VPHO: Joint Visual-Physical Cue Learning and Aggregation for Hand-Object Pose Estimation

Jun Zhou[1,2,3,4], **Chi Xu**[1,2,3*], **Kaifeng Tang**[1,2,3], **Yuting Ge**[1,2,3], **Tingrui Guo**[1,2,3], **Li Cheng**[4]

[1]School of Automation, China University of Geosciences, Wuhan 430074, China,
[2]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China,
[3]Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China,
[4]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada
{jchow, xuchi, tkf, gyting, guotingrui}@cug.edu.cn, lcheng5@ualberta.ca

Abstract

Estimating the 3D poses of hands and objects from a single RGB image is a fundamental yet challenging problem, with broad applications in augmented reality and human-computer interaction. Existing methods largely rely on visual cues alone, often producing results that violate physical constraints such as interpenetration or non-contact. Recent efforts to incorporate physics reasoning typically depend on post-optimization or non-differentiable physics engines, which compromise visual consistency and end-to-end trainability. To overcome these limitations, we propose a novel framework that jointly integrates visual and physical cues for hand-object pose estimation. This integration is achieved through two key ideas: 1) joint visual-physical cue learning: The model is trained to extract 2D visual cues and 3D physical cues, thereby enabling more comprehensive representation learning for hand-object interactions; 2) candidate pose aggregation: A novel refinement process that aggregates multiple diffusion-generated candidate poses by leveraging both visual and physical predictions, yielding a final estimate that is visually consistent and physically plausible. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches in both pose accuracy and physical plausibility.

Code — https://github.com/zhoujun-7/VPHO

Introduction

Hand-object pose estimation from single RGB images (Hasson et al. 2019) has broad applications across various fields, including virtual and augmented reality (Mueller et al. 2019; Chen et al. 2019), human-computer interaction (Ren and Bao 2020), and robotics (Billard and Kragic 2019). Most existing methods (Hampali et al. 2022; Liu et al. 2021; Lin et al. 2023; Wang, Mao, and Li 2023b) primarily rely on image-based visual cues to ensure that the 3D pose estimation is consistent with 2D observations (Zhou et al. 2024; Pavlakos et al. 2024). For example, segmentation consistency losses are widely adopted (Qi et al. 2024; Zhang et al. 2024; Xu et al. 2023) to align projected 3D meshes with 2D segmentation masks, while photometric consistency across frames is explored in (Hasson et al. 2020). However, these visual-based methods often neglect physical constraints, leading to physically implausible artifacts such as penetration or lack of contact. As illustrated in Figure 1, the result from HFL (Lin et al. 2023) appears visually reasonable in the original camera view but reveals improper grasping when rendered from a different viewpoint, as highlighted in the red circle.

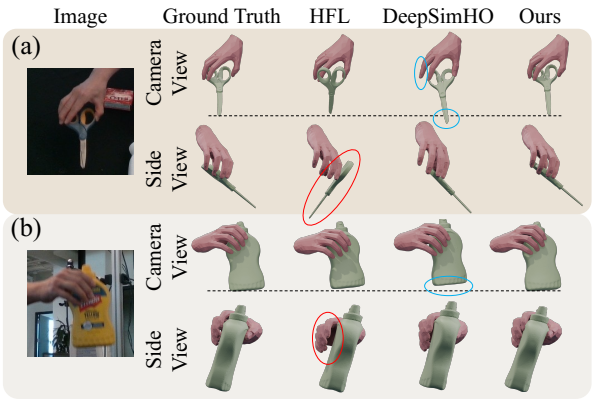

Figure 1: Visual comparison between a state-of-the-art visual-only method, HFL, and a method that incorporates physical cues, DeepSimHO. HFL yields visually aligned yet physically implausible results (red circles), while DeepSimHO improves physical plausibility at the cost of visual alignment (blue circles).

To improve physical plausibility, several works (Brahmbhatt et al. 2020; Chen et al. 2022; Yang et al. 2024a) introduce post-optimization strategies that incorporate physical constraints. While effective at reducing artifacts such as interpenetration or missing contacts, these methods are highly sensitive to the quality of the initial pose and often produce unstable or suboptimal outputs when the initialization is inaccurate. To better balance visual and physical fidelity, more recent approaches (Hasson et al. 2019; Ehsani et al. 2020; Yang et al. 2024b; Cho et al. 2023; Wang, Mao, and Li 2023a; Hu et al. 2024a) propose end-to-end training

*Corresponding author

methods that incorporate physical cues. While this integration improves physical plausibility, it often comes at the cost of degraded visual consistency. As shown in Figure 1, DeepSimHO (Wang, Mao, and Li 2023a) improves physical realism over its visual-only baseline (Yang et al. 2022); yet, the predicted object pose deviates from the image observation (blue circle), compromising pose accuracy.

To resolve these limitations, we propose a novel approach that effectively integrates visual and physical cues to ensure both visual consistency and physical plausibility. This is accomplished through two key ideas: 1) Joint visual-physical cue learning: Our model is trained to extract 2D visual cues (e.g., hand and object heatmaps) alongside 3D physical cues (e.g., force vectors), enabling richer representation learning for hand-object interactions. 2) Candidate pose aggregation: We propose a novel aggregation process that leverages both predicted visual and physical cues to aggregate multiple candidate poses generated by a diffusion model into a single, physically plausible and visually consistent estimate.

A core challenge lies in the prediction of 3D interaction forces due to: (1) their high dimensionality, (2) the complexity of contact dynamics and friction modeling, and (3) the lack of ground-truth force annotations. To address this, we introduce a **Force Prediction Module** that models local contact forces using friction cones and transforms them from local to global coordinates to compute the overall hand-object interaction force. The module is trained via a semi-supervised strategy using physical constraints and pseudo force labels, without requiring ground-truth annotations.

Pose aggregation also poses unique challenges due to the high degrees of freedom of the pose parameters (e.g., articulated hand joints) and the interdependence of hand-object interaction. To address this, we propose a two-stage aggregation scheme: 1) **Visual-based Aggregation**: Candidates are hierarchically aggregated along the kinematic chain using visual cues, which effectively reduces error accumulation and enhances visual consistency. 2) **Physics-based Aggregation**: Candidates are ranked and selected based on physical constraints, such as contact and torque balance, which improves contact quality and enhances physical plausibility.

In summary, the contributions of this work are as follows:

- We propose a novel hand-object pose estimation approach that integrates both visual and physical cues without compromising either.
- We introduce a force prediction module that models interaction forces via friction cones and local-to-global transformation, trained with physical constraints and pseudo force labels.
- We propose a two-stage pose aggregation strategy that leverages both visual and physical cues to yield accurate and physically plausible hand-object poses.

Extensive experiments on standard benchmarks demonstrate that our method achieves state-of-the-art performance in both pose accuracy and physical plausibility.

Related Works

Most existing efforts (Xu et al. 2023; Hasson et al. 2020; Tse et al. 2022a; Zhou et al. 2024; Potamias et al. 2025) estimate hand and object poses primarily by leveraging image-based visual cues, ensuring alignment between the 2D projections of the estimated 3D poses and the corresponding image observations. Several methods (Qi et al. 2024; Zhang et al. 2024; Xu et al. 2023) incorporate segmentation losses to enforce consistency between the projected 3D mesh and 2D image segmentations. Similarly, Park et al. (Park et al. 2022) employ a transformer module to inject hand information into occluded 2D regions. However, these methods predominantly focus on visual information and do not explicitly incorporate physical constraints, making them susceptible to physically implausible predictions such as interpenetration or lack of contact.

To address this limitation, several methods (Grady et al. 2021; Yang et al. 2024a; Zhao et al. 2024) propose post-optimization strategies that refine initial pose estimates by incorporating physical cues. For instance, Grady et al.(Grady et al. 2021) and Tse et al.(Tse et al. 2022b) infer desirable hand-object contact regions from initial poses and subsequently optimize both the hand and object to better conform to these inferred regions. Similarly, Hu et al.(Hu et al. 2022) refine hand-object interactions by adjusting fingertip forces and contact points based on initial motion trajectories. While these methods are effective in improving physical plausibility, they rely heavily on accurate initializations. Inaccurate initial poses can cause divergence from visual evidence, resulting in visually inconsistent estimations.

More recently, several studies have explored the unification of visual and physical reasoning within an end-to-end learning framework (Hasson et al. 2019; Cho et al. 2023; Wang, Mao, and Li 2023a; Hu et al. 2024a). Hu et al. (Hu et al. 2024b) model part-level and vertex-level contact probabilities to construct an implicit neural representation of the object, thereby facilitating object pose inference. Wang et al. (Wang, Mao, and Li 2023a) incorporate a physics engine into the training loop to supervise the learning of stability-aware poses based on simulated physical outcomes. While these methods represent progress toward unified visual-physical modeling, they often compromise visual fidelity in favor of physical plausibility. In contrast, our approach integrates visual and physical cues during inference by aggregating multiple candidate poses generated by a diffusion model, enabling the selection of solutions that are both visually consistent and physically plausible.

Method

The framework of our approach is illustrated in Figure 2.

Feature Extraction

Given an input RGB image, an enhanced ResNet50 backbone network (Lin et al. 2023) is employed to extract features for both the hand and the object. These features are subsequently processed by DeConv layers (Xiao, Wu, and Wei 2018) to generate the hand heatmap H^h and the object heatmap H^o, which serve as visual cues in pose aggregation module. Two residual blocks (He et al. 2016) further refine the hand and object features, preparing them for force prediction and candidate pose generation. The loss function

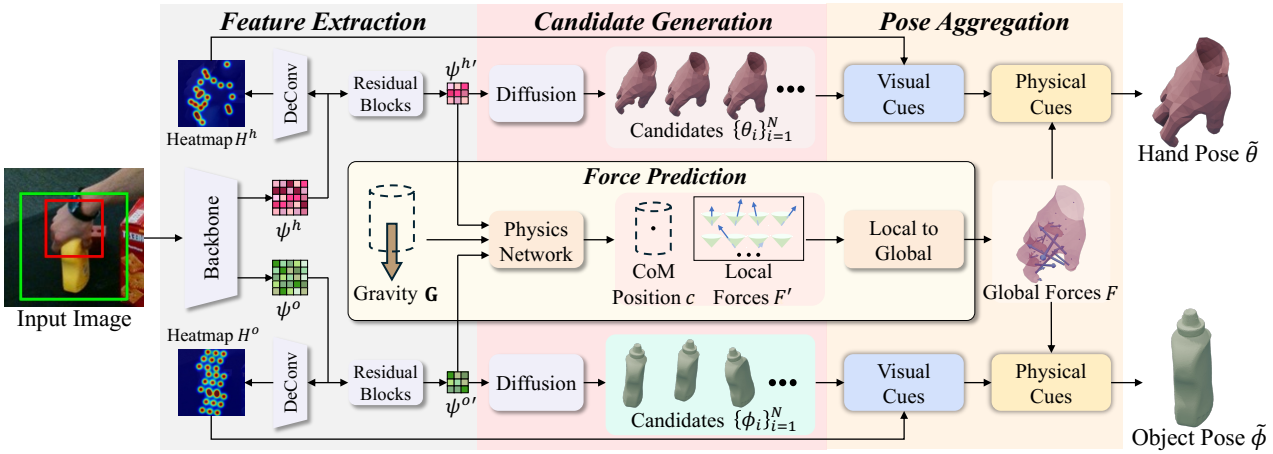

Figure 2: The framework of our approach, consisting of the following four modules: feature extraction, force prediction, candidate generation and pose aggregation.

used to supervise the heatmap predictions is defined as:

$$\mathcal{L}_{hm} = \lambda_{hm}\left(\mathcal{L}_{mse}(H^h, \bar{H}^h) + \mathcal{L}_{mse}(H^o, \bar{H}^o)\right), \quad (1)$$

where $\mathcal{L}_{mse}$ denotes the mean square error loss function, $H^h \in \mathbb{R}^{|J^h|\times h_m \times w_m}$ and $\bar{H}^h \in \mathbb{R}^{|J^h|\times h_m \times w_m}$ represent the predicted and ground truth heatmaps for the hand joints J^h, respectively. $H^o \in \mathbb{R}^{|J^o|\times h_m \times w_m}$ and $\bar{H}^o \in \mathbb{R}^{|J^o|\times h_m \times w_m}$ represent the predicted and ground truth heatmaps for the object keypoints J^o, respectively. h_m and w_m denote the height and width of the heatmap, and λ_{hm} is a hyperparameter that weights the heatmap loss. Detailed settings of all hyperparameters are provided in the extended version.

Force Prediction

We focus on hand-object pose estimation from a single image, where motion-related quantities such as acceleration cannot be inferred. Thus, we adopt the static equilibrium assumption as a necessary simplification, which is shown to be effective in our ablation studies.

Local Force Hand-object contact interactions are inherently complex (Hu et al. 2022). Following (Yang et al. 2024a), we approximate this complexity by representing contact forces through 32 sparse anchor points $\{O_k^a\}_{k=1}^{32}$ on the hand surface. Each anchor point O_k^a is associated with a local coordinate system, as depicted in Figure 3(a). According to Coulomb's friction law (Morin 2008), contact forces must lie within a friction cone determined by a friction coefficient μ. We model this cone using a set of base vectors $\{v_j\}_{j=1}^{N_v}$, where each vector is defined as:

$$v_j = \left(\mu \sin(\frac{2\pi j}{N_v}), \mu \cos(\frac{2\pi j}{N_v}), 1\right). \quad (2)$$

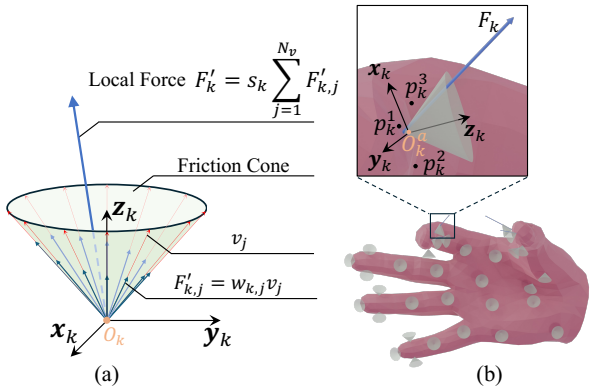

Figure 3: Friction cone and force representations in (a) local and (b) global coordinate frames.

The local force F'_k at anchor point O_k^a is expressed as a weighted sum of these base vectors:

$$F'_k = s_k \sum_{j=1}^{N_v} w_{k,j} v_j, \quad (3)$$

where $s_k \in \mathbb{R}$ and $w_{k,j} \in \mathbb{R}$ are the learned scaling and weighting coefficients, respectively.

Global Force As illustrated in Figure 3(b), the local force F'_k is transformed into a global force F_k using the MANO hand model (Romero, Tzionas, and Black 2017) through its linear blend skinning procedure. Each anchor point $\{O_k^a\}$ is attached to a triangle $\{\triangle p_k^1 p_k^2 p_k^3\}$ on the hand mesh vertices

V^h. The transformation is defined as:

$$\begin{cases} F_k = \mathbf{R}_k^{L2G} F_k', \\ O_k^a = \sum_{i \in \{1,2,3\}} \alpha_k^i p_k^i. \end{cases} \tag{4}$$

Here F_k' and F_k represent the local and global force for the anchor point O_k^a, respectively. $\mathbf{R}_k^{L2G} = [\mathbf{x}_k, \mathbf{y}_k, \mathbf{z}_k]$ denotes the rotation matrix, $\mathbf{x}_k$ is calulated by normalizing the vector $p_k^2 - p_k^1$, $\mathbf{z}_k$ is calulated by normalizing the vector $(p_k^2 - p_k^1) \times (p_k^3 - p_k^2)$, and $\mathbf{y}_k = \mathbf{z}_k \times \mathbf{x}_k$. O_k^a is the weighted sum of $\{p_k^i\}$, and α_k^i is the weight coefficient.

Physical Constraints Given the static equilibrium assumption (Morin 2008), all forces acting on the object must collectively satisfy Newtonian physical constraints. These constraints are used both to train the force prediction module and to guide the pose aggregation process.

Force balance: the sum of all external forces acting on it must be zero. This leads to the force balance constraint,

$$\mathcal{L}_{force} = \left\| \sum_{k=1}^{32} F_k + \mathbf{G} \right\|_2^2, \tag{5}$$

where $\mathbf{G}$ represents the gravity vector. Following (Hu et al. 2022), the magnitude of $\mathbf{G}$ is set to a relative value of 1N.

Torque balance: the sum of all torques acting on it must be zero. Thus, the torque balance constraint is formulated as,

$$\mathcal{L}_{torque} = \left\| \sum_{k=1}^{32} F_k \times r_k \right\|_2^2, \tag{6}$$

Here r_k refers to the position vector of the k-th anchor point, and $\times$ is a cross product between the two vectors.

Contact-force Relation: In hand-object interaction, an anchor point can exert force on an object only if it is in contact with the object's surface. Consequently, the magnitude of the contact force should be constrained based on the contact state between the anchor and the object. The distance between an anchor point and the object's surface serves as a key indicator of this contact state. Specifically, a shorter distance implies a higher likelihood of contact, whereas a larger distance suggests a lower probability of contact. Inspired by this, we approximate the contact-force constraint as

$$\mathcal{L}_{contact} = \sum_{k=1}^{32} \|F_k\|_2 \cdot |d_k|, \tag{7}$$

where $|d_k|$ denotes the distance between the anchor point and the object's surface, and $\|F_k\|_2$ represents the magnitude of the force. This constraint is an essential physical cue for our physics-based aggregation.

Physics Network As illustrated in Figure 4, our physics network takes 1) the hand feature $\psi^{h\prime}$, 2) the object feature $\psi^{o\prime}$, and 3) the gravity vector $\mathbf{G} \in \mathbb{R}^3$ as inputs. During training, $\mathbf{G}$ is aligned vertically downward with respect to the tabletop. During inference, we approximate $\mathbf{G}$ using the camera's y-axis, as in (Wang, Mao, and Li 2023a). The network outputs 1) the weight matrix $\mathbf{w} \in \mathbb{R}^{32 \times N_v}$, 2) the scaling vector $\mathbf{s} \in \mathbb{R}^{32}$, and 3) the object center-of-mass position

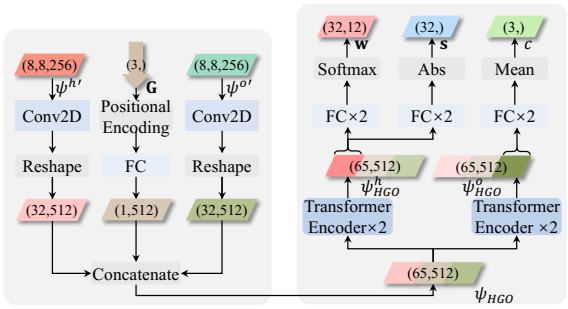

Figure 4: The architecture of our physics network.

$c \in \mathbb{R}^3$. The predicted local force F_k' is computed using $\mathbf{w}$ and $\mathbf{s}$ as defined in Equation 3. The network is supervised using the following loss:

$$\begin{aligned} \mathcal{L}_{phy} = \lambda_F \mathcal{L}_{mse}(F', \tilde{F}') + \lambda_c \mathcal{L}_{mse}(c, \bar{c}) \\ + \lambda_{force} \mathcal{L}_{force} + \lambda_{torque} \mathcal{L}_{torque}, \end{aligned} \tag{8}$$

where F' and $\tilde{F}'$ are the predicted forces and pseudo force labels, c and $\bar{c}$ are the predicted and ground-truth object center-of-mass positions, λ_F, λ_c, λ_{force}, and λ_{torque} are hyperparameters balancing the loss terms. The pseudo force labels $\tilde{F}'$ are precomputed via an optimization process constrained by the physical constraints, further details are provided in the extended version.

Pose Aggregation

Given a set of candidate hand poses $\{(\theta_i, \beta_{reg})\}_{i=1}^N$, where $\theta_i \in \mathbb{R}^{16\times3}$ are the MANO pose parameters of 16 hand joints and $\beta_{reg} \in \mathbb{R}^{10}$ is the MANO shape parameter, and a set of rigid object pose candidates, where $\{(R_i, T_i)\}_{i=1}^N$, $R_i \in \mathbb{R}^3$, $T_i \in \mathbb{R}^3$ represent rotation and translation respectively, the pose aggregation module is designed to refine these candidates through two sequential stages: visual-based aggregation and physics-based aggregation, as illustrated in Figure 2. The initial pose candidates are generated by a candidate generation module based on a score-based diffusion model (Song et al. 2021). Further details of this module are provided in the extended version.

Visual-based Aggregation In articulated systems such as the human hand, higher-level joint positions depend on lower-level ones due to kinematic dependencies. Consequently, errors tend to accumulate across levels, leading to increasingly inaccurate joint predictions at higher levels. To address this, we propose a level-by-level aggregation strategy that progressively refines joint parameters from lower to higher levels using heatmaps as visual guidance. As shown in Figure 5(a), hand joints with degrees of freedom are categorized into four hierarchical levels, with the index set of joints at level l denoted by L_l. The aggregation proceeds iteratively from level 1 to level 4 (Figure 5(b)). After aggregating each level, the refined joint parameters are used to overwrite the corresponding joints across all candidate poses. This reduces the propagation of errors to higher levels. For the j-th joint in level L_l, we compute a visual score

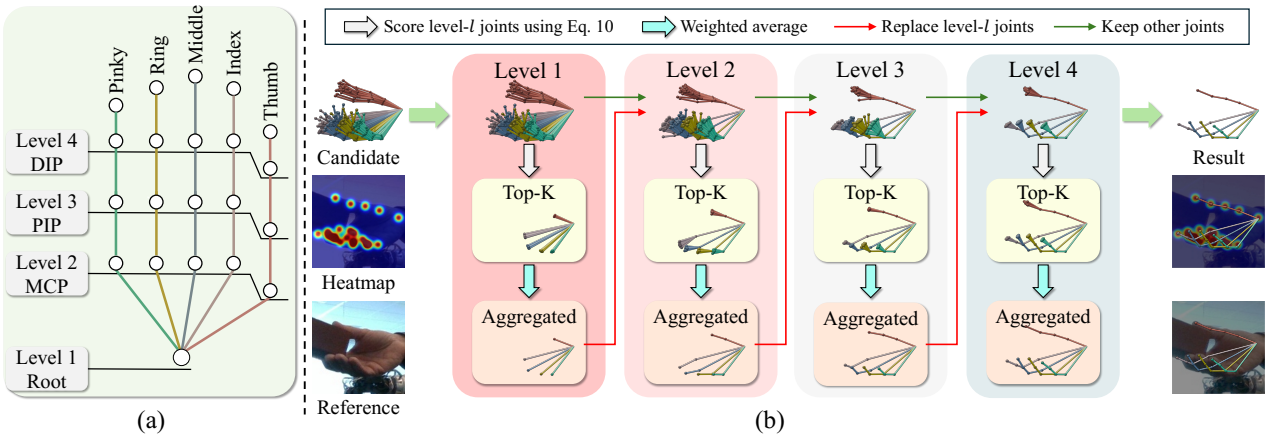

Figure 5: (a) The levels of hand pose parameters; (b) The visual-based aggregation hierarchically aggregate hand joints from lower to higher levels.

s_i^h for each hand pose candidate (θ_i, β_{reg}) using the following formula:

$$s_i^h = \sum_{c \in Children_j} H_c^h \left(\text{Proj2D}^h \left(c, \theta_i, \beta_{reg} \right) \right), \quad (9)$$

where $Children_j$ denotes the set of higher-level joints within the same kinematic chain as joint j, $\text{Proj2D}^h(\cdot)$ denotes the 2D projection of the c-th joint for candidate (θ_i, β_{reg}), and $H_c^h(\cdot)$ retrieves the heatmap value at the projected 2D location for joint c. Using these scores $\{s_i^h\}_{i=1}^N$, the top-K candidates $\{\theta_i\}_{i \in K^h}$ are selected. The aggregated pose parameter for joint j, denoted by $\tilde{\theta}[j,:]$, is computed as the weighted average of the selected candidates:

$$\tilde{\theta}[j,:] = \frac{\sum_{i \in K^h} s_i^h \theta_i[j,:]}{\sum_{i \in K^h} s_i^h}. \quad (10)$$

This aggregated value then replaces all $\theta_i[j,:]$ for $i = 1, \ldots, N$. This process is repeated for each level until all hand pose parameters are aggregated.

The same principle is applied to object pose aggregation. Object pose parameters are separated into two levels: 1) the translation parameter T_o, and 2) the rotation parameters R_o. The visual score for each object candidate is computed using:

$$s_i^o = \sum_{c=1}^{27} H_c^o \left(\text{Proj2D}^o \left(c, R_i, T_i \right) \right), \quad (11)$$

where $\text{Proj2D}^o(c, R_i, T_i)$ is the 2D projection of the c-th object keypoint after applying transformation (R_i, T_i) to the object model, and $H_c^o(\cdot)$ retrieves the heatmap value at the projected 2D location. Object translation parameters are aggregated first and used to overwrite the translation parameters of all candidates. Rotation parameters are subsequently aggregated using the updated translations.

Physics-based Aggregation To further enhance the physical plausibility of the hand-object interaction, we introduce a physics-based aggregation step that utilizes physical constraints to guide candidate pose aggregation. For the hand, we compute a physics-based score as:

$$s_{phy}^h = -\mathcal{L}_{force} \cdot \mathcal{L}_{contact}. \quad (12)$$

Method	Hand		Object		
	MJE	PA-MJE	MCE	OCE	ADDS
Liu et al. (2021)	15.2	6.58	-	-	-
HandOccNet (2022)	14.0	5.80	-	-	-
H2ONet (2023)	14.0	5.70	-	-	-
HandBooster (2024)	11.9	5.2	-	-	-
SimpleHand (2024)	12.4	5.5	-	-	-
HFL (2023)	12.6	5.47	48.0	42.7	33.8
HOISDF (2024)	10.1	5.13	35.8	27.6	18.6
Ours	**10.0**	**5.08**	**26.2**	**23.7**	**13.5**

Table 1: Comparison of pose accuracy on *DexYCB Full* (metrics are in mm).

Method	Hand		Object	
	PA-MJE	MJE	OCE	ADDS
Hasson et al. (2019)	11.0	-	67.0	22.0
Hasson et al. (2020)	11.4	-	80.0	40.0
Hampali et al. (2022)	10.8	25.5	68.0	21.4
Liu et al. (2021)	10.1	-	-	-
DMA (2023b)	10.1	23.8	45.5	20.8
HFL (2023)	8.9	28.9	64.3	32.4
HandBooster (2024)	**8.5**	21.1	-	-
LCP (2024)	**8.5**	21.5	-	-
HOISDF (2024)	9.2	**19.0**	35.5	14.4
Ours w/o pretraining	8.9	21.1	29.3	15.2
Ours	**8.5**	19.9	**27.1**	**14.3**

Table 2: Comparison of pose accuracy on *HO3Dv2 Full* (metrics are in mm).

Method	Hand MJE↓	Object SMCE↓	Physics CP(%)↑	PD↓	SD↓
Ground Truth	-	-	100	9.1	6.4
Hasson et al. (2020)	12.5	-	84.35	18.0	48.3
DMA (2023b)	11.5	-	89.16	15.7	35.3
ArtiBoost (2022)	10.7	16.0	94.23	15.0	27.8
DeepSimHO (2023a)	11.2	17.3	95.90	14.8	24.2
Ours	**8.5**	**15.1**	**98.85**	**13.4**	**17.3**

Table 3: Comparison of hand-object physics plausibility on *DexYCB Phy* (metrics, except CP, are in mm).

Method	Object SMCE↓	Physics CP(%)↑	PD↓	SD↓
Hasson et al. (2020)	5.35	78.52	2.02	6.40
CPF (2024a)	5.74	96.47	1.65	3.16
DMA (2023b)	4.79	93.07	1.88	3.47
ArtiBoost (2022)	4.86	94.47	1.27	2.83
DeepSimHO (2023a)	5.28	96.64	1.17	2.42
Ours	**3.12**	**98.80**	**0.96**	**2.21**

Table 4: Comparison of hand-object physics plausibility on *HO3Dv2 Phy* (metrics, except CP, are in cm).

We focus on refining the joints at the highest hierarchy level L_4. Let K_4^h denote the top-K hand pose candidates previously aggregated at level L_4. We collect joint parameters $\{\theta_i[j,:]\}_{i\in K_4^h, j\in L_4}$ and re-rank them based on their physics-based scores computed using Equation 12. The top-K joint parameters $\{\theta_n[j,:]\}_{n\in K_{\text{phy}}^h}$ are then averaged to produce the final hand pose estimates for level L_4. For the object, the physics-based score is defined as:

$$s_{phy}^o = -\mathcal{L}_{torque} \cdot \mathcal{L}_{contact}. \tag{13}$$

We retrieve the top-K translation candidates $\{T_i\}_{i\in K_T^o}$ and top-K rotation candidates $\{R_j\}_{j\in K_R^o}$ from the visual-based aggregation stage. These components are then combined to form a new set of object pose candidates $\{(T_i, R_j)\}_{i\in K_T^o, j\in K_R^o}$. Each pair is scored using Equation 13, and the top-K combinations $\{(T_n, R_m)\}_{(n,m)\in K_{\text{phy}}^o}$ are selected. The final object pose is obtained by averaging these top-ranked pose pairs. The analysis of the number of candidates and the top-K size is provided in the extended version.

Experiments

Our method is compared against state-of-the-art methods on two widely used benchmarks: DexYCB (Chao et al. 2021) and HO3D v2 (Hampali et al. 2020). Following (Lin et al. 2023; Qi et al. 2024), we focus on estimating both hand pose and object 6D pose from a single RGB image, with camera intrinsics and the object CAD model being available. Methods such as (Prakash et al. 2024), which reconstruct the mesh of the hand-held object without estimating either the hand pose or object's 6D pose, fall outside this problem setting and are therefore excluded from our comparison.

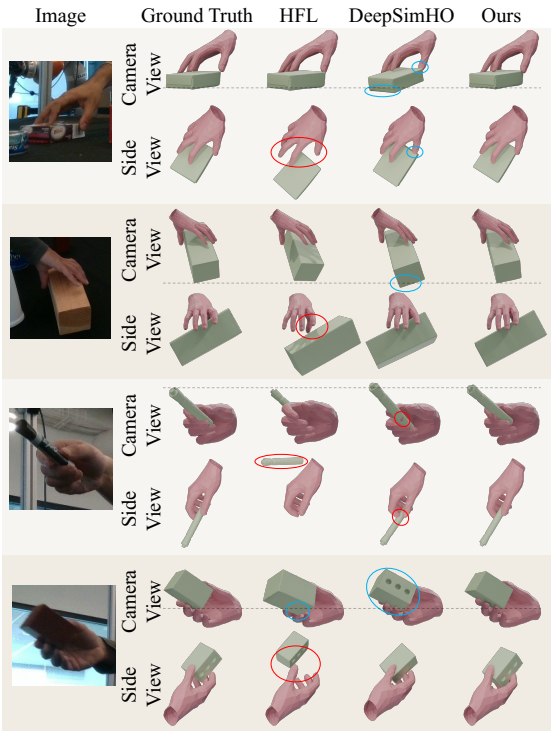

Figure 6: Qualitative results on *DexYCB Phy*. Red circles indicate incorrect hand object interaction. Blue circles point out the incorrect pose estimation. For more qualitative results, please refer to the extended version.

DexYCB: To train the model, the training set of official "S0" split (Chao et al. 2021) is employed. To evaluate the pose estimation accuracy, two common testing sets are used: *DexYCB Full*, the testing set of official "S0" split, includes scenarios of hands approaching objects as well as hands contacting objects; To evaluate the physical plausibility of the results, following (Wang, Mao, and Li 2023a), *DexYCB Phy* is employed as the testing set, in which the hands steadily hold the object.

HO3Dv2: To train the model, the standard training set is employed. To evaluate the pose estimation accuracy, standard testing set *HO3Dv2 Full* is used. To evaluate the physical plausibility of the results, following (Wang, Mao, and Li 2023a; Yang et al. 2024a), *HO3Dv2 Phy* is employed as the testing set, whose physics plausibility is manually verified by (Yang et al. 2024a). Some methods use additional data to train their models (Qi et al. 2024; Pavlakos et al. 2024; Xu et al. 2023; Wang, Mao, and Li 2023b). Following (Xu et al. 2023), we optionally pretrain on DexYCB for 5 epochs and report results both with and without this pretraining.

Evaluation Metrics

Pose Metrics: For hand, Mean Joint Error (MJE) and Procrustes aligned Mean Joint Error (PA-MJE) are reported.

Force Prediction	Visual-based Aggregation	Physics-based Aggregation	Hand		Object		Physics		
			MJE↓	PA-MJE↓	OCE↓	ADDS↓	CP(%)↑	PD↓	SD↓
			12.54	5.45	35.30	20.14	95.42	14.4	30.4
✓			12.16	5.33	29.81	16.66	96.51	14.1	25.9
✓	✓		10.05	5.09	26.12	15.21	97.95	13.9	20.0
✓	✓	✓	**10.01**	**5.08**	**23.72**	**13.47**	**98.85**	**13.4**	**17.3**

Table 5: Ablation study (metrics, except CP, are in mm). Hand and object metrics are evaluated on *DexYCB Full*. Physics metrics are evaluated on *DexYCB Phy*.

For object, Object Center Error (OCE), Mean Corner Error (MCE), Symmetry-aware Mean Corner Error (SMCE), and average closest point distance (ADD-S) are evaluated.

Physics Metrics: The contact percentage (CP)is calculated to assess the ratio of predictions with hand-object contact. The penetration depth (PD) is used to measure the maximum penetration distance between hand and object predictions. To evaluate the stability of hand holding object, the simulation displacement (SD) (Wang, Mao, and Li 2023a) is used to compute the average object center displacement after 200ms in the virtual physical simulator.

Pose Estimation Accuracy

The proposed method is compared with the state-of-the-art pose estimation methods, including both hand-object pose estimators (Hasson et al. 2019; Kuang, Ding, and Yao 2024; Hasson et al. 2021; Chen et al. 2023; Yang et al. 2022; Wang, Mao, and Li 2023b; Hasson et al. 2020; Liu et al. 2021; Lin et al. 2023; Qi et al. 2024) and hand pose estimators (Park et al. 2022; Hampali et al. 2022; Xie et al. 2024; Xu et al. 2023, 2024; Zhou et al. 2024).

The results on *DexYCB Full* are shown in Table 1. The proposed method outperforms the compared methods on both hand and object estimation. Especially in terms of object metrics, the proposed method significantly outperforms the second best with reducing the error of MCE, OCE and ADDS by 26.8%, 14.1%, and 27.4%, respectively.

The results on *HO3Dv2 Full* are shown in Table 2. Compared with the state-of-the-art methods, the proposed method achieves better performance on joint hand-object pose estimation. For the hand-related metrics, the proposed method is comparable to state-of-the-art methods. For object-related metrics, the proposed method achieves the best performance.

Physical Plausibility

The proposed method is compared with physics-based methods (Wang, Mao, and Li 2023a; Yang et al. 2024a; Hasson et al. 2021), and visual-based methods (Yang et al. 2022; Wang, Mao, and Li 2023b).

The results on *DexYCB Phy* are shown in Table 3. Among the existing methods, DeepSimHO (Wang, Mao, and Li 2023a) achieves the best physical performance, but its accuracy on hand and object is suboptimal. Comparing to DeepSimHO, ArtiBoost (Yang et al. 2022) has better pose estimation accuracy, but its physical plausibility is inferior. The existing methods cannot balance on both physics and pose estimation. In contrast, the proposed method significantly outperforms existing methods in both physical plausibility and pose accuracy.

The results on *HO3Dv2 Phy* are shown in Table 4. Our method achieves state-of-the-art performance across all physics metrics. Compared to the strongest baseline, DeepSimHO, our method further reduces PD and SD while improving contact rate and object accuracy, demonstrating the effectiveness of our approach in producing both physically plausible and accurate poses.

The qualitative results are shown in Figure 6. The visual-based method HFL (Lin et al. 2023) appears to have good visual consistency in the camera view. However, in the side view, incorrect hand-object interactions are observed. Compared to HFL, the physics-based method DeepSimHO (Wang, Mao, and Li 2023a) has more plausible hand-object interaction but less visual consistency in the front view. Comparing to these methods, the proposed method achieves better performance on both visual consistency and physical plausibility. For more qualitative results, please refer to the extended version.

Ablation Study

The ablation results in Table 5 show that the baseline without force prediction or aggregation performs the worst. Adding force prediction improves object pose accuracy, indicating that physical cues already enhance estimation quality. Visual based aggregation further reduces both hand and object errors. The full model with physics based aggregation achieves the best overall performance, with clear gains in both accuracy and physical plausibility. More analyses of pose aggregation module are provided in the extended version.

Conclusion

This paper presents a hand-object pose estimation approach that integrates visual cues and physical cues to address challenges in visual consistency and physical plausibility. The method learns both 2D visual features and 3D physical cues under a semi-supervised learning scheme combined with a local to global transformation process. These cues guide a candidate aggregation module that selects physically plausible and visually coherent hand-object poses. Experiments demonstrate state-of-the-art performance in both pose accuracy and physical plausibility. Future work includes incorporating temporal information to model dynamic equilibrium in hand-object interactions and using physical cues to improve object reconstruction.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the National Natural Science Foundation of China under Grant 62273318, the Science and Technology Project funds of Power Construction Corporation of China Ltd., and the Science and Technology Project of Sinohydro Bureau 8 Company Ltd. The project title is: the Key Technology Application Research on Green and Smart Mines (a major Science and Technology Special Project of Power Construction Corporation of China), Topic 11: Research on Key Technologies of Smart Terminals.

References

Billard, A.; and Kragic, D. 2019. Trends and challenges in robot manipulation. *Science*, 364(6446): eaat8414.

Brahmbhatt, S.; Tang, C.; Twigg, C. D.; Kemp, C. C.; and Hays, J. 2020. ContactPose: A Dataset of Grasps with Object Contact and Hand Pose. In *ECCV*.

Chao, Y.-W.; Yang, W.; Xiang, Y.; Molchanov, P.; Handa, A.; Tremblay, J.; Narang, Y. S.; Van Wyk, K.; Iqbal, U.; Birchfield, S.; Kautz, J.; and Fox, D. 2021. DexYCB: A Benchmark for Capturing Hand Grasping of Objects. In *CVPR*.

Chen, Y.; Wang, Q.; Chen, H.; Song, X.; Tang, H.; and Tian, M. 2019. An overview of augmented reality technology. *Journal of Physics: Conference Series*, 1237(2): 022082.

Chen, Z.; Chen, S.; Schmid, C.; and Laptev, I. 2023. gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction. In *CVPR*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *ECCV*.

Cho, H.; Kim, C.; Kim, J.; Lee, S.; Ismayilzada, E.; and Baek, S. 2023. Transformer-Based Unified Recognition of Two Hands Manipulating Objects. In *CVPR*.

Ehsani, K.; Tulsiani, S.; Gupta, S.; Farhadi, A.; and Gupta, A. 2020. Use the Force, Luke! Learning to Predict Physical Forces by Simulating Effects. In *CVPR*.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. ContactOpt: Optimizing Contact to Improve Grasps. In *CVPR*.

Hampali, S.; Rad, M.; Oberweger, M.; and Lepetit, V. 2020. HOnnotate: A Method for 3D Annotation of Hand and Object Poses. In *CVPR*.

Hampali, S.; Sarkar, S. D.; Rad, M.; and Lepetit, V. 2022. Keypoint Transformer: Solving Joint Identification in Challenging Hands and Object Interactions for Accurate 3D Pose Estimation. In *CVPR*.

Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-Object Reconstruction. In *CVPR*.

Hasson, Y.; Varol, G.; Schmid, C.; and Laptev, I. 2021. Towards Unconstrained Joint Hand-Object Reconstruction From RGB Videos. In *3DV*.

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 *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hu, H.; Yi, X.; Cao, Z.; Yong, J.-H.; and Xu, F. 2024a. Hand-Object Interaction Controller (HOIC): Deep Reinforcement Learning for Reconstructing Interactions with Physics. In *SIGGRAPH*.

Hu, H.; Yi, X.; Zhang, H.; Yong, J.-H.; and Xu, F. 2022. Physical Interaction: Reconstructing Hand-object Interactions with Physics. In *SIGGRAPH Asia*.

Hu, J.; Zhang, H.; Chen, Z.; Li, M.; Wang, Y.; Liu, Y.; and Sun, Z. 2024b. Learning Explicit Contact for Implicit Reconstruction of Hand-held Objects from Monocular Images. In *AAAI*.

Kuang, Z.; Ding, C.; and Yao, H. 2024. Learning Context with Priors for 3D Interacting Hand-Object Pose Estimation. In *MM*.

Lin, Z.; Ding, C.; Yao, H.; Kuang, Z.; and Huang, S. 2023. Harmonious Feature Learning for Interactive Hand-Object Pose Estimation. In *CVPR*.

Liu, S.; Jiang, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Semi-Supervised 3D Hand-Object Poses Estimation With Interactions in Time. In *CVPR*.

Morin, D. 2008. *Introduction to classical mechanics: with problems and solutions*. Cambridge University Press.

Mueller, F.; Davis, M.; Bernard, F.; Sotnychenko, O.; Verschoor, M.; Otaduy, M. A.; Casas, D.; and Theobalt, C. 2019. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics*, 38(4).

Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. HandOccNet: Occlusion-Robust 3D Hand Mesh Estimation Network. In *CVPR*.

Pavlakos, G.; Shan, D.; Radosavovic, I.; Kanazawa, A.; Fouhey, D.; and Malik, J. 2024. Reconstructing Hands in 3D with Transformers. In *CVPR*.

Potamias, R. A.; Zhang, J.; Deng, J.; and Zafeiriou, S. 2025. WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild. In *CVPR*.

Prakash, A.; Chang, M.; Jin, M.; Tu, R.; and Gupta, S. 2024. 3D Reconstruction of Objects in Hands without Real World 3D Supervision. In *ECCV*.

Qi, H.; Zhao, C.; Salzmann, M.; and Mathis, A. 2024. HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields. In *CVPR*.

Ren, F.; and Bao, Y. 2020. A Review on Human-Computer Interaction and Intelligent Robots. *International Journal of Information Technology & Decision Making*, 19(01): 5–47.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics*, 36(6).

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Tse, T. H. E.; Kim, K. I.; Leonardis, A.; and Chang, H. J. 2022a. Collaborative Learning for Hand and Object Reconstruction With Attention-Guided Graph Convolution. In *CVPR*.

Tse, T. H. E.; Zhang, Z.; Kim, K. I.; Leonardis, A.; Zheng, F.; and Chang, H. J. 2022b. S^2Contact: Graph-Based Network for 3D Hand-Object Contact Estimation with Semi-supervised Learning. In *ECCV*.

Wang, R.; Mao, W.; and Li, H. 2023a. DeepSimHO: Stable Pose Estimation for Hand-Object Interaction via Physics Simulation. In *NeurIPS*.

Wang, R.; Mao, W.; and Li, H. 2023b. Interacting Hand-Object Pose Estimation via Dense Mutual Attention. In *WACV*.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple Baselines for Human Pose Estimation and Tracking. In *ECCV*.

Xie, P.; Xu, W.; Tang, T.; Yu, Z.; and Lu, C. 2024. MS-MANO: Enabling Hand Pose Tracking with Biomechanical Constraints. In *CVPR*.

Xu, H.; Li, H.; Wang, Y.; Liu, S.; and Fu, C.-W. 2024. HandBooster: Boosting 3D Hand-Mesh Reconstruction by Conditional Synthesis and Sampling of Hand-Object Interactions. In *CVPR*.

Xu, H.; Wang, T.; Tang, X.; and Fu, C.-W. 2023. H2ONet: Hand-Occlusion-and-Orientation-Aware Network for Real-Time 3D Hand Mesh Reconstruction. In *CVPR*.

Yang, L.; Li, K.; Zhan, X.; Lv, J.; Xu, W.; Li, J.; and Lu, C. 2022. ArtiBoost: Boosting Articulated 3D Hand-Object Pose Estimation via Online Exploration and Synthesis. In *CVPR*.

Yang, L.; Zhan, X.; Li, K.; Xu, W.; Zhang, J.; Li, J.; and Lu, C. 2024a. Learning a Contact Potential Field for Modeling the Hand-Object Interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5645–5662.

Yang, Y.; Zhai, W.; Luo, H.; Cao, Y.; and Zha, Z.-J. 2024b. LEMON: Learning 3D Human-Object Interaction Relation from 2D Images. In *CVPR*.

Zhang, C.; Jiao, G.; Di, Y.; Wang, G.; Huang, Z.; Zhang, R.; Manhardt, F.; Fu, B.; Tombari, F.; and Ji, X. 2024. MOHO: Learning Single-view Hand-held Object Reconstruction with Multi-view Occlusion-Aware Supervision. In *CVPR*.

Zhao, Y.; Kwon, T.; Streli, P.; Pollefeys, M.; and Holz, C. 2024. EgoPressure: A Dataset for Hand Pressure and Pose Estimation in Egocentric Vision. *arXiv*.

Zhou, Z.; Zhou, S.; Lv, Z.; Zou, M.; Tang, Y.; and Liang, J. 2024. A Simple Baseline for Efficient Hand Mesh Reconstruction. In *CVPR*.