End-to-End HOI Reconstruction Transformer with Graph-based Encoding

Zhenrong Wang[1] Qi Zheng[1†] Sihan Ma[2] Maosheng Ye[3] Yibing Zhan[4] Dongjiang Li[4]
[1]Shenzhen University [2]University of Sydney [3]DeepRoute.AI [4]JD Explore Academy
2022280175@email.szu.edu.cn, qiz@szu.edu.cn
https://hoi-tg.github.io/

Abstract

With the diversification of human-object interaction (HOI) applications and the success of capturing human meshes, HOI reconstruction has gained widespread attention. Existing mainstream HOI reconstruction methods often rely on explicitly modeling interactions between humans and objects. However, such a way leads to a natural conflict between 3D mesh reconstruction, which emphasizes global structure, and fine-grained contact reconstruction, which focuses on local details. To address the limitations of explicit modeling, we propose the End-to-End HOI Reconstruction Transformer with Graph-based Encoding (HOI-TG). It implicitly learns the interaction between humans and objects by leveraging self-attention mechanisms. Within the transformer architecture, we devise graph residual blocks to aggregate the topology among vertices of different spatial structures. This dual focus effectively balances global and local representations. Without bells and whistles, HOI-TG achieves state-of-the-art performance on BEHAVE and InterCap datasets. Particularly on the challenging InterCap dataset, our method improves the reconstruction results for human and object meshes by 8.9% and 8.6%, respectively.

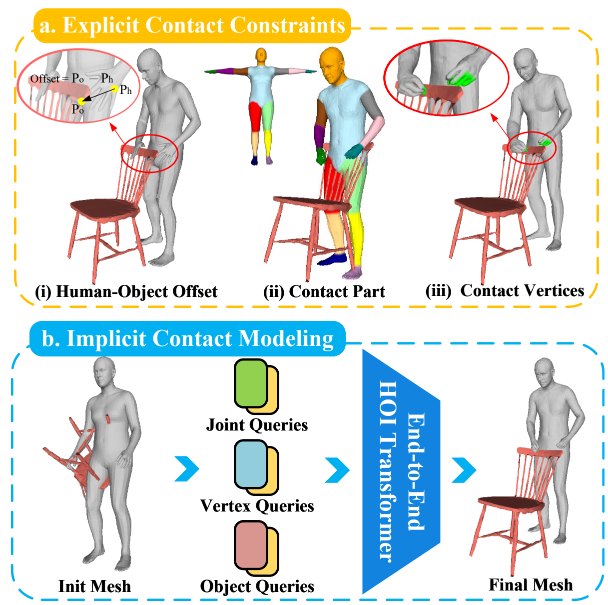


Figure 1. Comparison between existing explicit contact constraints for HOI reconstruction and our implicit contact modeling.

1. Introduction

Significant breakthroughs have recently advanced in reconstructing 3D human meshes from a single image [4, 24, 25, 30]. Such an advancement stimulates the study of joint reconstruction of 3D human and object meshes, which has numerous applications in areas such as augmented reality (AR) [5], virtual reality (VR) [1], and robotic manipulation [3]. Unlike reconstructing human meshes from a single image, Human-Object Interaction (HOI) reconstruction involves capturing the relative posture and interactions between humans and objects. Therefore, it presents considerable challenges, primarily because of the intricate articulated interactions and occlusions.

Recently, several studies [17, 31, 41] have explored leveraging interaction information for jointly modeling human and object reconstruction. These works generally incorporate interaction representations to impose hard constraints on the models. As shown in Fig. 1a, (i) StackFLOW [17] utilizes the human-object offsets between anchors, densely sampled from the surfaces of the human and object meshes, to represent their spatial relationships. (ii) CHORE [41] predicts a part correspondence field to identify which body parts are in contact with object points. (iii) CONTHO [31] estimates vertex-level human-object contact maps to prevent the learning of erroneous correlations between humans and objects.

Although leveraging contact constraints on reconstruction models seems suitable for the HOI task, a natural

†Corresponding author.

conflict appears. On the one hand, mesh reconstruction of humans and objects prioritizes an overall relative position between them. On the other hand, explicit interaction constraints—such as human-object offsets and contact maps—focus on local relationships. Achieving a good balance between these two types of constraints is not easy. For instance, StackFLOW [17] heavily relies on a time-consuming post-optimization process during inference to improve the overall reconstruction quality. To address this issue, as shown in Fig. 1b, we propose a straightforward framework that implicitly incorporates interaction-inclusive mesh reconstruction within a transformer architecture without additional explicit interaction constraints.

Given the remarkable success of transformers in human mesh recovery [4, 25–27], devising such a transformer-based framework for HOI reconstruction seems intuitive. However, three main challenges exist: 1) Using the same features across all 3D points would limit the ability to differentiate between various 3D modalities. 2) Learning interaction poses directly from static templates is much more intricate than that in human mesh recovery. 3) Simply employing self-attention would confuse the local boundaries between the human and the object, which blurs the distinction between the reconstruction of the vertices from these two relatively independent targets.

To address the first two challenges, we draw inspiration from previous works [14, 31] and provide transformers with more precise feature inputs. Specifically, our approach involves: 1) Applying grid sampling to the initial vertices, which allows us to assign distinct features to each vertex, thereby improving the discriminative power of the 3D models. 2) Utilizing the initial mesh rather than a template mesh to mitigate the difficulties associated with learning complex interactions within the model. For the third challenge, we propose integrating specific graph convolutional structures into the transformer architecture for both human and object components. Such structures have been proven successful in human mesh reconstruction [27]. In our case, the integration facilitates better fusion of local information, thereby improving the ability to differentiate between the reconstructions of the human and the object.

Therefore, we propose an End-to-End Human-Object Interaction Reconstruction Transformer with Graph-based Encoding, referred to as HOI-TG, to tackle the above challenges. First, we generate initial mesh vertices for the human and the object. These initial 3D coordinates are concatenated with the grid-sampled features to create a robust input, including the queries of joint, vertex, and object as the input to our transformer. The enhanced input improves the model's capacity to differentiate between 2D and 3D representations while simplifying the learning process for complex interactions via the attention mechanism. We reconstruct the coordinates using the output from a linear layer, derive the final human vertices based on a predefined upsampling matrix, and predict the object's pose—translation and rotation—through rigid body transformation using template-based methods. Owing to the transformer architecture, our model can effectively focus on global interactions between humans and objects. To further improve the ability to integrate local features of both entities, we devise and integrate the Human Graph Residual Block and the Object Graph Residual Block within the transformer module.

Compared with previous state-of-the-art results, our HOI-TG achieved 1) improvements of **8.0%** (human) and **5.0%** (object) in 3D reconstruction accuracy and enhanced precision by **3.4%** and recall by **5.8%** for reconstructed contact accuracy on the BEHAVE dataset [2]; 2) improvements of **8.9%** (human) and **8.6%** (object) in 3D mesh reconstruction, along with **3.9%** (precision) and **4.1%** (recall) for contact accuracy on the InterCap dataset [16]. Experiments underscore the effectiveness of our HOI-TG framework in accurately modeling interactions and complex structures. Regarding reconstruction visualization, we observe more accurate relative positions between humans and objects with reduced instances of model penetration.

In summary, we make the following contributions.

- We introduce a novel end-to-end HOI reconstruction transformer framework to model human and object meshes in interaction from a single image.
- We design the Human Graph Residual Block and the Object Graph Residual Block for the transformer encoder to enhance the distinction between two meshes and strengthen the capability of local feature fusion.
- HOI-TG achieves new state-of-the-art performance on the BEHAVE and the InterCap datasets, which highlights the effectiveness of the proposed framework.

2. Related Works

3D human reconstruction. 3D human reconstruction from a monocular camera has been an active research topic for a long time. It is difficult due to complex pose variations, occlusions, and limited 3D training data. Prior studies [10, 18, 19, 22, 23, 30, 32, 36] propose adopting the pre-trained parametric human models (*e.g.*, SMPL [28], MANO [35]). ExPose [9] directly predicts the 3D parameters of the body, face, and hands in the SMPL-X format, employing body-driven attention to localize the face and hand regions in the original image. PIXIE [10] utilizes SMPL-X's unified shape space for all body parts, allowing every expert to contribute to the overall model. Hand4Whole [30] presents a comprehensive framework for 3D human mesh estimation, reconstructing the body, hands, and face by leveraging features from 3D pose-guided grid sampling.

On the other hand, instead of adopting a parametric human model, non-parametric approaches regress vertices di-

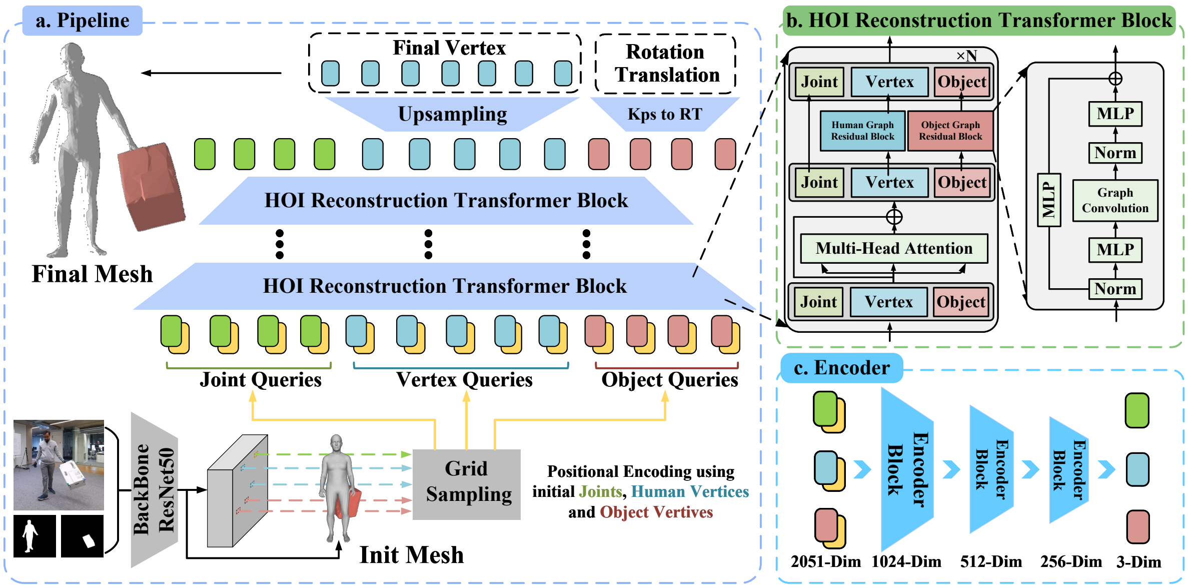


Figure 2. **Overview of our HOI-TG.** a. Pipeline draws the process of HOI reconstruction. Given the input image and human & object segmentations, we extract the image feature and generate an initial human mesh using the ResNet50 backbone. Then, we prepare joint queries, vertex queries, and object queries by concatenating grid sampling features and per-vertex 3D coordinates. Based on the queries, HOI reconstruction transformer blocks reconstruct human joints & vertices, and object mesh. Final HOI meshes are calculated by upsampling and rigid transformation. b. HOI Reconstruction Transformer Block contains a Human Graph Residual Block and an Object Graph Residual Block for separate encoding for humans and objects. c. Encoder shows the change of hidden dimensions throughout HOI-TG.

rectly from an image [7, 8, 20, 21, 26, 27, 29, 31, 43, 46]. Among the previous studies, graph convolutional neuralnetwork (GCN) is one of the most popular options as it can model the local interactions between neighboring vertices based on a given adjacency matrix [21, 27]. However, it is less efficient to capture global interactions between the human vertices and body joints. To overcome this limitation, transformer-based methods use a self-attention mechanism to attend vertices and body joints freely in the mesh and thereby encode the non-local relationship of a human mesh. For instance, METRO [26] leverages an encoder-based transformer to model non-local intersections among mesh vertices and joints. Graphormer [27] integrating GCN into the transformer architecture to improve mesh vertices local modeling. Deformer [43] proposes using new body-mesh-driven attention modules to extract multi-scale visual features, enhancing the efficiency of 3D mesh recovery.

Inspired by the success of transformers-based methods in human mesh recovery, we propose a novel transformer framework suitable for HOI reconstruction.

3D human and object reconstruction. Many early studies on human-scene contact estimation have focused on representing contact through 2D projections [6], 3D joint-level interactions [34, 37], or 3D patch-level contacts [11]. More recently, many researchers have begun to explore the estimation of dense vertex-level contact maps that are defined on the human body surface [12, 15, 31, 38, 39]. Advances in human-scene contact estimation facilitate the study of human-object interaction reconstruction. HODome [44] introduces a layer-wise neural network to reconstructe humans and objects from multi-view. HolisticMesh [40] and PHOSA [45] present optimization frameworks that are rooted in physics-based and contact-based approaches, respectively. CHORE [41] compute a neural reconstruction of human and object represented implicitly with two unsigned distance fields. StackFLOW [17] infers the posterior distribution of human-object spatial relations from the image, and optimizes the human body pose and object 6D pose. CONTHO [31] proposes contact-based refinement to prevent the learning of erroneous correlations between humans and objects.

The above methods impose explicit constraints for HOI, which causes a conflict between local contact prediction and global mesh reconstruction. Thus, we propose employing a transformer architecture that implicitly learns character interaction relationships.

3. Method

Given an RGB image depicting an HOI event and the 3D mesh template of the corresponding object, the objective of HOI reconstruction is to reconstruct the human mesh and determine the object's rotation and translation relative to the template. We propose a novel transformer-based framework, namely the HOI Reconstruction Transformer with Graph-based Encoding (HOI-TG). To effectively model complex interactions between humans and objects, HOI-TG combines the strengths of transformer architectures and GCN. To ensure the pipeline functions as intended, we first prepare joint queries, vertex queries, and object queries as inputs by grid sampling. Then, the Multi-Layer Transformer conducts HOI reconstruction. Fig. 2a shows the overall pipeline of HOI-TG.

3.1. Preparation of queries

Transferring a transformer-based human reconstruction method, such as METRO [26], directly to this task failed to yield satisfactory results. We identified two primary reasons for the failure. First, using the same 2D features at all 3D points diminishes the differentiation among various 3D modalities. Second, learning interaction poses from a static template is very challenging. Drawing inspiration from CONTHO [31], we develop an initialization strategy to provide the transformer with more effective input.

The input tensor $\mathbf{I}_{\text{input}} \in \mathbb{R}^{5\times H\times W}$ comprises an RGB image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, a segmentation mask $\mathbf{S}_h \in \mathbb{R}^{1\times H\times W}$ of the human, and a segmentation mask $\mathbf{S}_o \in \mathbb{R}^{1\times H\times W}$ of the object. H and W represent the height and width of the image, respectively. A ResNet50 backbone network [30] pre-trained for human mesh reconstruction is employed to extract features $\mathbf{F} \in \mathbb{R}^{2048\times H/32\times W/32}$ from $\mathbf{I}_{\text{input}}$. Meanwhile, the backbone also produces a rough estimation of body parameters $\theta_{\text{body}} \in \mathbb{R}^{76}$ and hand parameters $\theta_{\text{hand}} \in \mathbb{R}^{90}$. The predicted parameters are subsequently forwarded to the SMPLH [35] model, resulting in the human mesh $\mathbf{M}_h^{\text{init}*} \in \mathbb{R}^{6890\times 3}$ and joints $\mathbf{M}_j^{\text{init}} \in \mathbb{R}^{73\times 3}$. We also predict camera parameters from $\mathbf{F}$ for subsequent grid sampling and the 2D projection of joint loss calculations. To reduce computational cost and redundancy in the original mesh, we downsample $\mathbf{M}_h^{\text{init}*} \in \mathbb{R}^{6890\times 3}$ to obtain $\mathbf{M}_h^{\text{init}} \in \mathbb{R}^{431\times 3}$, following [13, 26, 27].

For the object mesh $\mathbf{M}_o^{\text{init}}$, we predict the init rotation $\mathbf{R}_{\text{init}}$ and translation $\mathbf{T}_{\text{init}}$ by adding a linear layer after the pre-trained feature extractor. Since the linear layer is randomly initialized, the initial stage does not incorporate interaction information between humans and objects. The initial results are represented by the "Init mesh" in Fig. 2a. This stage aims to provide more 3D recognizable input features for the subsequent transformer encoder.

The 3D queries features $\mathbf{Q}_j$, $\mathbf{Q}_{hv}$, and $\mathbf{Q}_{ov}$ consist of grid sampling features and per-vertex 3D coordinates as positional encoding. The grid sampling features are obtained by projecting the 3D vertices ($\mathbf{M}_j^{\text{init}}$, $\mathbf{M}_h^{\text{init}}$, and $\mathbf{M}_o^{\text{init}}$) into a 2D (*x,y*) coordinate system using the predicted camera parameters, followed by extracting the corresponding coordinate features from $\mathbf{F}$. We then form the 3D queries features ($\mathbf{Q}_j$, $\mathbf{Q}_{hv}$ and $\mathbf{Q}_{ov}$) by concatenating the grid sampling features with the per-vertex 3D coordinates from the initial meshes ($\mathbf{M}_j^{\text{init}}$, $\mathbf{M}_h^{\text{init}}$, and $\mathbf{M}_o^{\text{init}}$). Consequently, the final 3D features for human joints, human vertices, and objects vertices are represented as $\mathbf{Q}_j \in \mathbb{R}^{73\times(2048+3)}$, $\mathbf{Q}_{hv} \in \mathbb{R}^{431\times(2048+3)}$ and $\mathbf{Q}_{ov} \in \mathbb{R}^{64\times(2048+3)}$.

3.2. Transformer encoder for HOI reconstruction

Traditional transformer architectures have fixed hidden layers, which poses challenges for directly predicting 3D vertices from high-dimensional features. To address this issue, as shown in Fig. 2c, we follow [13, 26, 27] and adopt a decreasing hidden layer strategy to reconstruct 3D coordinates. Specifically, we use linear layers to reduce the dimensionality at each encoder layer. As illustrated in Fig. 2a, our HOI-TG reconstructs human joints, 3D human and object meshes, denoted as $\mathbf{M}_j$, $\mathbf{M}_h$ and $\mathbf{M}_o$, from the 3D queries features $\mathbf{Q}_j$, $\mathbf{Q}_{hv}$, and $\mathbf{Q}_{ov}$.

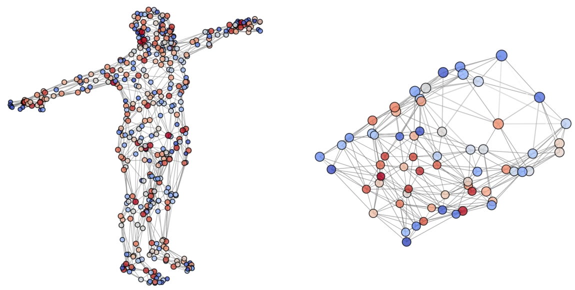
Given the 3D queries features ($\mathbf{Q}_j$, $\mathbf{Q}_{hv}$, and $\mathbf{Q}_{ov}$), we concatenate all 3D queries features to obtain $\mathbf{Q} \in \mathbb{R}^{(73+431+64)\times 2051}$. The multi-layer transformer-based encoder is employed to capture the attention relationships among humans, among objects, and between humans and objects. As shown in Fig. 2c, our encoder is composed of three HOI Reconstruction Transformer Blocks. The three encoder blocks maintain the same number of tokens but have different hidden dimensions. In this work, we follow the methodology outlined in [26], where the hidden dimensions of the three encoders are set to 1024, 512, and 256, respectively. We denote the encoded features of the joint, vertex, and object as $\mathbf{Q}_j^{'}$, $\mathbf{Q}_{hv}^{'}$, and $\mathbf{Q}_{ov}^{'}$, respectively.

Ultimately, we predict the 3D human joints $\mathbf{M}_j$, human mesh vertices $\mathbf{M}_h$, and object mesh vertices $\mathbf{M}_o$ from $\mathbf{Q}_j^{'}$, $\mathbf{Q}_{hv}^{'}$, and $\mathbf{Q}_{ov}^{'}$. For computational efficiency and redundancy reduction, we downsample the human vertices during the input stage and then upsample them back to the original scale in the prediction stage. We utilize the predefined upsampling and downsampling matrices [33] for these operations. For the prediction of $\mathbf{R}$ and $\mathbf{T}$, we calculate the rigid transform between the predicted object vertex $\mathbf{M}_o$ and the object template.

3.3. Graph-based encoding for human and object

Using self-attention to learn the global interaction relationships between humans and objects may disrupt the existing topological relationships. To address this issue, while employing the self-attention mechanism to capture global human interaction dynamics, we incorporate corresponding graph convolutional structures for both humans and objects to enhance the modeling of their respective topological structures.

As shown in Fig. 2b , we set four layers for each encoder block. For each layer, given the 3D query features, the input features $\mathbf{Q}$ are processed sequentially through the multi-



(a) Human graph (b) Object graph

Figure 3. **Graph adjacency of the human and a specific object.** The adjacency contains connectivity and distance information. Warmer colors indicate vertices with higher centrality.

head attention module and the Graph Residual Blocks. We design the Human Residual Block for the human vertices and the Object Residual Block for the object vertices.

Human Graph Residual Block. Given the human mesh features $\mathbf{Q}_{hv}^{\text{mid}} \in \mathbb{R}^{n\times d}$ generated by a multi-head attention module, we enhance local vertex interactions using graph convolution:

$$\mathbf{Q}_{hv}^{\text{mid}'} = \mathbf{GraphConv}(\bar{\mathbf{A}}, \mathbf{Q}_{hv}^{\text{mid}}; \mathbf{W}_G) = \sigma(\bar{\mathbf{A}}\mathbf{Q}_{hv}^{\text{mid}}\mathbf{W}_G). \tag{1}$$

Here, $\bar{\mathbf{A}}$ denotes the adjacency matrix of the graph as shown in Fig. 3a, and $\mathbf{W}_G$ represents the trainable parameters. The function $\sigma(\cdot)$ is the GeLU activation function. We adhere to the design principles outlined in [21, 27] to construct this block.

Object Graph Residual Block. The Object Graph Residual Block is analogous to the Human Graph Residual Block. However, due to different topological structures, the predefined $\bar{\mathbf{A}}$ in the graph convolution varies. We have defined different graph adjacency matrices for various object templates, where a specific example shows in Fig. 3b. We utilize the K-Nearest Neighbors Graph algorithm to extract the graph structure corresponding to distinct object templates. More details about the implementation can be found in the supplementary material.

3.4. Loss functions

Our proposed HOI-TG framework is trained in an end-to-end manner by minimizing the loss function $\mathcal{L}$ defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{human}} + \mathcal{L}_{\text{object}} + \mathcal{L}_{\text{hbox}}, \tag{2}$$

where $\mathcal{L}_{\text{human}}$ focuses on the reconstruction results of humans, $\mathcal{L}_{\text{object}}$ targets the reconstruction results of objects. We designed $\mathcal{L}_{\text{hbox}}$ loss function following CONTHO [31] and Hand4Whole [30]. $\mathcal{L}_{\text{hbox}}$ is the L1 distance between the predicted and ground truth (GT) bounding boxes of the hands.

The term $\mathcal{L}_{\text{human}}$ is defined as:

$$\mathcal{L}_{\text{human}} = \mathcal{L}_{\text{human}}^{\text{ms-vertex}} + \mathcal{L}_{\text{human}}^{\text{param}} + \mathcal{L}_{\text{joint}} + \mathcal{L}_{\text{edge}}, \tag{3}$$

where $\mathcal{L}_{\text{joint}}$ represents the L1 distance between the predicted and GT human joint coordinates, encompassing both initial and refined 3D and 2D coordinates. The term $\mathcal{L}_{\text{edge}}$ denotes the edge length consistency loss between the predicted and GT edges of the refined human meshes $\mathbf{M}_h$. Additionally, $\mathcal{L}_{\text{human}}^{\text{param}}$ quantifies the L1 distance between the predicted initial SMPLH parameters (θ_{body} and θ_{hand}) and their GT counterparts.

$\mathcal{L}_{\text{human}}^{\text{ms-vertex}}$ refers to the human multi-scale vertex loss. It involves a two-stage upsampling process: first, we upsample the predicted coarse mesh (431 vertices) to an intermediate resolution (1723 vertices), followed by further upsampling to the original mesh (6890 vertices for the SMPLH human mesh). The loss $\mathcal{L}_{\text{human}}^{\text{ms-vertex}}$ applies a three-scale vertex loss between the predicted mesh and the ground truth.

The term $\mathcal{L}_{\text{object}}$ is defined as:

$$\mathcal{L}_{\text{object}} = \mathcal{L}_{\text{object}}^{\text{vertex}} + \mathcal{L}_{\text{object}}^{\text{param}}, \tag{4}$$

where $\mathcal{L}_{\text{object}}^{\text{vertex}}$ represents the L1 distance between the predicted and GT per-vertex 3D coordinates of the refined object meshes $\mathbf{M}_o$, and $\mathcal{L}_{\text{object}}^{\text{param}}$ quantifies the L1 distance between the predicted and GT 3D object rotation $\mathbf{R}_o$ and translation $\mathbf{t}_o$.

4. Experiments

4.1. Datasets

We conduct experiments on the BEHAVE [2] and the InterCap [16] datasets. BEHAVE is an indoor dataset that captures seven subjects interacting with 20 diverse objects using a multi-view camera capture system. We follow CHORE [41] and CONTHO [31] for the split of BEHAVE for a fair comparison. InterCap is another indoor human-object interaction dataset containing ten human subjects with ten different objects. We split the dataset following prior works [31, 42].

4.2. Evaluation metrics

Chamfer distance ($\mathbf{CD}_{\text{human}}$, $\mathbf{CD}_{\text{object}}$). Following previous works [41, 42], we evaluate human and object reconstruction using Chamfer distance between predicted and GT meshes. With the aligned 3D human and object meshes, we measure the Chamfer distance from GT separately on 3D human SMPL and 3D object in centimeters.

Contact quality ($\mathbf{Contact_p}$, $\mathbf{Contact_r}$). In addition to employing chamfer distance to evaluate the quality of human and object reconstruction, we also assess the accuracy of reconstructed contact predictions to measure human and object interaction quality. We follow [31] to generate the contact map by identifying human vertices that are within 5 cm

	BEHAVE				InterCap			
Methods	$\text{CD}_{\text{human}}\downarrow$	$\text{CD}_{\text{object}}\downarrow$	$\text{Contact}_{\text{p}}\uparrow$	$\text{Contact}_{\text{r}}\uparrow$	$\text{CD}_{\text{human}}\downarrow$	$\text{CD}_{\text{object}}\downarrow$	$\text{Contact}_{\text{p}}\uparrow$	$\text{Contact}_{\text{r}}\uparrow$
METRO [26]	46.82	59.13	0.084	0.520	42.83	82.12	0.049	0.641
Graphormer [27]	42.40	36.60	0.047	0.348	38.82	46.18	0.031	0.332
PHOSA [45]	12.17	26.62	0.393	0.266	11.20	20.57	0.228	0.159
CHORE [41]	5.58	10.66	0.587	0.472	7.01	12.81	0.339	0.253
CONTHO [31]	4.99	8.42	0.628	0.496	5.96	9.50	0.661	0.432
HOI-TG (Ours)	**4.59**	**8.00**	**0.662**	**0.554**	**5.43**	**8.68**	**0.700**	**0.473**

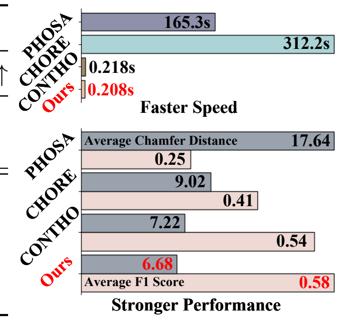


Table 1. Comparison of various methods on the BEHAVE [2] and InterCap [16] datasets. As a reference, we report the results of directly transferring the transformer-based human mesh recovery methods METRO [26] and Graphormer [27] to the HOI reconstruction task. The best results are marked as **bold**. We also compare the running time, average Chamfer distance, and average F1 score ($=2\times p\times r/(p+r)$).

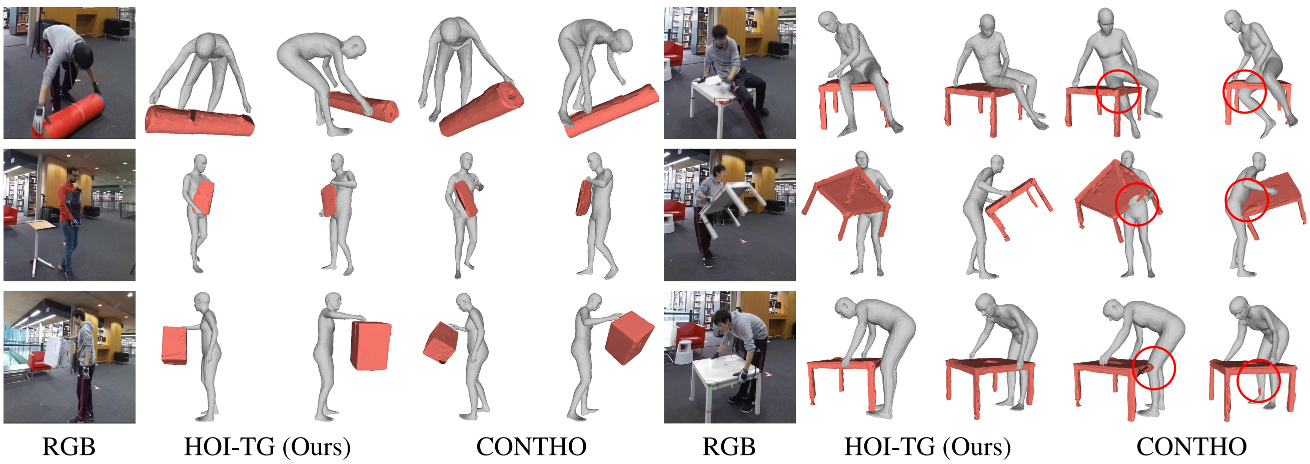


Figure 4. Qualitative comparison of 3D human and object reconstruction with CONTHO [31] on BEHAVE [2]. Our HOI-TG achieves higher accuracy regarding the relative poses between the human and the object while also reducing instances of mesh penetration.

of the object mesh. Then, we calculate precision and recall by comparing the human contact map with the ground truth.

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

We compare our method with previous approaches on BEHAVE [2] and InterCap [16] datasets. The counterparts include a classical optimization-based method PHOSA [45], a representative implicit neural field-based model CHORE [41], and the latest 3D vertex refinement-based model CONTHO [31]. According to Tab. 1, our method outperforms the previous state-of-the-art on both datasets. The results of PHOSA [45] on BEHAVE are not satisfactory, particularly regarding the prediction of reconstructed contact outcomes. It indicates the shortcoming of reliance on pre-set physical priors. CHORE [41] improves significantly over PHOSA, which shows the advantage of jointly modeling humans and objects via neural fields. CONTHO [31] outperforms the previous two approaches by devising an initialization-refinement pipeline.

Compared to CONTHO [31], our HOI-TG improves the reconstruction of humans and objects by 8.0% and 5.0%, respectively. For the reconstructed contact, our method promotes precision and recall by 3.4% and 5.8%, respectively. The results demonstrate a better reconstruction of global meshes and a higher quality of the contact areas. Since HOI-TG and CONTHO start from the same initial meshes, such improvements directly show the effectiveness of our straightforward transformer encoder and graph convolutional structures that implicitly learn the interactions between humans and objects.

The InterCap dataset contains complex topological structures (*e.g.*, umbrellas and skateboards) and is, therefore, more challenging. As a result, the performance of PHOSA and CHORE deteriorates heavily. It reveals the limitation of physical priors and unsigned neural fields. In contrast, vertex regression-based methods (*i.e.*, CHONTHO and our HOI-TG) achieve relatively stabler performance. Our method improves 3D mesh reconstruction by 8.9% and 8.6% for humans and objects, respectively. Meanwhile, it increases precision and recall by 3.9% and 4.1% for re-

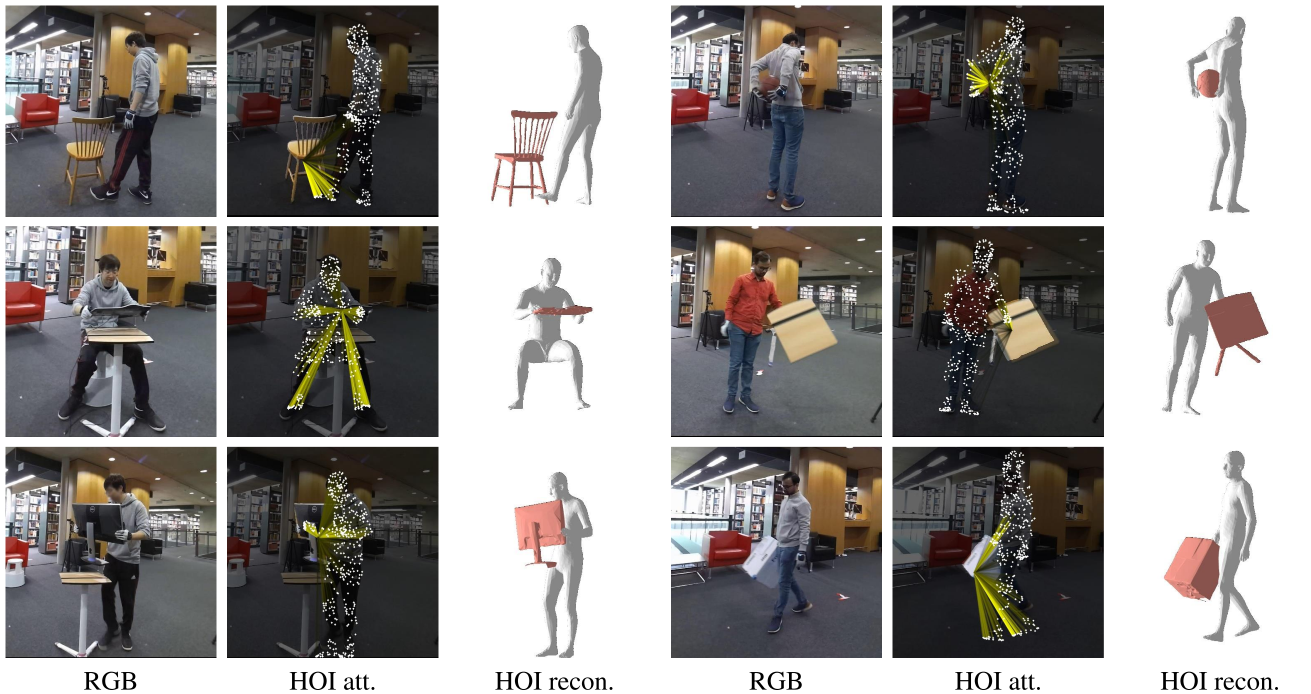


Figure 5. Visualization of the attention distribution (HOI att.) between human mesh vertices and the object and corresponding reconstruction results (HOI recon.) from our HOI-TG. The brighter color indicates more intensive attention.

constructed contact accuracy. These results highlight the advantages of HOI-TG in implicitly modeling both global posture and local interactions simultaneously.

Examples in the left half of Fig. 4 illustrate that, for relatively symmetrical graph structures, the previous state-of-the-art method CONTHO [31] fails to address the topological variations among different objects. Such deficiency prevents accurate pose predictions for those symmetrical entities. We introduced an object graph residual block to enhance the modeling of associations between object vertices. Along with the self-attention mechanism, the block allows our HOI-TG to model confusing symmetrical objects better and predict their poses more accurately. As demonstrated in the right half of Fig. 4, our model performs well in reconstructing complex human-object interactions. CONTHO [31] often struggles in identifying the relative positioning of humans and objects. It is likely due to the explicit constraints imposed on interactions, which can lead to mesh penetration. Benefiting from implicit interaction modeling, our HOI-TG produces a more robust human-object interaction reconstruction.

4.4. Visualization and ablation study

To thoroughly analyze HOI-TG, we conduct essential visualization to unveil the attention it learns and a series of ablation studies on the BEHAVE dataset. Please refer to the supplementary material for more results.

Human and object interactions: To further understand the effect of multi-layer HOI transformer encoder in learning interactions, we visualize the attention distribution (HOI att.) among human and object mesh vertices and also the reconstruction results (HOI recon.) in Fig. 5. We use the average of attention values from each human vertex to all object vertices as the attention value of that human vertex towards the object. Specifically, the dimension of the attention matrix between human and object is 431×64. The average operation produces a 431-dim vector for visualization. The brighter color indicates more intensive attention. For simple and static interactions, the positions of objects only relate to local body parts. For instance, the positions of a chair, table, monitor, and basketball solely depend on the locations of the interacting body parts. In contrast, for complex interactions, such as sitting on a chair while using a keyboard or moving with a suitcase, our model successfully attends to non-local body parts and leverages non-local vertices to predict the positions of the objects.

	$\text{CD}_{\text{human}}\downarrow$	$\text{CD}_{\text{object}}\downarrow$	$\text{Contact}_{\text{p}}\uparrow$	$\text{Contact}_{\text{r}}\uparrow$
Static	4.95	8.90	0.632	0.472
Initial	**4.59**	**8.00**	**0.662**	**0.554**

Table 2. Ablation study of 3D query features.

Representations of 3D query features: We study the behavior of our HOI-TG architecture when using two different 3D query features as inputs. 1) We utilize the global pooling of image features **F** along with static template coordinates (*Static*). 2) We employ the initial vertices for 2D projection, performing grid sampling of features in **F** and concatenate them with the initially estimated coordinates (*Initial*), which is our design in Sec. 3.1. As shown in Tab. 2, using the initial estimation significantly outperforms the *Static* variant. We consider two reasons for the phenomenon: 1) The effectiveness of 3D models relies on the ability to recognize and differentiate individual points belonging to different entities. If all points share the same descriptive features, their uniqueness in 3D space becomes obscured. 2) Interactions cover diverse poses between humans and objects, even for the same object. Learning such interactions directly from static templates becomes challenging because these templates do not capture dynamic variations or relative motions essential for understanding interactions.

Pipeline	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
Transformer	4.73	8.55	0.606	**0.559**
+H_g	4.61	8.11	0.651	0.539
+H_g+O_g	**4.59**	**8.00**	**0.662**	0.554

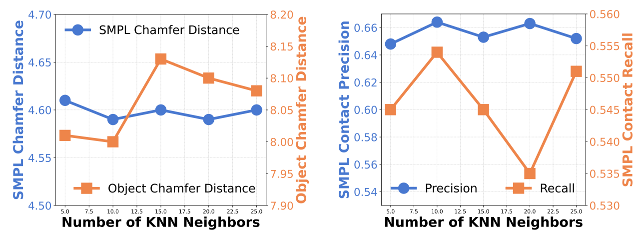
Table 3. Ablation study of graph residual blocks.

Graph residual blocks: We investigate the impact of Human Graph Residual Block (H_g) and Object Graph Residual Block (O_g) in Tab. 3. The first row corresponds to the transformer architecture that uses no graph convolution in the network. The second row shows the results of adding a human graph convolution layer to every encoder block, and the last row lists the results of adding both human and object graph convolution layers. We can conclude that: 1) H_g improves the reconstruction of humans and objects for both global meshes and local contact. It indicates that relying solely on transformers to model the interactions between them can cause ambiguity in independent topological structures. 2) Adding O_g on top of H_g further promotes the global reconstruction of humans and further increases the reconstruction accuracy of the contact area. Overall, the design of HOI-TG that involves H_g and O_g achieves good results for HOI reconstruction.

JR	MS	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
×	×	5.01	8.22	0.612	0.559
✓	×	4.76	8.30	0.631	0.563
×	✓	4.72	8.29	0.626	0.553
✓	✓	**4.62**	**8.05**	**0.644**	**0.573**

Table 4. Ablation study of $\mathcal{L}_{\text{joint}}^{\text{refine}}$ (JR) and $\mathcal{L}_{\text{human}}^{\text{ms-vertex}}$ (MS).

Loss functions: Compared with CONTHO [31], we use 1) a different $\mathcal{L}_{\text{joint}}$ containing an extra refinement of joints, which we denote as $\mathcal{L}_{\text{joint}}^{\text{refine}}$ (JR), and 2) a new $\mathcal{L}_{\text{human}}^{\text{ms-vertex}}$ (MS) to speed up convergence by constraining the quality of human reconstruction at different scales. Therefore, we investigate the impact of $\mathcal{L}_{\text{joint}}^{\text{refine}}$ and $\mathcal{L}_{\text{human}}^{\text{ms-vertex}}$ in Tab. 4. In Tab. 4, the proposed graph residual block is added in all encoders, while in Tab. 1, we only add the Object Graph Residual Block in the second transformer encoder module. The results indicate that adding one of them alone slightly hurts the reconstruction of objects, yet combining them improves global and local reconstruction for both humans and objects.



(a) Reconstruction distance. (b) Contact precision & recall

Figure 6. Ablation study of numbers of KNN neighbors.

KNN neighbors: In Fig. 6, we investigated the optimal number of KNN neighbors in the object graph residual block while constructing the adjacency matrix. Our results indicate that setting the number of neighbors to 10 leads to good performance in both the reconstructed Chamfer distance and the accuracy of contact predictions. We analyze that when a low number of neighbors is used, the model struggles to represent the adjacency relationships between nodes. Conversely, employing a high number of neighbors may result in the incorporation of redundant feature information from non-local adjacent nodes, thus obscuring the model's ability to capture these relationships accurately.

5. Conclusion

We have developed an end-to-end HOI Reconstruction Transformer with Graph-based Encoding, namely HOI-TG, for 3D human and object mesh reconstruction from a single image. By leveraging transformer architecture to implicitly model interactions and integrating graph convolutional networks for vertex interactions, our approach enhances the capture of both global and local interactions. This innovative methodology has demonstrated state-of-the-art performance on the BEHAVE and the InterCap datasets. The **limitations** of HOI-TG lie in dealing with lying poses and purely symmetric objects. A detailed analysis is provided in the supplementary material.

Acknowledgment Thanks for the support from Shenzhen Key Laboratory of Digital Creative Technology and Guangdong Province Engineering Laboratory for Digital Creative Technology. Thanks for the support of Grant 62422112, 2024YFB2808903, 2022B0101010001, and KJZD20230923114707016.

References

[1] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. State of the art of virtual reality technology. In *2016 IEEE aerospace conference*, pages 1–19. IEEE, 2016. 1

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2, 5, 6

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

[4] Zhongang Cai, Wanqi Yin, Ailing Zeng, CHEN WEI, Qingping SUN, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. Smpler-x: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, pages 11454–11468. Curran Associates, Inc., 2023. 1, 2

[5] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51:341–377, 2011. 1

[6] Yixin Chen, Sai Kumar Dwivedi, Michael J Black, and Dimitrios Tzionas. Detecting human-object contact in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17100–17110, 2023. 3

[7] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pages 342–359. Springer, 2022. 3

[8] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 769–787. Springer, 2020. 3

[9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 20–40. Springer, 2020. 2

[10] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 2

[11] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1343–1351, 2021. 3

[12] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 3

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 4

[14] Junxing Hu, Hongwen Zhang, Zerui Chen, Mengcheng Li, Yunlong Wang, Yebin Liu, and Zhenan Sun. Learning explicit contact for implicit reconstruction of hand-held objects from monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2220–2228, 2024. 2

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13285, 2022. 3

[16] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 2, 5, 6

[17] Chaofan Huo, Ye Shi, Yuexin Ma, Lan Xu, Jingyi Yu, and Jingya Wang. Stackflow: Monocular human-object reconstruction by stacked normalizing flow with offset. *arXiv preprint arXiv:2407.20545*, 2024. 1, 2, 3

[18] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 42–52. IEEE, 2021. 2

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[20] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is matter: Point-guided 3d human mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12880–12889, 2023. 3

[21] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4501–4510, 2019. 3, 5

[22] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 2

[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 *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 2

[24] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component

aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1

[25] Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21159–21168, 2023. 1, 2

[26] 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. 3, 4, 6

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2, 3, 4, 5, 6

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2

[29] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 3

[30] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2308–2317, 2022. 1, 2, 4, 5

[31] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10218–10227, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[32] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 2

[33] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, pages 704–720, 2018. 4

[34] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021. 3

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 2, 4

[36] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3d human pose and shape estimation in the wild. *arXiv preprint arXiv:2009.10013*, 2020. 2

[37] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 3

[38] 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 *European Conference on Computer Vision*, pages 516–533. Springer, 2022. 3

[39] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8001–8013, 2023. 3

[40] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. *arXiv preprint arXiv:2012.01591*, 2020. 3

[41] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pages 125–145. Springer, 2022. 1, 3, 5, 6

[42] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4757–4768, 2023. 5

[43] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2023. 3

[44] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8834–8845, 2023. 3

[45] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 3, 6

[46] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1611–1620, 2023. 3