Kinematics-Guided Reinforcement Learning for Object-Aware 3D Ego-Pose Estimation

Zhengyi Luo*, Ryo hachiuma*, Ye Yuan, Shun Iwase, Kris M. Kitani
Carnegie Mellon University

Abstract

We propose a method for incorporating object interaction and human body dynamics into the task of 3D ego-pose estimation using a head-mounted camera. We use a kinematics model of the human body to represent the entire range of human motion, and a dynamics model of the body to interact with objects inside a physics simulator. By bringing together object modeling, kinematics modeling, and dynamics modeling in a reinforcement learning (RL) framework, we enable object-aware 3D ego-pose estimation. We devise several representational innovations through the design of the state and action space to incorporate 3D scene context and improve pose estimation quality. We also construct a fine-tuning step to correct the drift and refine the estimated human-object interaction. This is the first work to estimate a physically valid 3D full body interaction sequence with objects (*e.g.*, chairs, boxes, obstacles) from egocentric videos. Experiments with both controlled and in-the-wild settings show that our method can successfully extract an object-conditioned 3D ego-pose sequence that is consistent with the laws of physics.

1 Introduction

From a video captured by a single head-mounted wearable camera (*e.g.*, smartglasses, action camera, body camera), we want to infer the wearer's 3D pose and interaction with objects in the scene, as shown in Figure 1. This is crucial for applications such as virtual/augmented reality, sports analysis, medical monitoring, etc., where third-person views are often unavailable and high-quality estimates of complex and dynamic human motion are needed. However, this task is challenging since the wearer's body is often unseen from a first-person view and the body motion needs to be inferred solely based on the visual context captured by the front-facing video. Furthermore, modeling physically realistic human-object interactions requires not only estimating the kinematic motion of the wearer, but also modelling the physical dynamics–the objects need to react to the forces applied by the human action in a physically realistic way. In this paper, we show that it is possible to infer accurate human motion and human-object interaction from a single, forward-facing wearable camera.

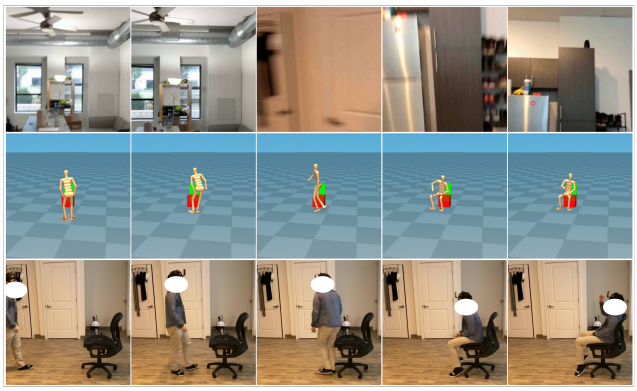

Figure 1: Given an in-the-wild egocentric video, our method can infer physically valid 3D human pose and human-object interaction. **Top**: input egocentric video. **Middle**: estimated 3D human pose and human-object interaction. **Bottom**: reference third person view.

First and foremost, we address the important issue of using either 1) kinematics or 2) dynamics-based human pose estimation approach. 1) Kinematics-based approaches directly output the joint angles of the human model based on videos and ignore physics (*e.g.,* how much force is needed at each joint to hold the pose). It can often achieve better pose estimates but produces results that may violate physical constraints (*e.g.,* joints bending the wrong way). Moreover, a kinematics-based model can not faithfully emulate human-object interactions–no physically realistic grasping, pushing, stepping, etc. can be performed without simulating physics. 2) Dynamics-based approaches represent methods that use a physics simulator and output joint torques to control the humanoid inside the simulator. The pose estimation is performed by extracting the simulated state of the humanoid. Thus, these approaches output physically realistic human poses and can lead to convincing human-object interaction (pushing an object will make it move accordingly). In this work, we argue that a hybrid approach is needed and propose using the output of the kinematics model as an additional signal to aid the training process and improve the performance of the dynamics model. Instead of training a Deep RL dynamics model to directly map from the visual context

to target joint angles like in prior works (Yuan and Kitani 2018, 2019), we propose to train an additional kinematics-based pose estimator and employ a novel action representation where the RL policy network is tasked to compute the residual pose against the output from a trained kinematics model. Such a formulation makes use of the accurate pose estimation from a kinematics model while using a dynamics model to refine the estimated pose to obey the laws of physics.

The second key missing piece is the scene context. Prior works (Yuan and Kitani 2018, 2019; Jiang and Grauman 2016) have long omitted the semantic scene context from the first-person view due to the complexity of modeling dynamically correct human-object interaction. However, the presence of objects in the first-person view can often provide a strong prior over the expected human behavior. For example, combining the presence of a chair and the motion of moving forward, turning around, and bending down, we can strongly infer the action of sitting down. Thus, it is imperative for learning models to draw upon this contextual information to make an educated guess about human motion and human-object interaction. To this end, we incorporate the 6 -degree-of freedom (DoF) pose of the main object of interest into the state representation of our RL model to make our model aware of the objects' states in the scene.

Another obstacle is the dynamics mismatch between the real world and simulation. As noted by prior work (Yuan and Kitani 2019), global position and orientation drifts can often be observed over a long horizon of simulation. This is not a significant issue if only the humanoid motion is considered, but such drift can lead to a catastrophic failure of human-object interaction; errors in global position may lead to completely missing the box to push or the chair to sit on. To make sure the correct human-object interaction can be simulated, we propose a fine-tuning step against video evidence to correct the drift in the root trajectories. We use a monocular camera tracking technique, such as Visual Inertial Odometry (VIO) (Wang et al. 2017; Engel, Sturm, and Cremers 2013), to extract the camera motion and fine-tune our learned policy to match against it. The fine-tuning step alleviates the mismatch between the estimated and real-world global trajectories, leading to a successful human-object interaction.

As there is no public available egocentric video dataset that contains the 3D full-body pose and the 3D object pose, we capture a large-scale motion capture (MoCap) dataset in which the person wears a head-mounted camera and interacts with various objects. We capture three types of interactions: 1)sitting on (and standing up from) a chair, 2)pushing a box, and 3)avoiding obstacles while walking. We also capture an in-the-wild dataset that contains the same set of interactions.

In summary, we tackle the challenging task of extrapolating 3D human motion and human-object interaction from egocentric videos. Our contributions are as follows: (1) We are the first to propose a DeepRL based method for physically valid 3D pose and human-object interaction estimation from egocentric videos. (2) We propose to use a hybrid of kinematics and dynamics approaches and employ a novel action representation in which the dynamics-based model outputs the residual of the action against a learned kinematics-based model. (3) We propose a fine-tuning step to reduce the drift and refine our estimation based on captured video evidence. (4) We experiment with a self-made large-scale MoCap dataset & an in-the-wild dataset, and show that our model outperforms other state-of-the-art methods on several pose-based and physics-based metrics, while generalizing to in-the-wild settings. Upon visual inspection, our framework can not only recover the 3D human motion from egocentric videos, but also simulate physically correct human-object interactions.

2 Related Works

This section is divided into two parts. First, we will discuss kinematics-based approaches for 3D human pose estimation from egocentric videos. Second, we will discuss recent advancements in dynamics-based humanoid control methods.

2.1 3D human pose estimation from egocentric videos

The task of estimating the 3D human pose from third-person videos is well researched in the computer vision community (Rogez, Weinzaepfel, and Schmid 2019; Pavllo et al. 2019; Habibie et al. 2019; Moon, Chang, and Lee 2019; Kolotouros et al. 2019; Kocabas, Athanasiou, and Black 2020; Luo, Golestaneh, and Kitani 2020). These methods are also all kinematics-based and often result in physically invalid motions such as foot sliding.

On the other hand, there are only a handful of attempts at estimating 3D full body poses from egocentric videos, due to the ill-posed nature of this task. Most existing methods still assume partial visibility of body parts in the image (Tome et al. 2019; Rhodin et al. 2016; Xu et al. 2019), often through a downward-facing camera. Among works where the human body is mostly not observable, (Jiang and Grauman 2016; Yuan and Kitani 2018, 2019; Ng et al. 2019), (Jiang and Grauman 2016) uses a kinematics-based approach where they construct a motion graph from the training motions and recover the pose sequence by solving the optimal pose path. (Ng et al. 2019) focuses on modeling person-to-person interactions from egocentric videos and infers the wearer's kinematic pose conditioning on the other person's pose. (Yuan and Kitani 2018, 2019; Isogawa et al. 2020), on the other hand, use dynamics-based approaches where a RL-based agent is tasked to perform physically valid human motions. In comparison, our work combines kinematics-based and dynamics-based approaches to achieve both accurate and physically valid pose estimation. In contrast to previous works, we also model human-object interactions such as sitting on a chair and pushing a box on the table. To the best of our knowledge, we are the first approach to estimate the 3D human poses from egocentric video while factoring in human-object interactions.

2.2 Humanoid control for object manipulation

Our work is also connected to controlling humanoids to interact with objects in a physics simulator (Peng et al.

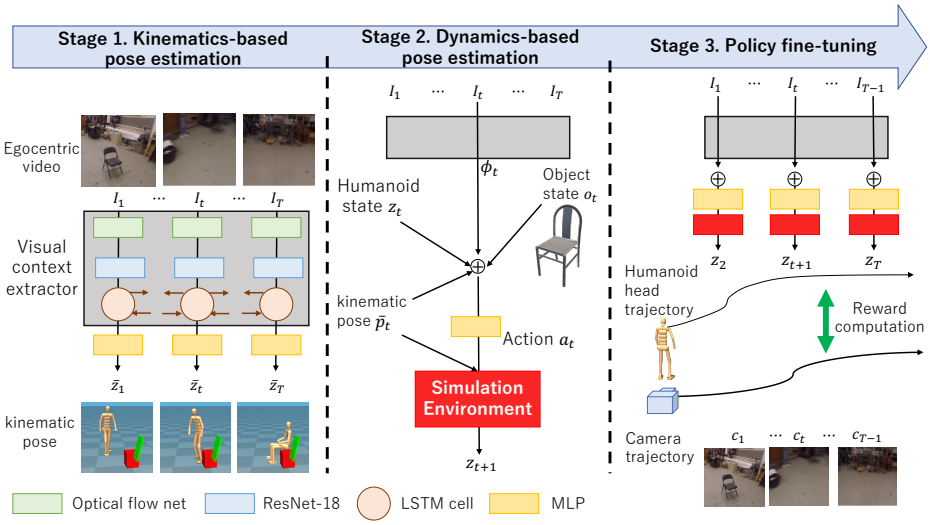

Figure 2: The overview of our proposed pipeline. At first, a pose regressor will estimate the kinematic pose from video evidence. Then an object-aware dynamics-based model will perform and correct the estimated kinematic pose in a physics simulator. Finally, a fine-tuning step corrects the root drifts and produces the final physically valid pose and human-object interaction.

2018a,b; Chao et al. 2019; Merel et al. 2019; Yuan and Kitani 2020). The core motivation of this line of work is to learn the necessary dynamics to imitate realistic human motion in a physics simulation.

(Chao et al. 2019) proposes a hierarchical reinforcement learning approach to generate realistic sitting motion. The authors manually decompose high-level actions (such as sitting) to low-level actions (such as walking, turning, and sitting) from which the proposed meta-controller can stochastically select. They do not aim to estimate the full body pose from egocentric videos, and their motion is limited to sitting. Our approach can not only replicate a diverse set of sitting motion, but also ground the sitting motion on video evidence. (Merel et al. 2019) proposes an approach for enabling humanoid full-body manipulation and locomotion in simulation. They use a *phased task* in which the task policy is trained to solve different stages of the task and show impressive results in human-object manipulation. Similarly, their generated motion is not grounded on video evidence. (Park et al. 2019) uses an action representation in which the target pose is the sum of the kinematic pose and the output of the policy network. Inspired by their work, we also employ this residual action representation to accelerate training and improve stability. Our task is also related to DeepMimic (Peng et al. 2018a) and its video variant (Peng et al. 2018b). DeepMimic has shown remarkable results in imitating human locomotion skills and is able to combine learned skills to achieve different tasks. However, human-object interaction and video grounding are not considered in their approach.

3 Method

The problem of 3D body pose estimation from egocentric videos can be formulated as follows: from a wearable camera footage $I_{1:T}$, we want to estimate the person's pose sequence $p_{1:T}$. In this paper, we propose a three-step method: 1) A kinematic pose is estimated by a pose regressor, followed by 2) object-aware pose correction using dynamics, and finally, 3) a fine-tuning step. The overview of our proposed method is depicted in Figure 2.

3.1 Pose Estimation using Kinematics

To recover the kinematic 3D human pose from egocentric videos, we train a regressor that predicts the pose $\bar{p}_{1:T}$ from the input video sequence $I_{1:T}$. Specifically, at first, a long -short-term-memory (LSTM) based visual context extractor is used to extract visual information from egocentric videos: $I_{1:T} \rightarrow \phi_{1:T}$. Then a multilayer perceptron (MLP) is used to produce the kinematic states from the visual context $\phi_{1:T} \rightarrow \hat{z}_{1:T}$. These states $\bar{z}_t$ consist of the human pose (position against the horizontal plane, orientation of the root, and the joint angles) and velocities (linear and angular velocities of the root and joint velocities). From the kinematic state, we can recover the full kinematic pose sequence $\bar{p}_{1:T}$ that consists of root position, root orientation, and joint angles. We employ mean squared error (MSE) as the loss function to train the regressor: $L(\xi) = \frac{1}{T}\sum_{t=1}^{T} \|\mathcal{F}(I_{1:T})_t - \hat{z}_t\|^2$, where ξ are the parameters of $\mathcal{F}$ and $\hat{z}_t$ is the ground truth kinematic state from MoCap. In general, this model does not consider the laws of physics like causal forces or actuation constraints, so the network is easier to train compared to dynamics-based models.

3.2 Object-Aware Pose Estimation using Dynamics

We formulate the task of estimating a physically-valid pose sequence $p_{1:T}$ from egocentric RGB images $I_{1:T}$ as a Markov Decision Process (MDP) defined as a tuple $\mathcal{M} =$

$\langle S, A, P, R, \gamma \rangle$ of states, actions, transition dynamics, reward function, and discount factor. The states S and the transition dynamics P are provided by the physics simulator, and the action A and the reward R are computed by the policy π. At each time step t, the agent in state s_t takes an action sampled from the policy $\pi(a_t|s_t)$ while the environment generates the next state s_{t+1} based on that action through physics simulation. Comparing the resulting state of the humanoid against the ground-truth, the agent will receive a reward r_t. This process repeats until some termination condition is triggered, such as when the time horizon is reached or the humanoid falls to the ground. We employ Proximal Policy Optimization (PPO) (Schulman et al. 2017) to calculate the optimal policy π^* that maximizes the expected discounted return $E[\sum_{t=1}^{T} \gamma^{t-1} r_t]$. At test time, we roll out the policy π^* to generate state sequence $s_{1:T}$ from which we extract the output pose sequences $p_{1:T}$.

To enable object-aware motion estimation from egocentric videos that abides by the laws of physics, we innovate on two key points upon prior dynamics-based models: (1) we factor in object pose into our RL agent's state representation, (2) we use the result from the pre-trained regressor from the previous step as an additional input to the RL model. Each of the MDP element is defined as follows:

State. The state s_t at the time step t consists of the humanoid state z_t, the visual context ϕ_t, the kinematic-pose state $\bar{q}_t$, and the object state o_t: $s_t =< z_t, \phi_t, o_t, \bar{q}_t >$. z_t consists of the humanoid pose q_t (position and orientation of the root joint, and joint angles) and velocity v_t (linear and angular velocities of the root, and joint velocities). $\bar{q}_t$ is the output of the kinematic pose regressor $\mathcal{F}$. Here, we factor in the 6DoF object position and orientation o_t as additional input to the control policy to enable object-aware 3D pose estimation.

Action. The action a_t specifies the target joint angles for the proportional-derivative (PD) controller controller (Tan, Liu, and Turk 2011) at each degree of freedom (DoF) of the humanoid joints except for the root (Hip). We use a novel residual action representation where:

$$q_t^d = \bar{q}_t + \Delta q_t^d, \tag{1}$$

q_t^d is the final PD target, Δq_t^d is the output (action a_t) of the control policy π, and $\bar{q}_t$ denotes the predicted pose of the kinematic pose regressor $\mathcal{F}$. For joint i, the torque to be applied is computed as $\kappa^i = k_p^i(q_t^d - p_t^i) - k_d^i v_t^i$ where k_p and k_d are manually-specified gains. Compared to directly estimating the target joint angles (Yuan and Kitani 2019), predicting the residual of target pose against the kinematic pose $\bar{q}_t$ offers a better starting point for the RL policy (if the kinematics-based regressor is perfect, the RL-based policy can output 0) and results in faster convergence and improved stability.

Policy. The policy $\pi_\theta(a_t|s_t) = \pi_\theta(a_t|z_t, \phi_t, o_t, \bar{q}_t)$ is represented by a gaussian distribution with a fixed diagonal covariance matrix Σ. We employ a MLP parametrized by θ as our policy network to map the state s_t to the mean μ_t of the distribution.

Reward function. The reward function is as follows:

$$r_t = w_p r_p + w_e r_e + w_{rv} r_{rv} + w_{rq} r_{rq} + w_{rp} r_{rp}, \tag{2}$$

where $w_p, w_e, w_{rv}, w_{rq}, w_{rp}$ are the weights of each reward. Our reward is similar to DeepMimic (Peng et al. 2018a), with the exception that we separate the root reward from the pose reward to better motivate the model to match the ground truth root trajectory. More importantly, unlike the conventional object manipulation control methods (Peng et al. 2018a; Merel et al. 2019; Peng et al. 2018b), we do not set any manually designed goal reward for each specific task and only use pose reward to match ground truth poses in order to handle multiple interactions. The pose reward r_p measures the difference between the generated pose q_t and the ground truth pose $\hat{q}_t$ in quaternion for each joint on the humanoid except for non-root joints. The end-effector reward r_e computes the distance between the estimated end-effector (feet, hands, head) position e_t and the ground truth position $\hat{e}_t$. The root velocity reward r_{rv} penalizes the deviation of the estimated root's linear l_t and angular ω_t velocity from the ground truth $\hat{l}_t$ & $\hat{\omega}_t$. The ground truth velocity is computed from the data via finite differences. The root position & orientation rewards r_{rp} & r_{rq} compute the difference between the generated 3D root position p_t & orientation q_t and the ground truth $\hat{p}_t$ & $\hat{q}_t$ in the world coordinate frame:

$$r_p = \exp\left[-5.0\left(\Sigma_j \|\bar{q}_t^j \ominus q_t^j\|^2\right)\right], \tag{3}$$

$$r_e = \exp\left[-4.5\left(\Sigma_e \|e_t - \hat{e}_t\|^2\right)\right], \tag{4}$$

$$r_{rv} = \exp\left[-\left\|l_t - \hat{l}_t\right\|^2 - 0.1\left\|\omega_t^r - \hat{\omega}_t^r\right\|^2\right], \tag{5}$$

$$r_{rq} = \exp\left[-40\left(\|q_t^r \ominus \hat{q}_t^r\|^2\right)\right], \tag{6}$$

$$r_{rp} = \exp\left[-45\left((p_t - \hat{p}_t)^2\right)\right]. \tag{7}$$

Initial state estimation. During training, we set the initial humanoid state z_1 and the object state o_1 to the ground truth $\hat{z}_1$, $\hat{o}_1$. At test time, the starting state of the humanoid and objects are given by the kinematic pose regressor and an off-the-shelf 6DoF object pose estimator, respectively.

3.3 Fine-tuning of the policy network

As mentioned previously, at test time, the output of the RL model will drift from the ground truth in terms of global position and orientation. causing human-object interaction to fail in the simulation. Moreover, as the training data cannot cover all of the object state (position & orientation), it is difficult for the policy to generalize against unseen states.

To overcome this problem, we propose to fine-tune the policy π_θ against test-time video evidence. While capturing the egocentric video, the 6DoF camera motion (position $\hat{h}_t^p$ and orientation $\hat{h}_t^q$) can be recovered using VIO techniques. Using the camera trajectory as an approximation to head motion, we can fine-tune the trained dynamics-based pose estimator to come up with physically valid pose estimates that conforms to the extracted camera trajectory from

the video. However, a naive fine-tuning step that only attempts to match humanoid motion with the tracked camera motion may lead to unnatural human poses since the policy may forget how to produce valid human poses. To overcome this issue, we introduce two types of regularization, 1) we regularize the pose to be similar to the kinematic pose, $\bar{p}_t$: the dynamics-based model must not deviate from the input kinematics result too drastically. 2) we regularize the action space by penalizing the difference of the output $\tilde{\mu}$ from the pre-trained policy network $\tilde{\pi}$ and μ from the fine-tuned policy π. In this way, our policy is forced to not deviate too much from its original behavior while trying to conform to the tracked camera trajectory.

The overall reward at the fine-tuning stage is as follows:

$$\hat{r}_t = w_{hp} r_{hp} + w_{hq} r_{hq} + w_{hv} r_{hv} + w_p \lambda_t r_p + w_a (1 - \lambda_t) r_a \tag{8}$$

where, $r_{hp}, r_{hq}, r_{hv}, r'_p, r_a$ are rewards for head position, head orientation, head linear and angular velocity, pose, and action, respectively. w_{hp}, w_{hq}, w_{hv}, w_p, w_a are the weighting factor. The head position r_{hp}, orientation r_{hq}, velocity r_{hv}, rewards are similar to their root counterparts r_{rp}, r_{rq}, and r_{rv}. They are computed similarly and here we use the extracted camera motion as an approximation to ground-truth head trajectory. The pose reward r'_p penalizes the difference of the generated pose p_t and the kinematic pose $\bar{p}_t$ for joints except for the root joint. The action reward penalizes the difference between the mean action of the pre-trained policy $\tilde{\mu}$ and the mean action of the current policy μ. Furthermore, we introduce an adaptive weighting factor λ_t on the action reward. The motivation is that if the kinematic regressor $\mathcal{F}$ generates unrealistic poses, the policy should avoid imitating that pose. To this end, we use a confidence value that puts more weight to the kinematic reward if the model has high confidence on the estimated kinematic pose (imitate estimated kinematic pose more) and leans toward the action reward if the kinematic pose is deemed unrealistic (stick to original action more). We calculate this value based on the head linear velocity in local coordinates: where $\hat{h}_t^{lv}$ denotes the linear velocity of camera and the $\hat{h}_t^{lv}$ denotes the linear velocity of the humanoid head in the local frame, respectively:

$$r_{hp} = \exp\left[-10.0\left(\|\hat{h}_t^p - h_t^p\|^2\right)\right], \tag{9}$$

$$r_{hq} = \exp\left[-10.0\left(\|\hat{h}_t^q \ominus h_t^q\|^2\right)\right], \tag{10}$$

$$r_{hv} = \exp\left[-0.1\left(\|\hat{h}_t^v - h_t^v\|^2\right)\right], \tag{11}$$

$$r'_p\left(\bar{q}_t, q_t\right) = \exp\left[-5.0\left(\Sigma_j \left\|\bar{q}_t^j \ominus q_t^j\right\|^2\right)\right], \tag{12}$$

$$r_a(\mu, \tilde{\mu}) = \exp\left[-1.0\left(\|\tilde{\mu} - \mu\|^2\right)\right], \tag{13}$$

$$\lambda_t = \exp\left[-0.1\left(\|\hat{h}_t^{lv} - h_t^{lv}\|^2\right)\right]. \tag{14}$$

4 Experimental Setup

4.1 Dataset

As there is no public dataset available containing the ground-truth full-body human pose and the object pose annotations where the person interacts with objects, we record a large-scale egocentric video dataset inside a Mocap studio. It includes three subjects and each subject is asked to wear a head-mounted camera and performs various complex human-object interactions for multiple takes. The actions consist of sitting on (standing up from) a chair, avoiding obstacles, and pushing a box. There are four categories of objects: chairs, tables, boxes, and obstacles. MoCap markers are attached to the camera wearer and the objects to get the 3D full-body human pose and 6DoF object pose. Each take is about six seconds long; 16 to 24 sequences exist for each combination of action and subject. As a result, our MoCap dataset consists of about 250 sequences. We use an 80–20 train-test data split on this MoCap dataset. To further showcase the generalization of our method, we also collect an in-the-wild dataset where an additional subject is tasked to perform similar tasks in an everyday setting wearing a head-mounted iPhone. Since monocular camera tracking and 6DoF object pose estimation are not the main focuses of this work, we use Apple's ARKit to provide camera position & orientation tracking as well as 6 DoF object pose estimation. The in-the-wild dataset has 15 takes in total each lasting about 6 seconds. As it is difficult to capture the ground truth 3D pose in-the-wild, we use a third-person camera to capture a synchronized side-view of the person and use an off-the-shelf 3D human pose estimator (Luo, Golestaneh, and Kitani 2020) to estimate 3D pose as pseudo-ground truth for evaluation.

4.2 Evaluation metrics

To quantitatively evaluate for 3D pose accuracy and its physical correctness, we employ the following metrics:

Root error ($\mathbf{E}_{root}$): a pose-based metric that measures the difference between the ground truth and generated root pose, both represented as the 4×4 transformation matrix $M_t(q_t^r|g_t)$ composed of the root orientation q^r and translation g. The error is calculated using the Frobenious norm: $\frac{1}{T}\sum_{t=1}^{T} \|I - (M_t \hat{M}_t^{-1})\|_F$ where I is the identity matrix.

Pose Error ($\mathbf{E}_{joint}$): a pose-based metric that measures the difference between the estimated pose sequence $p_{1:T}$ and the ground truth $\hat{y}_{1:T}$, both represented as euler angles in radian: $\frac{1}{T}\sum_{t=1}^{T} \|p_t - \hat{p}_t\|_2$.. This metric does not consider root position & orientation, and only compares the joint angles.

Mean Per Joint Position Error ($\mathbf{E}_{mpjpe}$): a pose-based metric used for our in-the-wild evaluation. As off-the-shelf pose estimator (Luo, Golestaneh, and Kitani 2020) from third-person view uses a different human model (Loper et al. 2015) than ours, we cannot directly compare the joints' angles. Thus, we find the common joints (15 in total) between the two human models and directly compare their 3D positions in global coordinate frame. Denote pseudo-ground truth joint positions and the estimated joint positions as $J_t, \hat{J}_t \in R^{15\times 3}$. The metric is calculated as $\frac{1}{T}\sum_{t=1}^{T} \|J_t -$

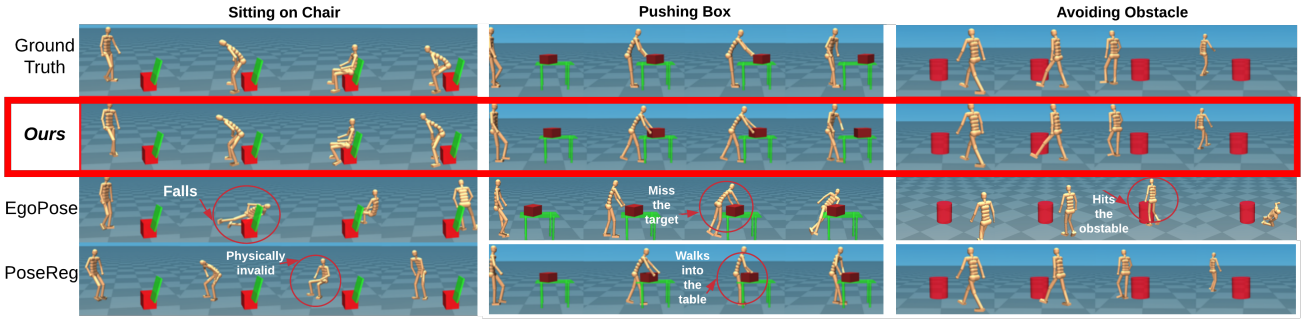

Figure 3: Results of 3D pose and human-object interaction estimation from egocentric videos.

$\hat{J}_t\|_2$, measured in millimeters. This metric factors in root position & orientation as well as the joint angles.

Velocity Error ($\mathbf{E}_{vel}$): a physics-based metric that measures the difference between the estimated joint velocity $v_{1:T}$ and the ground truth $\hat{v}_{1:T}$. It is calculated as $\frac{1}{T}\sum_{t=1}^{T}\|v_t - \hat{v}_t\|_2$. where v_t and $\hat{v}_t$ are computed by the finite difference. When ground-truth joint angles are available, this metric is calculated as angular velocity in radians; for in-the-wild evaluation, it is measured in linear velocity in millimeters/second.

Average Acceleration ($\mathbf{A}_{accel}$): a physics-based metric that uses the average magnitude of joint accelerations to measure the smoothness of the predicted pose sequence, calculated as $\frac{1}{T}\sum_{t=1}^{T}\|\dot{v}_t\|_1$ where $\dot{v}_t$ denotes the joint accelerations. Similar to velocity error, for experiments with ground-truth pose available, this metric is calculated in angular acceleration in radians; for in-the-wild evaluation, it is measured in linear acceleration in millimeters/second^2.

4.3 Baseline methods

To show the effectiveness of our framework, we compare against two baseline methods: (1) the previous state-of-the-art method in this task, **EgoPose** (Yuan and Kitani 2019), a dynamics-based approach that produces physically realistic motion but does not factor in object states, and (2) the previous best kinematics-based approach, also proposed in (Yuan and Kitani 2019), which we will call **PoseReg**.

More details about our network and the physics simulator (Mujoco) are provided in the supplementary material.

5 Results

Subject-Specific Evaluation. In this evaluation, we train an individual model for each subject, and evaluate on the test split of our MoCap dataset. From the quantitative results in Figure 3, we can clearly see that PoseReg can produce accurate pose estimation, but does not obey the laws of physics (such as sitting in mid-air, walks into the table). EgoPose, on the other hand, often fails to perform the action correctly (falls down or hits the obstacle). Overall, our method (second row) produces 3D human poses closer to the ground truth (top-row) than any other baselines, and successfully performs human-object interaction. To better visualize the quantitative results, please refer to our supplementary video.

Table 1 shows the quantitative comparison of our method against the two baseline methods. All of the results are the average across all three subjects. For all of the actions, we observe that our method outperforms two baselines across almost all actions and metrics, with occasional worse performance in joint angle estimation against kinematics-based method. This is expected as PoseReg disregards physics and can estimate poses without constraint, while our method requires estimating a physically valid pose. We find that the humanoid controlled by EgoPose often falls down to the ground and collide with the object, resulting in high pose error. This is expected as EgoPose does not factor in object states into its pose estimation, and suffers from error accumulation in global root positions. On the other hand, our object-aware state representation and our fine-tuning step ensure a correct human-object interaction, preventing falls and drifts. We can also observe the low velocity error and acceleration compared to the baseline method, which indicates that our residual action representation produces more stable and smoother pose estimation.

Cross-Subject Evaluation. To further test the robustness, we perform cross-subject experiments where we train our models on two subjects and test on the remaining subject. This is a challenging setting, since people have very unique styles and speeds for the same types of interactions. The quantitative results are summarized in Table 2. Our method again outperforms other baseline methods in almost all metrics. Especially for the smoothness of the pose ($\mathbf{A}_{accel}$), our method estimates much smoother (2.0x) pose sequences than those generated from other baseline methods.

In-the-wild Cross-subject Evaluation. To demonstrate the generalization of our method to real-world use cases, we further test our method on an in-the-wild dataset. Since there is no ground-truth 3D pose available, we evaluate against pesudo-ground truth 3D pose extracted from a third-person view camera. As shown in Table 3, our method out-performs the baseline methods by a large margin, especially on avoiding and pushing actions where root drifts are prominent. Notice that this dataset is captured in a real-world setting using

Table 1: ***Single-subject*** quantitative results for pose-based and physics-based metrics per action

	Sitting				Avoiding				Pushing			
	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$
PoseReg	1.151	**0.915**	6.431	12.609	0.682	**0.638**	6.152	12.126	0.909	0.825	6.621	13.058
EgoPose	1.444	1.3445	6.634	10.181	1.293	1.061	7.573	11.945	1.362	1.219	6.704	9.459
Ours	**0.607**	1.011	**5.084**	**5.489**	**0.372**	0.723	**5.889**	**7.081**	**0.377**	**0.820**	**5.257**	**5.986**

Table 2: ***Cross-subject*** quantitative results for pose-based and physics-based metrics per action

	Sitting				Avoiding				Pushing			
	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$
PoseReg	1.403	**1.775**	7.870	11.355	1.326	**1.469**	7.976	10.768	0.812	1.643	7.260	11.355
EgoPose	1.722	1.887	7.986	13.943	1.514	1.749	9.685	13.942	1.708	1.934	7.870	11.042
Ours	**0.756**	1.850	**6.085**	**5.880**	**1.112**	1.476	**7.368**	**5.446**	**0.449**	**1.631**	**6.407**	**5.605**

Table 3: ***In the wild &Cross-subject*** quantitative results, evaluated against pseudo-ground truth poses from a third person view. Notice that the unit (millimeters) in this table is different from the previous tables where ground-truth annotation is available.

	Sitting			Pushing			Avoiding		
	$\mathbf{E}_{mpjpe} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$	$\mathbf{E}_{mpjpe} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{E}_{accel} \downarrow$	$\mathbf{E}_{mpjpe} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$
PoseReg	453.40	26.91	26.48	576.41	23.29	13.55	1680.76	35.49	13.28
Egopose	551.91	28.5	20.47	518.19	26.48	22.73	1540.17	38.25	21.02
Ours	**313.76**	**18.50**	**4.79**	**248.85**	**16.65**	**4.17**	**440.08**	**23.12**	**5.69**

Table 4: Ablation study for the action representation, the fine-tuning step, and fine-tuning reward design.

Action representation	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$
(a) PD target	1.753	2.049	7.912	12.502
(b) Kinematic residual	**1.210**	**1.064**	**6.051**	**8.529**

Reward type	$\mathbf{E}_{root} \downarrow$	$\mathbf{E}_{joint} \downarrow$	$\mathbf{E}_{vel} \downarrow$	$\mathbf{A}_{accel} \downarrow$
(b) No fine-tuning	1.210	1.064	6.051	8.529
(c) head	0.366	0.920	5.227	6.477
(d) head, kinematic regularization	0.374	0.810	5.090	**6.291**
(e) head, action regularization.	0.361	0.882	5.136	6.436
(f) head, action + kinematic regularization.	0.340	0.805	5.074	6.351
(g) Full reward (Ours)	**0.322**	**0.788**	**5.063**	6.380

a different camera (iPhone) than used in the MoCap studio (GoPro), and our method is able to generalize and infer valid human poses and human-object interactions.

Quantitative Evaluation As motion is best seen in video, we refer readers to our supplementary video for a comprehensive quantitative analysis.

Ablation Study. To evaluate the importance of (1) our novel action representation and (2) our fine-tuning step and its reward design, we conduct an ablation study to quantify their benefits. This study is conducted for the first subject's sitting action. The results of the ablation study is presented in Table 4. To investigate the importance of our action representation (b), we compare the representation employed by (Yuan and Kitani 2019) (a) in which the policy outputs the target joint of PD control directly (instead of a residual of kinematics model). The prediction accuracy is improved greatly in all of the metrics (31%, 48%, 23%, and 31%, respectively) after using our residual action representation.

To investigate the importance of the fine-tuning step and each of its reward term, we train the comparison models with four types of rewards, (c) head (position, orientation and velocity) reward, (d) head and kinematic regularization reward, (e) head and action regularization reward, (f) head, and kinematic and action regularization reward. Model (g) has the additional adaptive weighting factor λ_t in addition to the reward (f). From Table 4, without the fine-tuning step, the model performs the worst (b). Our reward (g) outperforms all of the partial rewards (c-f) in the metrics except for the acceleration metric. Combining each reward term results in the best performance model across all spectrum, showcasing the benefit of our fine-tuning step and its respective rewards.

Failure cases and limitations Though our method can produce realistic human pose and human-object interaction estimation from only egocentric videos, we are still at the early stage of this challenging task. Our method performs well in the MoCap studio setting and out-performs state-of-the-art methods in-the-wild, but can still fail to produce natural pose-estimation for some in-the-wild videos. Due to the domain shift of the in-the-wild data, the result of the kinematics-base pose regressor is poor and can lead to unnatural motion such as severe foot sliding. Sometimes our dynamics-based RL agent is unable to correct for such sliding, and at the fine-tuning stage is forced to produce unnatural poses (e.g long steps, awkward turns) to prevent the humanoid from falling. Further, our method is still limited to pre-defined set of interactions where we have ample data to learn from. To enable pose and human-object estimation with arbitrary objects, much further investigation is needed.

6 Conclusion

In this paper, we tackle, for the first time, physically-valid 3D pose estimation from egocentric video while the person is interacting with objects. We collect a large scale motion capture dataset to develop and evaluate our method, and extensive experiments have shown that our method outperforms all prior methods. Real-world experiments using

an in-the-wild dataset further shows that our method generalizes well to real-world use-cases and can infer physically valid 3D human motion and human-object interaction from only a front-facing camera feed.

References

Chao, Y.-W.; Yang, J.; Chen, W.; and Deng, J. 2019. Learning to Sit: Synthesizing Human-Chair Interactions via Hierarchical Control. *ArXiv* abs/1908.07423.

Engel, J.; Sturm, J.; and Cremers, D. 2013. Semi-dense Visual Odometry for a Monocular Camera. In *IEEE International Conference on Computer Vision (ICCV)*, 1449–1456. Los Alamitos, CA, USA: IEEE Computer Society. ISSN 1550-5499.

Habibie, I.; Xu, W.; Mehta, D.; Pons-Moll, G.; and Theobalt, C. 2019. In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10897–10906. ISSN 1063-6919.

Isogawa, M.; Yuan, Y.; O'Toole, M.; and Kitani, K. M. 2020. Optical Non-Line-of-Sight Physics-based 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7013–7022.

Jiang, H.; and Grauman, K. 2016. Seeing Invisible Poses: Estimating 3D Body Pose from Egocentric Video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3501–3509.

Kocabas, M.; Athanasiou, N.; and Black, M. J. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 5252–5262.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 2252–2261.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34: 248:1–248:16.

Luo, Z.; Golestaneh, S.; and Kitani, K. M. 2020. 3D Human Motion Estimation via Motion Compression and Refinement. *ArXiv* abs/2008.03789.

Merel, J.; Tunyasuvunakool, S.; Ahuja, A.; Tassa, Y.; Hasenclever, L.; Pham, V.; Erez, T.; Wayne, G.; and Heess, N. 2019. Reusable neural skill embeddings for vision-guided whole body movement and object manipulation .

Moon, G.; Chang, J.; and Lee, K. M. 2019. Camera Distance-aware Top-down Approach for 3D Multi-person Pose Estimation from a Single RGB Image. In *IEEE Conference on International Conference on Computer Vision (ICCV)*, 10113–10142.

Ng, E.; Xiang, D.; Joo, H.; and Grauman, K. 2019. You2Me: Inferring Body Pose in Egocentric Video via First and Second Person Interactions. *CoRR* abs/1904.09882. URL http://arxiv.org/abs/1904.09882.

Park, S.; Ryu, H.; Lee, S.; Lee, S.; and Lee, J. 2019. Learning Predict-and-Simulate Policies from Unorganized Human Motion Data. *ACM Trans. Graph.* 38(6). ISSN 0730-0301.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Peng, X. B.; Abbeel, P.; Levine, S.; and van de Panne, M. 2018a. DeepMimic: Example-guided Deep Reinforcement Learning of Physics-based Character Skills. *ACM Trans. Graph.* 37(4): 143:1–143:14. ISSN 0730-0301.

Peng, X. B.; Kanazawa, A.; Malik, J.; Abbeel, P.; and Levine, S. 2018b. SFV: Reinforcement Learning of Physical Skills from Videos. *ACM Trans. Graph.* 37(6).

Rhodin, H.; Richardt, C.; Casas, D.; Insafutdinov, E.; Shafiei, M.; Seidel, H.-P.; Schiele, B.; and Theobalt, C. 2016. EgoCap: Egocentric Marker-Less Motion Capture with Two Fisheye Cameras. *ACM Trans. Graph.* 35(6). ISSN 0730-0301.

Rogez, G.; Weinzaepfel, P.; and Schmid, C. 2019. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms .

Tan, J.; Liu, K.; and Turk, G. 2011. Stable Proportional-Derivative Controllers. *IEEE Computer Graphics and Applications* 31(4): 34–44. ISSN 1558-1756.

Tome, D.; Peluse, P.; Agapito, L.; and Badino, H. 2019. xR-EgoPose: Egocentric 3D Human Pose from an HMD Camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 7728–7738.

Wang, S.; Clark, R.; Wen, H.; and Trigoni, N. 2017. DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2043–2050. ISSN null.

Xu, W.; Chatterjee, A.; Zollhoefer, M.; Rhodin, H.; Fua, P.; Seidel, H.-P.; and Theobalt, C. 2019. Mo^2Cap^2 : Real-time Mobile 3D Motion Capture with a Cap-mounted Fisheye Camera. *IEEE Transactions on Visualization and Computer Graphics* 1–1. ISSN 1077-2626.

Yuan, Y.; and Kitani, K. 2018. 3D Ego-Pose Estimation via Imitation Learning. In *The European Conference on Computer Vision (ECCV)*.

Yuan, Y.; and Kitani, K. 2019. Ego-Pose Estimation and Forecasting as Real-Time PD Control. In *IEEE International Conference on Computer Vision (ICCV)*, 10082–10092.

Yuan, Y.; and Kitani, K. 2020. Residual Force Control for Agile Human Behavior Imitation and Extended Motion Synthesis. In *Advances in Neural Information Processing Systems*.