Coarse-to-Fine Implicit Representation Learning for 3D Hand-Object Reconstruction from a Single RGB-D Image

Xingyu Liu*, Pengfei Ren*, Jingyu Wang†, Qi Qi, Haifeng Sun, Zirui Zhuang†, and Jianxin Liao

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
{liuxingyu, rpf, wangjingyu, qiqi8266, hfsun, zhuangzirui, liaojx}@bupt.edu.cn

Abstract. Recent research has explored implicit representations, such as signed distance function (SDF), for interacting hand-object reconstruction. SDF enables modeling hand-held objects with arbitrary topology and overcomes the resolution limitations of parametric models, allowing for finer-grained reconstruction. However, directly modeling detailed SDFs from visual clues presents challenges due to depth ambiguity and appearance similarity, especially in cluttered real-world scenes. In this paper, we propose a coarse-to-fine SDF framework for 3D hand-object reconstruction, which leverages the perceptual advantages of RGB-D modality in visual and geometric aspects, to progressively model the implicit field. Specifically, we model a coarse SDF for visual perception of overall scenes. Then, we propose a 3D Point-Aligned Implicit Function (3D PIFu) for fine-level SDF learning, which leverages both local geometric clues and the coarse-level visual priors to capture intricate details. Additionally, we propose a surface-aware efficient reconstruction strategy that sparsely performs SDF query based on the hand-object semantic priors. Experiments on two challenging hand-object datasets show that our method outperforms existing methods by a large margin.

Keywords: 3D hand-object reconstruction · RGB-D fusion · Implicit representation

1 Introduction

Modeling hand-object interaction [2, 9, 10, 15–18, 20, 27, 30, 51, 61] is critical for immersive applications of human-computer interaction and augmented reality. For instance, manipulating real-world objects, rather than relying solely on virtual objects or mechanical devices, can significantly promote the user experience and realism of immersive applications, such as virtual drawing and AR gaming.

* Equal contribution.
† Corresponding author.

Early mainstream efforts in the field assume known object templates, and use parametric models (e.g., MANO) with a fixed resolution, which hinders generalization and fine-grained reconstruction. Recently, some advancements, such as AlignSDF [10] and Grasping Field [27], introduce implicit representation for hand-object reconstruction, showcasing its capability to capture objects with arbitrary topology and overcome the resolution constraints of parametric models.

Existing methods primarily rely on 2D visual features to model implicit representations. However, cluttered real-world scenes are characterized by depth ambiguity and appearance similarity, making it ill-posed to directly model the fine-grained shape details based on such ambiguous visual information. For example, hand-object interaction in the wild often suffers from severe self-occlusion, occlusion of multiple overlapping objects, and cluttered backgrounds. These ambiguities hinder the elaborate modeling of SDF. Previous methods often produce implausible predictions in cluttered scenes, such as mesh collapse in occluded regions, and the inadvertent merging of meshes from adjacent objects.

Additionally, existing SDF-based approaches suffer from computationally expensive reconstruction procedures due to the dense evaluation of the network in 3D space. A common procedure involves mapping the densely sampled points in voxelized 3D space to signed distance values, followed by the employment of the Marching Cube algorithm [38] to extract the high-fidelity hand-object mesh. During this process, a substantial number of sampling points far from the surface of hands and objects leads to redundant SDF queries, thereby increasing the computational complexity during reconstruction.

With the widespread integration of depth sensors in mobile devices and head-mounted displays, many depth-based or RGB-D fusion-based methods have been widely studied in 3D hand pose estimation and reconstruction tasks [8, 12, 22, 31, 35–37, 49]. Depth modality provides fine-grained 3D geometric structure clues and can eliminate the depth ambiguity of visual features. Conversely, visual features can offer semantic and global context information that depth data lacks. By relying on global-aware visual information and local-aware geometric clues, we enable SDF learning as a gradual and steady process, which avoids the abrupt and simplistic direct modeling of SDF in existing methods.

Inspired by the above motivation, we decouple the implicit field modeling into a coarse-to-fine approach, by progressively leveraging the unique advantages of visual and geometric information. In particular, at the coarse level, we employ global visual features to discern the position and overall shape of the hand and object. Subsequently, at the fine level, we propose a 3D Point-Aligned Implicit Function (3D PIFu), utilizing fine-grained geometric clues of depth point cloud to capture intricate shape details. To establish a transition from coarse to fine, the coarse-level SDF enriches the 3D point cloud with the global prior knowledge of hand and object, such as point-level semantic information, which provides a holistic perception of global scenes for subsequent local fine-level SDF learning.

In addition, the point-level semantics convey the position information of the hand-object surface in 3D space, which can guide the elimination of redundant SDF computations in empty regions. We propose an efficient hand-object recon-

struction approach based on this surface-aware sparse query strategy. Leveraging hierarchical semantic information, we focus SDF queries exclusively on 3D regions containing the hand-object surface, to reduce redundant inference for sampling points located far from the hand and object surfaces. Code is available at `https://github.com/ru1ven/C2F-SDF`.

Our contributions can be summarized as follows:

1) We propose an RGB-D coarse-to-fine SDF framework for 3D hand-object reconstruction, which models implicit representations transitioning from global to local, by leveraging visual and geometric modalities in a progressive manner.

2) We propose a 3D Point-Aligned Implicit Function (3D PIFu) to aggregate multi-modal features in 3D point cloud space, mitigating the challenges posed by depth ambiguity in implicit representation learning.

3) We propose a Surface-Aware Efficient Reconstruction strategy (SER), which sparsely perform SDF queries guided by hierarchical semantic information, to reduce redundant inference of query points far from the surface.

4) Extensive experiments on synthetic and real-world hand-object datasets show that our method achieves state-of-the-art (SOTA) performance.

2 Related Works

2.1 Hand Pose Estimation and Reconstruction

In recent years, numerous model-based and model-free hand pose estimation and reconstruction methods have been proposed to predict the 3D positions of hand joints or reconstruct hand meshes from RGB and depth images. Model-based works [1, 7, 28, 32, 52, 60, 62, 64] reconstruct the hand mesh by estimating the parameters of MANO models, leveraging hand shape priors to obtain more plausible results. To address the highly non-linear mapping of directly regressing MANO parameters from images, several model-free approaches [11, 15, 25, 33, 34] use the Graph Convolutional Network (GCN) or Transformers to model the relationships between hand joints or vertices, and then predict their 3D coordinates. Alternatively, some works predict vertices and joints through other representations, such as heatmap [24, 39–41, 63], UV map [5], and implicit representations [21, 23]. However, due to the limited resolution of the parametric model and the predetermined number of hand vertices, existing methods have difficulty in recovering high-fidelity hand surfaces and cannot show vivid hand details in immersive interactive applications.

2.2 Hand-Object Reconstruction

Hand-object reconstruction plays an important role in virtual reality, augmented reality, and interactive applications, and is a challenging problem due to mutual occlusion [44, 55]. Since many hand-object interaction benchmarks have been released, numerous studies have been conducted to reconstruct the interacting hand-object mesh. Some works [15, 16, 57, 58] assume known instance-specific object templates during inference, and reduce the object reconstruction to 6DoF

pose estimation. In addition, several approaches [2,13,17,26] adopt optimization-based contact modeling to jointly fit hand-object meshes. To achieve the generalization of unknown object classes, Hasson et al. [18] use the MANO model to reconstruct 3D hand meshes, and simultaneously use AtlasNet [14] to deform object vertices from a sphere. However, the resolution limitations of parametric models and AtlasNet prevent fine-grained hand-object mesh. To address this problem, recent works introduce implicit representations, such as SDF, for high-fidelity hand-object mesh reconstruction. Karunratanakul et al. [27] propose a two-branch network to represent the hands and objects in a joint implicit field. However, implicit representation lacks shape priors and often produces disembodied and broken hands due to mutual occlusion. To mitigate this, several works incorporate prior knowledge into SDF learning explicitly. For instance, AlignSDF [10] combines the advantages of parametric models and implicit representation by encoding pose priors into SDF. Chen et al. [9] and Ye et al. [59] encode kinematic features into SDF to help reconstruct the hand-object mesh. Despite integrating the pose prior knowledge, previous methods model SDF heavily relying on 2D visual features, which are prone to depth and appearance ambiguity. Consequently, they often produce implausible predictions in cluttered scenes. In contrast, our method leverages visual information and geometric structure cues in a coarse-to-fine manner, enabling implicit representation learning as a progressive and steady process.

3 Method

Fig. 1 illustrates the overview of our framework. Given the input RGB and depth images, we employ an RGB-D encoder to extract multi-modal visual features, and lift them into 3D space to construct point cloud features through a 2D-3D projection module (Section 3.1). The SDF optimization adopts a coarse-to-fine strategy (Section 3.2 and 3.3). We utilize global image features for coarse-level SDF learning and point-level semantic refinement, thereby enhancing global context perception (Section 3.2). Then, we learn a 3D Point-Aligned Implicit Function (3D PIFu) for hand and object branches, respectively, enabling the fine-level SDF learning to perceive local geometric structure information (Section 3.3). To efficiently reconstruct hand and object meshes, we construct hierarchical semantic information, to perform sparse SDF queries exclusively in regions potentially containing surfaces (Section 3.4).

3.1 Multi-Modal Point Cloud Feature Extraction

Given the input RGB and depth images, we adopt two parallel hourglass networks [42] as the dual-branch encoder, to extract the RGB-D global image feature $F_{global} \in \mathbb{R}^C$ and the RGB-D visual feature $F_{visual} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent the height, width, and channel dimension of features. During feature extraction, we fuse RGB and depth features through a spatial attention mechanism [54], similar to previous RGB-D image-based fusion

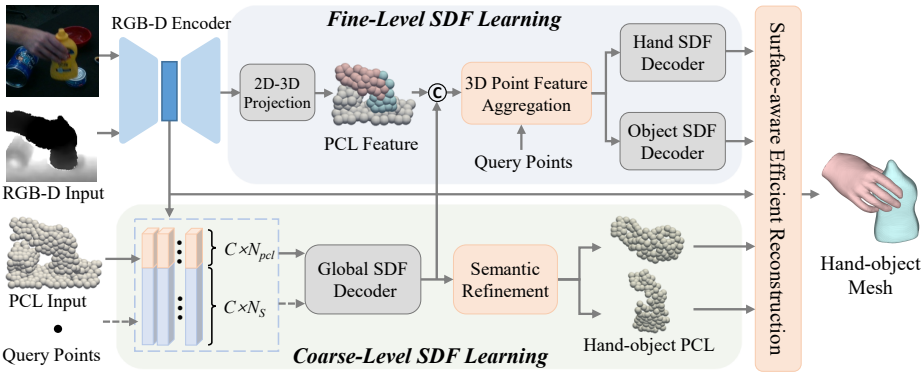


Fig. 1: Overview. *PCL* represents the point cloud. The dotted line represents the data flow only during training. We utilize an RGB-D encoder to extract global image features, followed by a coarse-level SDF modeling for point cloud semantic refinement. Then, we adaptively aggregate 3D point features from the coarse-level visual priors and geometric clues of the point cloud for fine-level SDF learning. Finally, we extract the hand-object mesh through the surface-aware efficient reconstruction strategy.

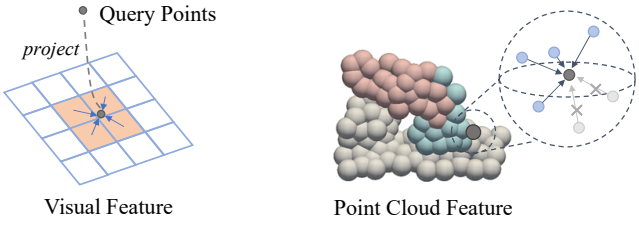
approaches [6, 19]. Instead of applying point-based networks such as PointNet [46, 47] to extract the point cloud feature, which is computationally expensive, we follow [49] to lift the RGB-D visual feature into 3D space through a 2D-3D projection. Specifically, based on the depth value of the corresponding image coordinate in the downsampled depth map, we calculate the 3D coordinates of each RGB-D feature pixel through the camera intrinsic parameters, to obtain the multi-modal point cloud feature $F^{3d} \in \mathbb{R}^{N_{pcl} \times C}$, where N_{pcl} indicates the number of depth points.

3.2 Point Cloud Feature Refinement

In this section, we demonstrate the utilization of coarse-level SDF learning for global context perception and point cloud feature semantic refinement.

Coarse-Level SDF Learning. For each query point $x \in \mathbb{R}^3$ sampled in 3D space, we employ a global SDF decoder to map its initial signed distances to the hand and object surface. Each query point is encoded with the global image feature F_{global}, enabling the network to better perceive global context information.

Point-Level Semantic Information Incorporation. Point clouds provide accurate depth information and capture the geometric properties in 3D space. However, the lack of semantic information in the point cloud prevents a more comprehensive understanding of the 3D scene. Several 3D object detection methods [53, 56] append point clouds with image-based semantics, which demonstrates the effectiveness of point-level semantic information for 3D scene understanding. Instead of employing the off-the-shelf semantic segmentation networks, we append semantic information on point clouds in an end-to-end manner



(a) 2D Pixel Feature Alignment (b) 3D Point Feature Aggregation

Fig. 2: Comparison of feature encoding operations for (a) Pixel-Aligned Implicit Function and (b) proposed 3D Point-Aligned Implicit Function. During aggregation, the background and noisy points (gray points) are masked based on their SDF values.

through the coarse-level SDF. Specifically, we concatenate the point cloud with query points and input them into the global SDF decoder. Then, by mapping the obtained signed distance value of each depth point to the hand-object semantic scores, we extract 3D points near the hand-object surface as hand and object point clouds, while points far from the surface are filtered out as noise and background. The refined point cloud feature as well as the point-level semantic information provide a comprehensive perception of 3D scenes, and can be utilized to promote subsequent fine-level SDF learning in cluttered scenes.

3.3 Fine-Level SDF Learning

Preliminary: Pixel-Aligned Implicit Function To reconstruct the underlining 3D geometry and texture of a clothed human, Saito et al. [50] introduce a Pixel-Aligned Implicit Function (PIFu), which defines a surface as the zero-level set of a function f:

$$f(F(\pi(x)), Z(x)) = 0, \tag{1}$$

where for a 3D query point x, $Z(x)$ represents the depth value in the camera coordinate space, $\pi(x)$ represents the 2D projection location of x, and $F(\pi(x))$ represents the pixel-aligned image feature obtained through the bilinear sampling. The pixel-aligned feature allows the learned functions to preserve the local detail present in the image. However, the 2D pixel-aligned visual feature lacks global context information and is prone to depth ambiguity.

3D PIFu: 3D Point-Aligned Implicit Function To alleviate the aforementioned problems, we propose a 3D Point-Aligned Implicit Function (3D PIFu), defining the surface as:

$$f(F_{PA}, x) = 0, \tag{2}$$

where for each query point x, F_{PA} represents the 3D point-aligned feature, which is obtained by adaptive weighted aggregation of point cloud features in the 3D neighborhood of the query point. Specifically, as shown in Fig. 2 (b), based

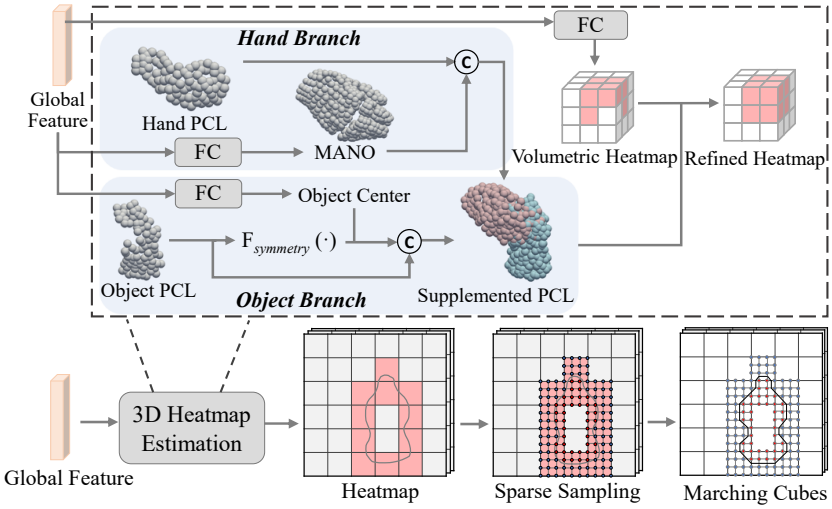


Fig. 3: Details of the surface-aware efficient reconstruction. For visualization purposes, we illustrate sparse sampling and Marching Cubes in 2D.

on the 3D coordinates of the query point, we select K closest point features from the point cloud feature F^{3d}, and fuse the spatial position information and signed distance information into the selected point features through channel de-differentiation operations [48]. The k-th point feature is donated as:

$$F_k^{point} = ReLU(BN(w_0 F_k^{3d} + w_1 P_k + w_2 D_k), \tag{3}$$

where $P_k \in \mathbb{R}^3$ represents k-th depth point, $D_k \in \mathbb{R}$ represents the signed distance value from k-th point to the surface of the hand or the object, w_0, w_1, and w_2 are learnable parameter matrices for point feature embedding, $ReLU$ and BN represent ReLU activation function and batch normalization layer, and K is set to 16 by default. Then, according to the distances from each point feature to the query point and to the hand-object surface, we adaptively aggregate K point features to generate the 3D point-aligned feature:

$$F_{PA} = \sum_{k=0}^{K-1} \frac{1}{d_{x,P_k} + \alpha D_k} F_k^{point}, \tag{4}$$

where d_{x,P_k} represents the 3D Euclidean distance between the k-th point feature and the query point, and α is a learnable parameter used to adjust the distance scale. In this way, 3D PIFu can effectively leverage local geometric structure information, thereby reducing depth ambiguity. Moreover, refined point cloud feature provides semantic priors of hands and objects, which enables fine-level SDF learning to perceive global context information and eliminate semantic ambiguity.

Table 1: Ablation study for the feature encoding of SDF learning on the DexYCB dataset. *Pixel* and *3D Point* represent pixel-aligned feature and 3D point-aligned feature. *Ref* represents point cloud feature refinement. *SER* represents the surface-aware efficient reconstruction strategy.

ID	Pixel (RGB)	Pixel (RGBD)	3D Point	Ref	SER	$CD_h\downarrow$	$F_h@1\uparrow$	$F_h@5\uparrow$	$CD_o\downarrow$	$F_o@5\uparrow$	$F_o@10\uparrow$
0	✓					0.334	0.163	0.782	2.04	0.391	0.660
1		✓				0.299	0.171	0.802	1.50	0.446	0.723
2			✓			0.280	0.180	0.815	1.56	0.456	0.732
3			✓	✓		0.272	0.183	0.819	1.26	0.485	**0.764**
4			✓	✓	✓	**0.267**	**0.185**	**0.823**	**1.24**	**0.488**	**0.764**

3.4 Surface-Aware Efficient Reconstruction

For hand-object reconstruction, existing SDF-based methods [9, 10, 27, 59] voxelize the 3D space based on specific resolution and map the densely sampled points to signed distance values. Then, they apply the Marching Cube algorithm [38] to extract high-fidelity hand-object meshes. However, this procedure frequently involves redundant calculations for numerous query points situated far from the hand-object surface, resulting in inefficient reconstruction.

To enhance the efficiency of reconstruction, we estimate a volumetric semantic heatmap by leveraging the hand-object point clouds and a learnable volumetric heatmap, and selectively sample query points in proximity to the hand-object surface, thereby reducing unnecessary computations for distant points. As shown in Fig. 3, we first use the global feature F_{global} to predict a low-resolution heatmap. Each voxel value corresponds to the density of hand vertices and object vertices within the 3D region. We apply a $3\times3\times3$ average filter to smooth out the outliers in the heatmap. Additionally, to enhance the robustness of our reconstruction results, we use the shape priors of hands and objects to supplement sparse point clouds. In particular, to address challenges posed by missing and occluded points, we predict MANO vertices to complete the hand point clouds. Simultaneously, by capitalizing on the inherent symmetry of the object, we optimize the object center and compute central symmetry points to supplement the object point cloud. Then, we merge the hierarchical surface position information from the supplemented point cloud and the volumetric heatmap, to obtain a volumetric semantic heatmap. Finally, guided by the semantic heatmap, we conduct sparse SDF queries on sampling points within the corresponding voxels.

4 Experiment

4.1 Datasets and Evaluation Metrics

DexYCB [4] is a hand-object dataset captured by multiple RGB-D cameras, containing 582K RGB-D frames over 1,000 sequences of 10 subjects grasping 20 different objects from 8 views. We follow the same dataset split in [9], filtering

Table 2: Comparison with more RGB-D fusion-based methods on DexYCB.

Methods	CD_h↓	F_h@1↑	F_h@5↑	CD_o↓	F_o@5↑	F_o@10↑
gSDF [9]	0.302	0.177	0.801	1.55	0.437	0.709
RGB-D PIFu	0.299	0.171	0.802	1.50	0.446	0.723
gSDF + SA-Gate [6]	0.279	0.178	0.809	1.28	0.478	0.757
Ours	**0.267**	**0.185**	**0.823**	**1.24**	**0.488**	**0.764**

Table 3: Efficiency analysis of the surface-aware efficient reconstruction strategy (SER) on DexYCB. RT and N_S represent the average reconstruction time (s) and average number of sampling points per frame on the DexYCB test set, respectively.

ID	Methods	Sparse sampling	CD_h↓	F_h@1↑	F_h@5↑	CD_o↓	F_o@5↑	F_o@10↑	RT	N_S
0	Ours w/o SER	×	0.272	0.183	0.819	1.26	0.485	0.764	3.44	2.1M
1	gSDF [9]	×	0.302	0.177	0.801	1.55	0.437	0.709	1.51	2.1M
2	Volumetric heatmap	✓	0.279	0.182	0.815	1.32	0.475	0.746	1.23	83.0K
3	Heatmap+PCL	×	0.267	0.185	0.823	1.23	0.490	0.768	3.45	2.1M
4	Heatmap+PCL	✓	0.267	0.185	0.823	1.24	0.488	0.764	1.26	89.8K

samples without hand-object interactions and downsampling the video data to 6 frames per second, which obtains 29,656 training samples and 5,928 testing samples. **ObMan** [18] is a large-scale synthetic image dataset of hands-grasping objects, containing 21K hand grasp poses for 2.7K objects of 8 categories from ShapeNet [3]. We follow previous works [9,27,43] to split the training and testing sets, and remove meshes that contain too many double-sided triangles, to obtain 87,190 training samples and 6,285 testing samples.

Evaluation metrics. Following previous works [9, 59], for the evaluation of hand reconstruction, we report Chamfer distance in cm^2 ($\mathbf{CD}_h$) and F-score evaluated at thresholds of 1mm and 5mm ($\mathbf{F}_h$**@1** and $\mathbf{F}_h$**@5**). For object reconstruction, we report Chamfer distance in cm^2 ($\mathbf{CD}_o$) and F-score at 5mm and 10mm thresholds ($\mathbf{F}_o$**@5** and $\mathbf{F}_o$**@10**). In the supplementary material, we additionally report the penetration depth, intersection volume, and contact ratio, to evaluate the physical quality of hand-object contact.

4.2 Implementation Details

Our experiments are conducted with an NVIDIA RTX 4090 GPU. The network is implemented based on PyTorch [45]. We use an AdamW optimizer [29] with an initial learning rate of 1e-4. The whole training process takes 800 epochs with a batch size of 32 on both DexYCB and ObMan datasets, and the learning rate is decayed by half at 600th epoch. For data augmentation, we crop the input RGB-D images to the size of 256×256, and perform random rotation and color jittering. In all conducted experiments, we reconstruct hand and object meshes

Table 4: Comparison with state-of-the-art methods on the ObMan dataset.

Methods	CD_h↓	F_h@1↑	F_h@5↑	CD_o↓	F_o@5↑	F_o@10↑
Hasson et al. [18]	0.415	0.138	0.751	3.60	0.359	0.590
Grasping Field [27]	0.261	-	-	6.80	-	-
Ye et al. [59]	-	-	-	-	0.420	0.630
DDF-HO [61]	-	-	-	-	0.550	0.670
AlignSDF [10]	0.136	0.302	0.913	3.38	0.404	0.636
gSDF [9]	0.112	0.332	0.935	3.14	0.438	0.660
Ours	**0.083**	**0.416**	**0.959**	**0.51**	**0.780**	**0.891**

Table 5: Comparison with state-of-the-art methods on the DexYCB dataset.

Methods	CD_h↓	F_h@1↑	F_h@5↑	CD_o↓	F_o@5↑	F_o@10↑
Hasson et al. [18]	0.537	0.115	0.647	1.94	0.383	0.642
Grasping Field [27]	0.364	0.154	0.764	2.06	0.392	0.660
AlignSDF [10]	0.358	0.162	0.767	1.83	0.410	0.679
gSDF [9]	0.302	0.177	0.801	1.55	0.437	0.709
Ours	**0.267**	**0.185**	**0.823**	**1.24**	**0.488**	**0.764**

with a resolution of 128×128×128. More details about the network architecture, data preparation, and training losses are provided in the supplementary material.

4.3 Ablation Study

Feature Encoding for SDF Learning To verify the effectiveness of the proposed 3D PIFu, we evaluate various feature encoding strategies for SDF modeling. As shown in Table 1, compared to the RGB and RGB-D baseline models (ID 0 and ID 1) that only utilize pixel-aligned visual features, the incorporation of 3D point-aligned features (ID 2) for optimizing the SDF results in superior performance across almost all metrics. Additionally, by performing the point cloud feature refinement through coarse-level SDF learning (ID 3) and the surface-aware reconstruction strategy (ID 4), the network can further bring a significant performance improvement.

Comparisons with More RGB-D Methods To further demonstrate the effectiveness of our method, we conduct a quantitative comparison with more RGB-D based methods. First, we re-implement an RGB-D based PIFu [50]. We utilize the RGB-D encoder of our model to generate the RGB-D visual features, and adopt the PIFu to model SDF. Second, we construct a strong RGB-D baseline by incorporating an existing RGB-D fusion method into single-modal hand-object reconstruction methods. Specifically, we adopt a common RGB-D image fusion method, SA-Gate [6], which utilizes channel-wise and spatial-wise

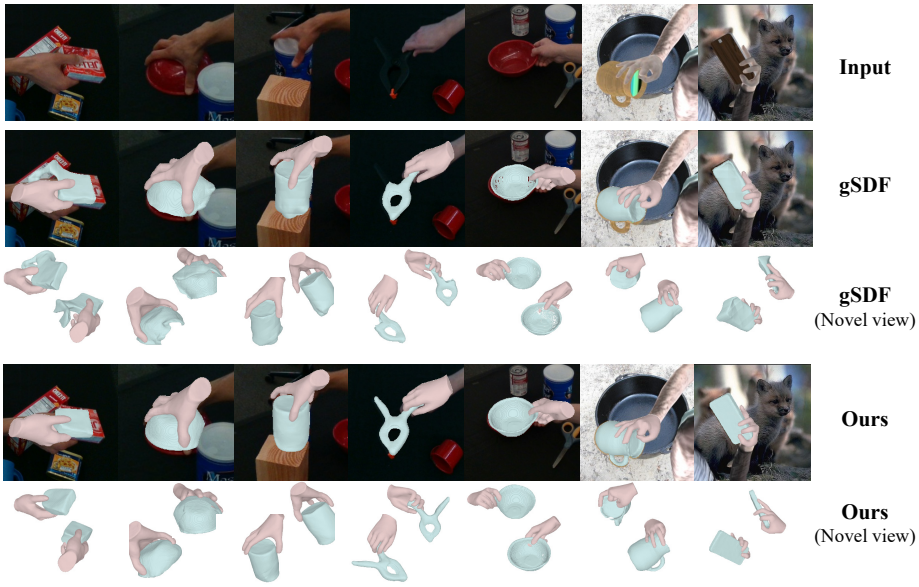


Fig. 4: Qualitative results of gSDF [9] and our method on DexYCB and ObMan.

soft attention to aggregate RGB-D features per pixel. We incorporate this fusion strategy into the current SOTA, gSDF [9], by replacing the single-modal image encoder with the RGB-D encoder. As shown in Table 2, our method achieves leading performance on all metrics.

Efficiency Analysis of Hand-Object Reconstruction To analyze the effectiveness of the proposed surface-aware efficient reconstruction strategy, we compare the performance and efficiency of different reconstruction strategies in Table 3. First, compared to our method without the surface-aware efficient reconstruction (ID 0), our full approach incorporating the sparse sampling (ID 4) can maintain the original reconstruction accuracy, while significantly reducing the average reconstruction time on the DexYCB test set, from 3.44 seconds to 1.26 seconds, and decreasing the number of sampling points from 128^3 to 89.8K. Second, despite the introduction of additional modalities, our method achieves faster reconstruction speed compared to the SOTA single-modal method (ID 1). Additionally, relying exclusively on heatmaps to perform sparse queries (ID 2) leads to reduced reconstruction accuracy, despite offering slightly lower latency. Furthermore, by ablating the sparse sampling process, we observe that the optimization of heatmaps and MANO boosts network performance (ID 3). We speculate that these components provide valuable prior knowledge about the shape and potential location of the surface.

Fig. 5: Results on consecutive DexYCB video frames.

4.4 Comparisons with State-of-the-arts

We present the performance comparison of hand-object reconstruction on the ObMan dataset. Since several works only focus on hand-held objects without performing hand mesh reconstruction, we only report their object reconstruction results. As shown in Table 4, our method outperforms existing hand-object reconstruction methods and hand-held object reconstruction methods by a large margin on F-score at various thresholds, and achieves better performance in terms of Chamfer distance, for both hand ($0.083cm^2$ vs. $0.112cm^2$) and object ($0.51cm^2$ vs. $3.14cm^2$).

The performance comparison with SOTA methods on the DexYCB dataset is shown in Table 5. Our method demonstrates outstanding performance across all metrics, outperforming the current SOTA method, gSDF [9], by a large margin in terms of F-score and Chamfer distance for both hand ($0.267cm^2$ vs. $0.302cm^2$) and object ($1.24cm^2$ vs. $1.55cm^2$).

4.5 Qualitative Results

Qualitative Comparison with SOTAs We present some qualitative comparisons on DexYCB and ObMan in Fig. 4. Compared with gSDF [9], our method

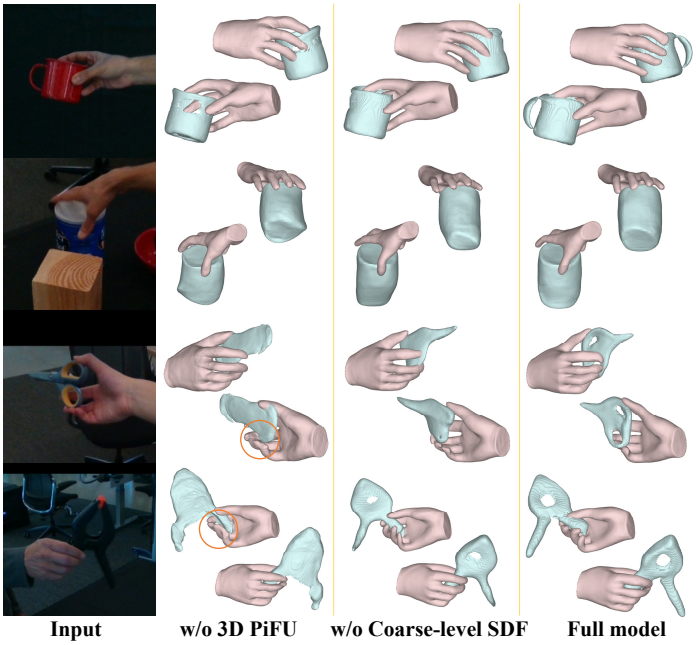


Fig. 6: Qualitative ablation study on DexYCB.

can better avoid mesh collapse for occlusion from stacked objects (columns 2 and 3) and mesh merging for adjacent objects with similar appearances (column 1). Additionally, our method reconstructs more complete meshes for objects with complex shapes (columns 4 and 6), and effectively avoids mesh holes for thin objects such as bowls (column 5). Meanwhile, ours achieves better hand-object mesh alignment in the contact area (columns 3 and 4) and reconstructs more plausible meshes for complex hand poses (columns 6 and 7).

Reconstruction Results of Consecutive Video Frames As shown in Fig. 5, in cluttered scenes with multiple stacked and obstructed objects (top three rows), and scenarios involving hands with contorted poses and objects with intricate shapes (bottom three rows), our method showcases superior generalization capabilities in consecutive frames.

Qualitative Ablation Study In Fig. 6, we show the importance of 3D PiFU for modeling high-frequency details, and coarse-level SDF for overall shape. First, abandon fine-level SDF modeling with 3D PiFU, some shape details such as fingers and holes cannot be recovered well (rows 3 and 4). Second, with coarse-level SDF learning, the overall shape of objects reconstructed through our full model is more regular.

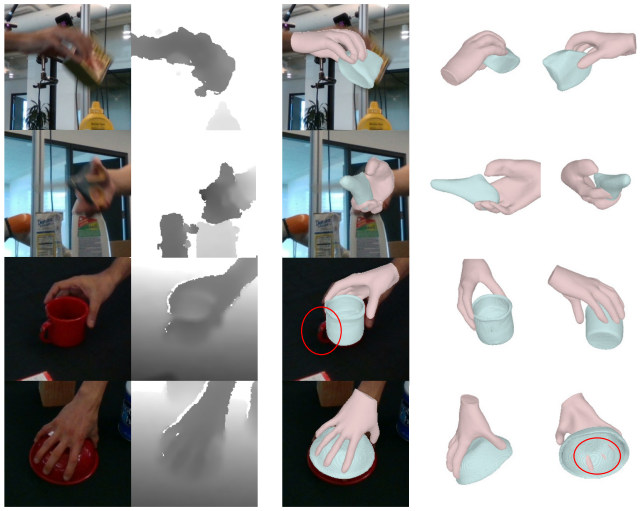


Fig. 7: Failure examples of our method on DexYCB.

Failure Examples As shown in Fig. 7, the blurring of hand and object motion leads to severe noise in depth modality, and our method produces irregular shapes (row 1) and collapses (row 2) in such cases. Second, highly geometrically complex objects may result in incomplete reconstruction, such as the handle of cups (row 3). Meanwhile, tightly interaction may result in reconstructed hands penetrating thin objects such as bowls (row 4).

5 Conclusion

In this paper, we propose a coarse-to-fine SDF framework for 3D hand-object reconstruction. The key insight is to employ implicit representation as a medium to progressively leverage the perceptual advantages of RGB-D modalities. We achieve this by using the visual-based coarse-level SDF networks to refine point clouds, and in turn, using geometric clues from point clouds to prevent fine-level SDF learning from global and local depth ambiguities. Additionally, based on the hand-object semantic priors, we reduce SDF queries far from the surface during inference, achieving reconstruction efficiency even surpassing that of single-modal methods. Qualitative and quantitative results on DexYCB and ObMan show that our method outperforms existing methods by a large margin.

Limitation. Our method does not explicitly model physical constraints such as penetration and contact between hands and objects, thus leading to accidental interpenetration and loose contact between hands and objects. A promising future direction is to use implicit functions to model differentiable physical constraints and optimize the hand-object contact to obtain more plausible grasping.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants (62101064, 62171057,62201072, U23B2001, 62001054, 62071067), the Ministry of Education and China Mobile Joint Fund (MCM20200202, MCM20180101), Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center, BUPT Excellent Ph.D. Students Foundation (CX20241014), in part by the Project funded by China Postdoctoral Science Foundation (2023TQ0039, 2024M750257), the Postdoctoral Fellowship Program of CPSF (GZC20230320).

References

1. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
2. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF International conference on computer vision (ICCV). pp. 12417–12426 (2021)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9044–9053 (2021)
5. Chen, P., Chen, Y., Yang, D., Wu, F., Li, Q., Xia, Q., Tan, Y.: I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In: Proceedings of the IEEE/CVF International conference on computer vision (ICCV). pp. 12929–12938 (2021)
6. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 561–577. Springer (2020)
7. Chen, X., Liu, Y., Ma, C., Chang, J., Wang, H., Chen, T., Guo, X., Wan, P., Zheng, W.: Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13283 (2021)
8. Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R., Yuan, J.: So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
9. Chen, Z., Chen, S., Schmid, C., Laptev, I.: gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12890–12900 (2023)
10. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In: European Conference on Computer Vision (ECCV). pp. 231–248. Springer (2022)

11. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3d hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10833–10842 (2019)
12. Ge, L., Ren, Z., Yuan, J.: Point-to-point regression pointnet for 3d hand pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 475–491 (2018)
13. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1471–1481 (2021)
14. Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3d surface generation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR). pp. 216–224 (2018)
15. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: 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 (CVPR). pp. 11090–11100 (2022)
16. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp. 571–580 (2020)
17. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: 2021 International Conference on 3D Vision (3DV). pp. 659–668. IEEE (2021)
18. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp. 11807–11816 (2019)
19. Hu, X., Yang, K., Fei, L., Wang, K.: Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1440–1444. IEEE (2019)
20. Huang, D., Ji, X., He, X., Sun, J., He, T., Shuai, Q., Ouyang, W., Zhou, X.: Reconstructing hand-held objects from monocular video. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
21. Huang, L., Lin, C.C., Lin, K., Liang, L., Wang, L., Yuan, J., Liu, Z.: Neural voting field for camera-space 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8969–8978 (June 2023)
22. Huang, W., Ren, P., Wang, J., Qi, Q., Sun, H.: Awr: Adaptive weighting regression for 3d hand pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11061–11068 (2020)
23. Huang, Z., Chen, Y., Kang, D., Zhang, J., Tu, Z.: Phrit: Parametric hand representation with implicit template. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14974–14984 (October 2023)
24. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5d heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
25. Jiang, C., Xiao, Y., Wu, C., Zhang, M., Zheng, J., Cao, Z., Zhou, J.T.: A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose

estimation from a single rgb image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8846–8855 (June 2023)
26. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE/CVF International conference on computer vision (ICCV). pp. 11107–11116 (2021)
27. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 2020 International Conference on 3D Vision (3DV). pp. 333–344. IEEE (2020)
28. Kong, D., Zhang, L., Chen, L., Ma, H., Yan, X., Sun, S., Liu, X., Han, K., Xie, X.: Identity-aware hand mesh estimation and personalization from rgb images. In: European Conference on Computer Vision (ECCV). pp. 536–553. Springer (2022)
29. Kulon, D., Wang, H., Güler, R.A., Bronstein, M., Zafeiriou, S.: Single image 3d hand reconstruction with mesh convolutions. arXiv preprint arXiv:1905.01326 (2019)
30. Leng, Z., Wu, S.C., Saleh, M., Montanaro, A., Yu, H., Wang, Y., Navab, N., Liang, X., Tombari, F.: Dynamic hyperbolic attention network for fine hand-object reconstruction. In: Proceedings of the IEEE/CVF international conference on computer Vision. pp. 14894–14904 (2023)
31. Li, L., Zhuo, L., Zhang, B., Bo, L., Chen, C.: Diffhand: End-to-end hand mesh reconstruction via diffusion models. arXiv preprint arXiv:2305.13705 (2023)
32. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2761–2770 (June 2022)
33. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR). pp. 1954–1963 (2021)
34. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 12939–12948 (2021)
35. Liu, X., Ren, P., Chen, Y., Liu, C., Wang, J., Sun, H., Qi, Q., Wang, J.: Safusion: Multimodal fusion approach for web-based human-computer interaction in the wild. In: Proceedings of the ACM Web Conference 2023. pp. 3883–3891 (2023)
36. Liu, X., Ren, P., Chen, Y., Liu, C., Wang, J., Sun, H., Qi, Q., Wang, J.: Sample-adapt fusion network for rgb-d hand detection in the wild. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5. IEEE (2023)
37. Liu, X., Ren, P., Gao, Y., Wang, J., Sun, H., Qi, Q., Zhuang, Z., Liao, J.: Keypoint fusion for rgb-d based 3d hand pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3756–3764 (2024)
38. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. In: Seminal graphics: pioneering efforts that shaped the field, pp. 347–353 (1998)
39. Moon, G., Chang, J.Y., Lee, K.M.: V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
40. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16. pp. 752–768. Springer (2020)

41. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 548–564. Springer (2020)
42. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 483–499. Springer (2016)
43. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
44. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1496–1505 (2022)
45. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
46. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
47. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
48. Ran, H., Liu, J., Wang, C.: Surface representation for point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18942–18952 (June 2022)
49. Ren, P., Chen, Y., Hao, J., Sun, H., Qi, Q., Wang, J., Liao, J.: Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
50. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
51. Tse, T.H.E., Kim, K.I., Leonardis, A., Chang, H.J.: 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. pp. 1664–1674 (2022)
52. Tu, Z., Huang, Z., Chen, Y., Kang, D., Bao, L., Yang, B., Yuan, J.: Consistent 3d hand reconstruction in video via self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9469–9485 (2023)
53. Vora, S., Lang, A.H., Helou, B., Beijbom, O.: Pointpainting: Sequential fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
54. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
55. Xu, H., Wang, T., Tang, X., Fu, C.W.: 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 (CVPR). pp. 17048–17058 (June 2023)

56. Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., Zhang, L.: Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). pp. 3047–3054. IEEE (2021)
57. Yang, L., Li, K., Zhan, X., Lv, J., Xu, W., Li, J., Lu, C.: 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 (CVPR). pp. 2750–2760 (2022)
58. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: Cpf: Learning a contact potential field to model the hand-object interaction. In: Proceedings of the IEEE/CVF International conference on computer vision (ICCV). pp. 11097–11106 (2021)
59. Ye, Y., Gupta, A., Tulsiani, S.: 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 (CVPR). pp. 3895–3905 (2022)
60. Zhang, B., Wang, Y., Deng, X., Zhang, Y., Tan, P., Ma, C., Wang, H.: Interacting two-hand 3d pose and shape reconstruction from single color image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11354–11363 (October 2021)
61. Zhang, C., Di, Y., Zhang, R., Zhai, G., Manhardt, F., Tombari, F., Ji, X.: Ddfho: Hand-held object reconstruction via conditional directed distance field. arXiv preprint arXiv:2308.08231 (2023)
62. Zhang, X., Huang, H., Tan, J., Xu, H., Yang, C., Peng, G., Wang, L., Liu, J.: Hand image understanding via deep multi-task learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11281–11292 (October 2021)
63. Zheng, X., Ren, P., Sun, H., Wang, J., Qi, Q., Liao, J.: Sar: Spatial-aware regression for 3d hand pose and mesh reconstruction from a monocular rgb image. In: 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 99–108. IEEE (2021)
64. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: Monocular real-time hand shape and motion capture using multi-modal data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)