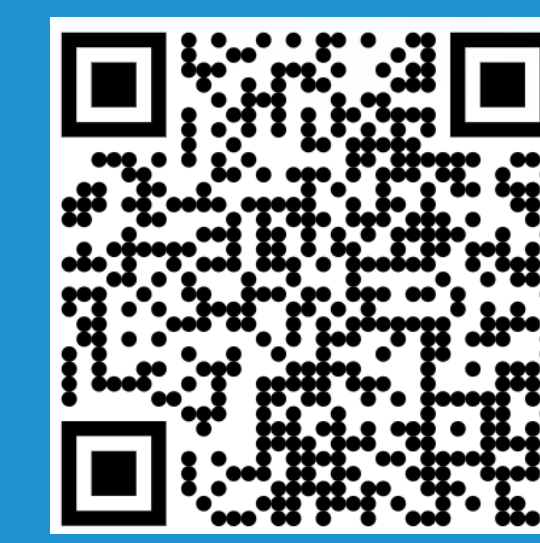
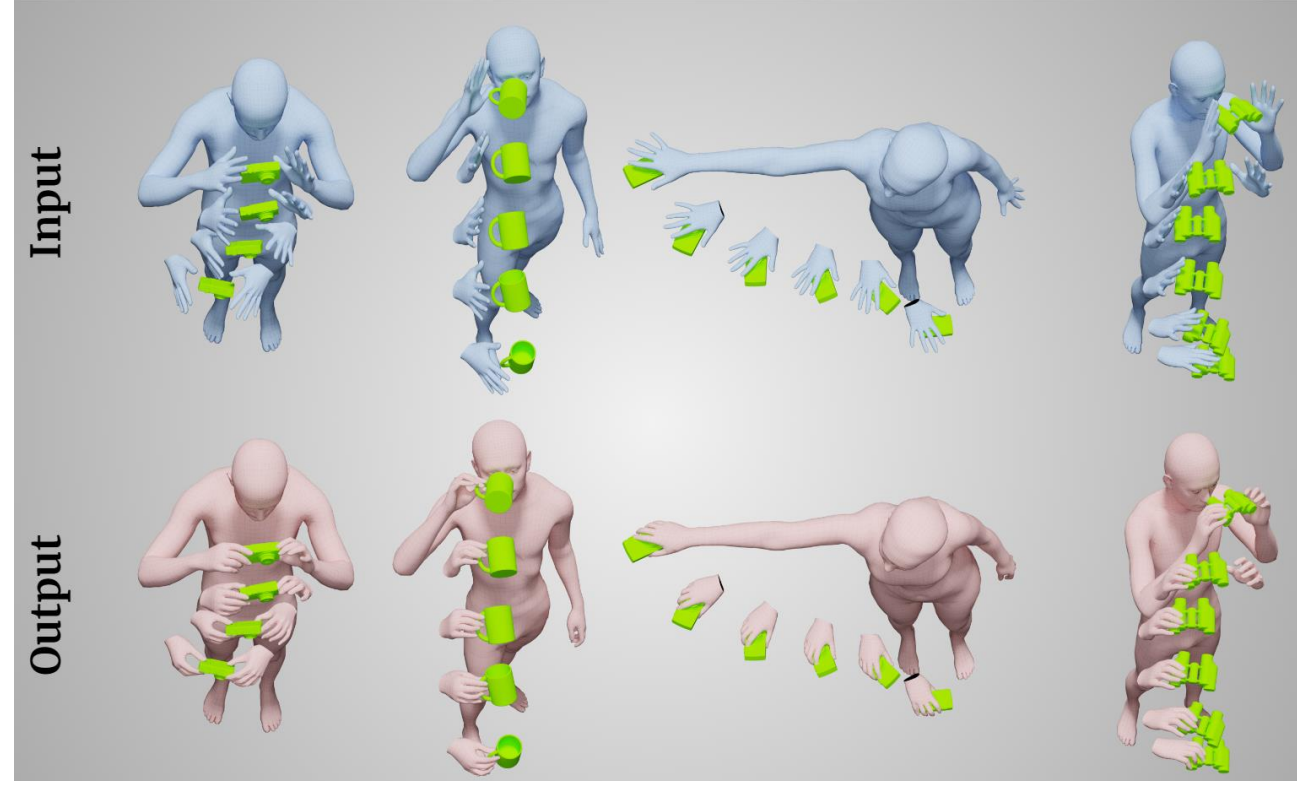




Objective

Given a sequence of body and object motion:
→ Accurately generate *interacting-hand poses*.



Why?

- ✓ Capture new datasets.
- ✓ Add hands to previous datasets.
- ✓ Refine generated/reconstructed hands.
- ✓ Retargeting grasp from one object to another.

Problem

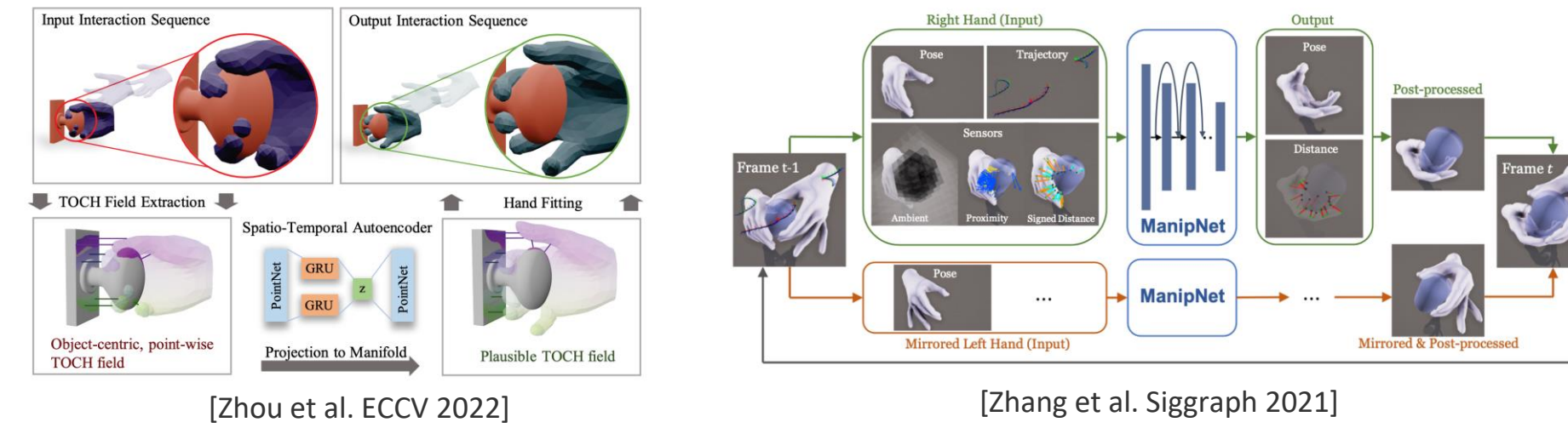
Generated Motions & Datasets:

- Bodies in “isolation” without objects.
- Only hands without the body.
- Not accurate hand grasping.
- Lack of accurate ground truth motions.



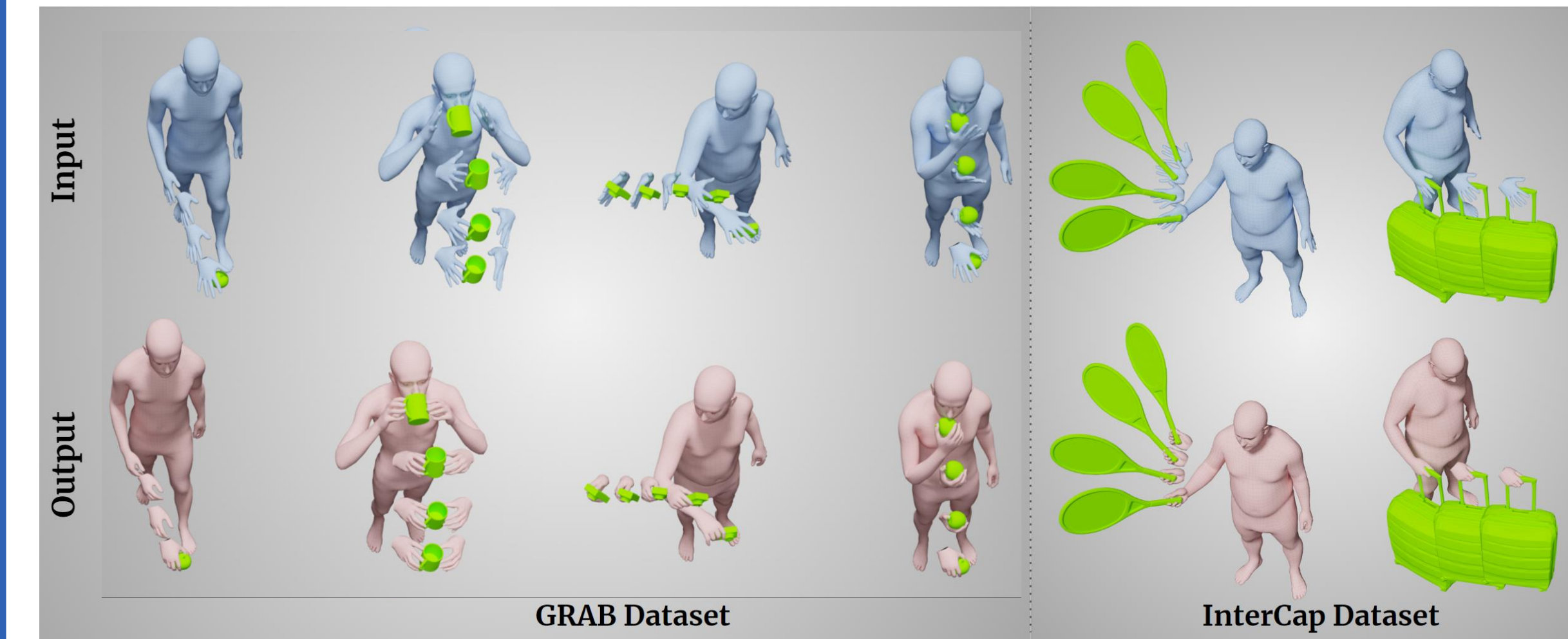
Limitations of Prior Work

- Not accurate hand grasping.
- Only hands without the body.
- Only refining hands, no generation.
- Slow (optimization).



Results

Fullbody Poses – Unseen Objects

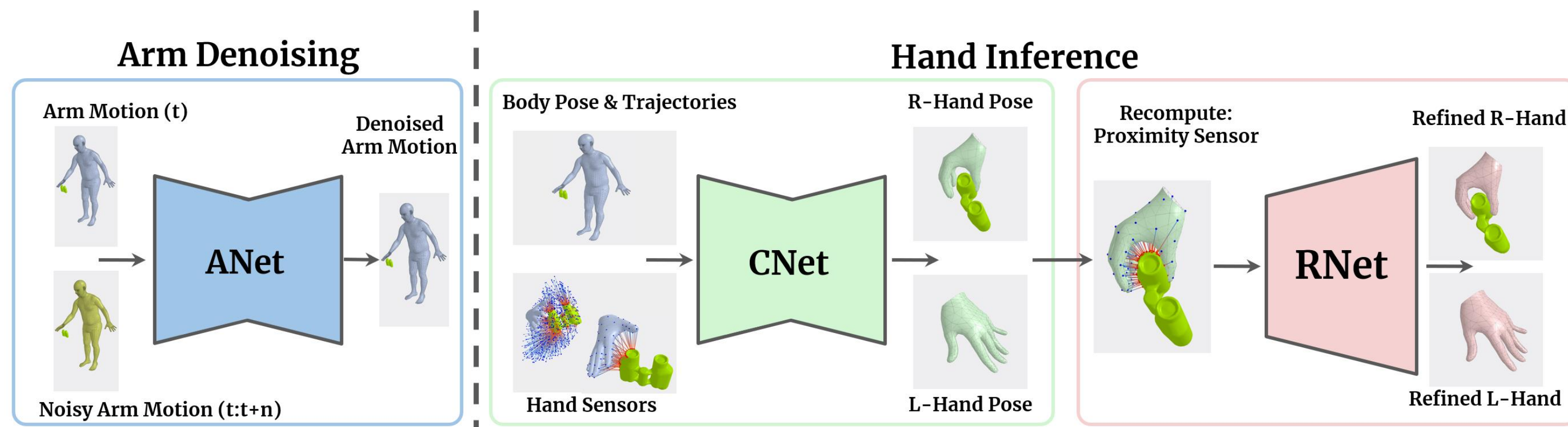


Single & Bi-manual Grasps – Unseen Objects



Method

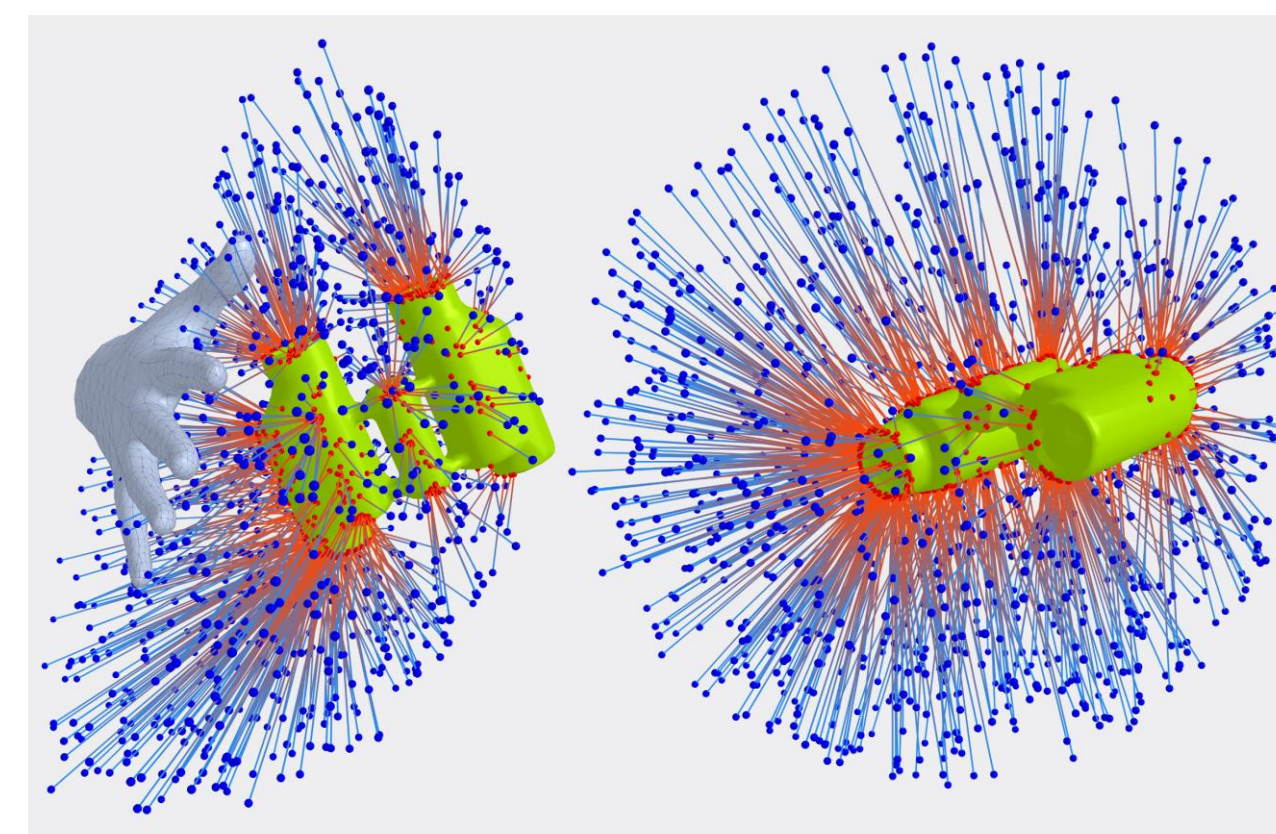
- Propose novel **Spatio-Temporal Virtual Sensors**, **Ambient & Proximity Sensor**.
- Novel **latent temporal consistency (LTC)** leads to smooth grasping motions.
- Two-stage inference for Arm and Hand.



Spatio-Temporal Virtual Sensors

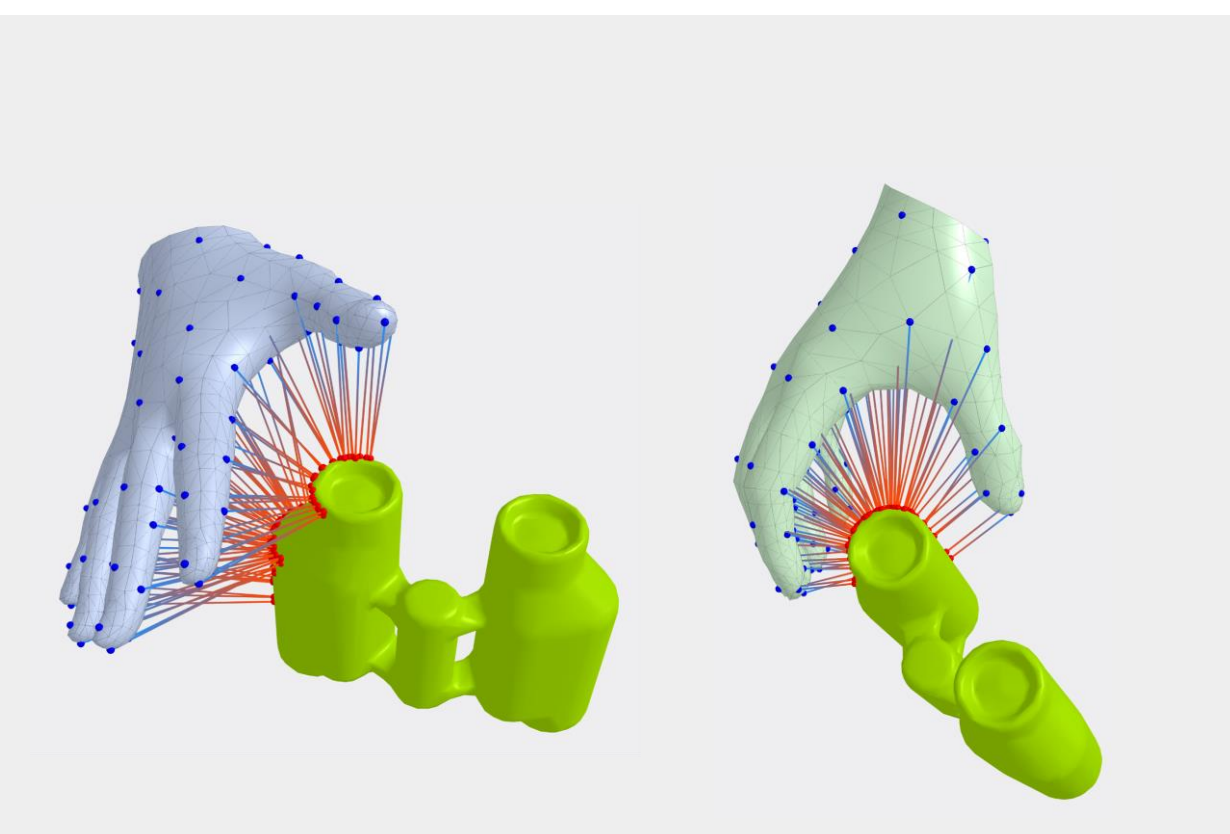
Ambient Sensor

- Continuous distance-based representation.
- Obtains the object’s geometric features.
- Spatial relation between the object & hands.



Proximity Sensor

- Fine-grained hand-object distance field.
- Captures hand to object correspondence.
- Helps with penetration and contact.

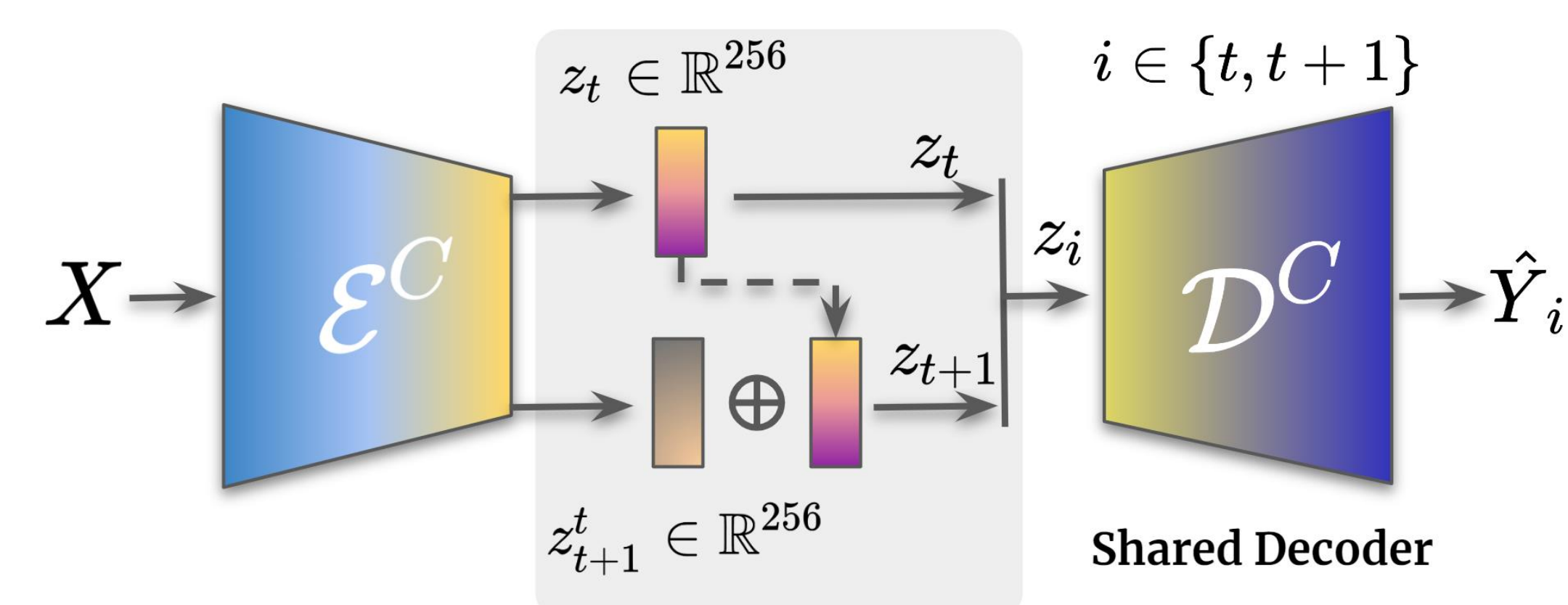


Consistency Network (CNet)

- Generates consistent hand interactions based on object motion.
- Infers both hands separately.
- Uses LTC to ensures realistic hand-object interactions.

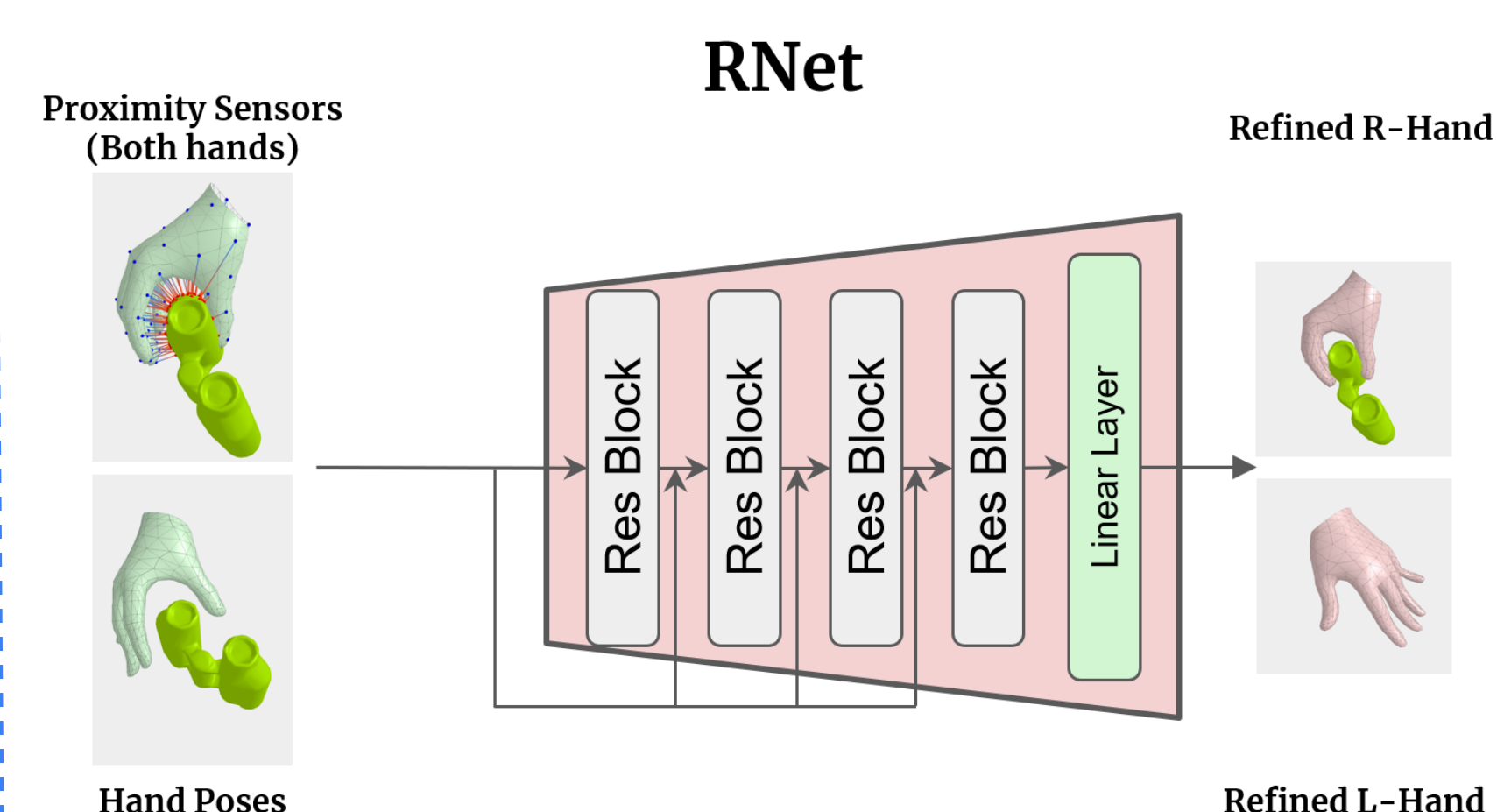
Latent Temporal Consistency (LTC)

- Defines consistency in the latent space for 2 successive frames.
 - A global latent code, z_t , and a relative one, z_{t+1}^t .
- Uses a shared decoder to further penalize the inconsistencies.



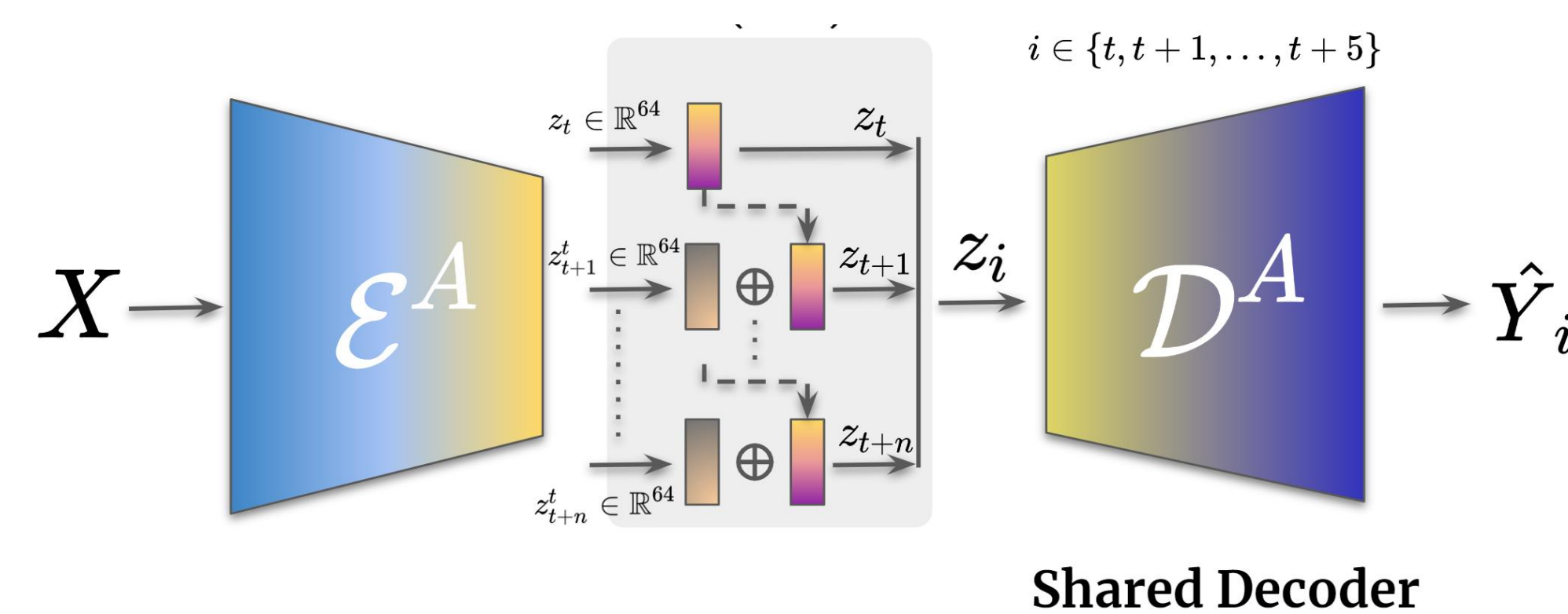
Refinement Network (RNet)

- Refines output of CNet to be more realistic.
- Preserves consistency generated by CNet.
- Recomputes proximity features to generate more accurate grasps and contact.



Arm Denoising Network (ANet)

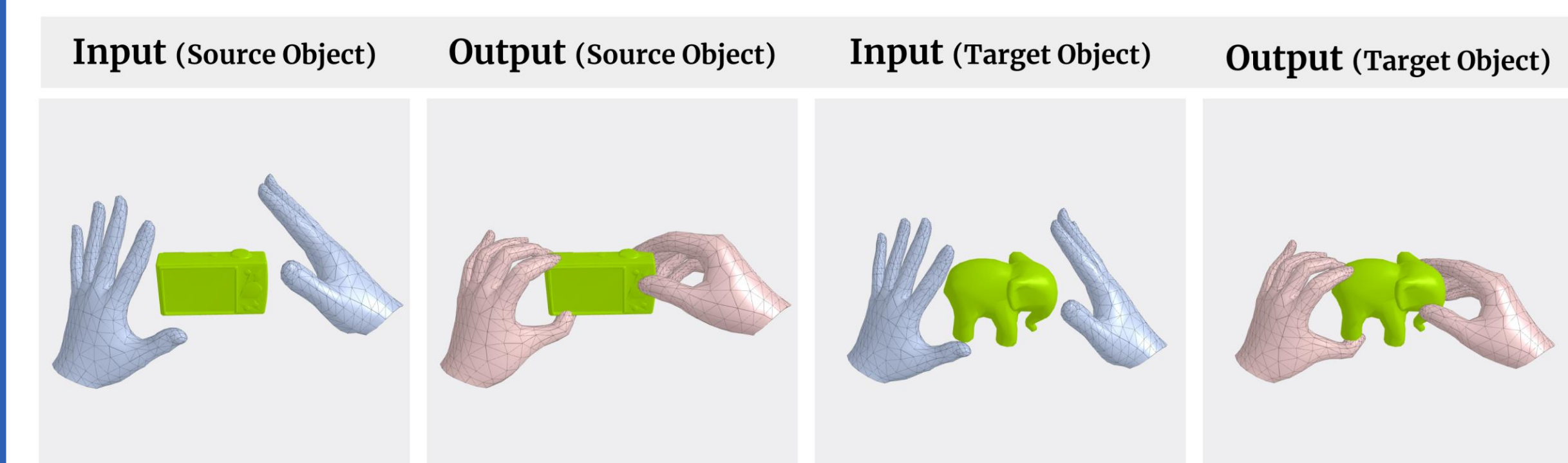
- Denoises arm motions before using as input to CNet.
- Exploits the LTC with 5 frames in the latent space.



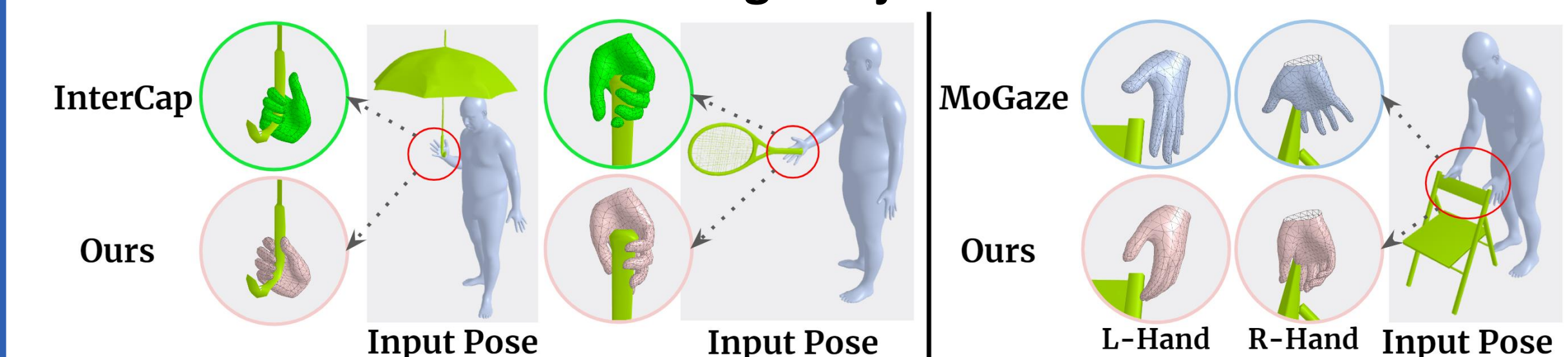
$X \rightarrow$ Future noisy poses + Current pose

$\hat{Y}_i \rightarrow$ Denoised pose for frame i

Application - Grasp Transfer



Large Objects



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] Zhang et al. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. | TOG 2021 |
| [4] Bhatnagar et al. Behave: Dataset and method for tracking human object interactions. | CVPR 2022 |
| [5] Pavlakos et al. Expressive body capture: 3D hands, face, and body from a single image. | CVPR 2019 |
| [6] Araujo et al. CIRCLE: Capture In Rich Contextual Environments | CVPR 2023 |
| [7] Huang et al. InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction | GCPR 2022 |

