Research Proposal

Yuming Chen

1 Introduction

My research interests and related experiences align with the VGX Group well. My research interests lie in **human-centric computer vision** (and reinforcement learning). I am particularly interested in generative models for interaction-rich problems, such as multi-body HOI generation and multi-person motion generation. Additionally, further applications in human-AI interaction in the real world, including VLA, and related topics also attract me.

2 Related Experiences

During my MSc, I conducted published work on 3D hand mesh reconstruction. Besides, I tried to train grasping policy for dexterous hand in simulators (MuJoCo and Raisim). Previously, I also have experience in *multi-agent reinforcement learning* and *opponent modelling*, which will be helpful for potential simulator-involved or interaction-rich researches.

3D Hand Reconstruction from Blurry Monocular Images

Most current hand mesh recovery approaches focus on the blurriness problem in video. However, blurry monocular images lack of temporal information, making it challenging. We insteadly estimate a hand mesh sequence from a single blurry image to utilise temporal information inherent in the image. To overcome the ambiguity caused by the blurriness, we make multiple estimations for one image in a generative manner, and select the plausible ones with a learned selection module.

3 Future Work

I am interested in all human-centric topics and currently (personally) studying on **Number-free person motion generation in a streaming / online manner**. Current researches focus on two-person generation. The motion of each character is concatenated and encoded into a uniform latent space (like *DuetGen*), or is fused with the other using a mechanism tailored for two people (like *InterGen*). Such methods are not able to trivially applied to scenarios including more than two people. On the other hand, most algorithms designed for number-free people generation (like *FreeMotion* and *SocialGen*) work in an offline manner, that is, the sequence is generated when the whole conditional input (like text) is given. while online generation algorithm (like *MotionStreamer*) only works for single person motion. The task of continuously generating interactive-rich motion sequence involving arbitrary number of people is still unexplored.

Besides, **Learning to generating physically plausible motion sequence with simulator** also attracts me. Existing works introduce the physical constraints with proxy loss network or heuristically designed loss as the physical simulators are not differentiable. Some works using simulators learn a manipulation policy in the simulator to produce the desired motion lack of generalisation. Such approaches are more like a controller instead of a generator. Naturally introduce the physical information into a generator with the help of simulator would have potential value to study.