# GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping

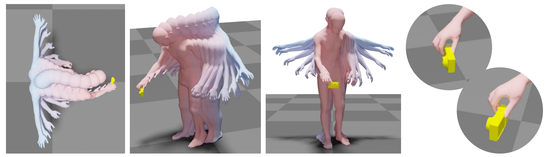
Omid Taheri[1] Vasileios Choutas[1,3] Michael J. Black[1] Dimitrios Tzionas[1,2]
[1]Max Planck Institute for Intelligent Systems, [2]University of Amsterdam, [3]ETH Zürich
[1]{otaheri, vchoutas, black}@tue.mpg.de, [2]d.tzionas@uva.nl

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

## Overview

**Objective**

- ☑ Generate full-body body motion to grasp 3D objects.
- ☑ Realistic hand grasps and head orientation.
- ☑ Natural foot-ground contact.

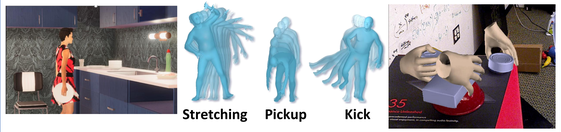


**Problem**

- High dimensional control problem
- Satisfy complex contact constraints

**Limitations of Prior Work**

- Non-realistic grasps.
- Bodies in "isolation" without objects.
- Not accurate hand grasping.
- Only hands without the body.




**Key Insights**

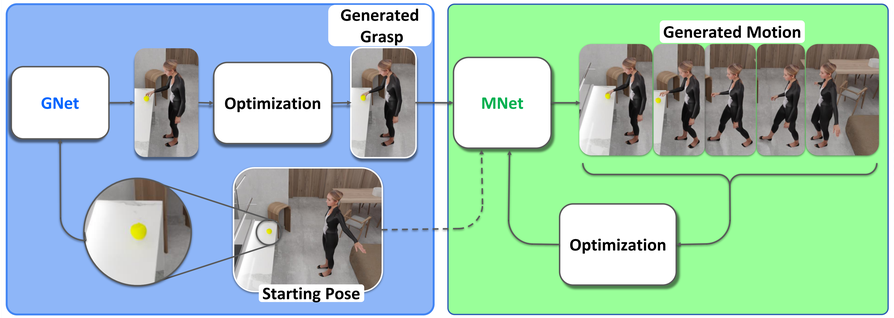
- Jointly inferring interaction features and body parameters.
- Using Interaction-Aware Attention representation.

## References

[1] Taheri et al. GRAB: A dataset of whole-body human grasping of objects. ECCV 2022
[2] Prokudin et al. Efficient learning on point clouds with basis point sets. CVPR 2019
[3] Rempe et al. HuMoR: 3D human motion model for robust pose estimation. ICCV 2021
[4] Starke et al. Neural state machine for character-scene interactions. TOG 2019
[5] Pavlakos et al. Expressive body capture: 3D hands, face, and body from a single image. CVPR 2019
[6] Zhang et al. We are more than our joints: Predicting how 3D bodies move. CVPR 2021

## Method: GOAL

- Use static grasps and dynamic motions from **GRAB** dataset.
- Generate a realistic **Grasping Pose** using GNet.
- Infill **the motion** between the start and Grasping Pose with MNet.
- Interaction-Aware attention representation improves grasps and motion.



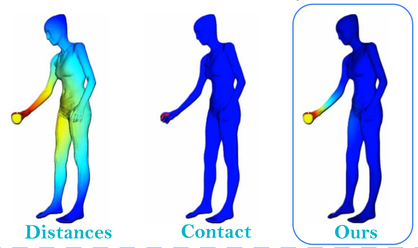

### Interaction-Aware Attention

**Novel representation for human-object Interaction**
Exponential transformation function on the distances.
As input to MNet:

- Reduces foot-sliding
- Improves the motion smoothness
- Results in better grasps

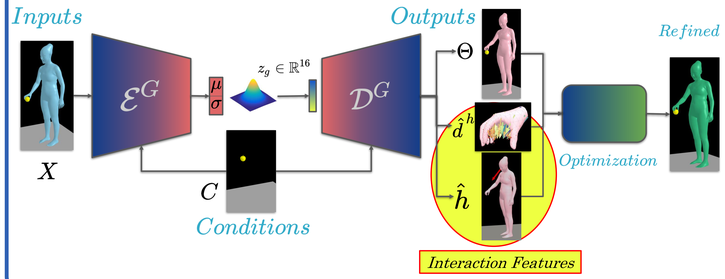
$$I_w(d) = \exp(-w \times d), \quad I_w : \mathbb{R}^D -> \mathbb{R}^D, \quad w > 0$$

$d \in \mathbb{R}^D \longrightarrow$ Distance Vector $w \longrightarrow$ Adjustable Parameter



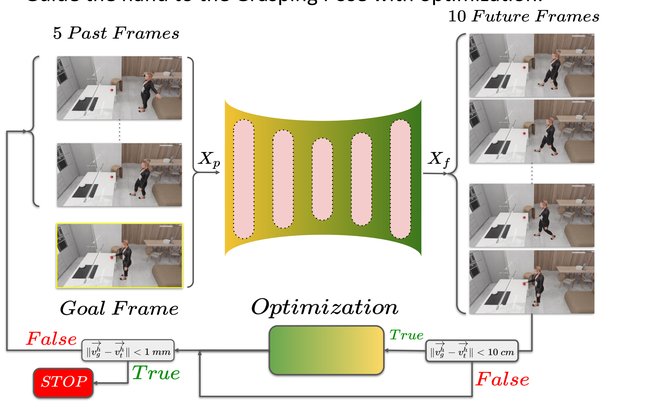

### GNet

- Generate interaction features in addition to SMPL-X parameters.
  - Hand-to-object offset vectors ($\hat{d}^h$)
  - Head direction vector ($\hat{h}$)
- SMPL-X predictions are good but only approximate.
- NNs learn more accurate Interaction features than SMPL-X parameters.
- We use the accurate interaction features to improve the Grasping Pose.

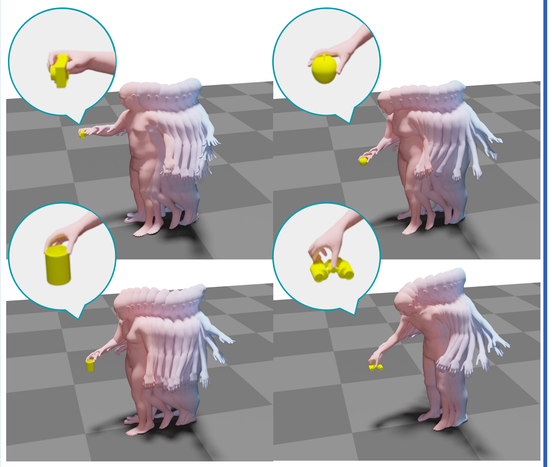



### MNet

- Autoregressively generates motion between start and Grasping Pose.
- Guide the hand to the Grasping Pose with optimization.




## Results



**Ratings:** **1** → Not Realistic **5** → Very Realistic

**GNet Evaluation – Before/After Optimization VS GRAB**

| Metric | GNet | GNet + Opt | Ground-truth [1] |
|---|---|---|---|
| Overall Grasping Pose ↑ | $3.89 \pm 0.93$ | $\mathbf{3.98} \pm 0.94$ | $3.78 \pm 1.06$ |
| Foot-Ground Contact ↑ | $3.98 \pm 1.06$ | $\mathbf{4.10} \pm 0.93$ | $3.82 \pm 1.11$ |
| Hand-Object Grasp ↑ | $2.70 \pm 1.37$ | $3.63 \pm 1.16$ | $\mathbf{3.98} \pm 1.04$ |
| Head Orientation ↑ | $3.83 \pm 1.01$ | $\mathbf{4.01} \pm 0.97$ | $3.84 \pm 1.07$ |
| Average ↑ | $3.60 \pm 1.22$ | $\mathbf{3.93} \pm 1.02$ | $3.86 \pm 1.07$ |

**MNet Evaluation – After Optimization VS GRAB**

| Metric | GOAL | Ground-truth [1] |
|---|---|---|
| Overall Body Motion ↑ | $3.74 \pm 0.97$ | $4.20 \pm 0.90$ |
| Foot-Ground Contact ↑ | $3.88 \pm 1.14$ | $4.18 \pm 1.05$ |
| Final Hand-Object Grasp ↑ | $3.66 \pm 1.05$ | $4.32 \pm 0.91$ |
| Head Orientation ↑ | $3.86 \pm 1.03$ | $4.18 \pm 1.00$ |
| Average ↑ | $3.79 \pm 1.05$ | $4.22 \pm 0.97$ |