

Hand-Centric Motion Refinement for 3D Hand-Object Interaction via Hierarchical Spatial-Temporal Modeling

Yuze Hao^{1,2,*}, Jianrong Zhang¹, Tao Zhuo³, Fuan Wen^{2,4}, Hehe Fan^{1,†}

¹Zhejiang University

²Beijing University of Posts and Telecommunications

³Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences)

⁴Beijing Key Laboratory of Network System and Network Culture

Abstract

Hands are the main medium when people interact with the world. Generating proper 3D motion for hand-object interaction is vital for applications such as virtual reality and robotics. Although grasp tracking or object manipulation synthesis can produce coarse hand motion, this kind of motion is inevitably noisy and full of jitter. To address this problem, we propose a data-driven method for coarse motion refinement. First, we design a hand-centric representation to describe the dynamic spatial-temporal relation between hands and objects. Compared to the object-centric representation, our hand-centric representation is straightforward and does not require an ambiguous projection process that converts object-based prediction into hand motion. Second, to capture the dynamic clues of hand-object interaction, we propose a new architecture that models the spatial and temporal structure in a hierarchical manner. Extensive experiments demonstrate that our method outperforms previous methods by a noticeable margin.

Introduction

Recently, hand-object interaction modeling has been widely used in many applications, such as VR/AR (Höll et al. 2018; Canales et al. 2019), and robotics (Xu et al. 2023; Qin et al. 2023; Bao et al. 2023). Recent methods can directly predict the hand-object pose from images (Chen et al. 2021; Doosti et al. 2020) or generate hand grasping pose given object information (Jiang et al. 2021; Brahmhatt et al. 2019; Zheng et al. 2023). However, due to the high degrees of freedom of hand articulation, self-occlusion of hands, and mutual occlusion of objects, the predicted results tend to be error-prone, limiting its application in downstream tasks.

Many approaches (Grady et al. 2021; Tse et al. 2022) are proposed to solve the aforementioned problems. To alleviate the potential anatomical irregularities in hand poses, Yang *et al.* (Yang et al. 2021) introduced a joint bending constraint to the parametric hand model (Romero, Tzionas, and Black

*This work was done when Yuze Hao was an intern at Zhejiang University. The code is publicly available at <https://github.com/Holiday888/HST-Net>.

†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

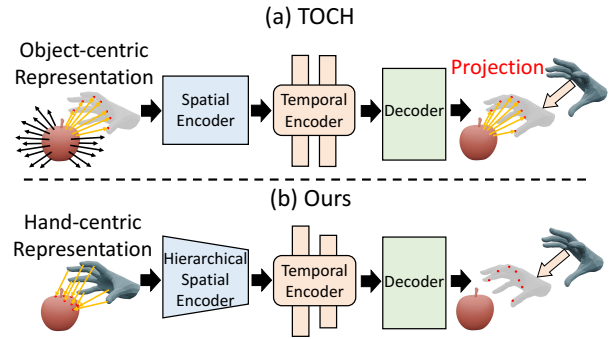


Figure 1: Comparing with TOCH (Zhou et al. 2022), our method has three advantages. (1) To capture the relation between hand and object, the existing method first emits rays from the object and then collects points that arrive at the hand. In contrast, we propose a straightforward hand-centric representation, which directly models the hand-object interaction. (2) Our hierarchical spatial-temporal architecture better captures dynamic information across different scales than the fixed-scale design in TOCH. (3) Due to the direct hand-object representation, our method does not require the additional post-process that converts object-centric representation into hand motions.

2017). To avoid the intersection between hands and objects or noncontact grasping, a heuristic repulsion loss and an attraction loss are employed (Hasson et al. 2019). However, these methods mainly consider the contacting moment (a single frame), instead of the entire hand-object interaction process (multiple frames) that includes both approaching and contacting stages.

In contrast, hand-object tracking, reconstruction (Hasson et al. 2021, 2020) and interaction synthesis (Taheri et al. 2022; Wu et al. 2022) can produce hand-object interaction motions, which involves pre-grasping and post-manipulation stages. Although these methods extend hand-object interaction from the contacting moment to the entire process, they introduce the jitter issue and inconsistent contact, leading to perturbed hand-object interaction motions. To refine the dynamic hand-object interaction mo-

tions, TOCH (Zhou et al. 2022) employs PointNet (Qi et al. 2017) for spatial modeling and Gated Recurrent Unit (GRU) (Chung et al. 2014) for temporal modeling. However, this fixed-scale spatial-temporal architecture is not efficient in capturing the spatial-temporal structure from local to global, thus limiting the ability of deep neural networks to learn the fine-coarse representations for hand-object interaction. Besides, as hand-object interaction is highly correlated with the object, TOCH employs an object-centric representation strategy that denotes hand motions from the object perspective, as shown in Figure 1(a). However, object-centric representation suffers from low-level information utilization because only a partial surface region of the object interacts with the hand. Moreover, employing such a representation necessitates an additional projection process, potentially introducing more ambiguity into the outcomes. To be specific, given the refined hand motions represented in the object-centric fashion, a projection process is required to align the hand pose with its corresponding representation. Nonetheless, when the contact area between the hand and object is small, it is difficult for the object-centric representation to correctly reflect the hand pose.

In this paper, we first introduce a novel hand-centric representation, based on hand poses and motions, as illustrated in Figure 1(b). The proposed representation captures the hand-object correspondence with the displacement between each hand vertex and the anchor points of the object. Compared to the previous object-centric method (Zhou et al. 2022), the hand-centric representation does not require information from the untouched, nontarget, or meaningless object areas. Moreover, our method is straightforward and eliminates the need for an additional projection process. Second, the previous fixed-scale architecture overlooks granularity in region-level contact, global-level pose, and short-term to long-term temporal dynamics. To address this issue, we introduce a hierarchical spatial-temporal encoder. The spatial encoder integrates regional details for a broader receptive field. The temporal encoder employs a transformer architecture across multiple frames, capturing both short and long-term dependencies. Extensive experiments on the GRAB and HO3D datasets demonstrate the effectiveness of the proposed method. In summary, our contributions are three folds:

- We propose a new hand-centric representation that effectively utilizes the hand-object interaction information. This representation avoids the ambiguity problem from the post-projection process.
- We propose a hierarchical spatial-temporal architecture to model hand-object interaction from the local level to the global level, which effectively captures the detailed clues and overall information of interactions.
- Experimental results demonstrate that our method outperforms the existing methods across various metrics, especially for longer-distant interaction sequences.

Related Work

Hand Object Interaction

Hand-object interaction is important in virtual reality and robotics. Recently, with the advent of datasets that con-

tain both hand and object annotations (Hampali et al. 2020; Taheri et al. 2020; Hasson et al. 2019; Fan et al. 2023), impressive progress has been made.

Early methods mainly focus on modeling static grasping interaction. Some approaches (Lin et al. 2023; Doosti et al. 2020; Hasson et al. 2019) directly estimate hand-object pose or reconstruct 3D mesh from a single image. Given the information of objects, several efforts (Jiang et al. 2021; Brahmbhatt et al. 2019; Turpin et al. 2022) generate the corresponding hand-grasping pose. Furthermore, other approaches focus on refining hand-object grasping state (Grady et al. 2021; Yang et al. 2021; Tse et al. 2022; Taheri et al. 2020). For instance, Grady *et al.* (Grady et al. 2021) proposed to predict the hand-object contact map and then align the hand with the predicted region. However, these methods only consider the static grasping stage.

Recently, there has been a growing emphasis on modeling dynamic hand-object interactions. Some methods (Hasson et al. 2021, 2020; Chen et al. 2023) can estimate/reconstruct interactions from video input. With the hand-object interaction motion datasets (Taheri et al. 2020; Hampali et al. 2020; Liu et al. 2022b; Fan et al. 2023), interaction motions can be generated (Zhang et al. 2021; Taheri et al. 2022; Wu et al. 2022; Zheng et al. 2023). As the first method to address refining perturbed 3D hand-object interaction sequence, Zhou *et al.* (Zhou et al. 2022) proposed an object-centric representation with a spatial-temporal modeling architecture. In this work, we share the same goal to refine the hand-object interaction sequence.

Modeling Human Motion Prior

Obtaining high-quality global motion prior is crucial for human motion modeling, by capturing the motion prior, tasks such as motion prediction (Zhao et al. 2023; Xu, Wang, and Gui 2022) and motion generation (Zhang et al. 2023; Tevet et al. 2023; Guo et al. 2022) can be benefited. Researchers have explored various approaches to obtain spatial-temporal motion prior through auto-encoder (Fragkiadaki et al. 2015; Rempe et al. 2021), generative adversarial network (Ferreira et al. 2021; Zhao, Su, and Ji 2020), and diffusion model (Zhang et al. 2022; Tevet et al. 2023).

For hand motion, similar to body motion modeling, by capturing the spatial-temporal motion prior, tasks like hand gesture prediction (Qi et al. 2023; Ng et al. 2021), dynamic hand pose estimation (Liu et al. 2022a) can be achieved. In terms of hand-object interaction, modeling interaction sequences in a spatial-temporal manner (Zhou et al. 2022) is also essential. In this work, we handle spatial-temporal information of hand-object interaction hierarchically.

Point Cloud Video Modeling

Hand-object interactions are represented as sequences of 3D point clouds or point cloud videos. While various studies have focused on modeling dynamic point clouds (Fan, Yang, and Kankanhalli 2021; Liu et al. 2022b; Fan et al. 2022; Shen et al. 2023; Fan, Yang, and Kankanhalli 2022; Sheng, Shen, and Xiao 2023; Fan et al. 2021; Liu et al. 2023) for classification and segmentation, our approach can be regarded as a point cloud video denoising method.

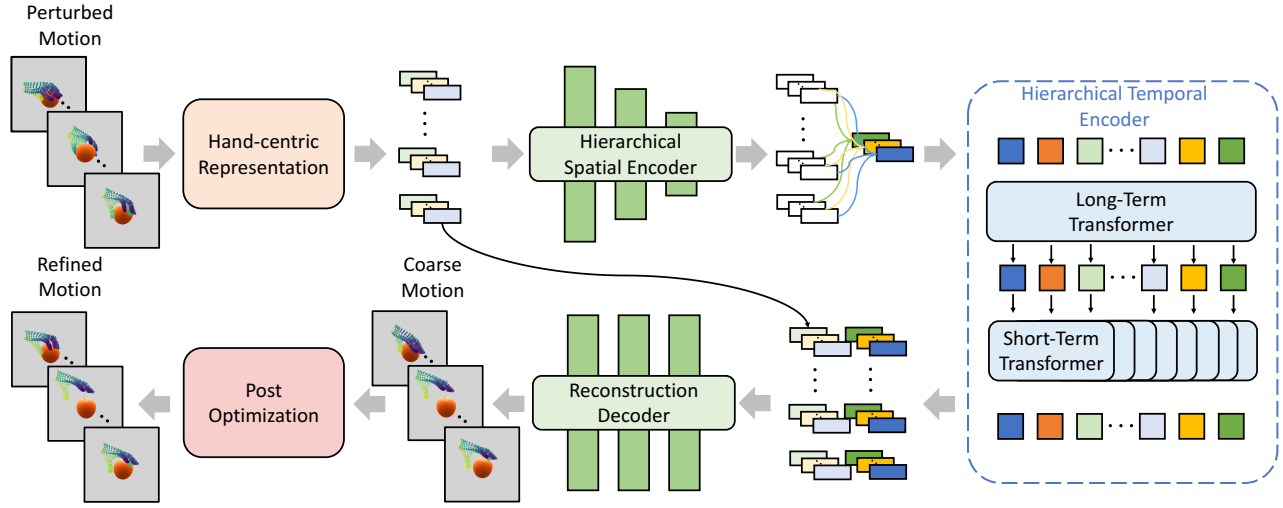


Figure 2: Overview of our framework. Given a perturbed interaction sequence, we first convert the sequence into our hand-centric correspondence representation. Then the representations are fed into a hierarchical spatial encoder to capture the local-global spatial features for each frame. Next, the features are passed through a hierarchical temporal encoder to extract long-term and short-term dependencies across frames. Lastly, the refined sequences are obtained from a reconstruction decoder followed by a post-optimization.

Methodology

Overview and Problem Setup

Our framework is illustrated in Figure 2. Given a perturbed hand sequence with corresponding object sequence $([\tilde{\mathbf{H}}^0; \mathbf{O}^0], [\tilde{\mathbf{H}}^1; \mathbf{O}^1], \dots, [\tilde{\mathbf{H}}^T; \mathbf{O}^T])$ where T is the number of frames for interaction sequence. For each frame t , $\tilde{\mathbf{H}}^t = \{\tilde{\mathbf{h}}_i^t\}_{i=1}^{N^h}$ and $\mathbf{O}^t = \{\mathbf{o}_i^t\}_{i=1}^{N^o}$ are point cloud representations for the hand and object, respectively. Besides, $\tilde{\mathbf{h}}_i^t \in \mathbb{R}^3$ and $\mathbf{o}_i^t \in \mathbb{R}^3$ are the i -th vertex for hand and object, and N^h and N^o are the respective point counts. Note that $\tilde{\mathbf{H}}^t$ is the hand mesh vertices that are generated from MANO (Romero, Tzionas, and Black 2017). Our goal is to refine the perturbed hand sequence and generate a more realistic hand interaction sequence $\hat{\mathbf{H}} = [\hat{\mathbf{H}}^0, \hat{\mathbf{H}}^1, \dots, \hat{\mathbf{H}}^T]$ where $\hat{\mathbf{H}}^t \in \mathbb{R}^{N^h \times 3}$. As we focus on enhancing the perturbed hand motions, we follow previous work (Zhou et al. 2022) to assume the ground truth object sequences $\mathbf{O} = [\mathbf{O}^0, \mathbf{O}^1, \dots, \mathbf{O}^T]$ are available during the training and testing stage.

To refine the perturbed hand motions, we first convert the perturbed hand-object interaction sequence pair into our hand-centric representation. Then we feed the motion representation into the designed hierarchical spatial-temporal architecture, generating coarse motions. Finally, a post-optimization process is applied to further refine the artifacts and smoothness.

Hand-Centric Representation

TOCH (Zhou et al. 2022) represents the hand-object interaction in an object-centric manner. This is achieved by emitting rays from the vertices of the object and subsequently gathering points that intersect with the hand. TOCH utilizes

both the distance of the ray and the positional information on the hand as distinctive features for interaction. However, such a representation method suffers from inefficient data utilization, because only a portion region of the object interacts with the hand (as shown in Figure 1(a)). Furthermore, the representation needs an additional projection process, potentially introducing more ambiguity into the outcomes, such as the visual results shown in Figure 3.

In this work, we propose a hand-centric representation to tackle the above problems. Specifically, for each frame in the sequence, given the perturbed hand $\tilde{\mathbf{H}}^t$ and object \mathbf{O}^t . The corresponding representation is defined as follows:

$$\mathbf{R}^t = \{(\vec{\mathbf{c}}_i^t, \tilde{\mathbf{h}}_i^t)\}_{i=1}^{N^h}, \quad (1)$$

where $\vec{\mathbf{c}}_i^t = \hat{\mathbf{o}}_i^t - \tilde{\mathbf{h}}_i^t$, $\hat{\mathbf{o}}_i^t \in \mathbf{O}^t$ is the closest vertex to $\tilde{\mathbf{h}}_i^t$. Based on such a hand-centric representation, our model integrates the relevant object information while effectively preserving the distinct characteristics of each hand vertex. Subsequently, the representation of the hand-object sequence is defined as $\mathbf{R} = [\mathbf{R}^0, \mathbf{R}^1, \dots, \mathbf{R}^T]$ with $\mathbf{R}^t = \{\mathbf{r}_i^t\}_{i=1}^{N^h} \in \mathbb{R}^{N^h \times 6}$.

Hierarchical Spatial-Temporal Motion Reconstruction

Unlike the previous method (Zhou et al. 2022) that focuses on fixed-scale hand-object interaction point cloud sequences, our approach employs a Hierarchical Spatial-Temporal Network (HST-Net). This network comprises hierarchical spatial and temporal encoders, along with a reconstruction decoder. Through HST-Net, our model captures region-level contact, global-level pose, and long-term and short-term dependencies, leading to a more comprehensive abstraction of the interaction sequence.

Hierarchical spatial encoder. Given the sequence $\mathbf{R} = [\mathbf{R}^0, \mathbf{R}^1, \dots, \mathbf{R}^T]$, the proposed hierarchical spatial encoder encodes the sequence to $\mathbf{S} = [s^0, s^1, \dots, s^T]$, where $\mathbf{S} \in \mathbb{R}^{d \times T}$ is the latent representation of the sequence with d is the dimension of the latent feature.

As our representation is a set of points attached to its relative feature, we propose to model the feature in a hierarchical local-global manner. Specifically, given the point cloud $\tilde{\mathbf{H}}^t$ with the corresponding feature \mathbf{R}^t . To get local level information, we first select several anchor points $\tilde{\mathbf{H}}^{t'}$ from $\tilde{\mathbf{H}}^t$ by using farthest point sampling. With the selected anchors, a continuous convolution layer is used to capture the local spatial information for each point:

$$\mathbf{f}_{(x,y,z)}^t = \underset{\|\delta_x, \delta_y, \delta_z\| \leq \gamma}{\text{MAX}} \text{MLP}(\mathbf{r}_{(x+\delta_x, y+\delta_y, z+\delta_z)}^t, \delta_x, \delta_y, \delta_z), \quad (2)$$

where $\mathbf{f}_{(x,y,z)}^t$ is the corresponding spatial feature of point (x, y, z) from $\tilde{\mathbf{H}}^{t'}$, and $(\delta_x, \delta_y, \delta_z)$ represents the displacement, $\mathbf{r}_{(x+\delta_x, y+\delta_y, z+\delta_z)}^t$ is the feature of position $(x+\delta_x, y+\delta_y, z+\delta_z)$ at time t , and γ is the spatial search radius.

After multiple layers of sampling and aggregation, the local receptive field is sufficiently enlarged. Then the global feature is obtained by applying the max-pooling operation on the output of the final spatial layer:

$$\mathbf{s}^t = \underset{(x,y,z) \in \tilde{\mathbf{H}}^{t'}}{\text{MAX}} \mathbf{f}_{(x,y,z)}^t. \quad (3)$$

Hierarchical temporal encoder. Based on the frame-wise spatial feature, we propose to use long-term and short-term transformers to obtain the inter-frame hierarchical dependencies. The former performs long-range information integration across all temporal dimensions, and the latter focuses more on the relationship within the local receptive field. Precisely, given the spatial features $\mathbf{S} = [s^0, s^1, \dots, s^T]$, we first take transformers with long-term self-attention to get global information which can be formulated as:

$$\mathbf{L} = \text{Self-Attention}(s^0, s^1, \dots, s^T), \quad (4)$$

where $\mathbf{L} = [l^0, l^1, \dots, l^T]$ and $\mathbf{L} \in \mathbb{R}^{d' \times T}$. Then the long-term representation \mathbf{L} is split into B frame bins, obtaining $\hat{\mathbf{L}} = \{\hat{\mathbf{L}}_b\}_{b=1}^B$, where $\hat{\mathbf{L}}_b = [\hat{l}_b^0, \hat{l}_b^1, \dots, \hat{l}_b^{T'}]$ denotes a frame bin with $T' = T/B$ frames. Then short-term self-attention is employed for local feature integration, the process can be computed as follows:

$$\tilde{\mathbf{S}}_b = \text{Self-Attention}(\hat{l}_b^0, \hat{l}_b^1, \dots, \hat{l}_b^{T'}), \quad (5)$$

where $\tilde{\mathbf{S}}_b = [\tilde{s}_b^0, \tilde{s}_b^1, \dots, \tilde{s}_b^{T'}]$ and $\tilde{\mathbf{S}}_b \in \mathbb{R}^{d'' \times T'}$. Finally, we concatenate B short-term representations for $\tilde{\mathbf{S}} = [\tilde{s}^0, \tilde{s}^1, \dots, \tilde{s}^T]$ as the input of the decoder, where $\tilde{\mathbf{S}} \in \mathbb{R}^{d'' \times T}$.

Reconstruction decoder. Given the frame-wise global feature $\tilde{\mathbf{S}}$ obtained by our hierarchical spatial-temporal encoder, the reconstruction decoder is employed to recover coarse hand motions $\hat{\mathbf{H}}'$. The feature is first repeated for each

point and then concatenated with the hand-centric representation, which can be formulated as $\tilde{\mathbf{R}}^t = \text{Concat}(\tilde{\mathbf{S}}^t, \mathbf{R}^t)$, where $\tilde{\mathbf{R}}^t \in \mathbb{R}^{N^h \times (d''+6)}$. Then, based on the input $[\tilde{\mathbf{R}}^0, \tilde{\mathbf{R}}^1, \dots, \tilde{\mathbf{R}}^T]$, we use several weight-shared MLPs as the decoder to reconstruct hand motion $\hat{\mathbf{H}}'$.

Given the ground truth of hand motion $\mathbf{H} = [\mathbf{H}^0, \mathbf{H}^1, \dots, \mathbf{H}^T]$ with $\mathbf{H}^t \in \mathbb{R}^{N^h \times 3}$. We minimize the distance between ground truth \mathbf{H} and prediction $\hat{\mathbf{H}}'$ during training, which can be written as:

$$\mathcal{L}_{recons} = \left\| \mathbf{H} - \hat{\mathbf{H}}' \right\|_2^2. \quad (6)$$

Post Optimization

Based on the predicted coarse hand motion $\hat{\mathbf{H}}'$, we further employ a post-optimization step to improve the overall smoothness. We use the MANO model (Romero, Tzionas, and Black 2017) with predicted hand vertices, to generate final refined hand motion $\hat{\mathbf{H}}$. Here, the MANO model serves as a linear blend skinning model that can produce the 3D hand vertices based on the shape and pose parameters. We optimize the shape parameter β , pose, orientation and translation parameter $\theta = \{\theta^t\}_{t=1}^T$ of the MANO model by minimizing:

$$\mathcal{L}(\beta, \theta) = \mathcal{L}_{consistent}(\beta, \theta) + \mathcal{L}_{smooth}(\beta, \theta), \quad (7)$$

where $\mathcal{L}_{consistent}$ ensures alignment with the predicted outcomes $\hat{\mathbf{H}}$ to maintain consistency, and \mathcal{L}_{smooth} aims to enhance overall smoothness. Importantly, we assume that the shape remains constant throughout the given interaction sequence, allowing us to utilize the same shape parameter β across the sequence.

Experiments

In this section, we apply our method to synthetic and real tracking errors for hand-object interaction and evaluate our method on various metrics. We first introduce the experiment settings including the datasets, evaluation metrics, and implementation details. Then we demonstrate that our method could achieve better performance compared with the state-of-the-art methods. Finally, we demonstrate the effectiveness of the proposed components with ablation studies.

Dataset

GRAB. GRAB (Taheri et al. 2020) is a Motion Capture dataset focusing on whole-body grasping. It comprises complete 3D shape and pose sequences, including 10 performers interacting with 51 objects. Following TOCH (Zhou et al. 2022), we select 47/4/6 objects for training, validation, and testing, and filter out frames where the hand wrist is more than 15cm away from the object. To simplify the problem, we use the right hand as the case of study. For left-hand cases, we mirror hands and objects. To evaluate our method on synthetic perturbations, we follow TOCH and adopt various perturbation strategies. These strategies include GRAB-T (translation-dominant perturbation), GRAB-R (pose-dominant perturbation), and GRAB-B (balanced

Methods		MPJPE (mm) ↓	MPVPE (mm) ↓	IV (cm^3) ↓	C-IoU (%) ↑
Task 1: GRAB-T (0.01)	Perturbation	16.0	16.0	2.48	16.24
	TOCH	9.93	11.8	1.79	23.25*
	Ours	6.90	7.75	1.70	28.00
Task 2: GRAB-T (0.02)	Perturbation	31.9	31.9	2.4	10.69
	TOCH	12.3	13.9	2.5	20.35*
	Ours	10.19	10.76	1.85	22.84
Task 3: GRAB-R (0.3)	Perturbation	4.58	6.30	1.88	23.21
	TOCH	9.58	11.5	1.52	23.39*
	Ours	5.85	6.92	1.61	28.61
Task 4: GRAB-R (0.5)	Perturbation	7.53	10.3	1.78	17.31
	TOCH	9.12	11.0	1.35	22.13*
	Ours	6.28	7.54	1.69	25.27
Task 5: GRAB-B (0.01 & 0.3)	Perturbation	17.3	18.3	2.20	12.14
	TOCH	10.3	12.1	1.78	23.10*
	Ours	7.17	8.09	1.69	26.36

Table 1: Quantitative results on the GRAB (Taheri et al. 2020) test set. Following TOCH (Zhou et al. 2022), we manually perturb the ground truth with GRAB-T (translation-dominant perturbation), GRAB-R (pose-dominant perturbation), and GRAB-B (balanced perturbation). The numbers inside the parentheses represent the magnitude of perturbation which is sampled from Gaussian noise. The symbol * denotes that we reproduced the results using the released code of TOCH (Zhou et al. 2022).

Methods		MPJPE	MPVPE	IV
Tracking	Hasson <i>et al.</i>	11.4	11.4	9.26
Static	RefineNet	11.6	11.5	8.11
	ContactOpt	9.47	9.45	5.71
Dynamic	TOCH	9.32	9.28	4.66
	Ours	9.18	9.21	4.52

Table 2: Quantitative evaluation on the HO3D dataset. In this experiment, we follow the method in TOCH (Zhou et al. 2022) to select interaction sequences from the predicted result of the tracking method (Hasson et al. 2021) and use different optimization methods (Taheri et al. 2020; Grady et al. 2021) to refine interaction sequences.

perturbation). Our model is only trained on the GRAB-B dataset and tested on the five perturbation tasks, as shown in Table 1.

HO3D. HO3D (Hampali et al. 2020) is a dataset capturing hand-object interaction. It consists of frame-wise annotation for hands and objects with corresponding RGB images. We follow the method in (Zhou et al. 2022) to select a subset of frames containing hand-object interaction. To evaluate our method on real tracking perturbations, we retrained a state-of-the-art hand-object tracking method (Hasson et al. 2021) and applied our method to the predicted results.

Evaluation Metrics

We follow TOCH (Zhou et al. 2022) to define evaluation metrics as:

- **Mean Per-Joint Position Error (MPJPE).** MPJPE is the average Euclidean distance between the refined hand joints and the ground truth.
- **Mean Per-Vertex Position Error (MPVPE).** MPVPE is similar to MPJPE which aims to measure the average Euclidean distance between the refined hand vertices and the ground truth vertices.
- **Solid Intersection Volume (IV).** This metric calculates hand-object interpenetration volume by voxelizing the object and hand mesh. Here, we set the voxel size as 2mm.
- **Contact IoU (C-IoU).** We report the Intersection-over-Union between the refined contact map and the ground truth contact map. To obtain the contact map, we calculate the distance between hand vertices and object mesh and threshold the distance within ± 2 mm.

Implementation Details

Our network consists of three consecutive modules: a hierarchical spatial encoder, a hierarchical temporal encoder, and a reconstruction decoder. The hierarchical spatial encoder consists of four identical blocks with a down-sampling rate of 2. The output dimension of our spatial encoder is $d = 256$. For the Hierarchical Temporal Encoder, we set $d' = d'' = 256$ and set $T = 30$ and $T' = 5$ as the length of input sequence for the long-term temporal transformer and short-term temporal transformer, respectively. Both transformer blocks have four layers and use fixed sine/cosine position encoding. For each layer, the attention head is set as 8, and the feed-forward dimension is set as 1024.

For the reconstruction decoder, four identical PointNet (Qi et al. 2017) akin blocks are used to map the latent

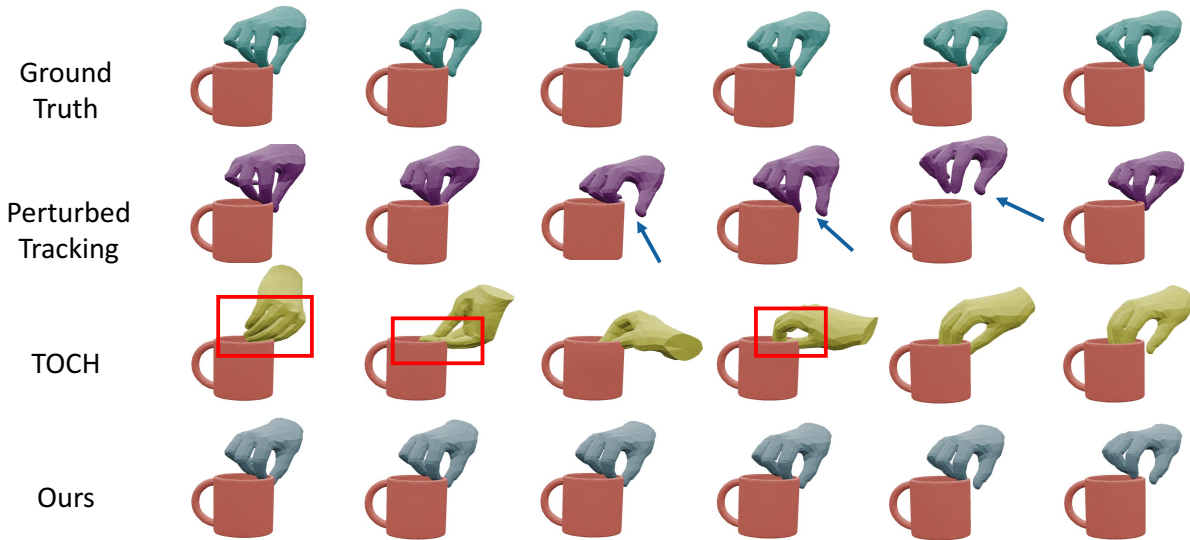


Figure 3: Qualitative results for refining inconsistency pose (highlighted with blue arrows) in perturbed tracking sequence. The refined sequence of TOCH (Zhou et al. 2022) exhibits improper grasping poses (highlighted with red boxes). In contrast, our reconstructions demonstrate a more plausible interacting pose.

Methods	MPJPE	MPVPE	IV	C-IoU
Hand Vertices	7.89	8.80	1.73	24.92
Object Center	7.89	8.82	1.73	24.71
Bounding Box	8.02	8.92	1.75	24.91
Closest Vertex (Ours)	7.17	8.09	1.69	26.36

Table 3: Quantitative evaluation with different hand-centric representations. For this experiment, we evaluate all the features using the GRAB-B (balanced perturbation) dataset. For a fair comparison, we only replace our representation and keep other settings untouched.

dimension into 16, and a fully connected layer is used to reconstruct the position of each point.

For all experiments, we set the batch size as 32, and use ADAM as the optimizer with an initial learning rate of $3e^{-4}$, and weight decay of $1e^{-6}$. Training our model takes about 16 hours on $4 \times$ NVIDIA Tesla V100S-32G GPUs.

Comparison With State-of-the-Arts

Our goal is to apply the designed method to real hand-object interaction tracking systems or synthesis methods to alleviate the erroneous. However, directly targeting specific tracking or synthesis methods may cause overfitting to specific errors making the performance difficult to quantify. Thus, we evaluate our method and compare it with TOCH (Zhou et al. 2022) on both synthetic tracking dataset and real tracking estimator (Hasson et al. 2021), where the results are based on their released source codes.

Quantitative results. We show the comparison results in Table 1 and Table 2 on GRAB (Taheri et al. 2020) test set and HO3D (Hampali et al. 2020) test set. For the synthetic GRAB dataset, Our model achieves a significant im-

provement in MPJPE and MPVPE for all perturbation strategies, which suggests the effectiveness and robustness of our method for refining the incorrect hand pose. For interaction metrics, though there was a marginal diminution in the performance of IV for the GRAB-R dataset, our method achieves notable improvements with C-IoU. For the HO3D dataset, we retrain the previous tracking method (Hasson et al. 2021) as our baseline model and applied both static and dynamic refining methods (Taheri et al. 2020; Grady et al. 2021; Zhou et al. 2022) on the predicted results. As demonstrated in Table 2, Our method outperforms previous approaches on all three metrics.

Qualitative results. The visualization results are presented in Figure 3 and Figure 4. It can be seen that the reconstruction results of TOCH (Zhou et al. 2022) suffer from unrealistic pose and incorrect interaction, while our method can generate hand-object interaction motion with better quality.

Ablation Study

For all ablation studies, we report MPJPE, MPVPE, IV, and C-IoU four metrics on the balanced perturbation dataset.

Investigating Hand-Centric representation. We investigate the effect of the proposed hand-centric representation by comparing it with different hand-object corresponding representations. For the object center representation, we replace the closest vertex with the mass center of the object to compute the interaction feature, while we use eight vertices of the object bounding box for bounding box representations. The comparative results of different representations are demonstrated in Table 3. It can be seen from the table that our representation achieves the best result on four metrics. The performance of different representations indicates

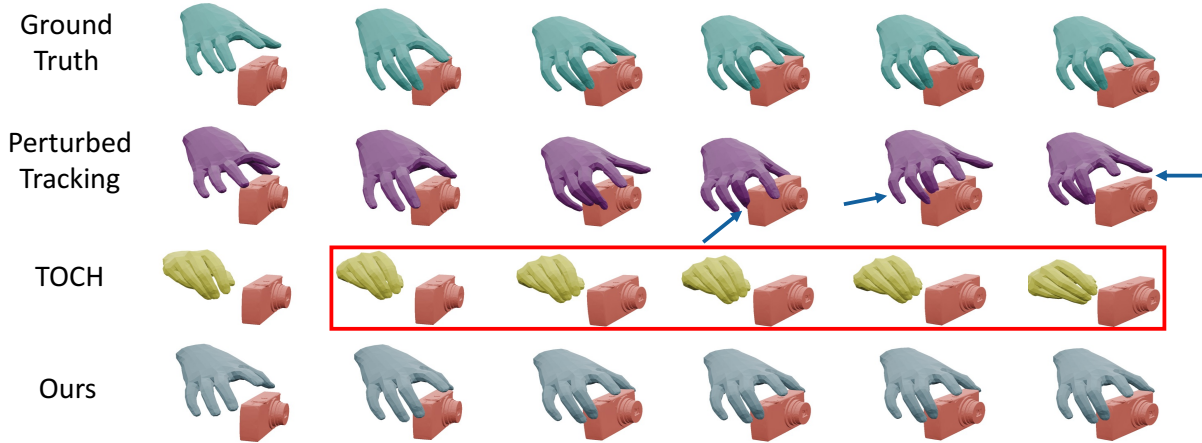


Figure 4: Qualitative results for refining inter-penetration (highlighted with blue arrows) in perturbed tracking sequence. The refined sequence of TOCH (Zhou et al. 2022) exhibits inadequate contact (highlighted with red boxes) while our results can achieve more realistic interaction.

Methods	MPJPE	MPVPE	IV	C-IoU
w/o hierarchical-spatial	7.35	8.22	1.94	25.25
short-long-term temporal	8.85	9.87	1.62	20.20
w/o hierarchical-temporal	7.47	8.39	1.72	25.70
w/o post-optimization	7.62	8.50	1.54	21.87
Ours	7.17	8.09	1.69	26.36

Table 4: Quantitative results with various baseline models are presented using the GRAB-B (balanced perturbation) dataset. In this experiment, we evaluate each baseline model by solely replacing the corresponding component while keeping all others unchanged.

Methods	MPJPE	MPVPE	IV	C-IoU
Perturbation	17.82	18.85	2.10	12.31
TOCH	14.11	15.87	1.45	21.44
Ours	7.80	8.71	1.60	26.14

Table 5: Quantitative results on the longer distance setting. In this experiment, we extended the hand-object distance to 30cm, which is twice the distance used in the previous setting. The evaluation was conducted directly using the network parameters trained on the short-distance setting.

the importance of integrating the relevant object information while effectively preserving the distinct characteristics of each hand vertex.

Investigating each component. We conducted ablation studies to evaluate the impact of each component in our model, and the results are summarized in Table 4. We investigate the effectiveness of the spatial encoder by replacing our hierarchical spatial encoder with a baseline model using four identical PointNet (Qi et al. 2017) layers. The effectiveness of the hierarchical temporal encoder is examined

by swapping long-term and short-term transformers, and by replacing short-term with long-term transformers to validate the hierarchy. Additionally, we also provide a comparison of whether there is post-optimization. Results show the superiority and essentiality of our method.

Experiments on a long hand-object distance. We also provide quantitative results for longer-distance hand-object interaction in Table 5. Specifically, we first filter out the motion frames with more than 30cm hand-object distance. Subsequently, we assess the model performance using the models initially trained for short-distance interactions. Our method outperforms TOCH (Zhou et al. 2022) on all evaluation metrics significantly. Our better performance demonstrates the importance of hand-centric representation. The proposed hand-centric representation can effectively capture hand motions, even when there is a considerable distance between the hand and the object.

Conclusion

In this paper, we tackle the problem of refining perturbed hand-object interaction sequences. To effectively represent the spatial-temporal interaction sequence, we explore various representations and propose a hand-centric approach that can avoid the ambiguous projection process in the previous object-centric representation. To obtain the dynamic clues of the interaction sequence, we design a hierarchical spatial-temporal architecture. This architecture effectively captures both local and global information. As a result, our method achieves state-of-the-art performance and demonstrates strong generalization capabilities on long-distant interaction sequences.

Acknowledgements

This paper is supported by Shandong Excellent Young Scientists Fund Program (Overseas) 2023HWYQ-114.

References

- Bao, C.; Xu, H.; Qin, Y.; and Wang, X. 2023. DexArt: Benchmarking Generalizable Dexterous Manipulation with Articulated Objects. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 21190–21200.
- Brahmbhatt, S.; Handa, A.; Hays, J.; and Fox, D. 2019. Contactgrasp: Functional Multi-Finger Grasp Synthesis from Contact. In *International Conference on Intelligent Robots and Systems (IROS)*, 2386–2393. IEEE.
- Canales, R.; Normoyle, A.; Sun, Y.; Ye, Y.; Luca, M. D.; and Jörg, S. 2019. Virtual Grasping Feedback and Virtual Hand Ownership. In *ACM Symposium on Applied Perception 2019*, 1–9.
- Chen, J.; Yan, M.; Zhang, J.; Xu, Y.; Li, X.; Weng, Y.; Yi, L.; Song, S.; and Wang, H. 2023. Tracking and Reconstructing Hand Object Interactions from Point Cloud Sequences in the Wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 304–312.
- Chen, Y.; Tu, Z.; Kang, D.; Chen, R.; Bao, L.; Zhang, Z.; and Yuan, J. 2021. Joint Hand-Object 3D Reconstruction from a Single Image with Cross-Branch Feature Fusion. *IEEE Transactions on Image Processing*, 30: 4008–4021.
- Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- Doosti, B.; Naha, S.; Mirbagheri, M.; and Crandall, D. J. 2020. Hope-Net: A Graph-Based Model for Hand-Object Pose Estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 6608–6617.
- Fan, H.; Yang, Y.; and Kankanhalli, M. 2021. Point 4D Transformer Networks for Spatio-Temporal Modeling in Point Cloud Videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 14204–14213.
- Fan, H.; Yang, Y.; and Kankanhalli, M. 2022. Point Spatio-Temporal Transformer Networks for Point Cloud Video Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2181–2192.
- Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; and Kankanhalli, M. 2022. PSTNet: Point Spatio-Temporal Convolution on Point Cloud Sequences. In *International Conference on Learning Representations (ICLR)*.
- Fan, H.; Yu, X.; Yang, Y.; and Kankanhalli, M. 2021. Deep Hierarchical Representation of Point Cloud Videos via Spatio-Temporal Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9918–9930.
- Fan, Z.; Taheri, O.; Tzionas, D.; Kocabas, M.; Kaufmann, M.; Black, M. J.; and Hilliges, O. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 12943–12954.
- Ferreira, J. P.; Coutinho, T. M.; Gomes, T. L.; Neto, J. F.; Azevedo, R.; Martins, R.; and Nascimento, E. R. 2021. Learning to Dance: A Graph Convolutional Adversarial Network to Generate Realistic Dance Motions from Audio. *Computers & Graphics*, 94: 11–21.
- Fragkiadaki, K.; Levine, S.; Felsen, P.; and Malik, J. 2015. Recurrent Network Models for Human Dynamics. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 4346–4354.
- Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. Contactopt: Optimizing Contact to Improve Grasps. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1471–1481.
- Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022. Tm2t: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 580–597.
- Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. Honnotate: A Method for 3D Annotation of Hand and Object Poses. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 3196–3206.
- Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging Photometric Consistency Over Time or Sparsely Supervised Hand-Object Reconstruction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 571–580.
- Hasson, Y.; Varol, G.; Schmid, C.; and Laptev, I. 2021. Towards Unconstrained Joint Hand-Object Reconstruction from RGB Videos. In *International Conference on 3D Vision (3DV)*, 659–668.
- Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning Joint Reconstruction of Hands and Manipulated Objects. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11807–11816.
- Höll, M.; Oberweger, M.; Arth, C.; and Lepetit, V. 2018. Efficient Physics-Based Implementation for Realistic Hand-Object Interaction in Virtual Reality. In *IEEE conference on virtual reality and 3D user interfaces (VR)*, 175–182.
- Jiang, H.; Liu, S.; Wang, J.; and Wang, X. 2021. Hand-Object Contact Consistency Reasoning for Human Grasps Generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11107–11116.
- Lin, Z.; Ding, C.; Yao, H.; Kuang, Z.; and Huang, S. 2023. Harmonious Feature Learning for Interactive Hand-Object Pose Estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 12989–12998.
- Liu, S.; Wu, W.; Wu, J.; and Lin, Y. 2022a. Spatial-Temporal Parallel Transformer for Arm-Hand Dynamic Estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 20523–20532.
- Liu, Y.; Chen, J.; Zhang, Z.; Huang, J.; and Yi, L. 2023. Leaf: Learning Frames for 4D Point Cloud Sequence Understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 604–613.

- Liu, Y.; Liu, Y.; Jiang, C.; Lyu, K.; Wan, W.; Shen, H.; Liang, B.; Fu, Z.; Wang, H.; and Yi, L. 2022b. HOI4D: A 4D Ego-centric Dataset for Category-Level Human-Object Interaction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 21013–21022.
- Ng, E.; Ginosar, S.; Darrell, T.; and Joo, H. 2021. Body2hands: Learning to Infer 3D Hands from Conversational Gesture Body Dynamics. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11865–11874.
- Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 652–660.
- Qi, X.; Liu, C.; Sun, M.; Li, L.; Fan, C.; and Yu, X. 2023. Diverse 3D Hand Gesture Prediction from Body Dynamics by Bilateral Hand Disentanglement. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 4616–4626.
- Qin, Y.; Huang, B.; Yin, Z.-H.; Su, H.; and Wang, X. 2023. Dexpoint: Generalizable Point Cloud Reinforcement Learning for Sim-to-Real Dexterous Manipulation. In *Conference on Robot Learning*, 594–605.
- Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3D Human Motion Model for Robust Pose Estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 11488–11499.
- Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics*, 36(6).
- Shen, Z.; Sheng, X.; Wang, L.; Guo, Y.; Liu, Q.; and Zhou, X. 2023. PointCMP: Contrastive Mask Prediction for Self-Supervised Learning on Point Cloud Videos. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 1212–1222.
- Sheng, X.; Shen, Z.; and Xiao, G. 2023. Contrastive Predictive Autoencoders for Dynamic Point Cloud Self-Supervised Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9802–9810.
- Taheri, O.; Choutas, V.; Black, M. J.; and Tzionas, D. 2022. GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 13263–13273.
- Taheri, O.; Ghorbani, N.; Black, M. J.; and Tzionas, D. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 581–600.
- Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations (ICLR)*.
- Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022. S2 Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-Supervised Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 568–584.
- Turpin, D.; Wang, L.; Heiden, E.; Chen, Y.-C.; Macklin, M.; Tsogkas, S.; Dickinson, S.; and Garg, A. 2022. Grasp’d: Differentiable Contact-Rich Grasp Synthesis for Multi-Fingered Hands. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 201–221.
- Wu, Y.; Wang, J.; Zhang, Y.; Zhang, S.; Hilliges, O.; Yu, F.; and Tang, S. 2022. Saga: Stochastic Whole-Body Grasping with Contact. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 257–274.
- Xu, S.; Wang, Y.-X.; and Gui, L.-Y. 2022. Diverse Human Motion Prediction Guided by Multi-Level Spatial-Temporal Anchors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 251–269.
- Xu, Y.; Wan, W.; Zhang, J.; Liu, H.; Shan, Z.; Shen, H.; Wang, R.; Geng, H.; Weng, Y.; Chen, J.; et al. 2023. Unidex-grasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 4737–4746.
- Yang, L.; Zhan, X.; Li, K.; Xu, W.; Li, J.; and Lu, C. 2021. CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 11097–11106.
- Zhang, H.; Ye, Y.; Shiratori, T.; and Komura, T. 2021. Manipnet: Neural Manipulation Synthesis with a Hand-Object Spatial Representation. *ACM Transactions on Graphics (ToG)*, 40(4): 1–14.
- Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023. T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 14730–14740.
- Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven Human Motion Generation with Diffusion Model. *arXiv preprint arXiv:2208.15001*.
- Zhao, M.; Tang, H.; Xie, P.; Dai, S.; Sebe, N.; and Wang, W. 2023. Bidirectional Transformer gan for Long-term Human Motion Prediction. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(5): 1–19.
- Zhao, R.; Su, H.; and Ji, Q. 2020. Bayesian Adversarial Human Motion Synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 6225–6234.
- Zheng, J.; Zheng, Q.; Fang, L.; Liu, Y.; and Yi, L. 2023. CAMS: Canonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 585–594.
- Zhou, K.; Bhatnagar, B. L.; Lenssen, J. E.; and Pons-Moll, G. 2022. Toch: Spatio-Temporal Object-to-Hand Correspondence for Motion Refinement. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 1–19.