Joint Reconstruction of 3D Human and Object via Contact-Based Refinement Transformer

Hyeongjin Nam[1,3*] Daniel Sungho Jung[2,3*] Gyeongsik Moon[4] Kyoung Mu Lee[1,2,3]
[1]Dept. of ECE&ASRI, [2]IPAI, Seoul National University, Korea
[3]SNU-LG AI Research Center, [4]Codec Avatars Lab, Meta
{namhjsnu28, dqj5182}@snu.ac.kr, mks0601@meta.com, kyoungmu@snu.ac.kr

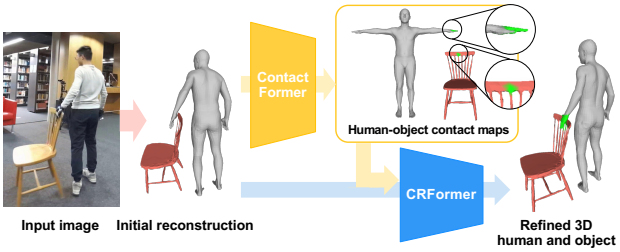
Abstract

Human-object contact serves as a strong cue to understand how humans physically interact with objects. Nevertheless, it is not widely explored to utilize human-object contact information for the joint reconstruction of 3D human and object from a single image. In this work, we present a novel joint 3D human-object reconstruction method (CONTHO) that effectively exploits contact information between humans and objects. There are two core designs in our system: 1) 3D-guided contact estimation and 2) contact-based 3D human and object refinement. First, for accurate human-object contact estimation, CONTHO initially reconstructs 3D humans and objects and utilizes them as explicit 3D guidance for contact estimation. Second, to refine the initial reconstructions of 3D human and object, we propose a novel contact-based refinement Transformer that effectively aggregates human features and object features based on the estimated human-object contact. The proposed contact-based refinement prevents the learning of erroneous correlation between human and object, which enables accurate 3D reconstruction. As a result, our CONTHO achieves state-of-the-art performance in both human-object contact estimation and joint reconstruction of 3D human and object. The code is publicly available[1].

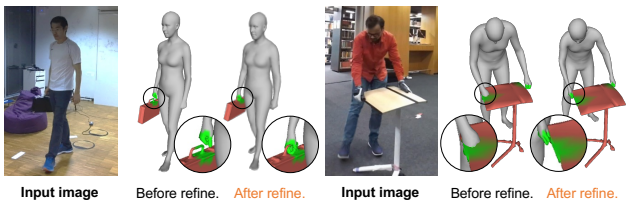
1. Introduction

Joint reconstruction of 3D human and object is an essential task for various applications of immersive experiences of AR/VR and robot manipulation of robotics. In essence, the task aims to learn a meaningful human-object interaction that further improves the reconstruction of humans and objects. Physical human-object contact is notably one of the most prevalent and basic interactions that humans make with objects.

Although human-object contact is a strong cue in joint reconstruction of 3D human and object, recent works of human-object interaction have been studied separately in two major tracks: 1) human-object contact estimation and 2) 3D human and object reconstruction. The recent research track for human-object contact estimation [13, 37, 40] predicts a contact map on the surface of a pre-defined human body model [24, 31] without reconstructing 3D human and object. Another research track for 3D human and object reconstruction [45, 46, 48, 52] does not yet sufficiently explore how to extract and utilize contact information for the reconstruction. For example, PHOSA [52] and D3D-HOI [48] heuristically pre-define contacting regions and follow the pre-defined regions as a hard constraint during their optimizations. The pre-defined contacting regions can be different from the authentic ones in the image, which



(a) Simplified pipeline of CONTHO



(b) Refinement results from CRFormer

Figure 1. **Overview of CONTHO.** Our proposed CONTHO estimates human-object contact maps through our proposed **ContactFormer** and exploits the contact maps for 3D human and object refinement with **CRFormer**. The green color indicates human-object contact regions estimated from **ContactFormer**.

[1] https://github.com/dqj5182/CONTHO_RELEASE

results in incorrect reconstructions of 3D human and object.

In this work, we integrate the two separate tracks with one unified framework, **CONTHO** (**CONT**act-based 3D **H**uman and **O**bject reconstruction) that estimates human-object contact maps and exploits the contact maps for 3D human and object reconstruction, as shown in Figure 1. In CONTHO, there are two core stages: 1) 3D-guided contact estimation and 2) contact-based 3D human and object refinement. In the first stage, our proposed contact estimation Transformer (**ContactFormer**) utilizes initially reconstructed 3D human and object meshes as 3D guidance on 3D positional relationships between human and object. In inferring contact, 3D positions of human and object surfaces provide valuable information about which parts of the human interact with the object. However, previous contact estimation methods [13, 37, 40] do not infer 3D geometric information of human and object surfaces during their estimation pipeline. Unlike these methods, our ContactFormer utilizes the 3D positions of human and object surfaces along with image evidence, enabling 3D geometric reasoning about the relationship between 3D human and object. In the end, 3D-guided contact estimation provides accurate human-object contact maps, which benefits the next stage, the contact-based 3D human and object refinement.

In the second stage, our proposed contact-based refinement Transformer (**CRFormer**) refines the initially reconstructed 3D human and object by effectively aggregating human and object features based on the estimated contact maps. In the CRFormer, human and object features are selectively forwarded based on human-object contact maps to learn human-object interaction. Such an approach has two advantages in 3D human and object refinement. First, the CRFormer makes the human-object contact maps the main decisive signals that indicate which features to focus on. While contact information is one of the most influential components for understanding relationships between humans and objects, human-object contact exists in small regions of the image and thus may often be neglected. Our CRFormer design explicitly spotlights the contact regions, making human-object contact a key signal for refinement. Second, the CRFormer alleviates the undesired human-object correlation by removing features unrelated to physical interaction (*i.e.*, contact). A refinement network can learn undesired human-object correlation by easily capturing a strong bias of human pose and object pose, different from the actual appearance in the image. When naively aggregating contact maps and image features (Transformer baseline in Figure 2), the refinement network reconstructs a monitor display always facing toward a human head, showing an undesired correlation between the object and the human. On the other hand, our CRFormer considers human-object relations solely from the features of contacting regions based on human-object contact captured in the image,

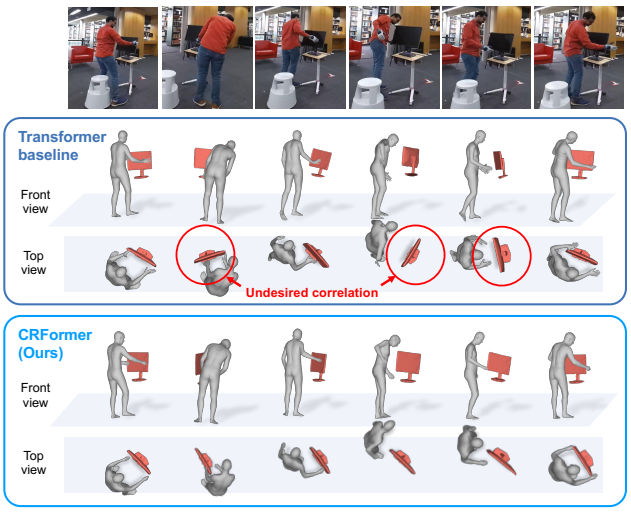


Figure 2. **Example of undesired human-object correlation.** Due to the undesired human-object correlation in the Transformer baseline, the monitor display always faces toward the human head, which should not move as in the images. Our proposed CRFormer effectively alleviates the undesired correlation, resulting in accurate reconstruction results.

preventing undesired human-object correlation. With these two strengths, our proposed CRFormer accurately refines 3D human and object with human-object contact maps.

As a result, we show that CONTHO achieves state-of-the-art performance in both human-object contact estimation and joint reconstruction of 3D human and object. Our contributions can be summarized as follows.

- We propose CONTHO, which jointly reconstructs 3D human and object by exploiting human-object contact as a key signal in reconstruction.
- To obtain precise human-object contact, we leverage intermediate 3D human and object reconstructions as explicit 3D guidance in contact estimation.
- To accurately reconstruct 3D human and object, our proposed CRFormer effectively aggregates human features and object features based on contact information, while preventing learning undesired human-object correlation.
- CONTHO largely outperforms previous methods in both human-object contact estimation and joint reconstruction of 3D human and object.

2. Related works

Human-object contact estimation. Most of the pioneering works on human-object contact estimation represent contact in the form of 2D contact [3], 3D joint-level contact [34–36, 49, 53], or 3D patch-level contact [9, 10, 27]. Recently, several works [11, 13, 37, 40] tackle the problem of estimating a dense vertex-level contact map, defined on the human body surface (*i.e.*, SMPL [24]). POSA [11] proposed

a conditional variational autoencoder (cVAE) [18] that outputs which vertices are likely to be in contact with objects, given a human pose without any use of image evidence. BSTRO [13] demonstrated a Transformer that learns contextual relationships among body vertices. DECO [40] proposed a cross-attention-based network that jointly leverages human body parts and scene contexts for contact estimation.

These methods are simply trained with cross-entropy loss between the predicted and ground-truth (GT) contact maps, without learning 3D geometry of human and object surfaces. On the other hand, our CONTHO jointly learns human-object contact maps along with reconstructing the 3D human and object meshes, which has two noticeable advantages in contact estimation. First, 3D human and object meshes provide guidance on where to focus on local image regions related to the human and object, with 2D vertex coordinates obtained by projecting the 3D meshes onto the input image. Second, the per-vertex 3D coordinates provide a 3D positional relationship between human and object surfaces that allows 3D geometric reasoning on contact between human and object. Under these two advantages, our 3D-guided contact estimation is much more effective in capturing human-object contact than the previous methods.

3D human and object reconstruction. Most of the recent works [2, 15, 45, 48, 52] of 3D human and object reconstruction are optimization-based approaches, which iteratively fit 3D human and object meshes to satisfy constraints of human-object interaction. Holistic++ [2] designed human-object interaction priors based on human actions. PHOSA [52] and D3D-HOI [48] each presented an optimization framework that fits human and object meshes based on pre-defined contact pairs to reason about human-object interaction. CHORE [45] proposed a two-stage approach, which first predicts distance fields and then optimizes 3D humans and objects based on the distance fields.

All of the above optimization-based methods only rely on optimization targets (*e.g.*, 2D silhouettes) without considering image context. One of the limitations of such an approach is vulnerability to imperfect optimization targets. Since their optimization targets are acquired by estimation, the targets contain estimation errors, and some optimization targets are ambiguous (*e.g.*, depth ambiguity of 2D silhouettes) for reconstruction. Accordingly, their optimization methods often fail by becoming biased toward the imperfect optimization targets. Different from these methods, our CONTHO is an end-to-end learning approach that is free from the above issues. This is because, in the inference stage, the system produces outputs based on the data-driven knowledge from the training data instead of being optimized towards imperfect targets. Despite such a strength, we found that the learning-based systems can be vulnerable to being biased to specific contexts within an image, which we call an *undesired human-object correlation*. In this work, we unveil the undesired human-object correlation in the joint learning of 3D humans and objects and address it with our CRFormer.

3D human reconstruction. Most of the 3D human reconstruction methods [5–7, 16, 19–23, 26, 28, 29, 51] are based on parametric 3D human model (*i.e.*, SMPL [24]). HMR [16] proposed an end-to-end learning framework that introduced adversarial loss to reconstruct a plausible 3D human mesh. PARE [20] used a part-guided attention network to ensure robustness in occlusions. Hand4Whole [25] proposed a whole-body 3D human mesh estimation framework to reconstruct 3D human body, hand, and face with features from the 3D positional pose-guided pooling. In our method, we bring the 3D body and hand reconstruction pipeline of Hand4Whole for initial 3D human mesh reconstruction.

3D object reconstruction. One of the main approaches [8, 32, 39, 42, 44] of 3D object mesh reconstruction is to predict 6DoF pose (*i.e.*, rotation and translation) of a given object mesh template after classifying the object category. PoseCNN [44] is a pioneering work that proposes a convolutional neural network for object pose estimation. SO-Pose [8] employs self-occlusion information to predict accurate object pose. ZebraPose [38] proposed a coarse-to-fine surface encoding technique for 6DoF object pose estimation. Our CONTHO also estimates the 6DoF object pose as an initial prediction and refines it considering human-object contact.

3. CONTHO

Figure 3 shows the overall pipeline of our CONTHO, which consists of three stages: initial reconstruction, 3D-guided contact estimation, and contact-based refinement.

3.1. Initial reconstruction

Given concatenated inputs $\mathbf{I}_{\text{input}} \in \mathbb{R}^{5\times H\times W}$ of image $\mathbf{I}$, human segmentation $\mathbf{S}_{\text{h}}$, and object segmentation $\mathbf{S}_{\text{o}}$, we obtain the initial 3D human and object meshes ($\mathbf{M}_{\text{h}} \in \mathbb{R}^{431\times 3}$ and $\mathbf{M}_{\text{o}} \in \mathbb{R}^{64\times 3}$), where H and W denote the height and width of the image, respectively. Following previous works [1, 45], human and object segmentations are obtained from DetectronV2 [43] for both training and inference. From the inputs $\mathbf{I}_{\text{input}}$, a backbone network (*i.e.*, ResNet-50 [12]) extracts an image feature $\mathbf{F} \in \mathbb{R}^{2048\times H/32\times W/32}$. To obtain the initial 3D human mesh $\mathbf{M}_{\text{h}}$, we predict human body parameters $\theta_{\text{body}} \in \mathbb{R}^{76}$ and hand parameters $\theta_{\text{hand}} \in \mathbb{R}^{90}$ of the SMPL+H model [24] from the image feature $\mathbf{F}$. Then, the predicted parameters are forwarded to SMPL+H model to obtain a 3D human mesh. To reduce computational burden, the obtained 3D human mesh is downsampled with a sampling algorithm [33]. To obtain initial 3D object mesh $\mathbf{M}_{\text{o}}$, we predict object rotation $\mathbf{R}_{\text{o}}$ and translation $\mathbf{t}_{\text{o}}$ from the image feature $\mathbf{F}$, given a 3D object mesh template as in prior works [45, 46]. The

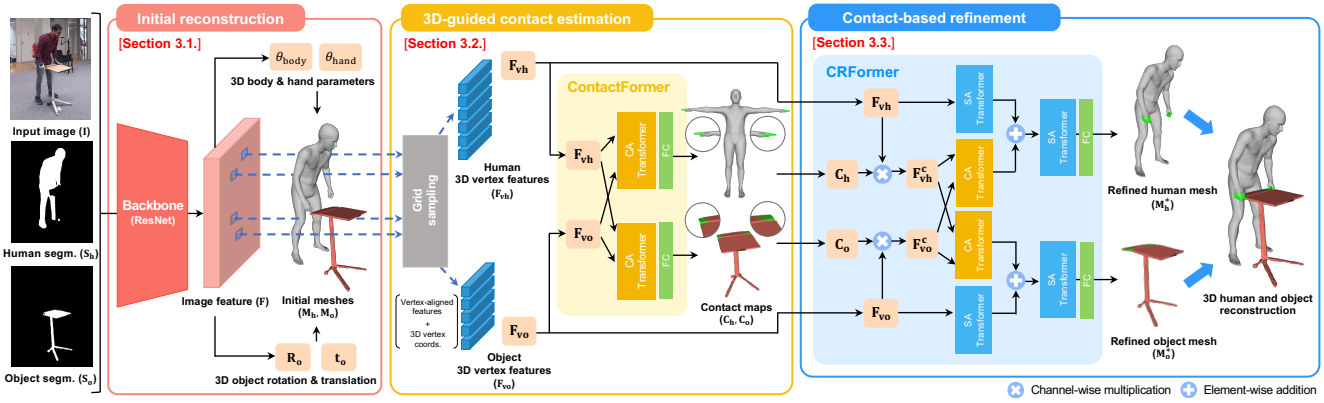


Figure 3. **Overall pipeline of CONTHO.** Our method first reconstructs 3D human and object meshes ($\mathbf{M}_h$ and $\mathbf{M}_o$). Then, the initial meshes are utilized to construct 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$). Subsequently, **ContactFormer** estimates human-object contact maps ($\mathbf{C}_h$ and $\mathbf{C}_o$) from the 3D vertex features. Lastly, **CRFormer** aggregates the 3D vertex features based on the estimated contact maps to provide refined human and object meshes ($\mathbf{M}_h^*$ and $\mathbf{M}_o^*$). The green color indicates the estimated contacting regions.

overall design of the initial reconstruction module follows a state-of-the-art whole-body 3D human mesh reconstruction method [25] with modifications to only predict the human body and hands. We provide a detailed description of the architecture in the supplementary material.

3.2. 3D-guided contact estimation

In this stage, **ContactFormer** predicts human and object contact maps ($\mathbf{C}_h \in \mathbb{R}^{431}$ and $\mathbf{C}_o \in \mathbb{R}^{64}$) from 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) extracted based on initially reconstructed 3D human and object meshes ($\mathbf{M}_h$ and $\mathbf{M}_o$).

3D vertex feature extraction. 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) consist of vertex-aligned features and per-vertex 3D coordinates. The vertex-aligned features are obtained by grid sampling of the image feature $\mathbf{F}$ with (x, y) positions of the projected 3D vertices of initial meshes ($\mathbf{M}_h$ and $\mathbf{M}_o$) to image space. After grid sampling, we apply a 1-by-1 convolution to the vertex-aligned features to reduce the channel dimension from 2048 to 256. Subsequently, we obtain 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) by concatenating the vertex-aligned features and per-vertex 3D coordinates of the initial meshes ($\mathbf{M}_h$ and $\mathbf{M}_o$). Therefore, the final dimensions of the 3D vertex features of the human and object are $\mathbf{F}_{vh} \in \mathbb{R}^{(256+3)\times 431}$ and $\mathbf{F}_{vo} \in \mathbb{R}^{(256+3)\times 64}$. The 3D vertex features contain rich contextual information around the 3D mesh vertices, allowing 3D guidance for human-object contact estimation. The 3D vertex features are passed to ContactFormer, the contact estimation Transformer.

Human-object contact estimation. Given the 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$), ContactFormer predicts human and object contact maps ($\mathbf{C}_h$ and $\mathbf{C}_o$). To encourage the ContactFormer to focus on relevant information across humans and objects, we perform a cross-attention operation between 3D vertex features of humans and objects with cross-attention (CA) Transformers [41]. Then, the contact maps ($\mathbf{C}_h$ and $\mathbf{C}_o$) are predicted with fully-connected (FC) layers, followed by a sigmoid activation function.

3.3. Contact-based refinement

In this stage, **CRFormer** provides refined 3D human and object meshes ($\mathbf{M}_h^*$ and $\mathbf{M}_o^*$) from the 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) and the contact maps ($\mathbf{C}_h$ and $\mathbf{C}_o$).

Contact-based masking. Based on the contact maps ($\mathbf{C}_h$ and $\mathbf{C}_o$), we mask a part of 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) that are not in contact with zero vectors, to remain only features corresponding to human-object contact. We denote the masked 3D vertex features of the human and object with $\mathbf{F}_{vh}^c$ and $\mathbf{F}_{vo}^c$, respectively. This contact-based masking feature aggregation technique is our core strategy to force the contact maps to be the main signal for CRFormer to indicate which features to focus on. Additionally, by removing features from non-contacting parts that are unrelated to human-object contact, we prevent learning undesired human-object correlation that is detrimental to accurate refinement. We provide further discussion about the effectiveness of contact-based masking in Section 5.3.

3D human and object refinement. The masked 3D vertex features ($\mathbf{F}_{vh}^c$ and $\mathbf{F}_{vo}^c$) and original 3D vertex features ($\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$) are processed with a combination of cross-attention (CA) and self-attention (SA) Transformers [41], to obtain refined 3D human and object meshes ($\mathbf{M}_h^*$ and $\mathbf{M}_o^*$). $\mathbf{F}_{vh}^c$ and $\mathbf{F}_{vo}^c$ are passed to CA Transformers to process relevant information across human and object. As $\mathbf{F}_{vh}^c$ and $\mathbf{F}_{vo}^c$ only contain features in contact, the CA Transformers mainly process contextual information related to human-object contact. $\mathbf{F}_{vh}$ and $\mathbf{F}_{vo}$ are passed separately to SA Transformers to infer each own 3D positional information, without considering human-object interaction. The SA Transformers mainly process contextual information related to non-contacting parts of the human and object. This com-

bination of CA and SA transformers prevents excessive bias or disregard of the contact information. Lastly, the outputs of CA and SA Transformers are added and processed with additional SA Transformers followed by FC layers to produce refined 3D human and object meshes ($\mathbf{M}_{\text{h}}^{*}$ and $\mathbf{M}_{\text{o}}^{*}$).

3.4. Loss functions

Our proposed CONTHO is trained in an end-to-end manner by minimizing loss function L, defined as follows:

$$L = L_{\text{contact}} + L_{\text{refine}} + L_{\text{init}}, \tag{1}$$

where L_{contact} is a binary-cross entropy loss between predicted and GT contact maps ($\mathbf{C}_{\text{h}}$ and $\mathbf{C}_{\text{o}}$). The L_{refine} is defined as

$$L_{\text{refine}} = L_{\text{vertex}} + L_{\text{edge}}, \tag{2}$$

where L_{vertex} is a L1 distance between predicted and GT per-vertex 3D coordinates of refined human and object meshes ($\mathbf{M}_{\text{h}}^{*}$ and $\mathbf{M}_{\text{o}}^{*}$), and L_{edge} is edge length consistency loss between predicted and GT edges of the refined human meshes ($\mathbf{M}_{\text{h}}^{*}$). The L_{init} is defined as

$$L_{\text{init}} = L_{\text{param}} + L_{\text{coord}} + L_{\text{hbox}}, \tag{3}$$

where L_{param} is a L1 distance between predicted and GT SMPL+H parameters (θ_{body} and θ_{hand}), 3D object rotation $\mathbf{R}_{\text{o}}$, and 3D object translation $\mathbf{t}_{\text{o}}$. L_{coord} is a L1 distance between the predicted and GT human joint coordinates, consisting of 3D and 2D joint coordinates. L_{hbox} is a L1 distance between the predicted and GT bounding boxes of the hands. We design L_{init} by modifying the loss function of Hand4Whole [25]. For a detailed explanation, please refer to the supplementary material.

4. Implementation details

PyTorch [30] is used for implementation. The backbone is initialized with pre-trained weights of publicly released Hand4Whole [25]. The weights are updated by Adam optimizer [17] with a mini-batch size of 16. The region of the reconstruction target is cropped using a GT box in both the training and the testing stages following previous works [45, 46]. Data augmentations, including scaling, rotation, and color jittering, are performed in training. The initial learning rate is set to 10^{-4} and reduced by a factor of 10 after the 30*th* epoch. We train the model for 50 epochs with an NVIDIA RTX 2080 Ti GPU.

5. Experiments

5.1. Datasets

BEHAVE [1, 47] and InterCap [14] datasets are used for our experiments. BEHAVE [1, 47] is a dataset that captures the interactions of 8 human subjects and 20 objects. We follow CHORE [45] for the split of BEHAVE [1, 47] for a fair comparison. InterCap [14] is another human-object interaction dataset containing 10 human subjects with 10 objects. Following the prior work [46], we split the dataset accordingly. For both datasets, we labeled contact maps on 3D human and object vertices with a 3D distance threshold of $5cm$ between human and object.

5.2. Evaluation metrics

Precision & recall for contact estimation ($\text{Contact}_{\text{p}}^{\text{est}}$, $\text{Contact}_{\text{r}}^{\text{est}}$). We evaluate human-object contact estimation with standard detection metrics: precision ($\text{Contact}_{\text{p}}^{\text{est}}$) and recall ($\text{Contact}_{\text{r}}^{\text{est}}$), following Huang *et al*. [13]. Unlike our CONTHO, which estimates both human and object contact maps, previous contact estimation methods [13, 40] only estimate human contact maps. Thus, for comparison, we report evaluations on human contact maps for all methods.

Chamfer distance (CD_{human}, $\text{CD}_{\text{object}}$). We evaluate 3D human and object reconstruction using Chamfer distance between predicted and GT meshes, following previous works of 3D human and object reconstruction [45, 46]. Specifically, given the predicted 3D human and object meshes, we apply Procrustes alignment on combined 3D human and object meshes with the GT 3D human and object meshes. With the aligned 3D human and object meshes, we measure the Chamfer distance from GT separately on 3D human (CD_{human}) and 3D object ($\text{CD}_{\text{object}}$), in centimeters.

Precision & recall for contact from reconstruction ($\text{Contact}_{\text{p}}^{\text{rec}}$, $\text{Contact}_{\text{r}}^{\text{rec}}$). To evaluate 3D human and object reconstruction, especially in terms of contact, we further adopt standard detection metrics for the reconstructed 3D human and object meshes. We obtain a contact map by classifying human vertices within $5cm$ of the object mesh. Then, we measure precision ($\text{Contact}_{\text{p}}^{\text{rec}}$) and recall ($\text{Contact}_{\text{r}}^{\text{rec}}$) between the human contact map and the GT counterpart.

5.3. Ablation study

We carry out the ablation study by training and evaluating all methods on BEHAVE [1].

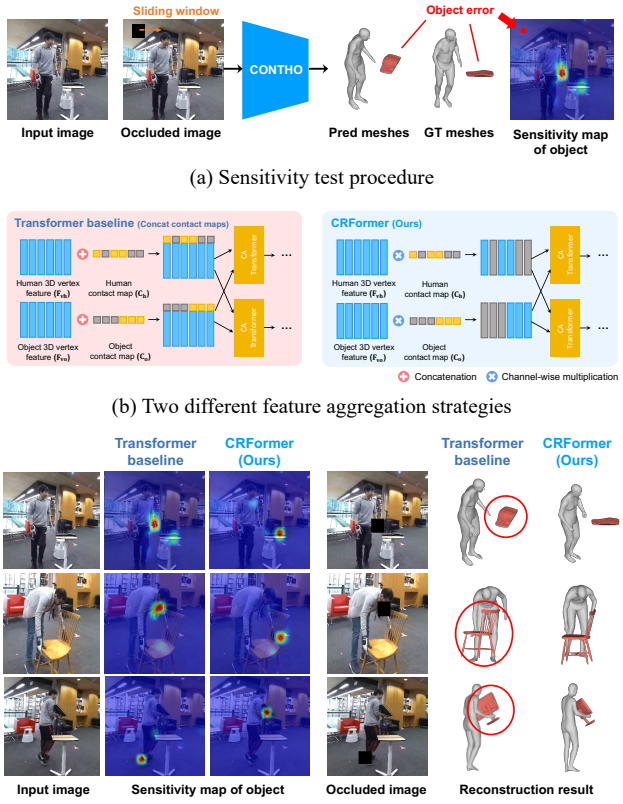
Effectiveness of 3D-guided contact estimation. In Table 1, we show the effectiveness of 3D-guided contact estimation by examining the following variations: 1) variations of ContactFormer inputs and 2) variations of ContactFormer design. The first block of Table 1 shows that using 3D vertex features as the ContactFormer input largely outperforms other input variants. The first variant uses a global average pooled (GAP) image feature. The second and third variants follow existing methods [13, 40] by implementing their feature extractors into our framework. Specifically, the second variant use a convolutional layer to extract per-vertex features from the image feature. The third variant

Methods	$\text{Contact}_{\text{p}}^{\text{est}}\uparrow$	$\text{Contact}_{\text{r}}^{\text{est}}\uparrow$
*** Variations of ContactFormer input**		
GAP feature	0.645	0.481
Per-vertex image feature [13]	0.716	0.539
Part-scene image feature [40]	0.719	0.556
3D vertex feature (Ours)	**0.754**	**0.587**
*** Variations of ContactFormer design**		
FC layers	0.639	0.471
SA Transformers	0.725	0.575
ContactFormer (Ours)	**0.754**	**0.587**

Table 1. **Ablation studies for 3D-guided contact estimation on BEHAVE [1].**

designs two encoders for human part and scene [4] and obtains features by applying cross-attention operation between two encoders' outputs. One major difference of the second and third variants from ours is that they do not extract localized features based on 3D positions of 3D human and object. Our 3D vertex feature contains localized contextual information around human and object regions by grid sampling on the image feature. Additionally, the 3D positional information of 3D vertex features enables 3D geometric reasoning of human-object contact. From these advantages, exploiting 3D vertex features outperforms other variants in contact estimation. The second block of Table 1 shows that our ContactFormer with CA Transformers achieves the best performance. Compared to other designs, the cross-attention operation of the CA Transformers encourages ContactFormer to capture meaningful contextual information within the image.

Effectiveness of contact-based refinement. In Table 2, we justify our proposed contact-based refinement by conducting ablation studies as follows: 1) variations of feature aggregation and 2) variations of CRFormer design. The first block of Table 2 validates the effectiveness of contact-based masking of CRFormer compared to other variants of feature aggregation strategies. Comparing the first and other variants shows the impact of contact maps in refinement. As the contact maps are strong cues for understanding human-object interactions, utilizing the contact maps significantly enhances refinement performance, especially in contact-related metrics. Among all variants, our proposed contact-based masking (the fourth variant) achieves the best performance, as the masking strategy explicitly highlights contact maps as a key signal for refinement, unlike other aggregation strategies. Additionally, the contact-based masking prevents learning undesired human-object correlation by removing unnecessary features unrelated to human-object interaction. Due to such reasons, our proposed contact-based masking outperforms other aggregation strategies by significant margins. The second block of Table 2 validates our CRFormer design as a combination of CA Transformers and SA Transformers. The CA Transformers have strength in encouraging attention to the relevant information across human and object. Differently, the SA Transformer has strength in learning positional relationships between human and object, separately. By combining the advantages of each Transformer, our CRFormer achieves the best performance.



(a) Sensitivity test procedure

(b) Two different feature aggregation strategies

(c) Sensitivity test for human-object correlation

Figure 4. **Analysis of undesired human-object correlation on BEHAVE [1].** We conduct a sensitivity test, inspecting which region is sensitive in reconstruction, for Transformer baseline and our CRFormer. In the Transformer baseline, the object errors are sensitive to human regions not actually related to human-object interaction, as a result of undesired correlation. In our CRFormer, the object errors are mostly sensitive around regions containing human-object contact.

Analysis of undesired human-object correlation. In this section, we provide an in-depth analysis of undesired human-object correlation, which is detrimental to plausible reconstruction. Human-object correlation is beneficial for learning 3D human and object reconstruction, in most cases. However, the reconstruction network can be biased by the strong correlation between human and object poses and marginalize evidence within the image. In the case of Figure 2, 3D pose of the object (*i.e.*, monitor) is primarily determined by the human head, ignoring image evidence. Although such an undesired correlation is detri-

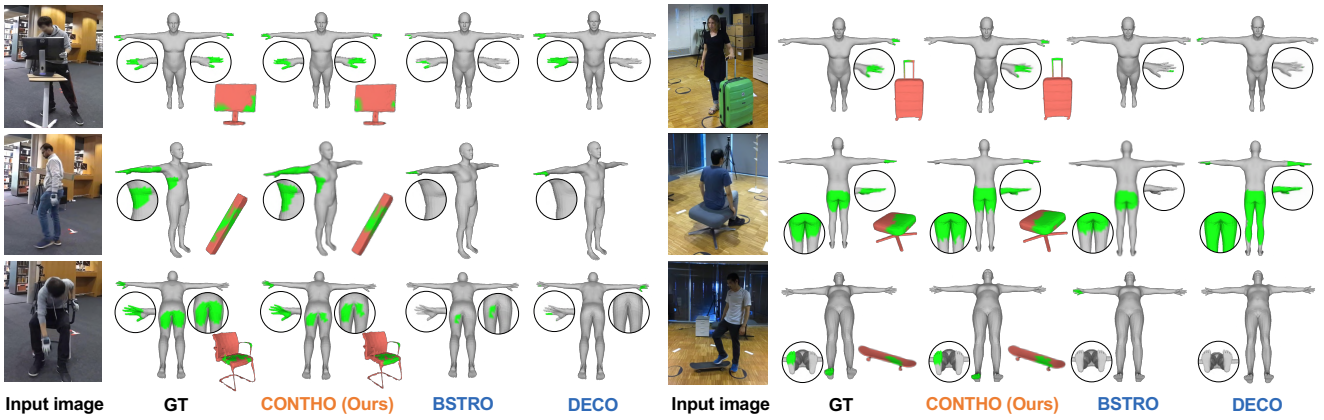


Figure 5. **Qualitative comparison of human-object contact estimation with BSTRO [13] and DECO [40], on BEHAVE [1] (left) and InterCap [14] (right).** The green color indicates the contacting regions.

Methods	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p^{rec}\uparrow$	$Contact_r^{rec}\uparrow$
Initial reconstruction	5.70	10.86	0.547	0.394
*** Variations of feature aggregation**				
Without contact maps	5.13	8.77	0.601	0.456
Add contact maps	5.12	8.54	0.616	0.483
Concat contact maps	5.18	8.65	0.618	0.477
Contact-based masking (Ours)	**4.99**	**8.42**	**0.628**	**0.496**
*** Variations of CRFormer design**				
CRFormer w/o CA Transformers	5.40	9.03	0.591	0.419
CRFormer w/o SA Transformers	5.49	8.88	0.598	0.473
CRFormer (Ours)	**4.99**	**8.42**	**0.628**	**0.496**

Table 2. **Ablation studies for contact-based refinement on BEHAVE [1].**

Datasets	Methods	$Contact_p^{est}\uparrow$	$Contact_r^{est}\uparrow$
BEHAVE	POSA [11]	0.514	0.299
	BSTRO [13]	0.615	0.527
	DECO [40]	0.638	0.337
	CONTHO (Ours)	**0.754**	**0.587**
InterCap	POSA [11]	0.561	0.333
	BSTRO [13]	0.506	0.427
	DECO [40]	0.635	0.479
	CONTHO (Ours)	**0.660**	**0.612**

Table 3. **Quantitative comparison of human-object contact estimation with state-of-the-art methods on BEHAVE [1] and InterCap [14].**

mental to plausible reconstruction, there has not been much discussion for 3D human and object reconstruction. Consequently, we analyze the undesired human-object correlation issue using a sensitivity test, motivated by pioneering works [20, 50]. Figure 4 (a) shows the procedure of the sensitivity test. Given an input image, we create an occluded image with an occluding patch for each pixel of the input image over a sliding window. Then, we measure object reconstruction error (*i.e.,* CD_{object}) from each occluded image. Repeating this process for all image regions yields a sensitivity map that indicates which regions in an image the object error is correlated with.

We conduct the sensitivity test for two different feature aggregation strategies of the contact-based refinement module, as shown in Figure 4 (b). The Transformer baseline naively aggregates 3D vertex features and contact maps with concatenation. As shown in Figure 4 (c), the sensitivity maps of the Transformer baseline are highly activated around human regions, which are not actually related to human-object interaction. This means that 3D object reconstructions are much more correlated with human features from human regions than object features, although the human regions do not contain reasonable human-object interaction. On the other hand, our CRFormer shows reasonable sensitivity maps, which are activated around object regions or human-object contacting regions. This means that 3D object reconstruction is mainly correlated with the object's own regions or human-object contacting regions, which is a result of desirable correlation between human and object features. Our CRFormer only considers human-object interaction among features from contacting regions through contact-based masking. Thus, by explicitly preventing the exploitation of features unrelated to human-object contact, the CRFormer alleviates the undesired human-object correlation, producing accurate reconstruction results. We provide more analysis examples in the supplementary material.

5.4. Comparison with state-of-the-art methods

We compare ours with previous state-of-the-art methods with two experimental protocols: 1) training & evaluating all methods on BEHAVE [1] and 2) training & evaluating all methods on InterCap [14].

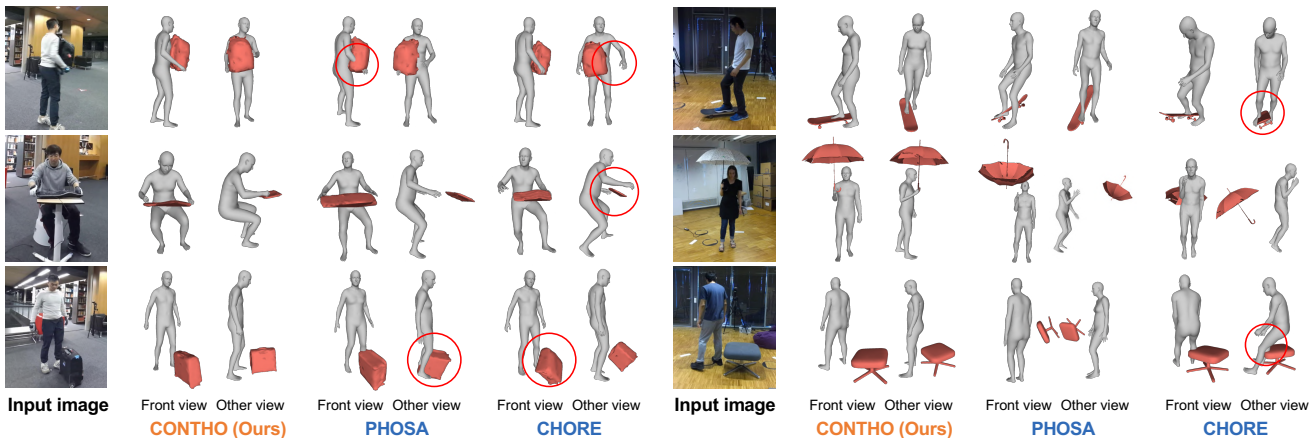


Figure 6. **Qualitative comparison of 3D human and object reconstruction with PHOSA [52] and CHORE [45], on BEHAVE [1] (left) and InterCap [14] (right).** We highlight their representative failure cases with red circles.

Datasets	Methods	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p^{rec}\uparrow$	$Contact_r^{rec}\uparrow$
BEHAVE	PHOSA [52]	12.17	26.62	0.393	0.266
	CHORE [45]	5.58	10.66	0.587	0.472
	CONTHO (Ours)	**4.99**	**8.42**	**0.628**	**0.496**
InterCap	PHOSA [52]	11.20	20.57	0.228	0.159
	CHORE [45]	7.01	12.81	0.339	0.253
	CONTHO (Ours)	**5.96**	**9.50**	**0.661**	**0.432**

Table 4. **Quantitative comparison of 3D human and object reconstruction with state-of-the-art methods on BEHAVE [1] and InterCap [14].**

Human-object contact estimation. Figure 5 and Table 3 show that our CONTHO largely outperforms the state-of-the-art methods: POSA [11], BSTRO [13], and DECO [40]. BSTRO [13] and DECO [40] often fail to capture human-object contact, especially in relatively small human parts (*e.g.,* hands), as human-object contact exists in a small area of the image. In such difficult scenarios, our CONTHO is superior in capturing the local human-object contact, with the proposed 3D-guided contact estimation. Under the 3D-guided contact estimation, 3D vertex feature provides explicit guidance on where to focus in local image regions, which allows the model to capture local context of human-object contact. Furthermore, whereas previous methods only estimate human contact map, CONTHO additionally estimates object contact map along with human contact map. This provides richer information about human-object interaction, showing which object parts are in contact with a human.

3D human and object reconstruction. Figure 6 and Table 4 show that our CONTHO produces much better reconstruction results than the state-of-the-art methods: PHOSA [52] and CHORE [45]. PHOSA [52] and CHORE [45] produce implausible reconstruction results, especially in terms of incorrect 3D object pose and human-object penetration. The previous methods also fail when human and object are not in contact (the last row in Figure 6), producing reconstructions with an excessively short human-object distance. This is largely due to their high reliance on imperfect optimization targets (*e.g.,* 2D silhouettes) during their optimization process. On the other hand, our CONTHO accurately reconstructs 3D human and object meshes in both contacting and non-contacting cases by the following reasons. First, our method reconstructs 3D human and object based on data-driven knowledge, instead of being optimized towards imperfect targets. Second, our CRFormer learns human-object correlation mainly based on contact maps; focusing on contact regions in contacting cases, while learning no correlation in non-contacting cases. As a consequence, our proposed method outperforms the previous reconstruction methods by a noticeable margin.

6. Conclusion

We propose CONTHO, a novel and powerful contact-based 3D human and object reconstruction method that utilizes human-object contact as the main driving signal in reconstruction. For both accurate contact estimation and 3D human and object reconstruction, we propose a 3D-guided contact estimation pipeline and a contact-based refinement Transformer. As a result, our CONTHO significantly outperforms previous methods in both human-object contact estimation and 3D human and object reconstruction.

Acknowledgements. This work was supported in part by the IITP grants [No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University), No.2021-0-02068, and No.2023-0-00156], the NRF grant [No.2021M3A9E4080782] funded by the Korean government (MSIT).

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *CVPR*, 2022. 3, 5, 6, 7, 8

[2] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *ICCV*, 2019. 3

[3] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *CVPR*, 2023. 2

[4] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6

[5] Hongsuk Choi, Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Beyond static features for temporally consistent 3D human pose and shape from a video. In *CVPR*, 2021. 3

[6] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *CVPR*, 2022.

[7] Hongsuk Choi, Hyeongjin Nam, Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Rethinking self-supervised visual representation learning in pre-training for 3D human pose and shape estimation. In *ICLR*, 2023. 3

[8] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. SO-Pose: Exploiting self-occlusion for direct 6D pose estimation. In *ICCV*, 2021. 3

[9] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *CVPR*, 2020. 2

[10] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3D human self-contact. In *AAAI*, 2021. 2

[11] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 2021. 2, 7, 8

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[13] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7, 8

[14] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *GCPR*, 2022. 5, 7, 8

[15] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *ICCV*, 2023. 3

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[19] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 3

[20] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 3, 7

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019.

[22] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019.

[23] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 3

[24] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015. 1, 2, 3

[25] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 3, 4, 5

[26] Gyeongsik Moon, Hyeongjin Nam, Takaaki Shiratori, and Kyoung Mu Lee. 3D clothed human reconstruction in the wild. In *ECCV*, 2022. 3

[27] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 2

[28] Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, and Kyoung Mu Lee. Cyclic test-time adaptation on monocular video for 3D human mesh reconstruction. In *ICCV*, 2023. 3

[29] JoonKyu Park, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Extract-and-adaptation network for 3D interacting hand mesh recovery. In *ICCVW*, 2023. 3

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 5

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1

[32] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *CVPR*, 2019. 3

[33] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3D faces using convolutional mesh autoencoders. In *ECCV*, 2018. 3

[34] Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *ECCV*, 2020. 2

[35] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. HuMoR: 3D human motion model for robust pose estimation. In *ICCV*, 2021.

[36] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. PhysCap: Physically plausible monocular 3D motion capture in real time. *ACM TOG*, 2020. 2

[37] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In *ECCV*, 2022. 1, 2

[38] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. ZebraPose: Coarse to fine surface encoding for 6DoF object pose estimation. In *CVPR*, 2022. 3

[39] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6D object pose prediction. In *CVPR*, 2018. 3

[40] Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *ICCV*, 2023. 1, 2, 3, 5, 6, 7, 8

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[42] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D object pose estimation by iterative dense fusion. In *CVPR*, 2019. 3

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 3

[44] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *RSS*, 2018. 3

[45] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *ECCV*, 2022. 1, 3, 5, 8

[46] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single RGB camera. In *CVPR*, 2023. 1, 3, 5

[47] Xianghui Xie, Xi Wang, Nikos Athanasiou, Bharat Lal Bhatnagar, Chun-Hao P Huang, Kaichun Mo, Hao Chen, Xia Jia, Zerui Zhang, Liangxian Cui, et al. RHOBIN Challenge: Reconstruction of human object interaction. *arXiv preprint arXiv:2401.04143*, 2024. 5

[48] Xiang Xu, Hanbyul Joo, Greg Mori, and Manolis Savva. D3D-HOI: Dynamic 3D human-object interactions from videos. *arXiv preprint arXiv:2108.08420*, 2021. 1, 3

[49] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 2

[50] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 7

[51] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 3

[52] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 1, 3, 8

[53] Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, and Jia-Bin Huang. Reducing footskate in human motion reconstruction with ground contact constraints. In *WACV*, 2020. 2