Nonrigid Object Contact Estimation With Regional Unwrapping Transformer

Wei Xie Zimeng Zhao Shiying Li Binghui Zuo Yangang Wang*

Southeast University, China

Abstract

Acquiring contact patterns between hands and nonrigid objects is a common concern in the vision and robotics community. However, existing learning-based methods focus more on contact with rigid ones from monocular images. When adopting them for nonrigid contact, a major problem is that the existing contact representation is restricted by the geometry of the object. Consequently, contact neighborhoods are stored in an unordered manner and contact features are difficult to align with image cues. At the core of our approach lies a novel hand-object contact representation called RUPs (Region Unwrapping Profiles), which unwrap the roughly estimated hand-object surfaces as multiple high-resolution 2D regional profiles. The region grouping strategy is consistent with the hand kinematic bone division because they are the primitive initiators for a composite contact pattern. Based on this representation, our Regional Unwrapping Transformer (RUFormer) learns the correlation priors across regions from monocular inputs and predicts corresponding contact and deformed transformations. Our experiments demonstrate that the proposed framework can robustly estimate the deformed degrees and deformed transformations, which makes it suitable for both nonrigid and rigid contact.

1. Introduction

Perceptions of hand-object contact patterns are crucial to advance human-computer interaction and robotic imitation [44]. The interactive objects in these applications, from mouse/keyboard to bottle/doll, are mostly nonrigid. Although impressive progress has been achieved towards monocular contact estimation between hands and 3D rigid objects [12, 41, 47, 35] or 2.5D cloth [37, 36, 1], it is still difficult to extend them to 3D nonrigid object. One important reason is that existing methods usually project the contact area of different objects onto their own surface (point cloud or mesh), which is represented by either unordered points or unregistered points and edges. As a result, it is challenging to store contact into a feature-aligned space.

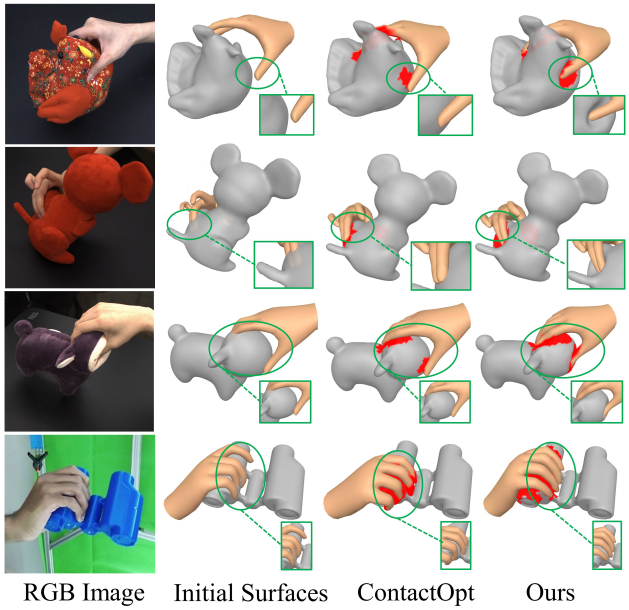

Figure 1: **Contact patterns estimated from monocular RGB images.** Since the deformed degrees of the contact areas are considered by our framework, both contact with nonrigid (Row1, Row2, Row3) and rigid objects (Row4) can be plausibly estimated.

To conquer this obstacle, our key idea is to **first represent regional 3D surface where hand-object contact may occur as regional 2D unwrapping profiles, then predict the nonrigid contact and deformation within/across regions according to monocular image cues through a Vision Transformer**. Considering that the mutual contact is caused by individual hand regions [17, 41, 47], our surface grouping is based on the 16 hand kinematic bones [30, 28, 41, 47] illustrated in Fig. 2(a). Each piece of object surface shown in Fig. 2(b) is divided into a cer-

*Corresponding author. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), the Natural Science Foundation of Jiangsu Province (No. BK20220127).

tain group when it can be directly intersected by a ray emanating from the region center associated with this group. Each subsurface is further mapped to the image plane according to the spherical unwrapping algorithm [46]. Consequently, the whole object surface is converted to 16 object *regional unwrapping profile*s (object-RUPs). Similarly, the hand surface is converted to 16 hand-RUPs, each of which records pixel-aligned ray intersections with the object-RUP in the same group. In contrast to object point clouds [25, 12, 35, 6], this novel representation preserves both the hand-object surface correlation and the contact point orderliness.

Numerous works [12, 35] only predicted plausible contact patterns according to data prior and ignore contact clues in the image. This may be applicable to rigid interaction. However, when the deformed degree is considered, multiple nonrigid contact patterns can be yielded from the same hand-object spatial relationship. Therefore, our framework crops the image patches of the corresponding 16 hand bones as extra visual cues to estimate nonrigid contact. Altogether, our RUFormer is tamed to take those 16 groups of hand-RUPs, object-RUPs, and visual cues as the inputs. It gradually estimates the contact and deformed features across RUPs, and finally predicts the deformed transformations of the object. To our best knowledge, this is the first framework that is applicable to reconstruct both rigid and nonrigid hand-object interaction from monocular images.

In summary, our main contributions are:

• A learning-based framework with the ambition to estimate the contact between hand and nonrigid objects from monocular images;

• A hand-object interaction representation to record hand-object surfaces into multiple pixel-aligned and fine-grained 2D profiles;

• A unwrapping-informed transformer to predict contact and deformation on the object according to both visual cues and data prior.

2. Related Work

Hand-object interaction reconstruction. Thanks to the creation of several hand-object interaction datasets [13, 5, 47, 2, 14, 26, 32] in recent years, monocular 3D hand-object interaction reconstruction has received extensive attention from researchers. Hasson *et al*. [19] proposed a two-branch network to reconstruct the hand and an unknown manipulated object. Subsequent works[7, 42] estimated hand-object pose and inferred implicit 3D shape of the object. Other works [34, 9, 17, 18, 3, 41, 47] assumed that the object template is known and reduce the object reconstruction to 6D pose estimation. They jointly regressed hand and object poses by reasoning about their interactions. However, all the existing work focuses on interactions between hands and rigid objects. Our framework attempts for the first time to reconstruct the interaction between hands and non-rigid objects from monocular images, while also being compatible with rigid ones.

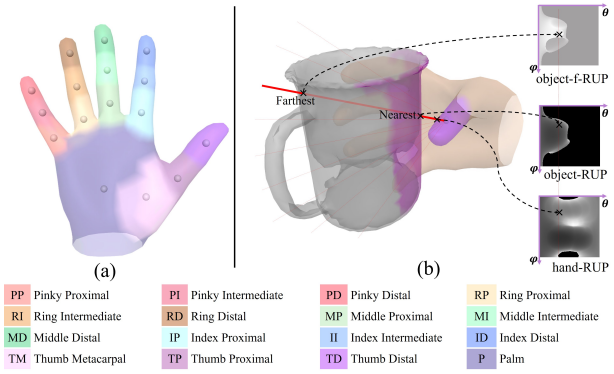

Figure 2: **Surface grouping strategy.** (a) Hand region division based on 16 kinematic bones of an LBS hand. Each region center is marked as a gray sphere. (b) The correlated hand-object sub-surfaces for each region are aligned according to rays emanating from the center and unwrapped to the 2D regional profiles (*i.e*. hand-RUP and object-RUP). An extra object-f-RUP is generated only for grid-wise sampling.

Hand-object contact pattern estimation. Inferring contact patterns is vital for 3D hand-object reconstruction. [19, 3] introduced contact losses which encourage contact surfaces and penalize penetrations between hand and object. However, these methods cannot enforce hand-object alignment at test time. Recently, Some works [12, 35, 43] used explicit contact inference and enforcement to achieve higher quality grasps. Grady *et al*. [12] estimated the contact pattern between hand and object based on PointNet [29]. Tse *et al*. [35] proposed a graph-based network to infer contact patterns. [12, 35] estimated hand-object contact patterns from sparse point clouds, which are unordered and challenging to store contact into a feature-aligned space. Yu *et al*. [43] proposed a dense representation in the form of a UV coordinate map, which only inferred the contact areas of the hand surface. All the existing works focus on contact with rigid objects and are not applicable to contact with non-rigid objects.

Vision transformer. Transformer and self-attention networks have revolutionized natural language processing [38, 8, 39] and are making a deep impression on visual-related tasks, such as object detection [4, 48], image classification [10], 3D pose estimation [24, 22, 27, 23, 15] and point cloud processing [33]. We refer the reader to [16] for a details survey of Vision Transformer. In our task, We use attention modules to exploit the visual and hand-object spatial correlations.

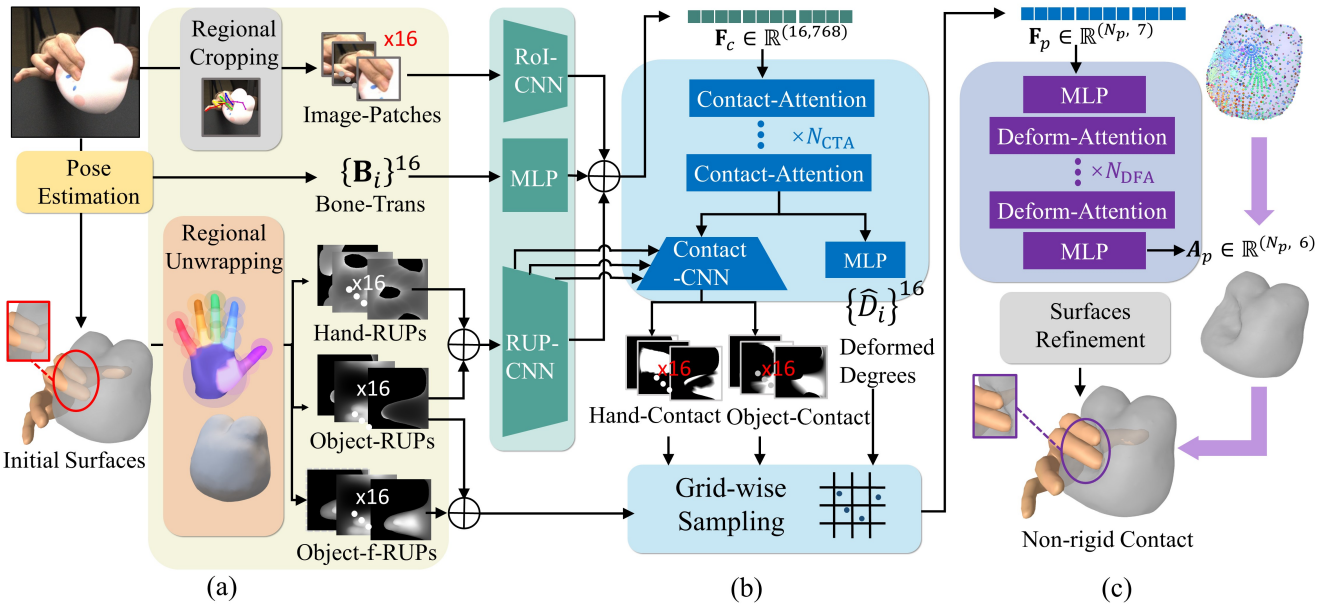

Figure 3: **Overview of RUFormer .** (a) The Preparation process of RUFormer input data, all of which are aligned to the hand 16 regions. (b) RUFormer Encoder estimates hand-object regional contact areas from image patches, hand bone transformation and RUPs. (c) RUFormer Decoder estimates fine-grained deformation from grid-wise sampling features.

3. Method

An overview of our pipeline for hand and nonrigid objects contact patterns estimation is shown in Fig. 3. It takes the image and the corresponding object mesh template as input. Through our RUFormer , it predicts the contact areas of the hand-object surface pair, as well as the deformation of the object. Initially, the hand-object surfaces are estimated and unwrapped into multiple high-resolution 2D regional profiles (Sec. 3.2). Then, RUFormer estimates contact according to region-aligned features (Sec. 3.3), and predicts deformed transformations of sampling points (Sec. 3.4). Hand-object surfaces refinement and deployment details are described in Sec. 3.5.

3.1. Preliminary

Surface representation. We represent the hand surface based on MANO [30]. It can be regarded as a differentiable function $M_h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\tau})$ parameterized by shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, pose $\boldsymbol{\theta} \in \mathbb{R}^{16\times3}$ and global translation $\boldsymbol{\tau} \in \mathbb{R}^3$ w.r.t. the camera coordinate system. For a left-hand case, the RGB images and hand-object surfaces are mirrored together in advance. We represent object 6D pose as its mesh template w.r.t. the right MANO hand coordinate system. It is noted that the plausible hand-object relationship is always the object in front of the hand, *i.e.* $y < 0$ in the vertex coordinates of the object. The deformation of a sampling point $\boldsymbol{p}$ on the nonrigid object is represented as an affine transformation $\mathbf{A}(\boldsymbol{p}) \in SE(3)$ w.r.t. to its template position.

Grouping strategy. Our hand-object surfaces grouping is shown in Fig. 2. We first divide the hand region based on the 16 kinematic bones of the posed MANO. Each piece of the object surface is divided into a certain group when it can be directly intersected by a ray emanating from the hand region center associated with this group. We further unwrap each subsurface to the image plane to obtain 16 hand-RUPs, 16 object-RUPs and 16 object-f-RUPs. The records are pixel-aligned in the same group. It should be noted that object-f-RUPs are only used for grid-wise sampling.

3.2. RUFormer Input: Region Alignment

Surface initialization. The hand-object pose estimation network is refer to [47]. We take RGB image and object template as inputs and predict the MANO parameters and object 6DoF pose. We integrate MANO as a differentiable network layer and use it to output the 3D hand surface.

Surface unwrapping. We unwrap the estimated hand-object surfaces into multiple fine-grained RUPs. RUPs define the projection to unwrap the hand-object surfaces into the image plane with the center of 16 hand sub-surfaces as the origin, respectively. We refer to [46] to map a surface point $\boldsymbol{p}(x, y, z)$ in the Cartesian coordinate system to the spherical coordinate $\boldsymbol{s}(\rho, \theta, \varphi)$. Specifically, the closest intersections between the hand surface and the rays emitted from i-th bone center are recorded in the i-th hand-RUPs, the closest intersections between the object surface and those same rays are recorded in the i-th object-RUPs, and the farthest intersections between the object surface

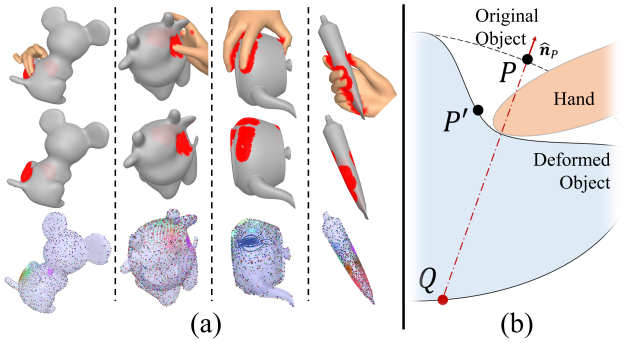

Figure 4: **(a) Grid-wise sampling** results after back-projecting on their object surfaces. Each column corresponds to an instance. **(b) The illustration of object deformation process.** The *deformation vector* of a surface point P is defined as its position difference before and after deformation (PP'). The *maximum deformation* of P is defined as the distance d_{PQ} between the point in its original state and the closest intersection point of the object projected from that point in the opposite normal direction. The *deformation degree* of P is $\boldsymbol{v}_P \triangleq d_{PP'}/d_{PQ}$.

and those same rays are recorded in the i-th object-f-RUPs. All rays emitted from a center can be parameterized as $\overrightarrow{O_i R}(\vartheta, \varphi)$, where $\vartheta \in [0, \pi], \varphi \in [0, 2\pi], \rho > 0$ are the spherical coordinates. Therefore, each RUP channel can be formulated as:

$$\mathbf{R}\left(\frac{\vartheta}{\pi}W_R, \frac{\varphi}{2\pi}H_R\right) \triangleq \arg\min_{\rho}\{\boldsymbol{s}(\rho) \mid \boldsymbol{s} \in (\overrightarrow{O_i R} \cap \partial \boldsymbol{S})\} \tag{1}$$

where $\mathbf{R}$ denotes as RUP and $\partial \boldsymbol{S}$ represents hand or object surface. The value ρ is set to zero if no interaction occurs. As a result, hand-RUPs $\{\mathbf{R}_i^H\}_{i=1}^{16}$, object-RUPs $\{\mathbf{R}_i^O\}_{i=1}^{16}$ and object-f-RUPs $\{\mathbf{R}_i^{OF}\}_{i=1}^{16}$ are all 16-channel image tensors. Furthermore, pixels with the same indices correspond to the intersections between the same ray and the hand bone surface / near object surface / far object surface.

Regional features. Our RUFormer utilizes regional aligned features to predict contact areas, deformed degree of contact areas and deformed transformations of the object. With the above estimation, the following features are aligned according to the region: (i) The image patches $\{\mathbf{I}_i\}_{i=1}^{16}$ belonging to each bone are cropped using the guidance of MANO joint image coordinates, where $\mathbf{I}_i \in \mathbb{R}^{(3, n_p, n_p)}$. (ii) The MANO pose is further converted as bone transformations $\{\mathbf{B}_i\}_{i=1}^{16}$ to measure the relative relationship across RUP groups, where $\mathbf{B}_i \in SE(3)$. (iii) $\{\mathbf{R}_i^H\}_{i=1}^{16}$, $\{\mathbf{R}_i^O\}_{i=1}^{16}$ and $\{\mathbf{R}_i^{OF}\}_{i=1}^{16}$ computed from the estimated hand-object 3D surface.

3.3. RUFormer Encoder: Contact Estimation

Contact attentions. Our contact area estimation process is shown in Fig. 3(b). Image patches, bone transformations, hand-RUPs and object-RUPs are embedded to the latent space through their respective feature extractors. We extract regional image features from 16 groups of image patches by the first four blocks of ResNet18 [21]. The regional unwrapping features are extracted from 16 groups of object-RUPs and hand-RUPs through the first four blocks of ResNet18. The bone transformations are sent to the MLP to encode features of the relative relationship. These features are later concatenated together as a regional feature embedding $\mathbf{F}_c \in \mathbb{R}^{16\times 768}$. After that, N_{CTA} cascaded ViT-based [10] attention modules are used to exploit the visual and hand-object spatial correlations within/across these 16 groups. It computes contact embeddings $\mathbf{F}_{c+}$ with the same size as $\mathbf{F}_c$.

Contact representation. We represent hand contact as 2D maps $\{\mathbf{C}_i^H\}_{i=1}^{16}$ on hand-RUPs and object contact as 2D maps $\{\mathbf{C}_i^O\}_{i=1}^{16}$ on object-RUPs. Each pixel on the contact map indicates the contact probability of the point recorded on RUP with the same indices. To estimate these image-like tensors, an extra CNN decoder with a symmetrical structure with RUP encoder is further adopted. The contact embeddings $\mathbf{F}_{c+}$ are up-sample back to the RUP space again.

Deformed degree. Besides hand-object contact maps, we further estimate regional deformed degree $\{D_i\}_{i=1}^{16}, D_i \in [0, 1]$ from $\mathbf{F}_{c+}$ by an MLP. Each value in $\{D_i\}_{i=1}^{16}$ represents the deformed degree of the contact area in the 16 object-RUPs.

Loss terms. During the training, we supervised the hand-object contact maps and the deformed degree of contact areas. The RUFormer encoder loss L_C can be expressed as follows:

$$L_C = L_M + \lambda_1 L_D \tag{2}$$

where L_D is the standard binary cross-entropy loss for deformed degrees, and $L_D = 1000$. The ground truth of the deformed degree for each RUP is the average deformed degree of all contact points within the region. The process of obtaining the deformation degree of the contact points is shown in Fig. 4(b). L_M is the MSE loss between the prediction and the ground truth of hand-object contact maps, which can be defined as:

$$L_M = \sum_{i=1}^{16}(\|\mathbf{C}_i^H - \hat{\mathbf{C}}_i^H\|_2^2 + \|\mathbf{C}_i^O - \hat{\mathbf{C}}_i^O\|_2^2) \tag{3}$$

where $\mathbf{C}_i^H$ and $\mathbf{C}_i^O$ are ground truth. $\hat{\mathbf{C}}_i^H$ and $\hat{\mathbf{C}}_i^O$ are predicted contact maps.

3.4. RUFormer Decoder: Deformation Estimation

Coarse deformation acquiring. The focus of the previous sections is on the contact area. However, fine-grained de-

formation should be described from point perspective. For each pixel on the object RUPs, it corresponds to a point on the object surface. With the help of deformed degrees of contact areas, the coarse deformation of contact points can be obtained. As illustrated in Fig. 4(b), the ray is emitted from the point to the object surface for intersection detection, and its maximum deformation is defined as the distance between the point from the closest intersection point of the object. The ray direction is set to the negative normal direction of the point. The coarse deformation of the contact point is the maximum deformation multiplied by the predicted deformed degree. However, the deformation obtained in this way does not consider the deformation priori of local geometries and lacks the understanding of the global deformation behavior.

Points sampled from RUPs. Therefore, we sample points from the object and utilize the RUFormer decoder to aggregate deformation features from these sampling points, ultimately predicting the deformation transformations of the object. Existing practices utilize the farthest point sampling to acquire point candidates, or iteratively optimize them through geodetic distance. By contrast, because the surface points of the object have been divided into 32 groups ($\{\mathbf{R}_i^O\}_{i=1}^{16}$ and $\{\mathbf{R}_i^{OF}\}_{i=1}^{16}$) based on their distance from each hand bone, we select sampling points from RUPs in an orderly manner. Specifically, we divide RUP into $n_g \times n_g$ grids and sample one point within a grid with maximum value. The coordinates of sampled points are converted back to Cartesian coordinates. For a grid with all-zero pixels, we use the one mask embedding $\boldsymbol{p}_{[M]} \in \mathbb{R}^3$ [20] as a replacement:

$$\boldsymbol{p} = \begin{cases} \Pi^{-1}(\rho, \theta, \varphi) & \rho \neq 0 \\ \boldsymbol{p}_{[M]}, & \rho = 0 \end{cases} \tag{4}$$

where $\boldsymbol{p}$ and ρ are the 3D point and pixel value corresponding to pixel (θ, φ) in a RUP. We obtain ordered point candidates. For a contact point, its deformation feature is set as its coarse deformation. For a non-contact point, its deformation feature is set as the learnable embedding $\boldsymbol{d}_{[M]} \in \mathbb{R}^3$. As shown in Fig. 4(a), the points sampled according to RUP grids emphasize more on contact area compared with other general sampling strategies. With the above conversion, $\frac{H_R}{n_g} \times \frac{W_R}{n_g} \times 32$ points are sampled.

Deformation attentions. We inherit the idea of the deformation graph [31] that represents the deformation of arbitrary points on the surface as a combined deformation of nearby nodes:

$$\tilde{\boldsymbol{p}} = \sum_{m=1}^{k} \omega_m \left[\mathbf{A}_m \left(\boldsymbol{p} - \boldsymbol{g}_m\right) + \boldsymbol{g}_m\right] \tag{5}$$

where $\boldsymbol{p}$ is original position of the point and $\tilde{\boldsymbol{p}}$ is its deformed position. ω_m is the weight of node $\boldsymbol{g}_m$ to $\boldsymbol{p}$. The weight calculation is referred to [31]. Therefore, the RUFormer decoder is designed to select N_q nodes from N_p points, and predict their affine transformations $\{\boldsymbol{A}_k\}_{k=1}^{N_p}$ according to input features $\mathbf{F}_p$. In practice, we use the farthest point sampling to select N_q nodes from N_p points, where $N_p = \frac{H_R}{n_g} \times \frac{W_R}{n_g} \times 32$. The input features $\mathbf{F}_p = \{\boldsymbol{p}_j \oplus \boldsymbol{d}_j \oplus c_j\}_{j=1}^{N_p} \in \mathbb{R}^{N_p \times 7}$, where c_j indicates whether the point is selected as the node. If the point is selected as the node, it is 1, otherwise, it is 0. Our deformation transformations of the object estimation process are shown in Fig. 3(c). We first utilize an MLP to encode $\mathbf{F}_p$ to the latent space and extract deformation embeddings $\mathbf{F}_d \in \mathbb{R}^{N_p \times 256}$. After that, N_{DFA} cascaded attention modules are used to enhance the understanding of global deformation behavior and aggregate the deformation features. It computes embeddings $\mathbf{F}_{d+}$ with the same size as $\mathbf{F}_d$. Finally, the deformation transformations are obtained through an MLP. We train RUFormer decoder in a semi-supervised manner (transformations of points not selected as nodes are not supervised). To reduce dimensionality, each rotation is represented as an axis-angle.

3.5. Implementation Details

Surface refinement. Based on hand-object contact maps and deformation transformations, we refine the hand and object surfaces. We first perform object surface deformation. The vertices in the object are deformed by Eqn. 5. Afterward, hand and object pose are refined based on hand contact maps $\{\hat{\mathbf{C}}_i^H\}_{i=1}^{16}$ and object contact maps $\{\hat{\mathbf{C}}_i^O\}_{i=1}^{16}$. We convert back to points in the surface by querying pixels in hand-RUPs and object-RUPs, then obtain the contact information of the hand and object surface vertices through interpolation, respectively. Then we follow the method in [12] and optimize the hand-object poses to achieve the target contact.

Parameter settings. The hand-object RUPs size is set to $H_R = W_R = 64$ and the image patch size is set to $n_p = 64$. The grid size for point sampling is set to $n_g = 4$. The depth of contact attentions modules and deformation attentions modules are set to $N_{\mathrm{CTA}} = 6, N_{\mathrm{DFA}} = 5$, respectively. We use Pytorch to implement our networks and train them on a computer configured with NVIDIA GeForce RTX 3090. RUFormer encoder and RUFormer decoder are trained separately. To train RUFormer encoder, we use SGD optimizer with a learning rate 1e-4. RUFormer decoder is trained with Adam optimizer with a learning rate 1e-4. The total training epochs for RUFormer encoder and RUFormer decoder are both 100.

4. Experiments

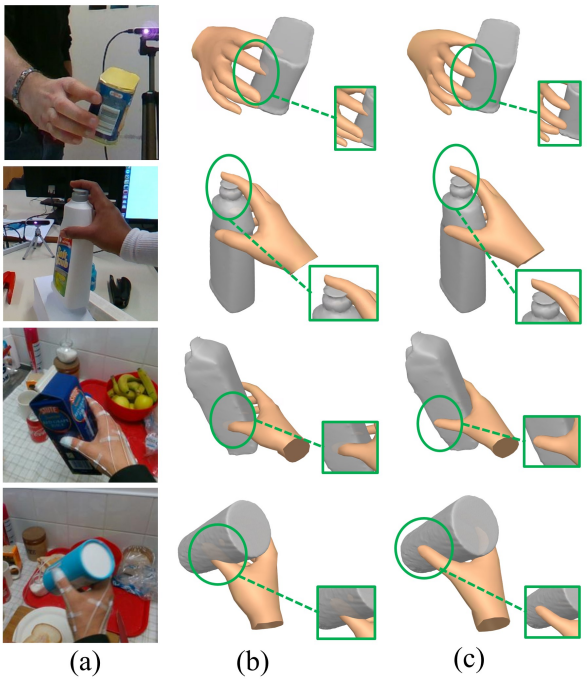

Figure 5: **Comparisons on monocular reconstruction.** (a) RGB images. (b) Reconstruction results from [17]. (c) Ours.

4.1. Datasets

Our experiments are performed on the HMDO [40], HO3D [13], FPHB [11] and ContactPose [2] datasets. The HMDO dataset records the interaction between hands and nonrigid objects. We split it with 4:1 for training and testing. HO3D and FPHB is the dataset of hands in manipulation with rigid objects. We follow the official dataset split for HO3D and adopt the action split following the protocol given by [17] for FPHB. ContactPose is the dataset of hand-object contact paired with hand-object pose. HMDO, HO-3D, and FPHB datasets are used to test our entire pipeline. ContactPose dataset is used to evaluate our hand and object contact estimation. To reduce the ambiguity in the selection of interacted objects, we filter these datasets with the 3D distance between the hand-object not exceeding 2mm as the threshold.

4.2. Metrics

Hand-object error. We use the mean per-point position error (*MPJPE*) of 21 hand joints to evaluate the 3D reconstruction error. The mean per-vertex position error (*MPVPE*) is adopted to evaluate the object error.

Contact quality. We adopt *max penetration* (denoted as Max Pene. in tables) and *intersection volume* (denoted as Inter. in tables) proposed in [19] to evaluate the hand-object geometric relationship.

Methods	Initial State	[12]	[35]	Ours
$\text{MPJPE}_H(\text{mm})\downarrow$	14.67	14.92	14.81	**14.78**
Max Pene.$(\text{mm})\downarrow$	11.82	9.46	9.75	**9.24**
Inter.$(\text{cm}^3)\downarrow$	10.69	7.51	7.58	**7.39**

Table 1: **Evaluations for rigid interactions** under HO3D [13] dataset.

Methods	Initial State	[12]	[35]	Ours
$\text{MPJPE}_H(\text{mm})\downarrow$	18.54	19.84	19.76	**18.97**
$\text{MPVPE}_O(\text{mm})\downarrow$	21.42	21.45	21.40	**21.04**
Max Pene.$(\text{mm})\downarrow$	19.46	13.19	12.78	**10.42**
Inter.$(\text{cm}^3)\downarrow$	14.25	8.74	9.02	**8.56**

Table 2: **Evaluations for nonrigid interactions** under HMDO [40] dataset.

4.3. Comparisons

Monocular hand-object reconstruction. In the task of reconstructing the hand-object from the monocular image, our method is compared with the hand-object reconstruction network from Hasson *et al*. [17]. The quantitative results on HO3D [13] and FPHB [11] datasets are shown in Tab. 3. Our method achieves better performance in hand-object interaction datasets. This demonstrates that our method can achieve explicit contact patterns inference and effective hand-object contact optimization, which can help us reconstruct hand-object interaction with higher quality. We show our qualitative results in Fig. 5. Our methods can achieve more plausible reconstructions with fewer penetrations than [17]. More qualitative results of our method are shown in Fig. 7.

Hand-object contact estimation. We take the result from our pose estimation network as the initial state and compare our RUFormer with ContactOpt [12] and S^2Contact [35]. We retrain DeepContact network in ContactOpt [12] and GCN-Contact network in S^2Contact [35] on the HMDO [40] dataset. As shown in Fig. 1 and Fig. 6, we show the qualitative results compared with [12, 35] under HMDO [40], ContactPose [2] and HO3D [13] datasets. We evaluate the contact patterns optimization results of 3D rigid hand-object interaction between [12, 35] and ours in Tab. 1. Our method achieved better performance than other methods. The contact optimization results of nonrigid hand-object interaction are shown in Tab. 2, and our method achieves higher quality grasping of hand and object. These demonstrate that our method can achieve better contact patterns optimization in both rigid and nonrigid interactions compared to other methods. Since RUFormer can estimate the deformed degree of the contact areas and the deformed transformations of the object, it allows our method to suitable for both contact optimization with nonrigid and rigid

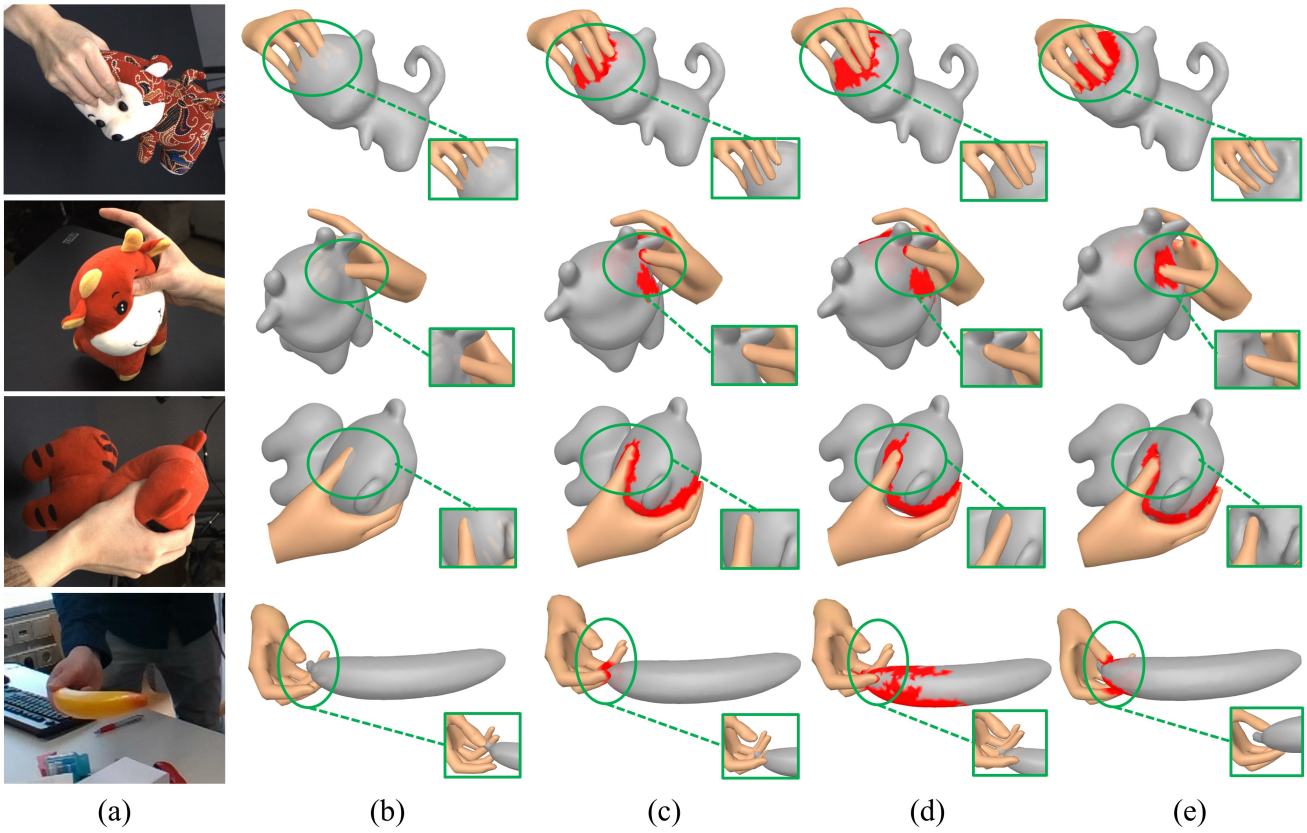

Figure 6: **Comparisons on contact patterns optimization.** (a) RGB images. (b) Initial hand-object surfaces. (c) Results from ContactOpt [12]. (d) Results from S^2Contact [35]. (e) Ours. Row1, Row2 and Row 3 are 3D nonrigid interactions. Row4 is 3D rigid interaction.

objects. From Tab. 1 and Tab. 2, it can be seen that although our method and [12, 35] did not improve hand pose estimation, they significantly reduced the intersection volume and max penetration, and improved the hand-object contact quality. This may be due to the optimization affected the hand region that did not interact with the object, as shown in the row 2 of Fig. 6.

4.4. Ablation Study

Baseline. We take our hand-object pose estimation network as our baseline. Since monocular estimation is ill-posed, there may be mutual penetration or no contact between the reconstructed hand and the object. In addition, the contact areas of the object may be deformed due to its non-rigidity and the force exerted by the subject. Since HMDO [40] is the dataset that records the 3D nonrigid interactions, most of our ablation experiments are based on this dataset. Where the results of completely using our entire pipeline are in the last row of Tab. 4.

Surface unwrapping. We explore the effects of the size of hand-object RUPs. As shown in row 3 to row 4 of Tab. 4, we compared the impact of different sizes of RUP on hand-object interaction reconstruction. Considering both efficiency and reconstruction quality, it is appropriate to set the size of RUPs to 64×64. RUP of $n \times n$ size is denoted as "RUP-n" in Tab. 4.

Datasets	FPHB [11]		HO3D [13]	
Methods	[17]	Ours	[17]	Ours
$\text{MPJPE}_H(\text{mm}) \downarrow$	18.23	**17.86**	14.74	**14.78**
$\text{MPVPE}_O(\text{mm}) \downarrow$	21.45	**21.22**	19.42	**19.27**
Max Pene.$(\text{mm}) \downarrow$	18.64	**13.35**	11.43	**9.24**
Inter.$(\text{cm}^3) \downarrow$	13.57	**8.28**	10.26	**7.39**

Table 3: **Comparisons for monocular reconstruction** under FPHB [11] and HO3D [13] datasets.

Contact estimation. The impact of contact attention modules is ablated as shown in row 5 of Tab. 4. We replace contact attention modules with an MLP architecture, resulting in a decrease in the quality of hand object reconstruction.

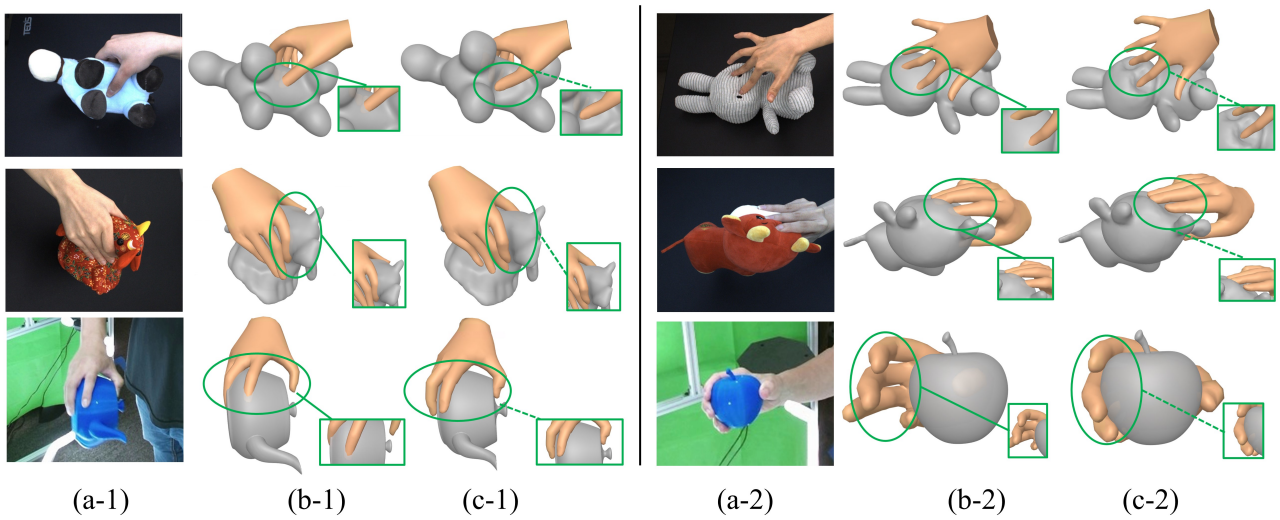

Figure 7: **More qualitative results.** (a) RGB images. (b) Initial hand-object surfaces. (c) Ours. High-quality reconstruction results certify the effectiveness of our framework.

Method	MPJPE$_H$(mm) ↓	MPVPE$_O$(mm) ↓	Max Pene.(mm) ↓	Inter.(cm^3) ↓
Baseline	18.54	21.42	19.46	14.25
w/ RUP-32	19.12	21.16	11.04	8.82
w/ RUP-128	18.94	21.08	10.25	8.64
w/o Con-Att	19.73	21.59	11.62	9.27
w/o Img-Pat	19.96	21.65	12.57	9.19
w/ Point-Tran	19.04	21.09	10.75	8.49
w/ grid-8	19.25	21.24	11.16	8.74
Ours	**18.97**	**21.04**	**10.42**	**8.56**

Table 4: **Ablation study of our method.** Our RUFormer , surface unwrapping and point sampling are evaluated.

The main reason may be that the contact attention modules can better explore the visual and hand-object spatial correlation, which can help better to predict the mutual contact areas and the object deformation. We ablate the effect of introducing image patches on the results. As shown in row 6 of Tab. 4, visual cues can improve the quality of hand-object reconstruction. Because the image patches contain contact and deformation information, which can guide our RUFormer to better predict contact areas and deformed degrees on region-aligned features. We denote the contact attention modules as "Con-Att" and image patches as "Img-Pat" in Tab. 4.

Deformation estimation. As shown in row 7 of Tab. 4, we ablate the deformation estimation block. We replace our deformation estimation block with Point-Transformer [45]. Compared with [45], our deformation estimation block can give consideration to both efficiency and accuracy. We do not need to constantly query and build the neighborhood. Our deformation estimation block can benefit from the ordered sampling points and achieve effective aggregation of deformation features. In addition, we ablate the grid size for ordered point sampling from object-RUPs and object-f-RUPs, as shown in row 8 of Tab. 4. Since the deformed transformations aggregated from fewer grid-wise sampling features can not well represent the object surface deformation and more points calculations are expensive, setting grid size to 4×4 is more appropriate. We denote Point-Transformer as "Point-Tran" and $n \times n$ grid size as "grid-n" in Tab. 4.

5. Conclusion

This paper proposes a learning-based framework to estimate the contact patterns between hand and nonrigid objects from monocular images. A hand-object interaction representation is proposed to record the hand-object surfaces into multiple fine-grained 2D regional unwrapping profiles. Based on this representation, the roughly estimated hand-object surfaces are first unwrapped into 2D regional profiles, then a Vision Transformer is tamed to predict contact areas and deformed transformations within/across regions according to region-aligned features. Finally, hand-object surfaces are refined based on contact areas and deformed transformations.

Limitations and Future Work. Due to the influence of 2D hand joints on image patch cropping, our method relies on reliable 2D pose estimation. Our method can be extended to RGBD input and multi-view RGB input. By introducing depth and multi-view information, we can improve the quality of contact patterns and hand-object reconstruction.

References

[1] Miguel Aranda, Juan Antonio Corrales Ramon, Youcef Mezouar, Adrien Bartoli, and Erol Özgür. Monocular visual shape tracking and servoing for isometrically deforming objects. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7542–7549. IEEE, 2020.

[2] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020.

[3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[4] 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.

[5] 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.

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

[7] 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.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

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

[13] 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.

[14] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Handsformer: Keypoint transformer for monocular 3d pose estimation ofhands and object in interaction. *arXiv preprint arXiv:2104.14639*, 2021.

[15] 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.

[16] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2(4), 2020.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Yihui He, Rui Yan, Katerina Fragkiadaki, and Shoou-I Yu. Epipolar transformers. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 7779–7788, 2020.

[23] Buzhen Huang, Yuan Shu, Jingyi Ju, and Yangang Wang. Occluded human body capture with self-supervised spatial-

temporal motion prior. *arXiv preprint arXiv:2207.05375*, 2022.

[24] Lin Huang, Jianchao Tan, Ji Liu, and Junsong Yuan. Hand-transformer: non-autoregressive structured modeling for 3d hand pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 17–33. Springer, 2020.

[25] 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.

[26] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021.

[27] 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.

[28] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021.

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

[30] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17, 2017.

[31] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007.

[32] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[33] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. *arXiv preprint arXiv:2210.05616*, 2022.

[34] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019.

[35] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 568–584. Springer, 2022.

[36] Aggeliki Tsoli, Antonis Argyros, et al. Patch-based reconstruction of a textureless deformable 3d surface from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[37] Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018.

[38] 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.

[39] Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*, 2019.

[40] Wei Xie, Zhipeng Yu, Zimeng Zhao, Binghui Zuo, and Yangang Wang. Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models*, 127:101178, 2023.

[41] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021.

[42] 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.

[43] Ziwei Yu, Linlin Yang, You Xie, Ping Chen, and Angela Yao. Uv-based 3d hand-object reconstruction with grasp optimization. *arXiv preprint arXiv:2211.13429*, 2022.

[44] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5628–5635. IEEE, 2018.

[45] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021.

[46] Zimeng Zhao, Ruting Rao, and Yangang Wang. Supple: Extracting hand skeleton with spherical unwrapping profiles. In *2021 International Conference on 3D Vision (3DV)*, pages 899–909. IEEE, 2021.

[47] Zimeng Zhao, Binghui Zuo, Wei Xie, and Yangang Wang. Stability-driven contact reconstruction from monocular color images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2022.

[48] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.