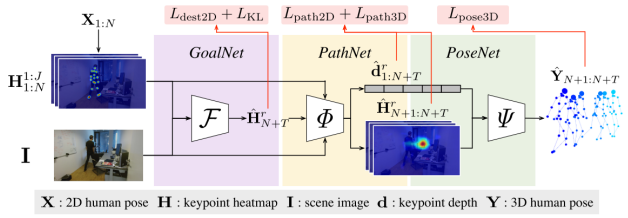

What: Predict future 3D human poses and trajectories using scene context.

Why: Past methods cannot predict long-term motions because they don't consider scene contexts.

Human motion prediction has many application areas such as autonomous driving, living assistance, or robotics. Because humans move towards some goal in the environment, it is necessary to take the surrounding scene into account for long-term motion prediction. No past method, however, exploit scene contexts for human motion prediction. They can only predict short future motions (within 1 sec.), or spatially local trajectories.

How: Learn three networks: 1. *GoalNet* generates multiple possible goals in the scene, 2. *PathNet* estimates 3D paths to each goal, and 3. *PoseNet* predicts future 3D poses in each frame (Fig. 2).

Fig. 3 shows the architecture. The proposed model consists of these three networks stacked on top of each other. The model receives past 2D poses, trajectories, and the scene image as input, and predicts future 3D poses and trajectories. The three networks are separately trained.

GoalNet is a CVAE that receives past 2D poses and trajectories as input, and the scene image as a condition. This network generates a 2D heatmap that points to a possible goal in the scene image. The network is trained with two losses: L1 norm between true and predicted goal positions (L_{dest2D}), and KL-divergence to make the latent vector follow Gaussian distribution (L_{KL}). *GoalNet* enables stochastic predictions, but it can be removed if we want deterministic predictions.

PathNet predicts 3D paths to the goals from the scene image, past 2D poses, and the output of *GoalNet*. Instead of directory predicting 3D coordinates, *PathNet* separately predicts 2D heatmap in the x-y coordinate and depth of the path (Fig. 3 middle). The network is trained with the L1 loss between true and predicted 2D path heatmaps (L_{path2D}) and a smoothness regularization term in the 3D coordinate (L_{path3D}).

PoseNet predicts future 3D poses and trajectories from the past 2D poses, trajectories, depth images, and the output of *PathNet*. Firstly, the past 3D poses were obtained by lifting 2D poses with depth and camera intrinsics. The future 3D poses are initialized with the current 3D poses. The past and initial future 3D poses are fed into a transformer-based network, and the network refines them to estimate future 3D poses. This network is trained with an L1 loss between true and predicted future 3D poses (L_{pose3D}).

Created a simulated GTA indoor motion (GTA-IM) dataset, in which we randomly set goal destination in the scene and move game characters to the goal. To increase the diversity of the data, we randomized characters' walking styles, tasks, scene types, lightning, and camera poses. In total, we collected one million RGB-D images and ground-truth 3D poses.

Results: Outperforms past methods; Path prediction is essential; Pretraining on synthetic data is important; Stochastic prediction is effective for long-term prediction.

Evaluated on the GTA-IM and the PROX dataset. PROX contains RGBD images of real humans moving in 12 different scenes. Mean per joint position error (MPJPE) is used for evaluating 3D pose and 3D path prediction. The future 2-seconds 3D poses were estimated from past 1-second 2D poses.

Compared with three baseline models that do not consider scene contexts. "TR" is a transformer that directly predicts future 3D poses from past 2D poses. "LTD" is an RNN that predicts future 3D poses from past 3D poses. "VP" is a 2D-to-3D pose estimation method used to convert past 2D poses to 3D.

Tab. 1 and 2 shows the results with the GTA-IM and PROX datasets, respectively. Both deterministic (w/o *GoalNet*) and stochastic (w/ *GoalNet*) model was evaluated. For the stochastic model, the error of the sample that is closest to ground-truth was reported. The proposed model outperforms past methods. Ablation studies show that each of the path prediction module, heatmap representation for the *PathNet*, scene images, and pretraining on GTA-IM contributes to accuracy.

Fig. 5 shows qualitative results. While the proposed model can predict physically plausible human trajectories, baseline methods cannot (*e.g.* penetrating with objects), suggesting the importance of scene contexts for motion prediction.

In addition, we tried more challenging longer (3-second-long) prediction. The prediction error of the deterministic model becomes greatly larger, but the stochastic model achieves lower error in the best case by producing diverse predictions (Fig. 6). Fig.9 shows the output examples, and we can see that the proposed model generates diverse human motions in each scene.

Thoughts: The accuracy is still not good; Scene semantics or gaze will be helpful.

There is still a gap between estimated and true goals. Scene image alone may not be sufficient for goal prediction. Scene semantics or gaze will be helpful.