HOISDF: Constraining 3D Hand-Object Pose Estimation with Global Signed Distance Fields

Haozhe Qi Chen Zhao Mathieu Salzmann Alexander Mathis
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[first name].[surname]@epfl.ch

Abstract

Human hands are highly articulated and versatile at handling objects. Jointly estimating the 3D poses of a hand and the object it manipulates from a monocular camera is challenging due to frequent occlusions. Thus, existing methods often rely on intermediate 3D shape representations to increase performance. These representations are typically explicit, such as 3D point clouds or meshes, and thus provide information in the direct surroundings of the intermediate hand pose estimate. To address this, we introduce HOISDF, a Signed Distance Field (SDF) guided hand-object pose estimation network, which jointly exploits hand and object SDFs to provide a global, implicit representation over the complete reconstruction volume. Specifically, the role of the SDFs is threefold: equip the visual encoder with implicit shape information, help to encode hand-object interactions, and guide the hand and object pose regression via SDF-based sampling and by augmenting the feature representations. We show that HOISDF achieves state-of-the-art results on hand-object pose estimation benchmarks (DexYCB and HO3Dv2). Code is available at https://github.com/amathislab/HOISDF.

1. Introduction

Pose estimation during hand-object interaction from a single monocular view can contribute to widespread applications, e.g., in augmented reality [10], robotics [2, 15], human-computer interaction [42], and neuroscience [36]. Many excellent 3D hand [32, 45, 51, 58] and object [8, 25, 41] pose estimation algorithms have been developed. However, due to severe occlusion, they can easily fail during hand-object interactions. This has led to the emergence of dedicated hand-object interaction datasets [7, 18, 20, 35], and subsequently joint hand-object pose estimation has drawn increasing attention. Despite much progress, most methods still struggle when the hand or object is heavily occluded [11, 19, 22, 30, 39, 47, 49]. We argue that this limitation is rooted in the way 3D shape information is embedded in these algorithms.

In essence, existing methods can be classified into two approaches: Direct lifting and coarse-to-fine methods (see Figure 1). Direct lifting methods first filter 2D image features according to the pixel positions of the hand and object and then use the remaining features to make predictions [11, 19, 30, 33, 39]. These methods do not utilize explicit 3D intermediate representations and rely entirely on the network to learn the mapping from 2D image to 3D pose. Coarse-to-fine techniques make an initial prediction from the 2D image and improve upon it with a refinement network [12, 13, 22, 47, 49]. The intermediate representations can either be hand joints [12, 13] or hand vertices [22, 47, 49], which can be interpreted as explicit shape representations. Although these representations can incorporate 3D shape information, we argue that implicit shape representations in the form of signed distance fields (SDFs) offer more effective 3D shape information for subsequent computations.

To achieve this, we introduce HOISDF (a Hand-Object Interaction pose estimation network with Signed-Distance Fields), which uses SDFs to guide the 3D hand-object pose estimation in a global manner (Figure 1). HOISDF consists of two sequential components: a module learning to predict the signed distance field, and a module that performs pose regression that is field-guided (Figure 2). The signed distance field learning module regresses the hand and object signed distance fields based on the image features. The module is encouraged to focus on capturing global information (e.g., rough hand/object shape, global rotation and translation) by regressing signed distances in the original camera space, since we believe global plausibility is more important in the intermediate stage, while fine-grained details can be recovered in the later stages. To effectively leverage the dense field information, our field-guided pose regression module effectively uses the learned field information to (i) sample informative query points, (ii) augment the image features for those points, (iii) gather cross-target (i.e., hand-to-object or object-to-hand) cues to reduce the influence of mutual occlusion, and (iv) combine the point features together to estimate the hand and object poses.

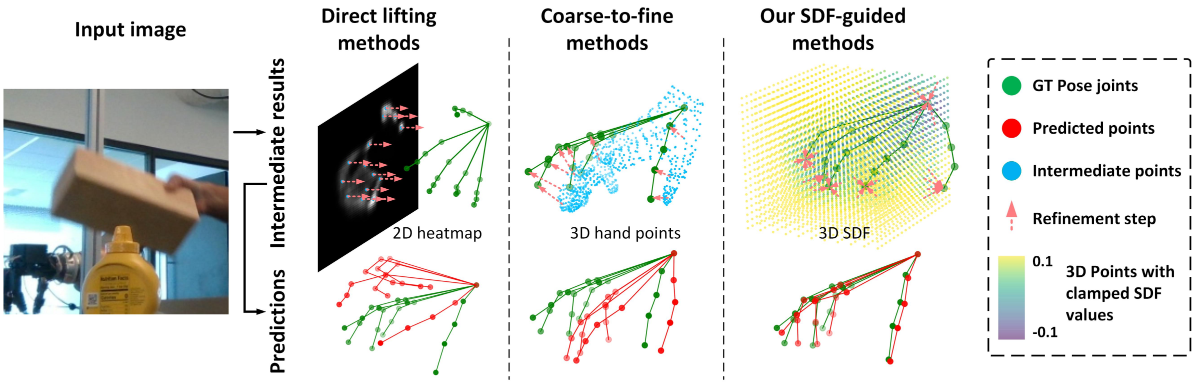

Figure 1. **Conceptual advantage of the SDF-guided model over existing approaches.** Our model utilizes Signed Distance Fields (SDF) to provide global and dense constraints for hand-object pose estimation. In contrast to direct lifting and coarse-to-fine methods, which struggle to refine poor initial predictions, the distance field yields global cues not limited to areas near an initial prediction.

Overall, HOISDF can be trained in an end-to-end manner. We achieve state-of-the-art results on the DexYCB and HO3Dv2 datasets, corroborating the benefits of using SDFs as global constraints for hand-object pose estimation and the effectiveness of our approach to exploiting the field information. Altogether, our main contributions are:

- We introduce a hand-object pose estimation network that uses signed distance fields (HOISDF) to introduce implicit 3D shape information
- We develop a new signed-distance field-guided pose regression module to effectively integrate the relevant parts of the global field information for hand and object pose estimation.

2. Related Work

2.1. 3D Hand-Object Pose Estimation

Recently, joint hand-object pose estimation has drawn increasing research interest [29], and many hand and object interaction datasets have been developed [7, 18, 20, 35]. The current methods can be divided into direct lifting techniques and coarse-to-fine strategies. Among the former, Chen *et al.* [11] fused hand and object features with sequential LSTM models. Hampali *et al.* [19] extracted 2D keypoints and sent them to a transformer architecture to find the correlation with the 3D poses. Li *et al.* [30] proposed a data synthesis pipeline that can leverage the training feedback to enhance hand object pose learning. Lin *et al.* [33] proposed to learn harmonious features by avoiding hand-object competition in middle-layer feature learning. For the coarse-to-fine methods, Hasson *et al.* [22] obtained initial hand and object meshes and optimized them with interaction constraints. Tse *et al.* [47] used an attention-guided graph convolution to iteratively extract features from the previous hand-object estimates. Wang *et al.* [49] designed a dense mutual attention module to explore the relations from the initial hand-object predictions. We build on those methods but, in contrast, focus on implicit 3D shape information by learning SDFs, which provide global, dense constraints to guide the pose predictions.

2.2. Distance Fields in Hand-Object Interactions

Unlike explicit representations such as point clouds and meshes, neural distance fields provide a continuous and differentiable implicit representation that encodes the 3D shape information into the network parameters. Given a 3D query point, a neural distance field outputs the signed or unsigned distance from this point to the object surface. Neural distance fields have been widely used in 3D shape reconstruction and representation [1, 16, 38, 40, 55]. Recently, SDFs have also been exploited in the context of hand-object interaction. In particular, Karunratanakul *et al.* [26] proposed to jointly model the hand, the object, and contact areas using an SDF. Ye *et al.* [54] used an SDF and the predicted hand to infer the shape of a hand-held object. Chen *et al.* [12] pre-aligned the 3D space with hand-object global poses to support the SDF prediction. Chen *et al.* [13] further used entire kinematic chains of local pose transformations to obtain finer-grained alignment. However, those methods mainly use SDF as the endpoint of the model to directly reconstruct 3D meshes instead of using SDF as an intermediate representation. Here we explore how SDFs as an intermediate representations can guide subsequent pose estimation. Our experiments clearly demonstrate the benefits of our approach.

2.3. Attention-based Methods

Attention mechanisms [48] have been wildly successful in machine learning [4, 6, 14, 17, 24] due to their effectiveness at exploiting long-range correlation. In the context of modeling hand-object relationships, Hampali et al. [19] propose modeling correlations between 2D keypoints and 3D hand and object poses using cross attention. Tze *et al.* [47] design an attention-guided graph convolution network to cap-

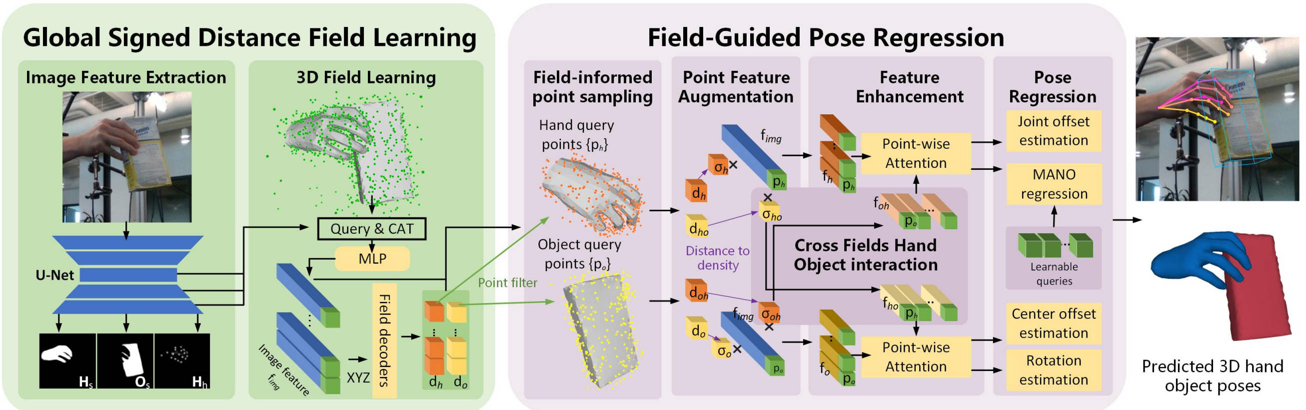

Figure 2. **Overall pipeline of HOISDF.** HOISDF has two parts: A global signed distance field learning module and a field-guided pose regression module. The global signed distance field learning module regresses the hand object signed distances as the intermediate representation and encodes the 3D shape information into the image backbone through implicit field learning. The field-guided pose regression module uses global field information to filter and augment the point features as well as guiding hand-object interaction. Those enhanced point features are then sent to regress hand and object poses using point-wise attention.

ture hand and object mesh information dynamically. Wang *et al.* [49] propose to exploit mutual attention between hand and object vertices to learn interaction dependencies. By contrast, our HOISDF applies attention across field-guided query points to mine the global 3D shape consistency context and cross-attend between hand and object.

3. HOISDF

We propose Hand-object Pose Estimation with Global Signed Distance Fields (HOISDF), a joint hand-object pose estimation model that leverages global shape constraints from a signed distance field. HOISDF comprises two components: A global signed distance field learning module and a field-guided pose regression module (Figure 2). Both components benefit from the robust 3D shape information modeled with the SDF and the whole architecture is trained end-to-end.

3.1. Global Signed Distance Field Learning

We simultaneously learn hand and object signed distance fields (SDFs) with the following rationale: i) An SDF implicitly represents 3D shape with the model parameters; the implicit learning procedure can thus propagate 3D shape information to the feature extraction module. ii) Jointly learning hand and object fields allows the model to encode their mutual constraints. Meanwhile, since we predict hand and object signed distances in the initial stage as intermediate representations, we encourage our SDF learning module to focus more on global plausibility rather than local fine-grained details. Below, we describe the image feature extraction and the SDF learning in detail.

Image Feature Extraction. For extracting hierarchical features $\mathbf{F}$, we use a standard encoder-decoder architecture, specifically a U-Net [19, 23, 46]. Following standard practice [19, 33, 49], we regress 2D predictions (a single channel heatmap [19] and hand/object segmentation masks with loss $\mathcal{L}_{img}$, see Supp. Mat. for details) to enable the model to represent hand-object interaction at the 2D image level.

3D Signed Distance Field Learning. With the extracted image features, the SDF module learns the continuous mapping from a 3D query point $\mathbf{p} \in \mathbb{R}^3$ to the shortest signed distances between $\mathbf{p}$ and the hand/object surfaces. Compared to [12, 13], we directly learn SDFs in the original space without rotating to canonical spaces using pose predictions. Our SDF module will consequently focus on the global information (e.g., general shape, location and global rotation) of the hand and object.

Specifically, given a 3D query point $\mathbf{p} \in \mathbb{R}^3$, we project it to the 2D image space to compute the pixel-aligned image features [13, 19, 49, 50] extracted by the U-Net decoder $\{\mathbf{F}^i_{dec}\}$, where $i \in \mathcal{X}$ indexes over the hierarchical decoder levels of the U-Net. We then concatenate the queried image features and pass them to a Multilayer Perceptron (MLP) to obtain a feature vector

$$\mathbf{f}_{img} = \mathrm{MLP}(\oplus_{i \in \mathcal{X}} \mathbf{F}^i_{dec}(\pi_{3D \to 2D})), \tag{1}$$

where $\pi_{3D \to 2D}$ represents the projection and interpolation operation, $\oplus$ indicates the concatenation of all the hierarchical pixel-aligned image features, and $\mathcal{X}$ is the set of hierarchical features.

To emphasize the importance of $\mathbf{p}$, we expand the coordinate representation by a Fourier Positional Encoding [37] into a vector $\mathbf{f}_{pos}$. We then concatenate the triplet $\mathbf{p}$, $\mathbf{f}_{pos}$ and $\mathbf{f}_{img}$ together and pass them to the hand SDF decoder $\mathbb{SDF}_h$ and the object SDF decoder $\mathbb{SDF}_o$. This can be ex-

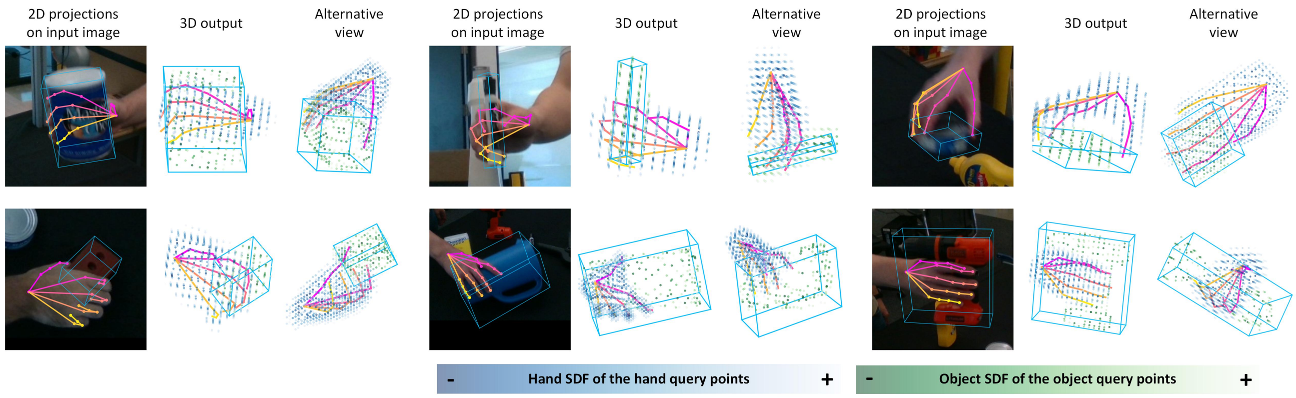

Figure 3. **Visualization of the intermediate query points on DexYCB testset.** The darkness of the query points reflects the predicted distance from the query point to the hand (in blue) and object (in green) surfaces. The intermediate SDF representations can capture the GT 3D hand and object shapes. HOISDF effectively uses the robust global clues from SDFs to deal well with various objects and hand movements as well as their mutual occlusions.

pressed as

$$\mathbf{f}_{sdf} = \mathbf{p} \oplus \mathbf{f}_{pos} \oplus \mathbf{f}_{img}, \tag{2}$$
$$d_h = \mathbb{SDF}_h(\mathbf{f}_{sdf}), \tag{3}$$
$$d_o = \mathbb{SDF}_o(\mathbf{f}_{sdf}). \tag{4}$$

Here, d_h is the shortest distance from $\mathbf{p}$ to the hand mesh surface, and d_o is the shortest distance from $\mathbf{p}$ to the object mesh surface; d_h and d_o will be positive if they are outside the surface and negative otherwise. The field decoders $\mathbb{SDF}_h$, and $\mathbb{SDF}_o$ are all 3-layer MLPs with *tanh* activation in the last layer [26].

During training, we sample N_s 3D query points, ensuring that most points are sampled near the hand and object mesh surfaces. We pre-compute the ground-truth distances from the query point to the hand and object surfaces and use the smooth-L1 loss [43] to supervise the learning of d_h and d_o. We sum the losses together and refer to the resulting loss as $\mathcal{L}_{sdf}$.

3.2. Integrating Field Information: Field-guided Pose Regression

After the field learning module, we aim to use the learned fields to predict the hand and object poses. However, effectively using the field information is non-trivial: i) The field information is implicitly encoded in the model parameters; we can only read the field information at a specific location by sending a query point into the network; ii) The resulting signed distance at a certain query point is just a scalar distance, which on its own provides only a weak link with the pose prediction; iii) How to explicitly model the hand-object interaction using SDF is unclear. To address these challenges, we hence introduce the field-guided pose regression module described below.

3.2.1 Field-informed Point Sampling

To address the first problem, we propose a point-sampling strategy that aims to extract the most helpful field information while querying only a few points. It builds on the assumption that the query points near the ground-truth surface are the most informative ones. As such, during inference, we voxelize the 3D space with N_v bins, which gives us N_v^3 query points. We first use the hand and object bounding boxes to filter the points in 2D space. Then, we send the remaining points into $\mathbb{SDF}_h$ and $\mathbb{SDF}_o$ and sort them according to the obtained hand and object signed distances separately. We sample N_v^2/n_h hand query points and N_v^2/n_o object query points with the lowest absolute hand distance and object distance, respectively. Here, n_h and n_o are two positive hyperparameters controlling the number of samples. Since we can access the ground-truth mesh during training, we directly sample N_h hand query points near the hand mesh and N_o object query points near the object mesh (with an absolute distance smaller than 4cm) for speed and memory optimization (2x faster). Towards the end of training, we also sample points with the same strategy as during testing to learn the point distribution. We will show the effectiveness of our proposed sampling strategy in Sec. 4.4.

3.2.2 Field-based Point Feature Augmentation

To address the second problem, given a sampled hand query point $\mathbf{p}_h$, we convert d_h to the volume density $\sigma_h = \alpha^{-1} sigmoid(-d_h/\alpha)$, where α is a learnable parameter to control the tightness of the density around the surface boundary. This is motivated by the strategy used in StyleSDF [38] for image rendering, but here we use it for the purpose of feature augmentation. We then multiply σ_h with $\mathbf{f}_{img}$. The field information will thus influence the whole feature representation. $\mathbf{p}_h$ and its positional encoding $\mathbf{f}_{pos}$ discussed in Sec. 3.1 are also concatenated to further augment the point feature. The final hand query point feature $\mathbf{f}_h$ is obtained as

$$\mathbf{f}_h = \mathbf{p}_h \oplus \mathbf{f}_{pos} \oplus (\mathbf{f}_{img} \cdot \sigma_h). \tag{5}$$

For a sampled object query point $\mathbf{p}_o$, the object query point feature $\mathbf{f}_o$ is obtained in an analogous way (i.e., augmenting the feature by the volume density σ_o based on object SDF d_o).

3.2.3 Cross Fields Hand-Object Interaction

Since we use the shared image backbone to learn the hand-object SDFs jointly, hand-object relations can be implicitly modeled during implicit field learning. Here, we aim to model the hand-object interaction explicitly to better deal with the mutual occlusions. Intuitively, the hand-object contact areas are highly informative about the object/hand pose. Therefore, we augment the hand/object query points with the object/hand SDFs, respectively, to serve as interaction cues (Fig. 2). Specifically, for a sampled object query point $\mathbf{p}_o$, we send it to the hand SDF decoder $\mathbb{SDF}_h$ to obtain the cross-hand signed distance d_{oh}. d_{oh} is then converted to the volume density σ_{oh} and used to augment the queried image feature $\mathbf{f}_{img}$ similarly to Sec. 3.2.2. The final cross-hand query point feature $\mathbf{f}_{oh}$ is obtained as

$$\mathbf{f}_{oh} = \mathbf{p}_o \oplus \mathbf{f}_{pos} \oplus (\mathbf{f}_{img} \cdot \sigma_{oh}). \tag{6}$$

$\mathbf{f}_{oh}$ will serve as object cues for hand pose estimation. A $\mathbf{p}_o$ with smaller d_{oh} will play a bigger role in helping the hand pose estimation. Similarly, a hand query point $\mathbf{p}_h$ is also sent to object SDF decoder $\mathbb{SDF}_o$ and used to generate a cross-object query point feature $\mathbf{f}_{ho}$.

3.2.4 Feature Enhancement with Point-wise Attention

As the pixel-aligned feature $\mathbf{f}_{img}$ mainly contains local information, the local query point features $\mathbf{f}_h$ and $\mathbf{f}_o$ could be misled and thus make wrong predictions in the presence of severe occlusion. To address this problem, we propose to use an attention mechanism [27, 48] to exploit reliable dependencies in the global context. In contrast to existing approaches that either perform attention over 2D features [19] or over 3D mesh vertex features [47, 49], our point-wise attention explores the global field information and the local image information with the aim of finding global 3D shape consistency between the sampled query points. Specifically, the extracted N_h hand query point features $\{\mathbf{f}_h^i\}_{i\in(0,N_h)}$ are sent into a hand attention module, which consists of six Multi-Head Self-Attention (MHSA) layers [27, 48].

Meanwhile, to leverage object cues inside the cross-hand query point features $\{\mathbf{f}_{oh}^i\}_{i\in(0,N_o)}$, we also send them to the MHSA layers $\mathbb{SA}$ to conduct cross attention with $\{\mathbf{f}_h^i\}_{i\in(0,N_h)}$. The resulting enhanced hand point features are computed as

$$(\{\mathbf{f}_{eh}^i\}_{i\in(0,N_h)}, *) = \mathbb{SA}(\{\mathbf{f}_h^i\}_{i\in(0,N_h)}, \{\mathbf{f}_{oh}^i\}_{i\in(0,N_o)}), \tag{7}$$

where $*$ denotes that we ignore the output from the N_o cross-hand query tokens. Analogously, the enhanced object point features $\{\mathbf{f}_{eo}^i\}_{i\in(0,N_o)}$ can be obtained by processing object query point features $\{\mathbf{f}_o^i\}_{i\in(0,N_o)}$ and cross-object query point features $\{\mathbf{f}_{ho}^i\}_{i\in(0,N_h)}$ with an object attention module.

3.2.5 Point-wise Pose Regression

With attention, we incorporate globally consistent information and cross-target cues into the hand point features $\{\mathbf{f}_{eh}^i\}$ and object point features $\{\mathbf{f}_{eo}^i\}$. Those points thus have enough global-local shape context information to regress hand-object poses. We apply asymmetric designs for hand and object pose estimation. Since the hand is non-rigid, flexible, and typically occluded when grasping an object, regressing the hand pose requires gathering richer information inside the $\{\mathbf{f}_{eh}^i\}$. We hence follow [19] to use Cross-Attention layers $\mathbb{CA}$ with the learned hand pose queries $\{\mathbf{q}^i\}$. We supervise the learning of hand pose queries with MANO parameters [44] to obtain both hand joints and a hand mesh. Sixteen hand pose queries regress 3-D MANO joint angles, and one more hand pose query regresses the 10-D mano shape parameters β. This can be expressed as

$$\begin{aligned}(\{\boldsymbol{\theta}^i \in \mathbb{R}^3\}_{i\in(0,16)}, \beta) = \\ \mathbb{CA}(\{\mathbf{f}_{eh}^i\}_{i\in(0,N_h)}, (\{\mathbf{q}^i\}_{i\in(0,16)}, \mathbf{q}^{16})).\end{aligned} \tag{8}$$

We use a smooth-L1 loss [43] to supervise the learning of the MANO parameters, referred to as $\mathcal{L}_{\text{mano}}$. Similarly to [19], we also regress the intermediate hand pose objective to guide the final predictions. However, since our $\{\mathbf{f}_{eh}^i\}$ already contains rich 3D information, we directly regress 3D hand joints instead of 2D joints as in [19]. We use $\{\mathbf{f}_{eh}^i\}$ as dense local regressors [31, 51] to predict the offsets $\{\mathbf{o}_h^{ij}\}$ from each hand query point $\mathbf{p}_h^i$ to every pose joint as well as the prediction confidence. The corresponding loss is denoted as $\mathcal{L}_{\text{off}}$. Note that the design of the hand pose regressor is not identical. Our field-guided query points already include rich global-local shape context information and yield satisfactory pose estimation results with various regressors (see Sec. 4.5).

Compared with the hand, the object is more rigid. Therefore, we simply regress rotation vectors $\{\mathbf{r}^i\}$ and translation vectors $\{\mathbf{t}^i\}$ with all the enhanced object point features $\{\mathbf{f}_{eo}^i\}$ and use a smooth-L1 loss [43] $\mathcal{L}_{\text{obj}}$ to supervise them. During inference, we average the predictions from all the object points to obtain the final object translation and orientation.

4. Experiments

We first introduce the hand-object benchmarks, describe implementation details and compare HOISDF with state-of-the-art (SOTA) methods. We finally detail ablations.

4.1. Datasets and Evaluation Metrics

We evaluate HOISDF on DexYCB [7] and HO3Dv2 [18] datasets containing, respectively, 582K and 77K images of human interacting with YCB objects [5].

DexYCB Dataset. We use the default S0 train-test split defined by DexYCB [7]. Some methods [33, 34] use the full DexYCB dataset by flipping the left-hand images (denoted as DexYCB Full), while other methods [12, 13, 20, 22, 47, 49, 53] select input frames in which the right hand and the object are in close interaction to ensure the physical contact (denoted as DexYCB). In general when we refer to DexYCB we mean this latter split. To broadly compare, we train HOISDF on both settings. Since most of the methods use the data only with the right hand, we conduct our ablations under the DexYCB split.

For hand pose estimation, we report Mean Joint Error (MJE) and Procrustes Aligned Mean Joint Error (PAMJE) [57]. We also report Mean Mesh Error (MME), area under the curve of the percentage of correct vertices (VAUC) the F-scores (F@5mm and F@15mm), and corresponding Procrustes Aligned version following [52] to measure hand mesh reconstruction performance. For object 6D pose estimation, we report Object Center Error (OCE) following [12, 13], Mean Corner error (MCE) following [49], and standard pose estimation average closest point distance (ADD-S) following [20, 22, 49] to measure performance in center, corner, and vertex levels.

HO3Dv2 Dataset. We use the standard train-test splitting protocol and submit the test results to the official website to report performance. Since the HO3Dv2 is relatively small-scale, some methods [49, 53] render synthetic hand object images to enhance learning. Therefore, apart from training the model only with the original data in the HO3Dv2 training set, we also train another model (denoted with '*' in Table 4) by including synthetic images. We follow the render pipeline of Wang et al. [49].

For hand pose estimation, we use the HO3Dv2 evaluation metrics: Mean Joint Error (MJE), Scale-Translation aligned Mean Joint Error (STMJE) [58], and Procrustes aligned Mean Joint Error (P-MJE) [57]. For object 6D pose estimation, we report mean Object Mesh Error (OME) and standard pose estimation average closest point distance (ADD-S) following [20, 22, 49].

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S	Object
Lin *et al.* [32]	15.2	6.99	-	-	-	No
Spurr *et al.* [44]	17.3	6.83	-	-	-	No
Liu *et al.* [34]	15.2	6.58	-	-	-	Yes
Park *et al.* [39]	14.0	5.80	-	-	-	No
Chen *et al.* [9]	14.2	6.40	-	-	-	No
Xu *et al.* [52]	14.0	5.70	-	-	-	No
Lin *et al.* [33]	12.6	5.47	42.7	48.0	33.8	Yes
HOISDF (ours)	**10.1**	**5.13**	**27.6**	**35.8**	**18.6**	Yes

Table 1. Quantitative comparison on the DexYCB dataset. Trained and tested on the DexYCB Full split. HOISDF reaches lower hand and object pose estimation errors. The metrics are represented in millimeters. The last column indicates whether a method performs the object 6D pose estimation.

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S	Object
Hasson *et al.* [20]	17.6	-	-	-	-	Yes
Hasson *et al.* [22]	18.8	-	-	52.5	-	Yes
Tze *et al.* [47]	15.3	-	-	-	-	Yes
Li *et al.* [53]	12.8	-	-	-	-	Yes
Chen *et al.* [12]	19.0	-	27.0	-	-	Yes
Chen *et al.* [13]	14.4	-	19.1	-	-	Yes
Wang *et al.* [49]	12.7	6.86	27.3	32.6	15.9	Yes
Lin *et al.* [33]	11.9	5.81	39.8	45.7	31.9	Yes
HOISDF (ours)	**10.1**	**5.31**	**18.4**	**27.4**	**13.3**	Yes

Table 2. Same as Table 1, but for DexYCB split, see Sec. 4.1.

4.2. Implementation and Training Details

We adopt ResNet-50 as the U-Net backbone [23, 46]. All the point features: the image $\mathbf{f}_{img}$, the hand $\mathbf{f}_{eh}$, and object $\mathbf{f}_{eo}$ are of size 256. We employ a transformer [48] encoder as our point-wise attention module and a transformer decoder as our MANO regressor [44]. We follow the standard practice [19, 33, 49] to train a unified model for all the objects in the dataset. The overall loss is a weighted sum of all individual loss functions,

$$\mathcal{L} = \lambda_1 \mathcal{L}_{img} + \lambda_2 \mathcal{L}_{sdf} + \lambda_3 \mathcal{L}_{\text{mano}} + \lambda_4 \mathcal{L}_{\text{off}} + \lambda_5 \mathcal{L}_{\text{obj}}, \quad (9)$$

where λ_1 to λ_5 are used to balance all the loss terms to the same scale. During training, the network parameters are optimized with Adam [28] with a mini-batch size of 32. The initial learning rate is 1e-4 and decays by 0.7 every 5 epochs. HOISDF typically converges to a satisfying result after about 40 epochs.

For query points sampling, during training, we sample $N_s = 1000$ query points for 3D field learning. During inference, we empirically found that with a discretization size of $N_v = 64$, sampling $N_v^2/n_h = 600$ hand query points and $N_v^2/n_o = 200$ object query points was enough for good performance.

4.3. Comparisons with State-of-the-Art Methods

Quantitative comparisons on DexYCB. We evaluate HOISDF on the DexYCB test sets (Tables 1 and 2) and compare it with (SOTA) methods. Among the best mod-

Metrics	MME↓	VAUC↑	F@5↑	F@15↑	PAMME↓	PAVAUC↑	PAF@5↑	PAF@15↑	Object
Park *et al.* [39]	13.1	76.6	51.5	92.4	5.5	89.0	78.0	99.0	No
Chen *et al.* [9]	13.1	76.1	50.8	92.1	5.6	88.9	78.5	98.8	No
Xu *et al.* [52]	13.0	76.2	51.3	92.1	5.5	89.1	80.1	99.0	No
Lin *et al.* [33]	11.6	77.6	53.0	93.3	5.2	89.6	79.8	99.2	Yes
HOISDF (ours)	**9.9**	**80.5**	**60.1**	**94.9**	**4.9**	**90.2**	**81.8**	**99.3**	Yes

Table 3. Quantitative comparison with hand mesh metrics on the DexYCB Full testset. MME and PAMME are in millimeters.

els, [49] is best at object estimation while [33] is best at hand pose estimation. However, HOISDF outperforms prior methods by a substantial margin for both hand and object metrics (Table 2). It is worth mentioning that HOISDF beats the methods that perform just hand pose estimation [32, 39, 44, 52]. Furthermore, we also compare HOISDF with SDF-based hand object interaction methods [12, 13]. As mentioned in Sec. 3.1, both of them use SDFs to regress the (output) hand meshes, while we use SDFs as intermediate representations and for field-guided inference. HOISDF significantly outperforms these methods.

As HOISDF also predicts a MANO mesh, we compare with the SOTA methods for hand mesh reconstruction performance on the DexYCB Full test set (Table 3). We observe a consistent improvements with HOISDF.

Qualitative comparisons on DexYCB. Here, we compare our HOISDF qualitatively with two SOTA hand object pose estimation methods on the DexYCB test set (Fig. 4). We can see HOISDF outperforms [33, 49] under various objects and different types of hand object interactions.

Quantitative comparisons on HO3Dv2. As further evidence of the effectiveness of HOISDF, we also evaluate it on the HO3Dv2 dataset. Again, HOISDF consistently beats the current SOTA methods on almost all the hand and object metrics both with and without synthetic data (Table 4). Lin et al. [33] obtains slightly better performance with regard to PAMJE, but performs very poorly in the other metrics, while HOISDF is more balanced.

On both datasets, especially HO3Dv2 with fewer data, HOISDF yields a larger improvement on the metrics that exploit more global information (MJE, STMJE and object metrics). We attribute this advantage to the fact that SDFs, as intermediate representations, capture global information effectively to guide the subsequent pose estimations. We will first visualize query points (Fig. 3) and then validate our design choices.

Visualization of the learned SDFs. We visualize the pose predictions and the intermediate hand-object query points on the DexYCB testset (Fig. 3). We can see that the remaining query points after the field-informed point sampling already reveal the general hand object shape.

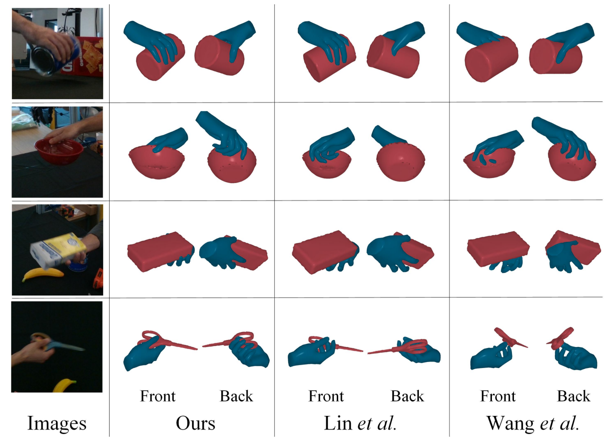

Figure 4. **Qualitative comparisons between HOISDF and [33, 49] on DexYCB testset.** HOISDF effectively uses robust global clues near the hand and object to deal well with various objects and severe occlusions.

Metrics in [mm]	MJE	STMJE	PAMJE	OME	ADD-S
Hasson *et al.* [20]	-	31.8	11.0	-	-
Hasson *et al.* [21]	-	36.9	11.4	67.0	22.0
Hasson *et al.* [22]	-	26.8	12.0	80.0	40.0
Liu *et al.* [34]	-	31.7	10.1	-	-
Hampali *et al.* [19]	25.5	25.7	10.8	68.0	21.4
Lin *et al.* [33]	28.9	28.4	**8.9**	64.3	32.4
HOISDF (ours)	**23.6**	**22.8**	9.6	**48.5**	**17.8**
Li *et al.** [53]	26.3	25.3	11.4	-	-
Wang *et al.** [49]	22.2	23.8	10.1	45.5	20.8
HOISDF* (ours)	**19.0**	**18.3**	**9.2**	**35.5**	**14.4**

Table 4. Quantitative comparison on the HO3Dv2 dataset. The metrics are represented in millimeters.'*' denotes models that were co-trained with synthetic data.

4.4. Ablation for Intermediate Representations

Since using SDF as a global intermediate representation is the key component of HOISDF, we analyze the role of the SDF here, comparing it with other intermediate representations, and analyzing the query points.

Comparison different intermediate representations. Here, to elucidate the role of the SDF, we build several baselines that use different intermediate representations while trying to keep the remaining model components (e.g., image backbones, feature dimensions, pose regressors, etc.) the same as in our model. We replace the 3D field learning module (Sec. 3.1) with 2D keypoint learning, 2D segmen-

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S
2D Keypoint	14.9	7.13	34.2	45.3	22.9
2D Segmentation	14.1	6.88	31.3	43.1	21.0
3D Vertices	12.7	6.57	24.1	35.3	16.5
3D SDFs (ours)	**10.1**	**5.31**	**18.4**	**27.4**	**13.3**

Table 5. Comparison between different intermediate representations on DexYCB testset. The SDF-based representation outperforms other representations because it encodes 3D shape information, is direct to regress, and has less joint cumulative error.

Metrics in [mm]	Mean	MCP	PIP	DIP	Tip
Wang *et al.* [49]	7.67	7.63	6.36	6.29	10.4
HOISDF (ours)	**6.16**	**6.02**	**5.27**	**5.40**	**7.95**

Table 6. Sampled point distributions. Using SDF as global guidance for point sampling gathers query points closer to the GT pose joints. MCP, PIP, DIP, and Tip are different finger parts.

tation learning, and 3D mesh learning (see Supp. Mat.). We found that utilizing intermediate 2D representations is much worse, and that 3D vertices are also significantly less powerful than SDFs (Table 5). Next, we provide further evidence for the effectiveness of using SDF as an intermediate representation by analyzing the sampled query points during inference.

Analysis of the sampled points. We argued that the SDF representation better captures global shape information across the capture volume (Figure 1). We analyze the point distributions of HOISDF's hand query points sampled using our proposed point sampling strategy and the intermediate hand mesh vertices extracted by the initial stage of Wang et al. [49]. Indeed, our model samples closer points to the hand joints, particularly for the most challenging finger joints like the finger tips (Table 6).

4.5. Ablations for the Field-Guided Pose Regression Module

The field-guided pose regression module is the other key component to let HOISDF effectively leverage the SDF information. To verify that, we conduct ablations for different parts. Firstly, we showed that our field-guided sampling method is efficient and robust by comparing it with other sampling ways (Table 7). Secondly, we assessed the role of the point feature augmentation method by comparing it with different variations; altering various parts gracefully reduced the performance (Table 8). Next, the mutual hand-object feature enhancement method proposed in Sec.3.2.3 is also proven to be effective by removing the cross attention or replacing with other non-augmented features (Table 9). Finally, we show that HOISDF is robust to changes in regression targets (Table 10) since our hand/object query points already capture enough global-local context with our field-guided module. Overall, the ablations validate our design choices.

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S
Random	25.8	13.5	48.4	53.7	29.6
Signed distance	13.3	6.58	19.7	30.7	15.9
Field gradient	**10.1**	**5.29**	18.5	27.7	13.5
Absolute distance (ours)	**10.1**	5.31	**18.4**	**27.4**	**13.3**

Table 7. Comparison between different sampling strategies on DexYCB testset. Our field-informed point sampling can achieve the best performance. See Supp. Mat. for details on the alternative sampling strategies.

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S
w/o SDF augmentation	11.5	6.05	23.6	31.2	15.7
w density concatenation	11.0	5.71	22.7	30.5	15.3
w distance concatenation	11.5	6.07	23.3	30.9	15.6
w SDF augmentation	**10.8**	**5.68**	**22.2**	**30.0**	**15.1**

Table 8. Effects of field-based point feature augmentation on the DexYCB test set. Our SDF feature augmentation best enhances features for the subsequent pose estimations. See Supp. Mat. for details on the alternative augmentations.

Metrics in [mm]	MJE	PAMJE	OCE	MCE	ADD-S
w/o cross feature enhancement	10.8	5.68	22.2	30.0	15.1
w cross image feature	11.1	5.74	20.2	28.6	14.2
w cross target feature	11.3	5.81	23.7	31.8	15.9
Cross feature enhancement (ours)	**10.1**	**5.31**	**18.4**	**27.4**	**13.3**

Table 9. Effects of hand-object feature enhancement on the DexYCB testset. HOISDF's cross feature enhancement gave the best results. See Supp. Mat. for details on the alternative feature computations.

Metrics in [mm]	MJE	PAMJE
w/o intermediate joint regression	10.4	5.49
w/o MANO regression	10.5	5.65
w MANO shape & inverse kinematics	**10.0**	5.35
MANO regression (Ours)	10.1	**5.31**

Table 10. Robustness to different pose regressors on the DexYCB testset. Benefiting from the rich global-local context information inside the enhanced features, HOISDF can obtain great performance even with simple pose regression targets. See Supp. Mat. for details on the alternative regression targets.

5. Conclusion

We proposed a novel 3D hand-object pose estimation algorithm that takes advantage of jointly learned signed distance fields. It achieves strong results and inference is fast (see Sup. Mat.) We believe this paradigm could also be applied to other pose estimation problems, e.g., [2, 10, 15, 36, 42].

Acknowledgments: Our work was funded by EPFL and Microsoft Swiss Joint Research Center (H.Q., A.M.) and a Boehringer Ingelheim Fonds PhD stipend (H.Q.). We are grateful to Niels Poulsen for comments on an earlier version of this manuscript. We also sincerely thank Rong Wang, Wei Mao and Hongdong Li for sharing the hand-object rendering pipeline [49].

References

[1] Ma Baorui, Han Zhizhong, Liu Yu-Shen, and Zwicker Matthias. Neural-pull: Learning signed distance functions from point clouds by learning to pull space onto surfaces. In *International Conference on Machine Learning (ICML)*, 2021. 2

[2] Aude Billard and Danica Kragic. Trends and challenges in robot manipulation. *Science*, 364(6446):eaat8414, 2019. 1, 8

[3] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing: Algorithms, architectures and applications*, pages 227–236. Springer, 1990. 1

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[5] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22 (3):36–52, 2015. 6

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 2

[7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 1, 2, 6

[8] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2781–2790, 2022. 1

[9] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 6, 7

[10] Yunqiang Chen, Qing Wang, Hong Chen, Xiaoyu Song, Hui Tang, and Mengxiao Tian. An overview of augmented reality technology. In *Journal of Physics: Conference Series*, page 022082. IOP Publishing, 2019. 1, 8

[11] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30: 4008–4021, 2021. 1, 2

[12] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 231–248. Springer, 2022. 1, 2, 3, 6, 7

[13] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-Driven signed distance functions for 3D hand-object reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 1, 2, 3, 6, 7, 4

[14] Alberto Silvio Chiappa, Alessandro Marin Vargas, and Alexander Mathis. Dmap: a distributed morphological attention policy for learning to locomote with a changing body. *Advances in Neural Information Processing Systems*, 35: 37214–37227, 2022. 2

[15] Alberto Silvio Chiappa, Pablo Tano, Nisheet Patel, Abigail Ingster, Alexandre Pouget, and Alexander Mathis. Acquiring musculoskeletal skills with curriculum-based reinforcement learning. *bioRxiv*, pages 2024–01, 2024. 1, 8

[16] Julian Chibane, Gerard Pons-Moll, et al. Neural unsigned distance fields for implicit function learning. *Advances in Neural Information Processing Systems*, 33:21638–21652, 2020. 2

[17] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 2

[18] 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. 1, 2, 6

[19] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 1, 2, 3, 5, 6, 7

[20] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1, 2, 6, 7

[21] 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

[22] 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. 1, 2, 6, 7

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceed-*

ings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 3, 6

[24] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[25] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020. 1

[26] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2, 4

[27] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 5

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015. 6

[29] Vincent Lepetit. Recent advances in 3d object and hand pose estimation. *arXiv preprint arXiv:2006.05927*, 2020. 2

[30] Kailin Li, Lixin Yang, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. *arXiv preprint arXiv:2109.05488*, 2021. 1, 2

[31] Shile Li and Dongheui Lee. Point-to-pose voting based hand pose estimation using residual permutation equivariant layer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11927–11936, 2019. 5, 1

[32] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 1, 6, 7

[33] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12989–12998, 2023. 1, 2, 3, 6, 7, 4

[34] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 6, 7

[35] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 1, 2

[36] Mackenzie Weygandt Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current opinion in neurobiology*, 60:1–11, 2020. 1, 8

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[38] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2, 4

[39] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 1, 6, 7

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[41] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 1

[42] Fuji Ren and Yanwei Bao. A review on human-computer interaction and intelligent robots. *International Journal of Information Technology & Decision Making*, 19(01):5–47, 2020. 1, 8

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 4, 5, 1

[44] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 5, 6, 7

[45] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1749–1759, 2021. 1

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3, 6

[47] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 1, 2, 5, 6

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5, 6

[49] Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 1, 2, 3, 5, 6, 7, 8, 4

[50] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 3

[51] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 1, 5

[52] Hao Xu, Tianyu Wang, Xiao Tang, and Chi-Wing Fu. H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17048–17058, 2023. 6, 7

[53] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2750–2760, 2022. 6, 7

[54] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 2

[55] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021. 2

[56] Junsheng Zhou, Baorui Ma, Yu-Shen Liu, Yi Fang, and Zhizhong Han. Learning consistency-aware unsigned distance functions progressively from raw point clouds. *Advances in Neural Information Processing Systems*, 35: 16481–16494, 2022. 3

[57] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 6

[58] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 1, 6