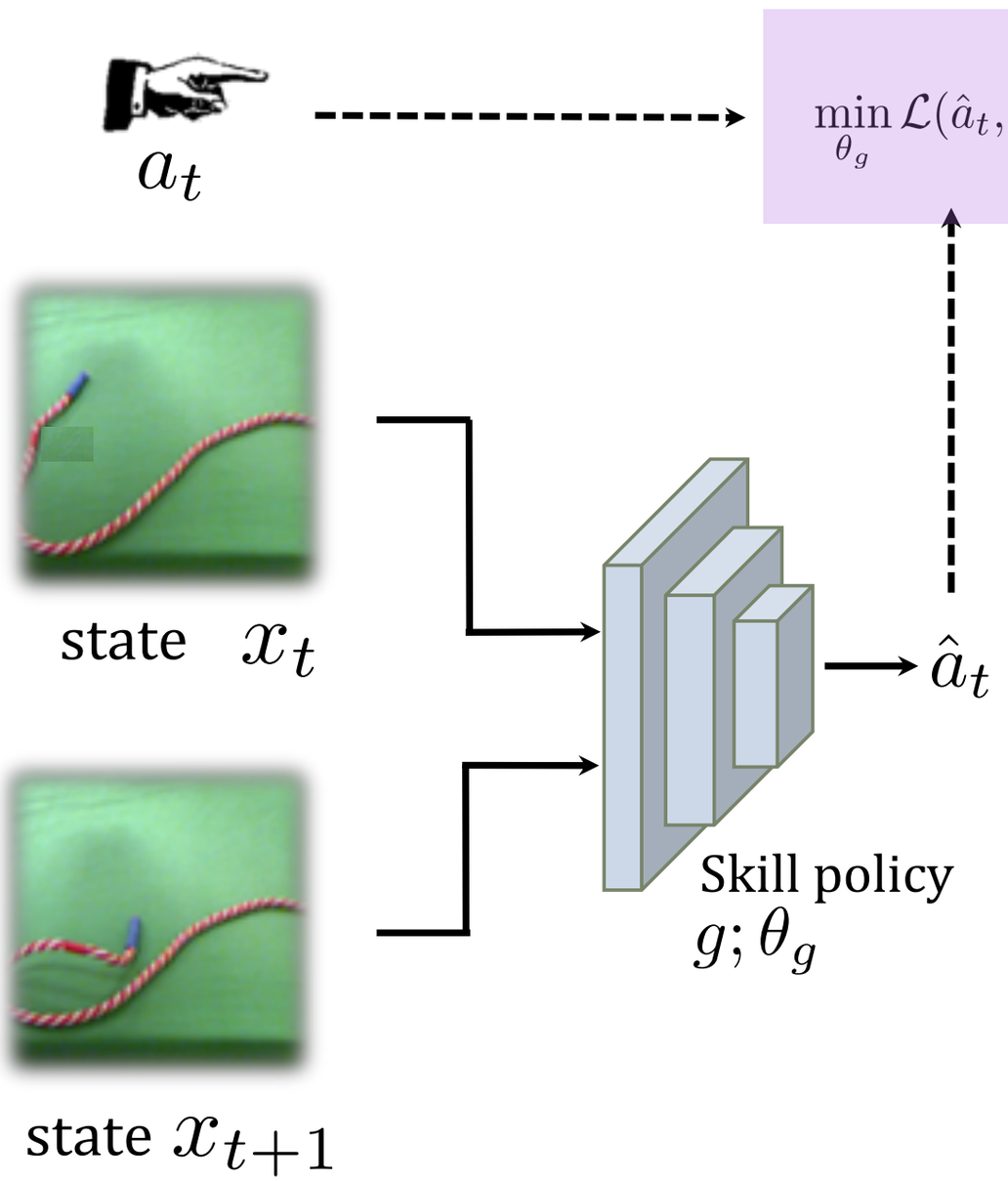
handle via VAEs or auto-regressive models?

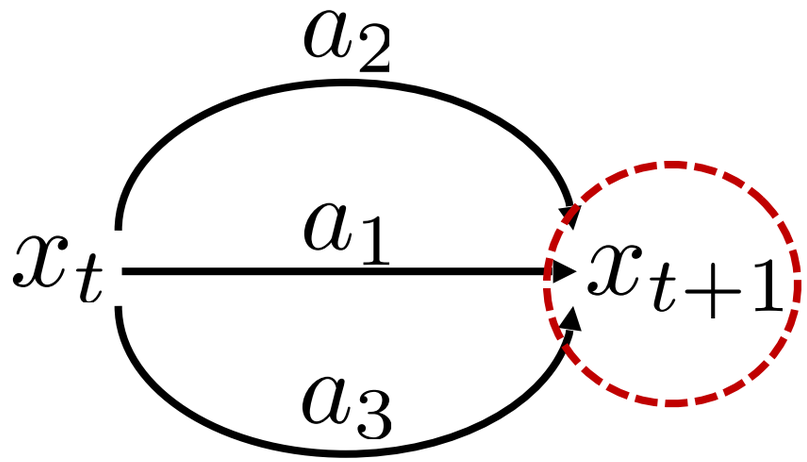


multi-modality in action space

Learning Inverse Models by penalizing resulting state

We do not care what actions to choose as long as it takes us to the right goal state-> let's penalize the resulting state as opposed to the action.





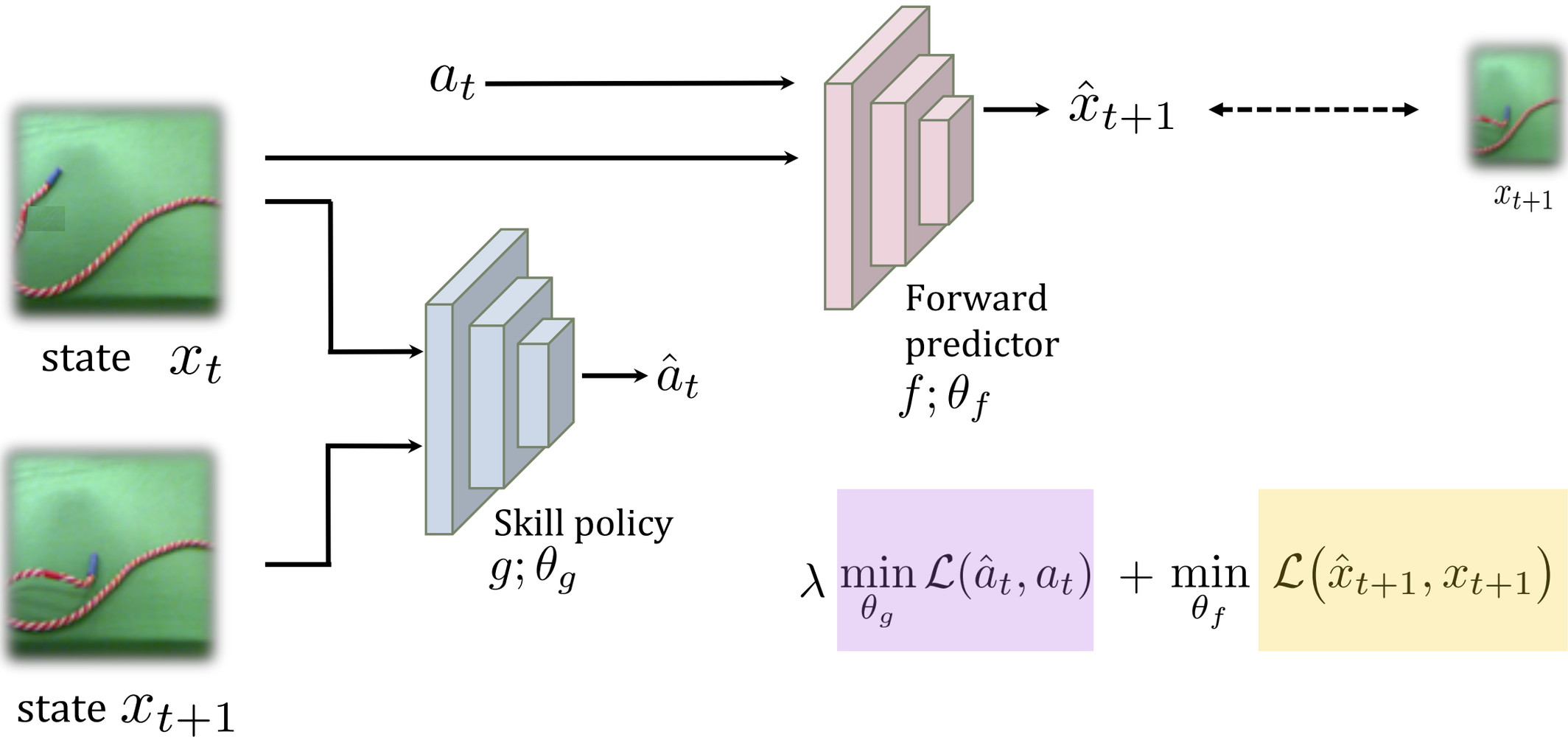
Penalize the "effect" of actions

How to operationalize it?

Learning Jointly Inverse and Forward Models

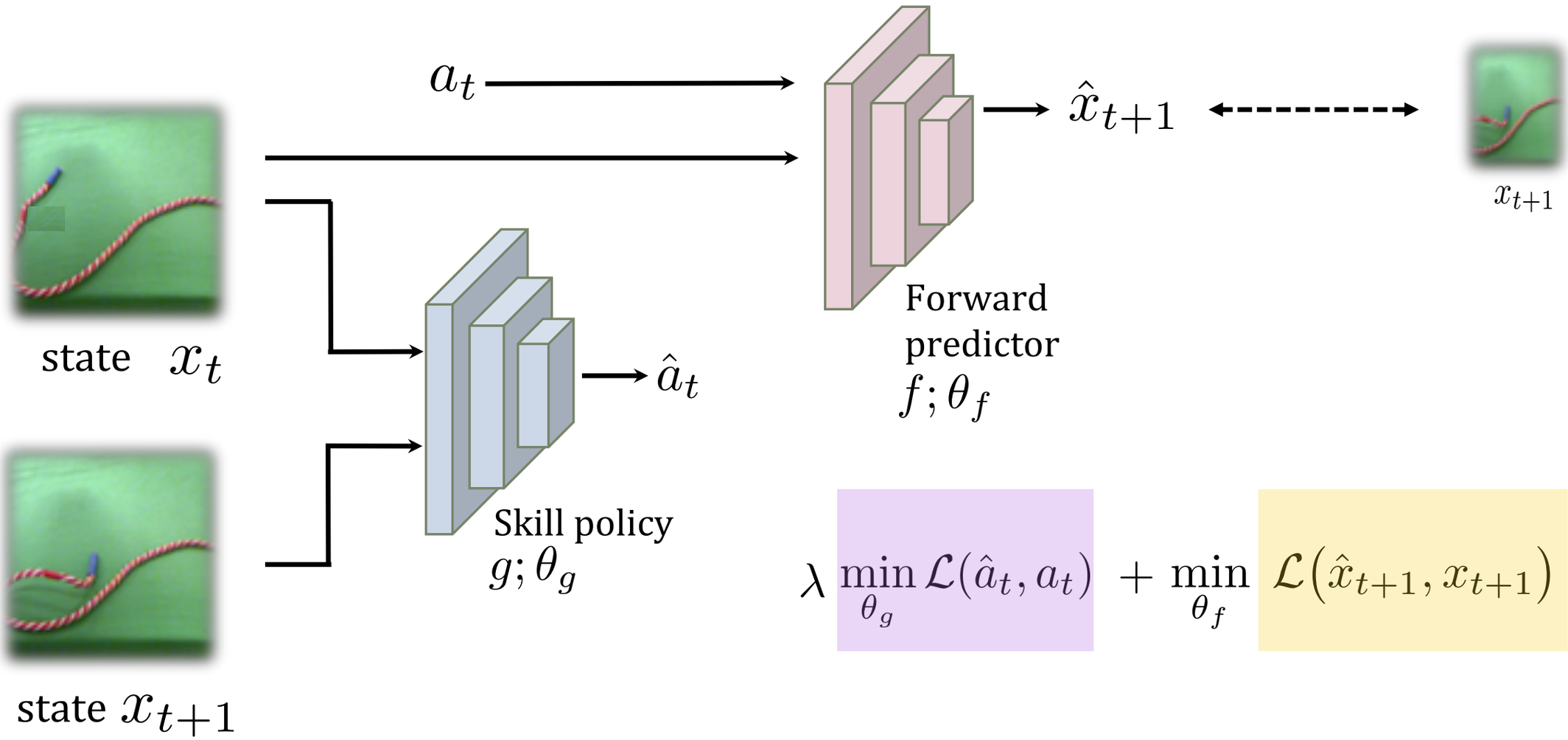
Train jointly a forward and an inverse model by regressing to the next state: next state will also be multimodal, but less multimodal than the action!

$$x_{t+1} \approx \hat{x}_{t+1} = f(x_t, a_t; \theta_f)$$



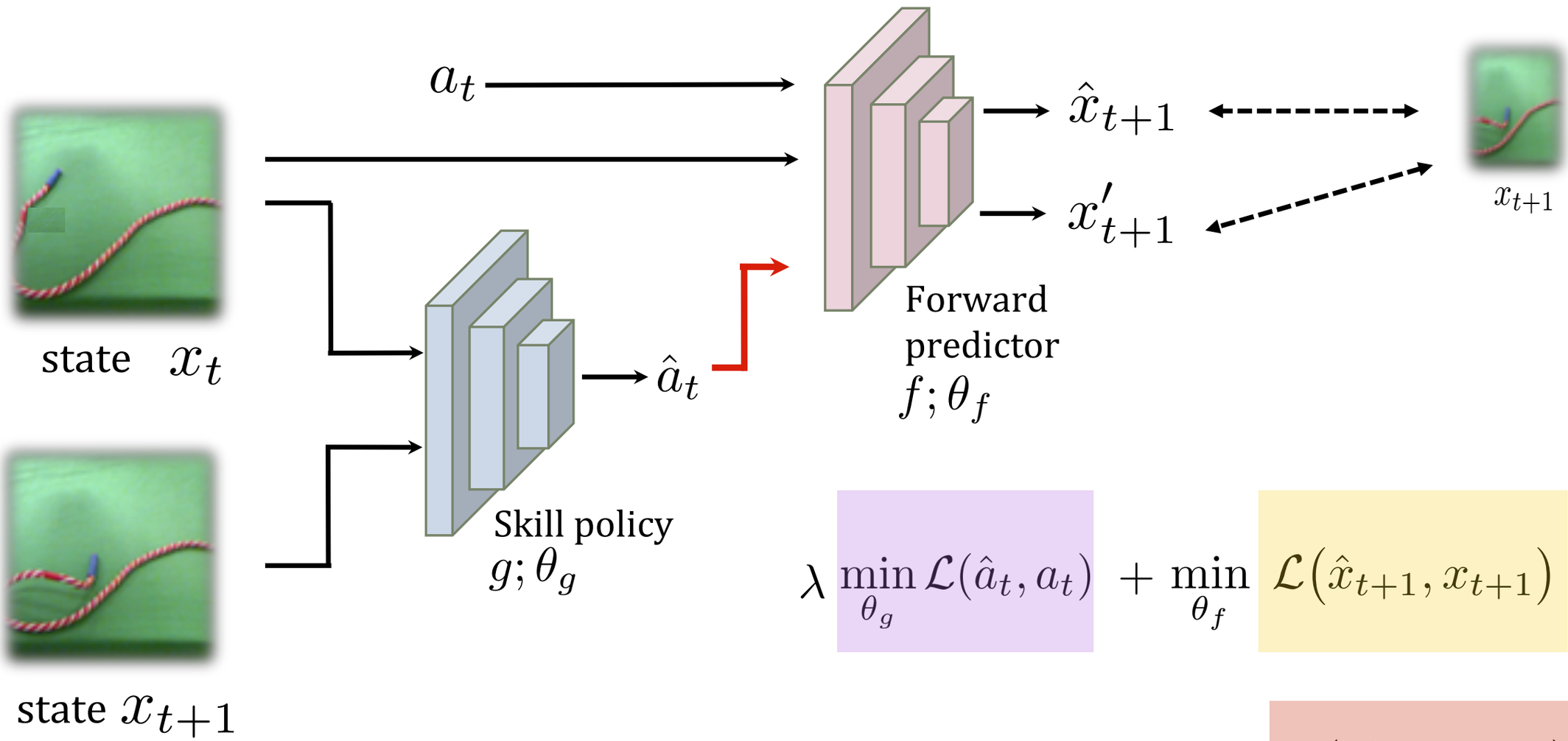
$$\lambda \min_{\theta_g} \mathcal{L}(\hat{a}_t, a_t) + \min_{\theta_f} \mathcal{L}\left(\hat{x}_{t+1}, x_{t+1}\right)$$

Handling Multimodality with Forward Consistency



$$\lambda \min_{\theta_g} \mathcal{L}(\hat{a}_t, a_t) + \min_{\theta_f} \mathcal{L}\left(\hat{x}_{t+1}, x_{t+1}\right)$$

Handling Multimodality with Forward Consistency

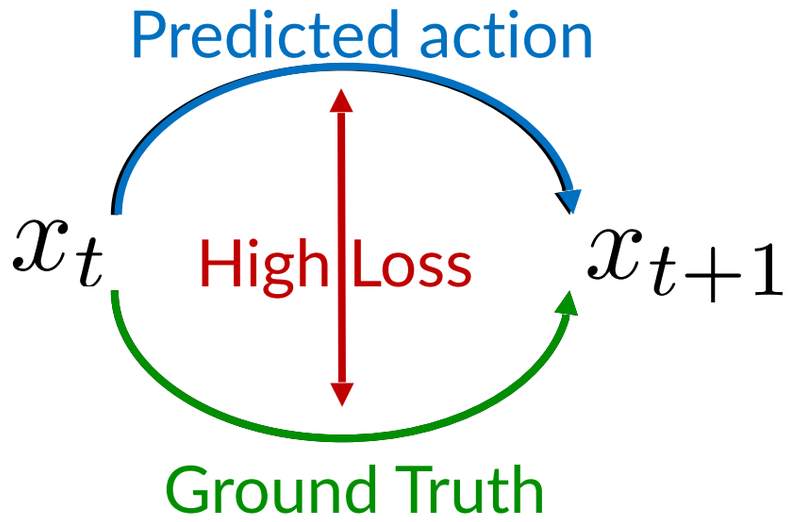


$$\lambda \min_{\theta_g} \mathcal{L}(\hat{a}_t, a_t) + \min_{\theta_f} \mathcal{L}\left(\hat{x}_{t+1}, x_{t+1}\right)$$

$$+ \min_{\theta_g} \mathcal{L}\left(x'_{t+1}, x_{t+1}\right)$$

In practice: Multi-step planning

multi-modality gets even worse with multi-step policy



Forward Consistency in Multi-step Planning