This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Physics-aware Hand-object Interaction Denoising

Haowen Luo[1], Yunze Liu[1,3], Li Yi[1,2,3]
[1]Tsinghua University, [2]Shanghai Artificial Intelligence Laboratory, [3]Shanghai Qi Zhi Institute

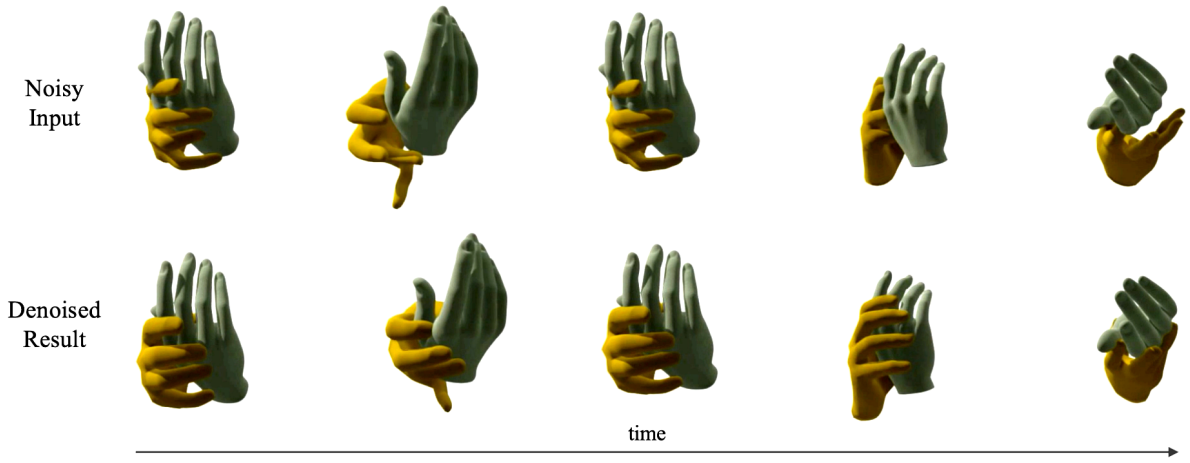

Figure 1. Given a noisy hand-object interaction sequence, our method produces de-noised hand poses conditioning on the object trajectory, mitigating physically-implausible artifacts such as erroneous contact and hand-object penetration. In this example of a human hand (yellow) manipulating a hand model (grey), the de-noised result demonstrates higher physical plausibility. Please see our supplementary material for more annimated results.

Abstract

The credibility and practicality of a reconstructed hand-object interaction sequence depend largely on its physical plausibility. However, due to high occlusions during hand-object interaction, physical plausibility remains a challenging criterion for purely vision-based tracking methods. To address this issue and enhance the results of existing hand trackers, this paper proposes a novel physically-aware hand motion de-noising method. Specifically, we introduce two learned loss terms that explicitly capture two crucial aspects of physical plausibility: grasp credibility and manipulation feasibility. These terms are used to train a physically-aware de-noising network. Qualitative and quantitative experiments demonstrate that our approach significantly improves both fine-grained physical plausibility and overall pose accuracy, surpassing current state-of-the-art de-noising methods.

1. Introduction

Hand pose tracking during hand-object interaction is a crucial task for various applications such as gaming, virtual reality, and robotics. Vision-based hand tracking methods have made significant progress in recent years by estimating hand poses from vision data sequences. However, heavy occlusions often occur during hand-object interaction, leading to ambiguity for vision-based trackers. Consequently, even state-of-the-art tracking methods still produce obvious errors and generate physically implausible artifacts such as inter-penetrating hand-objects and unrealistic manipulation. It is critical to remove such artifacts, increase the physical plausibility, and ensure the usefulness of tracking results.

Several previous works have proposed to post-process the estimations generated by vision-based trackers using de-noising techniques. Some works have attempted to optimize hand poses by minimizing low-level penetration and attraction energy [5]. However, such optimization methods may struggle to handle severe noise, as they easily get stuck

at local optima. Other works instead train neural networks for de-noising purposes [9][31], leveraging data-driven pose or contact priors to correct potentially significant pose errors. Nevertheless, purely data-driven approaches rely on the quality of the training dataset labels to achieve satisfactory visual effects and physical plausibility, and may be susceptible to overfitting and over-smoothing. Additionally, there is no guarantee that the resulting neural network is physically-aware.

In an effort to overcome the limitations of existing methods, we propose to combine data-driven de-noising with explicit modeling of the physical plausibility. In particular, such modeling needs to cover two essential aspects: i) grasp credibility, which demands that the hand pose in each frame be realistic given the object's geometry, avoiding interpenetration in particular; ii) manipulation feasibility, which considers the object's movement and requires proper hand-object contact that can plausibly explain the object's trajectory through hand manipulation, complying with physical laws.

Incorporating physics-based constraints into data-driven de-noising necessitates reshaping the loss landscape, enabling the network to learn to de-noise hand motions while adhering to physical constraints. While this idea appears straightforward, achieving it is difficult due to the intricate and non-differentiable process of verifying the plausibility of hand motions. Furthermore, when the de-noising algorithm violates physical constraints, it is necessary to provide a suitable path to guide it back to feasible motions. Meeting this requirement is even more challenging.

To address the aforementioned challenges, we introduce neural physical losses for assessing grasp credibility and manipulation feasibility, respectively. These losses are differentiable and can approximate non-differentiable and computationally intensive physical metrics effectively. Furthermore, they not only differentiate physically invalid hand motions from valid ones but also offer good projection directions to correct physically implausible hand motions. We integrate these neural physical losses into a novel hand motion denoising framework. Specifically, we design a denoising auto-encoder that operates on a dual hand-object-interaction representation, along with a two-stage training process that effectively balances physical constraints and data priors.

To demonstrate the effectiveness of our designs, we conduct experiments on both data with synthetic errors and actual errors caused by trackers and achieve both qualitative and quantitative improvements over the previous state of the arts. To sum up, our main contributions are:

First, we propose a physically-aware hand-object interaction de-noising framework which nicely combines data priors and physics priors.

Second, we introduce differentiable neural physical losses to model both grasp credibility and manipulation feasibility to support end-to-end physically-aware de-noising.

Third, we demonstrate the generalization of our neural physical losses regarding object, motion, and noise pattern variations, and show their effectiveness on different benchmarks.

2. Related work

Hand reconstruction and tracking The problem of reconstructing 3D hand surfaces from RGB or depth observations has garnered considerable attention in research. The existing body of work can be broadly classified into two distinct paradigms. Discriminative approaches focus on directly estimating hand shape and pose parameters from the observation, employing techniques such as 3D CNNs and volumetric representations [1, 6, 8, 17, 30]. In contrast, generative approaches adopt an iterative optimization process to refine a parametric hand model, iteratively aligning its projection with the observed data [23, 25, 26]. While recent advancements have explored more challenging scenarios, such as reconstructing two interacting hands [18, 22, 28], these approaches often overlook the presence of objects, leading to decreased reliability in interaction-intensive scenarios. The absence of object-awareness greatly limits their ability to accurately reconstruct hand surfaces in complex and dynamic environments.

Hand pose denoising The goal of hand pose denoising is to improve the reliability and accuracy of the hand pose estimation or tracking system, enabling more robust and realistic hand motion analysis and interaction in applications such as virtual reality, augmented reality, robotics, and human-computer interaction. TOCH [31]improves motion refinement by establishing spatio-temporal correspondence between objects and hands. GraspTTA [15] utilizes contact consistency reasoning to generate realistic and stable human-like grasps. D-Grasp [7] generates physically realistic and dynamic grasps for interactions between the hand and objects. Grabnet [24] aims to create accurate and visually realistic hand mesh models while interacting with previously unseen objects.

Physical plausibility in hand-object interaction The physical plausibility during hand-object interaction has been studied by many previous works. [29] uses a neural network to learn from human motion and synthesize physics-plausible manipulation. [7] and [27] generate a physics-based hand control policy with deep reinforcement learning to reach specific grasping or moving goals. While these works focus on synthesizing hand motions, some works, such as [2, 9, 14, 16, 19], explore reconstructing or refining hand-object interaction leveraging physics priors,

which is more relevant to our work. However, these works have limitations such as being limited to static grasps, relying on purely data-driven approaches and assuming only finger tips as source of forces.

3. Method

In this section, we describe our method for physically-aware hand pose de-noising. We begin by introducing the problem and outlining our approach. Given a potentially noisy hand pose trajectory during human-object interaction, represented by a sequence of hand meshes over T frames denoted by $\widetilde{H} = (\widetilde{H}^i)_{1\leq i\leq T}$ with $\widetilde{H}^i \in \mathbb{R}^{K\times 3}$, our method conditions on the object's geometry and dynamic information to refine the input and get more physically-plausible results $\hat{H} = (\hat{H}^i)_{1\leq i\leq T}$. We consider hand-object interaction sequences containing a single hand and a single rigid object. Let $O = (O^i)_{1\leq i\leq T}$ with $O_i \in \mathbb{R}^{L\times 3}$ denote object vertices over T frames.

We choose to only de-noise the hand poses in our setting while assuming accurate object trajectory because tracking rigid objects is much easier than tracking articulated rigid body with many joints like hands, especially for marker-based motion capture systems popularly used in laboratories and industry. What's more, beyond refining tracking results for traditional hand object interaction reconstruction tasks, our setting is also critical for broader applications such as virtual object manipulation and interaction retargeting in VR/AR and gaming, where clean object trajectory is usually accessible.

To combine data priors and physics priors in the de-noising process, we introduce a dual representation of hand-object interaction $F = (F^i)_{1\leq i\leq T} = (\mathcal{F}(H^i, O^i))_{1\leq i\leq T}$ that enables physical reasoning besides capturing data priors about hand-object interaction. We elaborate on this representation in Section 3.1. Using F as an intermediate representation, we refine hand poses via mapping noisy representation $\widetilde{F}$ to its correct version $\hat{F}$ with a de-noising auto-encoder. Using the refined representation $\hat{F}$, the corrected hand pose sequence $\hat{H}$ is fitted. Details regarding the architecture of our de-noising network and the de-noising framework can be found in Section 3.1.

To produce more physically plausible results, during the training of the de-noising network, besides traditional data losses, we propose two neural physical loss terms for assessing grasp credibility and manipulation feasibility. They are capable of conducting physical reasoning given the HOI representation F and producing assessment scores differentiable to F. What's more, with carefully designed training scheme and training targets, these loss terms form smooth landscape, and therefore yield contributive signals that effectively guide the de-noising network to gain physical-awareness and produce more physically-plausible results. Section 3.2 and 3.3 describe the construction of these two loss terms. And we describe their usage in the training process in Section 3.1.

3.1. Training framework

Dual HOI representation Instead of directly using hand vertices as the representation in the de-noising process, we propose a dual hand-object interaction representation that combines holistic hand pose information and fine-grained hand-object relation. This intermediate representation bridges explicit hand mesh vertices and physics-related knowledge learned by the proposed neural loss terms, allowing knowledge to flow between them and eventually improve the physical plausibility of the refined hand pose. On the one hand, by emphasizing hand-object relation at two different granularity, this dual representation enables the physical loss terms to reason about contact and deep penetration cases and better evaluate the physical plausibility of a given hand pose. On the other hand, with its focus on the physics-related characters of hand poses, when used to fit the explicit hand pose representation, this proposed representation can effectively improve vital aspects of the hand pose regarding physical plausibility.

Given hand vertices $H = (H^i)_{1\leq i\leq T}$ of hand MANO [21] meshes and object vertices $O = (O^i)_{1\leq i\leq T}$ over T frames, the whole HOI representation $(F^i)_{1\leq i\leq T}$ we use is denoted as:

$$F^i = (S^i, \mathcal{T}^i)$$

We regress the 21 hand key points $(S^i)_{1\leq i\leq T}$ with $S^i \in \mathbb{R}^{21\times 3}$ from hand vertices and model object-centric hand-object correspondence with the implicit field proposed by [31], that is, $(\mathcal{T}^i)_{1\leq i\leq T}$ where $\mathcal{T}^i = \{\mathcal{C}^i_j\}^N_{j=1} = \{(m^i_j, d^i_j, p^i_j)\}^N_{j=1}$. With N points randomly sampled on the object surface in the i-th frame, $\mathcal{C}^i_j$ represents the corresponding hand point of the j-th object point. To find hand-object correspondence, we cast rays in the object surface's normal directions from the sampled points and consider hand points the rays first hit, as the corresponding points. m^i_j is set to 1 if the j-th object point has corresponding hand point and is 0 otherwise. d^i_j denotes the distance between the j-th object points and its corresponding hand point, while $p^i_j \in \mathbb{R}^3$ encodes the semantic information of the hand point encoded as its position in the canonical space. After obtaining the refined representation, we can fit hand mesh to the representation and obtain the denoised hand poses. We show the inference process in Figure 2.

De-noising anto-encoder We use an auto-encoder as our backbone, which consumes noisy data in the form of our dual HOI representation $\widetilde{F}$ and produces the corrected version $\hat{F}$. Our de-noising network adopts the PointNet [20] architecture to process the the TOCH field part $\mathcal{T}^i$ of the input representation into a global object feature for the i-

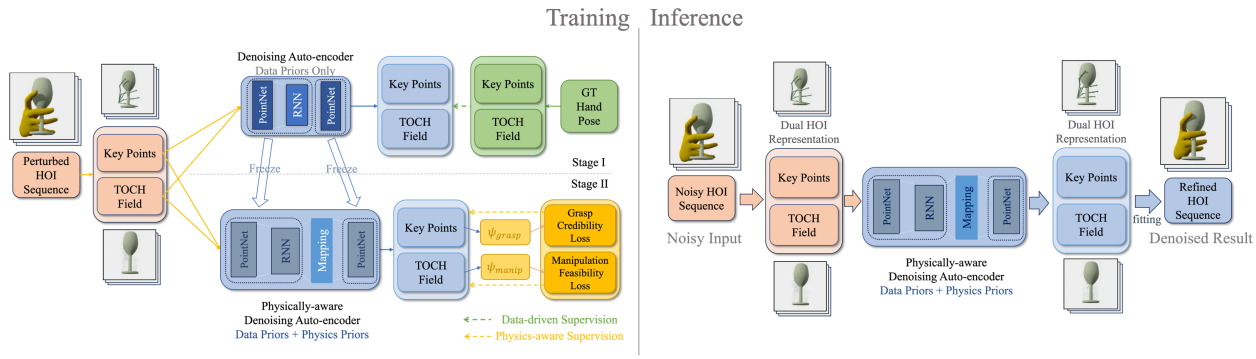

Figure 2. Overview of the training and inference frameworks.

th frame, which is then concatenated with the hand-centric part S^i in our representation to form the HOI frame feature x^i. $(x^i)_{1\leq i\leq T}$ is considered as time series data and further processed by a RNN, whose output is used to decode the refined representation $\hat{F}$.

Training process To obtain a physically aware de-noising network, we train the auto-encoder in two stages as shown in Figure 2. In the stage I, we train the model with supervision from ground truth data, learning a projection mapping from noisy hand pose representation to the clean data manifold. The training loss of stage I can be expressed as:

$$\mathcal{L}_{\mathrm{I}} = \alpha_{\mathcal{T}}\mathcal{L}_{\mathcal{T}}(\hat{\mathcal{T}}, \mathcal{T}_{\mathrm{GT}}) + \alpha_{S}\mathcal{L}_{S}(\hat{S}, S_{\mathrm{GT}})$$

where $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{S}$ measure the difference between the refined representation $\hat{F}$ and the ground truth representation F_{GT}. To reinforce the physical awareness of the de-noising network, we introduce the two physical loss terms in stage II. Specifically, we freeze the auto-encoder trained in stage I, but plug in a mapping layer ϕ between the encoder and the decoder, which is trained with the neural physical losses in stage II. ϕ is a MLP with residual connection that consumes latent vectors produced by the encoder, together with the linear accelerations of the object obtained with finite difference method. As the dual representation is in object-frame, it is necessary to introduce of linear acceleration to enable ϕ to reason about manipulation feasibility. We enable ϕ to capture the physical reasoning enforced by the two losses, therefore reshaping the clean hand pose manifold learned in stage I. The training loss of stage II integrates three components:

$$\begin{aligned}\mathcal{L}_{II} = \alpha_{\text{grasp}}\mathcal{L}_{\text{grasp}}(\hat{S}, O) + \alpha_{\text{manip}}\mathcal{L}_{\text{manip}}(\hat{\mathcal{T}}) \\ + \alpha_{\text{reg}}||\phi(v, a) - v||_2 + \alpha_{\text{GT}}\mathcal{L}_{\text{I}}\end{aligned}$$

where $\mathcal{L}_{\text{grasp}}$ and $\mathcal{L}_{\text{manip}}$ are the learned physical losses to be explained in the following sections, while v denotes the latent vector produced by the encoder and $a = (a^i)_{1\leq i\leq T}$ denotes the linear accelerations of the object. The regulation term ensures proximity between the mapped and the original latent vector. Together with supervision from ground truth hand pose, it helps to maintain the data priors learned in the first stage.

3.2. Grasp credibility loss

Given a single frame from a HOI sequence, with no knowledge about the movement of the object, the geometry relation between hand and object is the most important clue for humans to evaluate the plausibility of such a frame. To be specific, the hand pose must conform to the object's geometry, forming a plausible grasp and avoiding penetration or collision. While many recent works, such as [13], [10] and [4], focus on mitigating hand-object penetration during hand pose estimation, they tend to be helpless when dealing with deep penetration cases where the hand penetrates through the object completely instead of just entering the object surface, which are common when dealing with thin and delicate object parts.

When deep penetration happens, the direction to move the vertices to resolve penetration becomes ambiguous as the hand mesh lies on both side of the object, and it is very challenging to have a differentiable term producing correct gradients. While our method understands how to handle deep penetration through data prior learned from a novel penetration depth metric PD.

PD quantifies the severity of penetration of a noisy hand pose $\widetilde{H}^i$, which is paired with the corresponding clean hand pose H^i, with respect to the object mesh vertices O^i. The metric is computed by comparing $\widetilde{H}$ with the ground truth hand mesh vertices.

To compute this penetration depth metric, we focus on the hand vertices in $\widetilde{H}$ whose counterparts in H are in contact with the object. Specifically, we consider hand points with distances less than a threshold $c_{\text{contact}} = 2\text{mm}$ to the object surface as contact points. For C contact points $\{h_i\}_{i=1}^{C}$ on the ground truth hand mesh, whose counterparts on the evaluated mesh are denoted as $\{\widetilde{h}_i\}_{i=1}^{C}$, we compute their shift between two meshes as $\{\vec{h}_i = \widetilde{h}_i - h_i\}_{i=1}^{C}$. At

the object contact points $\{o_i\}_{i=1}^{C}$, we compute the normal $\{\vec{n}_i\}_{i=1}^{C}$ of the object surface. Then the metric is computed as:

$$PD(\widetilde{H}^i, H^i \mid O^i) = (\max_{\substack{j=1,2,\ldots,C \\ ||\vec{h}_j - (\vec{n}_j \cdot \vec{h}_j)\vec{n}_j|| < c_{\text{tangent}}}} (-\vec{n}_j \cdot \vec{h}_j))^+$$

where x^+ denotes $\max(0, x)$ and c_{tangent} is set to 1cm empirically so that only hands vertices with small shift in directions orthogonal to the object surfaces are considered for penetration evaluation, avoiding false positive cases where the hand vertices shift so much that their positions with respect to the object change completely. This definition can be considered as an approximation of the penetration depth at the most severe penetration position.

Compared with previous works that measure the intersection volume or count hand vertices inside the object mesh to evaluate hand-object inter-penetration, our metric better reflects the severity of deep penetration cases commonly encountered with thin and delicate objects, while existing loss terms only leverage local geometry information and tend to get stuck at local minima when exploited to remedy penetration.

This metric, however, isn't differentiable to our intermediate hand-object interaction representation and requires the ground truth hand pose to be computed, hence can't be directly used for improving the geometry credibility of results produced by our de-noising network. We train a physics-aware neural network ψ_{grasp} with a PointNet-like backbone that consumes hand skeletons and object point clouds to produce prediction results p between 0 and 1, where 0 indicates no severe penetration and 1 indicates the opposite. We introduce a threshold c_{PD} use the comparison result of $PD > c_{\text{PD}}$ as a hard target, as well as the original PD as a soft training target to encourage smooth prediction output. The loss for training $\mathcal{L}_{\text{grasp}}$ can be defined as:

$$\begin{aligned}\mathcal{L}_{\text{grasp train}}(p, PD) &= \alpha_{\text{grasp hard}}\text{BCE}(p, b_{\text{hard}}) \\ &\quad + \alpha_{\text{grasp soft}}\text{BCE}(p, b_{\text{soft}})\end{aligned}$$

$$b_{\text{hard}} = \mathbf{1}_{PD \geq c_{\text{PD}}}, \quad b_{\text{soft}} = 1 - e^{-c_{\text{soft}} * PD}$$

We set $c_{\text{soft}} = \frac{ln(2)}{c_{\text{PD}}}$ such that $b_{\text{soft}} = 0.5$ when $PD = c_{\text{PD}}$ to make sure that b_{hard} and b_{soft} are consistent. We set $c_{\text{PD}} = 1.5\text{cm}$ empirically. $\text{BCE}(\cdot, \cdot)$ denotes the binary cross entropy function.

As shown in Figure 3, the combination of soft target and hard target during the training allows $\mathcal{L}_{\text{grasp}}$ to distinguish deep penetration cases smoothly, improving its usability as a loss term.

We also attempt to improve the generalization ability of $\mathcal{L}_{\text{grasp}}$ and the loss landscape smoothness by assuring data variation in its training dataset. Using a HOI dataset containing ground truth data $\{(H^i, O^i)\}_{i=1}^{D}$, we first get its perturbed version $\{(\widetilde{H}^i, O^i)\}_{i=1}^{D}$ by adding Gaussian noise to the MANO parameters. To assure variation in noise magnitude, we further conduct linear interpolation between the MANO parameters of the ground truth data and the perturbed data to get m hand poses from each perturbed-clean hand pose pair. The dataset we finally obtain can be expressed $\bigcup_{i=1,2,\ldots,D}\{(\breve{H}_j^i, O^i)\}_{j=1}^{m}$ where $\breve{H}_j^i$ denotes the j-th interpolation result between $\breve{H}_1^i = H^i$ and $\breve{H}_m^i = \widetilde{H}^i$.

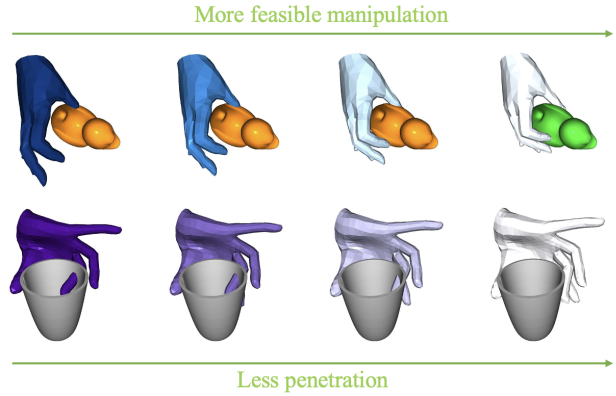

Figure 3. The proposed grasp credibility loss and manipulation feasibility loss can help quantify the physical plausibility of hand pose estimation results in a smooth way. The darkness of hand colors indicates the value of our proposed neural losses on frames with different noise levels. In this example, as the noise level of hand poses decreases, our proposed neural losses also decrease gradually, despite the discrete nature of the hand-object interaction (contact vs non-contact), providing smooth guidance for training of physics-aware de-noising network.

3.3. Manipulation feasibility loss

For a HOI sequence to be realistic, besides grasp credibility that only focuses on the single frame, whether the object's movement seems feasible also matters. To this end, we propose two manipulation feasibility metrics, force error (FE) and manipulation expense (ME), that evaluate whether the given hand-object contact can feasibly move the object along its trajectory. The two metrics are used as hard target and soft target respectively to train the neural physical loss term $\mathcal{L}_{\text{manip}}$. The two metrics, as well as the neural loss term $\mathcal{L}_{\text{manip}}$, take the hand-object correspondence part $\mathcal{T}$ of our representation and the object's linear acceleration a as input.

In the first force error metric, we measure to what extent can the object's movement be explained by forces applied at the the contacts within the corresponding friction cones.

Given an implicit field $\mathcal{T}^i$, we consider the M contact points among the randomly sampled N points on the object surface, and denote the object surface normal at these contact points as $\{\vec{n}\}_{j=1}^{M}$. Let $\vec{F} = m_0(-\vec{g} + \vec{a})$ denote the force needed for the object with mass m_0 to achieve its acceleration $\vec{a}$ when subject to gravity $m_0\vec{g}$. We ob-

tain $\vec{a}$ with finite difference method. And let $\{\vec{f_j}\}_{j=1}^{M}$ with $f_j \in \mathbb{R}^3$denotes the set of contact forces applied to the object that we solve for. We require the forces to lie in the corresponding friction cones specified by the coefficient of static friction μ. Then the force error metric can be expressed as:

$$FE(\{\vec{n}_j\}_{j=1}^{M}, \vec{F}) = \min_{\frac{\vec{f_j}}{||\vec{f_j}||}\cdot(-\vec{n}_j) \geq \sqrt{\frac{1}{1+\mu^2}}} ||\sum_{j=1}^{M} \vec{f_j} - \vec{F}||$$

Notice that FE is always between 0 and 1, and ideally, for a physically plausible frame, FE should be 0. Since the object mass is only a relative value for solving forces, therefore doesn't affect resultant force distribution of the optimization process, and we only care about the relative force error instead of the absolute force value, m_0 can be set to any non-zero constant. We set $m_0 = 1\text{kg}$. μ is set to 0.8 empirically following the common practice in previous works such as [14, 29]. In practice, we find that as long as the selected friction coefficient isn't too off, the result of force error and manipulation expense can align well with human perception concerning the manipulation feasibility.

While this force error metric can be used as a binary result to verify whether the movement of the object is feasible given a certain hand pose, a major drawback is that its value doesn't correctly reflect how infeasible a hand pose is. However, to obtain a loss term with smoother landscape, a metric with continuous result indicating the degree of manipulation feasibility would be more favorable. Therefore, we propose the manipulation expense metric that evaluates the distance between the given hand pose and the closest feasible hand pose.

Intuitively, this manipulation expense metric considers all the plausible force distribution maps that yield the required total force, and find the one that best match the current contact map, in that least forces are applied at object points which are actually far from the hand. The difference between the current contact map and its best match found in the above manner can reflect the quality of the current contact map regarding manipulation feasibility.

In this metric, we consider all N sampled object points in $\mathcal{T}^i$, and d_j denotes the signed distance between the j-th sampled object point and its corresponding hand point. Let $\{\vec{f_j}\}_{j=1}^{N}$ with $\vec{f_j} \in \mathbb{R}^3$ denote potential forces exerted at the N sampled points, the manipulation expense metric can be expressed as:

$$\begin{aligned} & ME(\{\vec{n}_j\}_{j=1}^{N}, \{d_j\}_{j=1}^{N}, \vec{F}) \\ = & \min_{\substack{\frac{\vec{f_j}}{||\vec{f_j}||}\cdot(-\vec{n}_j) \geq \sqrt{\frac{1}{1+\mu^2}} \\ \sum_{j=1}^{N} \vec{f_j} = \vec{F}}} \sum_{j=1}^{N} ||\vec{f_j}||\cdot(|d_j| - c_{\text{contact}})^+ \end{aligned}$$

We calculate FE and ME by solving for $\{\vec{f_j}\}_{j=1}^{M}$ and $\{\vec{f_j}\}_{j=1}^{N}$ through optimization processes respectively. Please refer to our supplementary material for details. We train a PointNet-like neural predictor ψ_{manip} that produces a output q between 0 and 1, to form the manipulation feasibility loss. The training loss of ψ_{manip} is formed as:

$$\begin{aligned} \mathcal{L}_{\text{manip_train}}(q, FE, ME) = & \ \alpha_{\text{manip hard}} \, \text{BCE}(q, s_{\text{hard}}) \\ & + \alpha_{\text{manip soft}} \, \text{BCE}(q, s_{\text{soft}}) \end{aligned}$$

$$s_{\text{hard}} = \mathbf{1}_{FE \geq c_{FE}}, \quad s_{\text{soft}} = \mathbf{1}_{FE \geq c_{FE}} \cdot (0.5 + \frac{arctan(ME)}{\pi})$$

$\alpha_{\text{manip hard}}$ and $\alpha_{\text{manip soft}}$ are constant weights. We use the same training dataset for $\mathcal{L}_{\text{manip}}$ as the one used to train $\mathcal{L}_{grasp}$.

4. Experiments

In this section we evaluate the proposed method on data with synthetic noise and actual tracking noise. We first introduce the datatsets (Section 4.1) and the evaluation metrics (Section 4.2). Results of our approach on correcting synthetic tracking error and refining results from vision-based trackers are in (Section 4.3) and (Section 4.4) respectively. Finally, we conduct ablation studies to evaluate the advantages of our physically-aware loss design in Section 4.5.

4.1. Datasets

GRAB. We train our de-noising network and the two neural loss terms on GRAB [24], a MoCap dataset for whole-body grasping of objects containing 51 objects from [3]. We follow the recommended split and select 10 objects for evaluation and testing.

HO-3D. HO-3D [11] is a dataset of hand-object interaction videos captured by RGB-D cameras, paired with frame-wise annotations of 3D hand poses and object poses. We use HO-3D to evaluate how well our pipeline trained with synthetic pose errors can generalize to real tracking errors produced by vision-based trackers. We use the second official release of HO-3D.

4.2. Metrics

Mean Per-Joint Position Error (MPJPE) and **Mean Per-Vertex Position Error (MPVPE)**. We report the average Euclidean distances between refined and ground truth 3D hand joints and vertices. These metrics measure the accuracy of hand poses and shapes.

Intersection Volume (IV). This metric measures volume of the intersection between the hand mesh and the object mesh. It reflects the degree of hand-object interpenetration.

Table 1. Quantitative results on refining GRAB dataset with sythetic noise. T-0.01 denotes the dataset with translation noise complying the distribution $\mathcal{N}(0, 0.01)$, θ-0.3 denotes the dataset with pose noise complying the distribution $\mathcal{N}(0, 0.3)$, other noise patterns are denoted similarly. We additionally add $\Delta r \sim \mathcal{N}(0, 0.05)$ to the global orientation in all the datasets. We train the network only on the "T-0.01, θ-0.3" dataset and test it on datasets with different noise patterns. Our method is robust to different noise patterns and magnitudes, especially in its ability to produce physically plausible results.

	T-0.01	T-0.02	θ-0.3	θ-0.5	T-0.01 θ-0.3	T-0.02 θ-0.5
MPJPE	16.01 → **6.91**	27.83 → **10.01**	7.52 → **5.61**	8.74 → **6.38**	18.86 → **7.39**	31.49 → **11.97**
MPVPE	16.32 → **6.24**	28.43 → **10.46**	**6.23** → 6.31	8.85 → **6.97**	19.13 → **6.78**	33.25 → **11.24**
contact IoU	3.31 → **24.82**	2.39 → **21.77**	5.45 → **25.14**	4.46 → **24.18**	3.62 → **23.94**	2.47 → **20.50**
IV	0.91 → **0.90**	1.43 → **1.10**	**0.88** → 0.92	1.77 → **0.94**	**0.87** → 1.13	2.35 → **1.12**
PD	0.77 → **0.32**	0.85 → **0.43**	0.74 → **0.44**	1.22 → **0.47**	0.80 → **0.44**	1.56 → **0.45**
plausible rate	0.42 → **0.95**	0.33 → **0.92**	0.45 → **0.93**	0.43 → **0.93**	0.42 → **0.91**	0.31 → **0.90**

Penetration Depth (PD). To better measure cases of deep penetration, we also report the penetration depth metric proposed in Section 3.2. Since the ground truth hand poses of HO-3D are withheld, we cannot calculate PD on HO-3D and only report it on GRAB.

Contact IoU. This metric assesses the Intersection-over-Union between the ground truth contact map and the predicted contact map. The contact maps are obtained by considering object vertices within $2\,mm$ from the hand as in contact. This metric is also reported on GRAB only.

Plausible Rate. To consider both grasp credibility and manipulation feasibility for a holistic assessment regarding physical plausibility, we use this metric that reflects both aspect. For a given frame of hand pose to be plausible, i) the PD metric should be less than $1.5\mathrm{cm}$ (this threshold is chosen following [29]), ii) the force error metric proposed in 3.3 should be less than 0.1. Since PD is not available for HO-3D, we only consider ii) for evaluation on HO-3D.

4.3. Refining sythetic Error

To employ our de-noising method in more realistic applications and achieve best potential performance, it would be ideal to train the model on the predictions of the target tracking method to be augmented. Yet this might lead to overfitting to tracker-specific noise, which is undesirable for generalization. Hence we trained the network on dataset with Gaussian noise of different composition and magnitude, which synthesizes the hand errors that tracking methods induce. Quantitative results are shown in Table 1 and qualitative results are presented in Figure 4.

4.4. Refining vision-based hand tracker

We also use our network to refine the results produced by state-of-the-art vision-based models on HO-3D dataset. To evaluate how well our method generalize to actual tracker error, both the de-noising auto-encoder and the two neural loss terms are trained only on GRAB with synthetic errors, and then used on HO-3D without further fine-tuning.

Figure 4. Qualitative results on GRAB dataset. We can see that TOCH produces physically implausible results such as hand-object penetration when dealing with thin and delicate parts of objects, while our results are more realistic.

Hasson *et al.* [12], a RGB-based hand pose tracker is used to predict hand poses on the test split of HO-3D dataset. TOCH [31], Physical Interaction [14] and our method are used to refine the results it produces. Physical Interaction is an optimization-based method that solves for contact forces at finger tips and the refined hand pose simultaneously. The results are shown in Table 2 and Figure 5. Physical plausibility is significantly improved, which is indicated by IV and plausibility rate.

Table 2. Quantitative results on HO-3D dataset. The hand joint and mesh errors are obtained after Procrustes alignment following the official evaluation protocol of HO-3D.

	MPJPE	MPVPE	IV	plausible rate
Hasson *et al.*	11.4	11.4	8.75	0.71
TOCH	10.9	11.3	7.24	0.73
Physical Intercation	11.4	11.6	8.33	0.83
Ours	10.7	11.2	5.95	0.85

Table 3. Ablation on the physical neural losses. The best result is highlighted in red, while the second-best result is highlighted in blue.

	MPJPE	MPVPE	contactIoU	IV	PD	plausible rate
GT	-	-	-	0.75	-	0.92
TOCH	13.17	12.24	21.83	2.24	0.55	0.78
Physical Interaction	13.5	14.6	7.15	1.94	1.27	0.89
Ours(w/o $\mathcal{L}_{\text{manip}}$)	7.41	6.78	23.96	0.96	0.41	0.82
Ours(w/o $\mathcal{L}_{\text{grasp}}$)	7.28	6.65	23.51	2.43	0.59	0.81
Ours	7.39	6.78	23.94	1.13	0.44	0.91

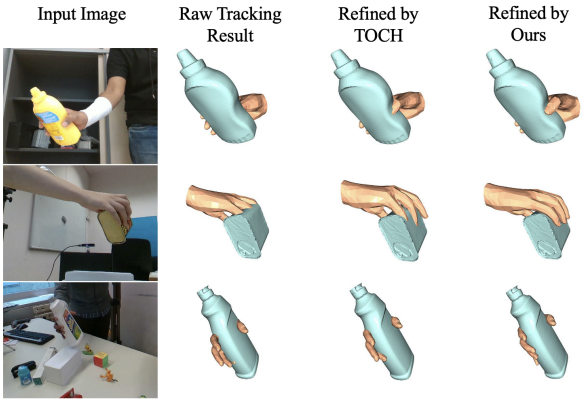

Figure 5. Qualitative results on HO-3D dataset. Our method effectively denoises tracking result, and produces more physically plausible hand-object interaction than TOCH.

4.5. Ablation Studies

Physically-aware loss. To demonstrate the advantages of our proposed physically-aware loss terms in modeling and improving physical plausibility, in stage II, we remove signals from the two neural loss terms perspectively and train two baseline de-noising networks. Comparison between them and our complete method on the GRAB dataset is presented in Table 3.

It can be observed that the highest overall plausible rate is achieved when combining the two neural losses druing training. Without the grasp credibility loss, the network tends to more actively adhere the hand to the object surface, increasing the contact area so that FE and ME can likely be reduced. However, the increased manipulation feasibility is achieved at the expense of more severe hand-object inter-penetration, indicated by IV and PD. Therefore, the overall plausible rate which considers both aspects is limited. Removing the manipulation feasibility loss induces the opposite result. While lowest IV and PD are attained, cases where the object is hovering in the air appear more frequently. This result reflects the network's tendency of avoiding inter-penetration regardless of the manipulation feasibility, when only signals from the grasp credibility is present during the training process.

Soft target for training neural losses. When training the physically-aware neural loss terms, we use both soft target and hard target for smooth loss landscape. To demonstrate the advantages of this design, we train baseline neural losses with hard target only and compare them with neural losses trained with both targets, and present comparisons in Table 4 regarding their classification performance on test set and ability of improving the de-noising network when exploited.

Table 4. We use the losses to discern implausible frames on the test split of Perturbed GRAB and report their F-scores. While the classification performance of losses trained with two types of target choices are close, neural losses trained with both targets yield better results when exploited.

(a) Target used to train grasp credibility loss.

	F-score	PD	IV
hard	0.85	0.49	1.52
hard + soft	0.93	0.44	1.13

(b) Target used to train manipulation feasibility loss.

	F-score	plausible rate
hard	0.90	0.84
hard + soft	0.88	0.91

5. Conclusion

We propose a physically-aware hand-object interaction de-noising framework which combines data priors and physics priors to generate plausible results.In particular, our differentiable neural physical losses effectively assess grasp credibility and manipulation feasibility of given hand poses and form smooth loss landscape for hand poses with different noise level, enabling physically-aware de-noising. Experiments demonstrate that our method generalizes well to novel objects, motions and noise patterns.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019. 2

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 6

[4] Thomas Buffet, Damien Rohmer, Loic Barthe, Laurence Boissieux, and Marie-Paule Cani. Implicit untangling: A robust solution for modeling layered clothing. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 4

[5] Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. *arXiv preprint arXiv:2209.12009*, 2022. 1

[6] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Mvhm: A large-scale multi-view hand mesh benchmark for accurate 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 836–845, 2021. 2

[7] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 2

[8] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2

[9] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 2

[10] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. Garnet: A two-stream network for fast and accurate 3d cloth draping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8739–8748, 2019. 4

[11] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3196–3206, 2020. 6

[12] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 7

[13] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 4

[14] Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 6, 7

[15] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. *arXiv preprint arXiv:2104.03304*, 2021. 2

[16] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9–16, 2013. 2

[17] Jameel Malik, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7113–7122, 2020. 2

[18] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019. 2

[19] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 2

[20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[21] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

[22] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[23] Srinath Sridhar, Helge Rhodin, Hans-Peter Seidel, Antti Oulasvirta, and Christian Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *2014 2nd International Conference on 3D Vision*, pages 319–326. IEEE, 2014. 2

[24] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 581–600. Springer, 2020. 2, 6

[25] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[26] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Transactions on Graphics (TOG)*, 36(6):1–12, 2017. 2

[27] Zeshi Yang, Kangkang Yin, and Libin Liu. Learning to use chopsticks in diverse gripping styles. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 2

[28] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021. 2

[29] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 2, 6, 7

[30] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2478–2482. IEEE, 2020. 2

[31] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 1–19. Springer, 2022. 2, 3, 7