CrossHOI: Learning Cross-View Representations for Monocular 3D Human-Object Interaction Reconstruction

Pei Geng[1], Shanshan Zhang[1,2]*, Jian Yang[1,2]
[1]PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]PCA Lab, School of Intelligence Science and Technology, Nanjing University
{peigeng,csjyang,shanshan.zhang@njust.edu.cn}

Abstract

Reconstructing 3D human-object interaction (HOI) from monocular images is highly challenging especially when human and object are mutually occluded. Existing methods primarily rely on single-view inputs, which fundamentally limit their ability to recover occluded regions and accurately estimate contact areas. To address these challenges, we for the first time, consider to introduce novel-view feature priors to enhance monocular 3D HOI reconstruction. We first design a cross-view generator that learns to infer novel-view image features from a single-view input, enriching spatial geometry at the feature level without requiring extra inputs during inference. Guided by both real and generated view features, a spatial cross-view feature fusion module adaptively aggregates complementary cues to enhance the initial reconstruction of human and object meshes. Built upon this reconstruction, we sample 3D vertex features from both views and introduce a bidirectional cross-view Transformer to integrate multi-view vertex representations for accurate contact estimation. Finally, the predicted contact maps are leveraged to refine human-object meshes, yielding geometrically consistent and physically plausible reconstructions. Experiments on BEHAVE and InterCap show that our proposed CrossHOI surpasses state-of-the-art methods in both reconstruction accuracy and contact prediction, especially under severe occlusions. Code is available at https://github.com/peigeng99/CrossHOI.git

1. Introduction

With the remarkable progress in reconstructing 3D human meshes from single-view images [7, 9, 18, 27], researchers have been motivated to focus on the more challenging task of monocular 3D human-object interaction (HOI) reconstruction. This task aims to jointly reconstruct the human and object meshes from a single-view image while accurately recovering their interaction relationships. It has broad applications in embodied intelligence [12, 31], virtual reality (VR) [2], augmented reality (AR) [6], and related fields.

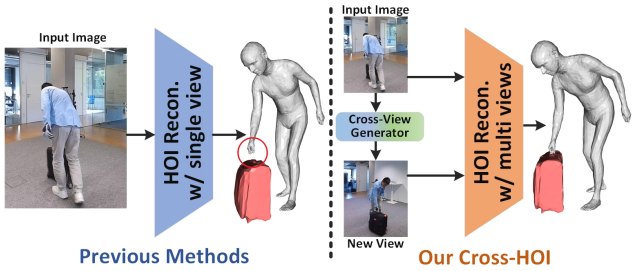

Figure 1. Compared with previous methods, our approach improves reconstruction accuracy in human-object occlusion scenarios by integrating cross-view features.

Unlike methods [3, 10, 11, 20, 24, 34, 44] that reconstruct humans or objects independently, HOI reconstruction must further predict accurate human-object contact regions. However, these contact regions are often partially or fully hidden under human-object occlusion, making it highly challenging to accurately infer their spatial locations and extents from the visible image alone. Early studies [42, 43] heuristically define contact regions and use them as hard constraints to optimize the reconstruction process, but such predefined regions often fail to match the true contact distribution in the image, leading to noticeable discrepancies between the reconstructed results and the actual contacts. Therefore, recent works [15, 28, 36, 38, 40] explore a two-stage reconstruction paradigm, where an initial reconstruction of human-object meshes is first obtained from a single-view image, and the reconstructed results are then used as 3D guidance to refine the human-object contact regions. For example, HOI-TG [36] introduces graph-aware Transformer to implicitly fuse graph structural information, enhancing the modeling capability of human-object contacts, while CONTHO [28] explicitly estimates contact re-

*Corresponding author.

gions and leverages the initially reconstructed meshes together with contact priors to refine the results. Despite these advances, relying solely on single-view features fundamentally limits their ability to resolve occluded contact areas, making the interaction reconstruction still ambiguous in complex human-object occlusion scenarios.

Inspired by the imagination ability of human recognition systems, we consider to derive novel-view image features given the single-view image as input, thereby supplementing spatial geometric information at the feature level (as shown in Fig. 1). We further leverage these cross-view features to improve HOI reconstruction. Since the quality of the initial reconstruction directly affects subsequent contact prediction, we fuse image features from two views to improve the quality of the initial reconstruction. The enhanced reconstructed mesh is combined with multi-view vertex features to jointly optimize contact regions. By doing so, the enhanced initial reconstruction facilitates contact estimation, while refined contact estimation further improves reconstruction integrity. Even under challenging occlusions, our method achieves more complete and physically consistent HOI reconstruction.

In more detail, we propose **CrossHOI**, a novel framework that generates novel-view representations from single-view inputs to enhance monocular 3D HOI reconstruction. We first design a cross-view generator and pre-train it using images from two distinct views. With only a monocular input during inference, it can infer image features of a novel view, thereby supplementing spatial geometric information at the feature level. After that, we utilize the image features from both viewpoints for HOI reconstruction. Guided by both original and generated view features, we introduce a spatial cross-view feature fusion module to aggregate complementary information adaptively, producing a more reliable initial reconstruction results. Based on this reconstruction, we further perform vertex feature sampling on the image features from both views, obtaining two sets of view-related 3D vertex representations. We then design a bidirectional cross-view Transformer to fuse the two sets of 3D vertex features, enabling more accurate estimation of human-object contact maps. Finally, the contact maps and fused 3D vertex features are jointly leveraged to refine the reconstruction, thereby improving the estimation accuracy of contact regions and the overall reconstruction quality.

In summary, our contributions are as follows:

- We, for the first time, consider to generate novel-view image features to address the human-object mutual occlusion problem and enhance monocular 3D HOI reconstruction.
- We design a cross-view generator that infers complementary view features from single-view inputs, enhancing spatial reasoning in monocular reconstruction.
- We develop a bidirectional cross-view fusion strategy that jointly optimizes reconstruction and contact estimation, leading to more accurate and realistic HOI reconstruction.
- On the BEHAVE and InterCap datasets, our CrossHOI achieves state-of-the-art performance in both reconstruction accuracy and contact estimation, showing the effectiveness of our method.

2. Related Work

2.1. 3D Human Reconstruction

Monocular 3D human reconstruction has long been a core research topic in the field of computer vision. Most 3D human reconstruction methods [1, 5, 8, 9, 17, 19, 29] typically rely on a parametric human model (e.g., SMPL [25]) to represent the 3D shape and pose of the human body. SPIN [19] integrates regression and optimization in a self-improving loop for more accurate 3D human reconstruction. TCMR [8] leverages temporal context from past and future frames to achieve temporally consistent and accurate 3D human motion reconstruction from videos. Hand4Whole [26] improves whole-body 3D human mesh estimation by modeling hand-specific kinematics and using dedicated hand features for accurate 3D wrist and finger reconstruction. In our method, we adopt the Hand4Whole [26] reconstruction manner for initial 3D human mesh reconstruction.

2.2. 3D Object Reconstruction

A common paradigm in recent single-view 3D object mesh reconstruction [23, 30, 32, 33, 35] is to first classify the object and then predict the 6DoF pose (i.e., rotation and translation) of the corresponding mesh template. PVNet [30] introduces a pixel-wise voting scheme for keypoint localization, enabling robust 6DoF pose estimation from a single RGB image under severe occlusion and truncation. ZebraPose [33] designs a discrete surface descriptor with hierarchical binary grouping, enabling more accurate 6DoF pose estimation via a coarse-to-fine strategy. Consistent with previous methods, our method also performs initial mesh reconstruction by predicting the object's 6DoF pose.

2.3. 3D Human-Object Interaction Reconstruction

Reconstructing 3D HOI from a single image is more challenging than modeling 3D humans or objects individually, as it requires not only high-quality reconstruction of both humans and objects but also precise estimation of their contact relationships. Early works [37, 42, 43] optimize the reconstruction process using predefined contact regions or physical constraints. PHOSA [43] jointly reconstructs globally consistent 3D human-object scenes by leveraging scale, occlusion-aware, and interaction constraints. HolisticMesh [37] introduces a joint optimization process with comprehensive physical priors to ensure plausible predictions. However, such hard-constraint approaches struggle

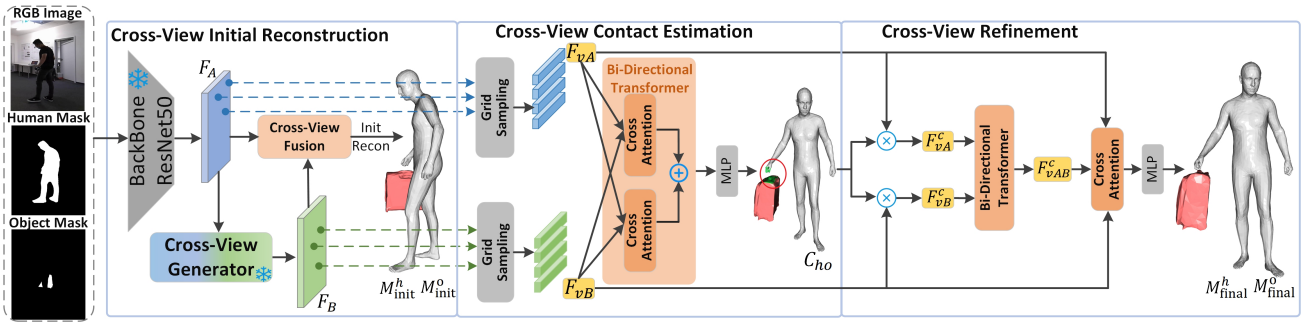

Figure 2. Pipeline of the proposed CrossHOI. Image features F_A are extracted from the original view, and features F_B are obtained using the cross-view generator. Then, a cross-view fusion module integrates F_A and F_B for the initial reconstruction. Next, 3D vertex features from both views are constructed via grid sampling and fused using a bidirectional Transformer to estimate contact maps. Finally, the predicted contact maps and dual-view vertex features are jointly leveraged by the bidirectional Transformer to refine the reconstruction.

to adaptively learn contact-related semantics, especially in occluded areas. Some studies [15, 28, 36, 38–40] explore optimization based on reconstruction results, i.e., utilizing the initial reconstruction to guide the model to optimize contact estimation. For instance, CHORE [38] jointly reconstructs humans and objects using implicit surface representations for robust 3D interaction recovery. CONTHO [28] and several recent approaches [39–41] improve human-object interaction reconstruction under a multi-stage optimization framework. While these methods have demonstrated promising results, the reliance on single-view features limits their ability to capture fine-grained interaction semantics under severe occlusion. In contrast, we explicitly introduce cross-view features to complement geometric information at the feature level and emphasize their effective integration to enhance human-object contact modeling. This method enables more accurate modeling of relative positions and contact regions between humans and objects.

3. Methodology

For a given RGB image of human-object interaction, HOI reconstruction aims to recover the 3D geometry and relative pose of the human and the object in space. In this section, we first provide an overview of the overall framework, followed by detailed descriptions of the cross-view generator and cross-view fusion for HOI reconstruction.

3.1. Overview

In this subsection, we provide an overview of the proposed CrossHOI, with the overall framework illustrated in Fig. 2. First, we train the cross-view generator in an offline manner using image pairs with the largest view differences (see Fig. 3), enabling it to generate target-view image features from a single-view input. We then employ a spatial cross-view fusion module to aggregate image features from both views and perform the initial reconstruction. Next, we sample features for 3D vertices for the human and object from the image features of both views, yielding two sets of view-dependent vertex representations for subsequent contact estimation. We then introduce a bidirectional cross-view Transformer to fuse the two sets of vertex features, fully integrating complementary cross-view information and predicting the human-object contact maps. Finally, the predicted contact maps and the 3D vertex features from both views are leveraged to refine the reconstruction, improving both contact estimation accuracy and overall reconstruction quality.

In the following subsections, we provide detailed descriptions of the cross-view generator pre-training, the HOI reconstruction process, and the loss functions.

3.2. Cross-View Generator

To obtain cross-view features, we independently train the cross-view generator. Specifically, we leverage multi-view RGB images from reconstruction datasets such as BEHAVE [4] and InterCap [14]. For each training sample, we select two images captured from the most distinct views (e.g., frontal vs. back view) to form a cross-view pair (I_A, I_B), encouraging the generator to learn robust view transformation under large viewpoint differences. Moreover, to reduce the noise from backgrounds, we employ human and object segmentation masks to extract only the interaction-relevant areas. The training pipeline is illustrated in Fig. 3.

Given an input RGB image I_A, we first extract its intermediate visual features using a ResNet-50 [13] backbone. Specifically, we use the feature map from the C3 stage, denoted as $F_A \in \mathbb{R}^{H\times W\times C}$, as it provides a good trade-off between spatial resolution and semantic information. Similarly, for another view I_B, we extract the corresponding feature map $F_B \in \mathbb{R}^{H\times W\times C}$.

To enhance geometric consistency across viewpoints, we incorporate the intrinsic camera parameters $K_A \in \mathbb{R}^{3\times 3}$ into the visual feature F_A. Specifically, We project K_A into the same embedding space as F_A via linear transformation:

$$E_{K_A} = \text{MLP}(\text{Flatten}(K_A)), \tag{1}$$

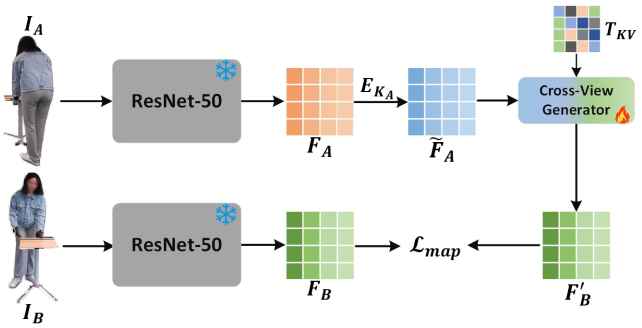

Figure 3. The training process of the cross-view generator. The image I_A is encoded into features F_A, with camera parameters added as positional bias to obtain $\tilde{F}_A$. Then, $\tilde{F}_A$ is fed as queries into the cross-view generator to generate F'_B, which is aligned with the target view features F_B. The learnable keys/values T_{KV} are initialized from view I_B and its camera parameters.

where $E_{K_A} \in \mathbb{R}^C$ denotes the camera embedding. The feature map F_A is then flattened into $F_A^t \in \mathbb{R}^{N \times C}$. To learn geometry-aware representations without disrupting spatial structure, the camera embedding E_{K_A} is added as a positional bias to each token, ensuring dimensional consistency:

$$\tilde{F}_A = F_A^t + E_{K_A}. \tag{2}$$

The geometry-aware feature $\tilde{F}_A$ serves as the query input to the cross-view generator, which maps it into the feature F'_B. We represent the key and value as learnable tokens T_{KV}, which are initialized by the features and camera parameters from view I_B. This design aligns features across views in a shared embedding space, thereby enforcing cross-view geometric consistency and enabling the generator to infer target-view features without using camera parameters from view I_B during inference. The process is defined as:

$$F'_B = \text{CrossAttn}(\tilde{F}_A, T_{KV}), \tag{3}$$

where CrossAttn$(\cdot)$ denotes a lightweight cross-attention module with a compact set of learnable tokens. This design enables effective cross-view feature mapping while keeping the low computational cost.

To further encourage cross-view feature consistency, we constrain the generated feature F'_B to align with the target feature F_B through both numerical and directional supervision:

$$\mathcal{L}_{\text{map}} = \lambda_1 \left\| F'_B - F_B \right\|_2^2 + \lambda_2 \left(1 - \cos\left(F'_B, F_B\right)\right), \tag{4}$$

where the MSE loss promotes value-level alignment, and the cosine loss enforces directional consistency in feature space. During inference, the cross-view generator can infer the feature representation of a novel view given only a single-view input. The quality of the generated target-view features will be evaluated in Section 4.4.

3.3. Cross-View Fusion for HOI Reconstruction

In this subsection, we detail how cross-view image features are integrated into the HOI reconstruction process (Fig. 2).

Cross-View Initial Reconstruction. Given an input RGB image $I_A \in \mathbb{R}^{3 \times H \times W}$ and its corresponding human and object masks, we concatenate them along the channel dimension to form the input $I_{in} \in \mathbb{R}^{5 \times H \times W}$. It is fed into a ResNet-50 backbone to extract visual features $F_A \in \mathbb{R}^{2048 \times H/32 \times W/32}$. Then, we employ the pre-trained cross-view generator (trained offline) to obtain the target-view features F_B. To effectively fuse features from the two views while avoiding redundancy, we introduce a spatial cross-view feature fusion module. Instead of directly summing F_A and F_B, we treat the spatial tokens of F_A as queries, and the tokens of F_B as keys/values, performing a cross-spatial attention to selectively inject informative cues from view B into A. The fusion process is formulated as:

$$F_{AB} = \text{Softmax}\left(\frac{\left(Q_A K_B^\top\right)}{\sqrt{d}}\right) V_B + F_A, \tag{5}$$

where Q_A, K_B, and V_B are the projected query, key, and value embeddings, and F_{AB} denotes the fused features. The residual design ensures that the fused features complement rather than override the original representation.

Based on the fused feature F_{AB}, we predict both human and object reconstruction parameters. For the human, the model estimates the 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 [25], which are then fed into the parametric model to generate the initial 3D human mesh $M^h_{\text{init}} \in \mathbb{R}^{431 \times 3}$. It is downsampled from the original 6890 vertices to reduce computational cost. For the object, we regress its 6DoF pose parameters, including rotation R_{init} and translation T_{init}, to construct the initial 3D object mesh $M^o_{\text{init}} \in \mathbb{R}^{64 \times 3}$. For fair comparison, the initial reconstruction module is implemented upon the established Hand4Whole [26] framework, following prior works [28, 36].

Cross-View Contact Estimation. After obtaining the initial 3D meshes of the human and object, M^h_{init} and M^o_{init}, we extract 3D vertex features from them to predict the human-object contact map. Each 3D vertex feature consists of two components: (1) the grid sampling features, obtained by grid sampling of the image feature; and (2) the corresponding per-vertex 3D coordinates.

To capture cross-view geometric cues, we project the vertices of the initial meshes onto both feature maps F_A and F_B to sample per-vertex visual features from each view. Subsequently, we concatenate the cross-view grid sampling features with the initial mesh's vertex coordinates to construct the 3D vertex features $F_{vA}, F_{vB} \in \mathbb{R}^{(256+3) \times (431+64)}$. This cross-view vertex representation allows the model to integrate multi-view geometric infor-

mation, enhancing the discrimination of local contact regions under occluded scenarios.

Based on the 3D vertex features from two views, we further predict human-object contact maps for contact estimation. To fully leverage their complementary information, we introduce a bidirectional fusion mechanism, which promotes the model to learn more comprehensive and consistent geometric and semantic representations. Given vertex features F_{vA} and F_{vB}, we perform bidirectional cross-attention operations:

$$\hat{F}_{vA} = \text{Softmax}\left(\frac{Q_{vA}K_{vB}^{\top}}{\sqrt{d_v}}\right)V_{vB} + F_{vA}, \quad (6)$$

$$\hat{F}_{vB} = \text{Softmax}\left(\frac{Q_{vB}K_{vA}^{\top}}{\sqrt{d_v}}\right)V_{vA} + F_{vB}. \quad (7)$$

Then, the fused vertex features $\hat{F}_{vA}$ and $\hat{F}_{vB}$ are aggregated and fed into an MLP to predict the human-object contact map $C_{ho} \in \mathbb{R}^{431+64}$. The bidirectional attention mechanism facilitates mutual feature refinement across views, while the residual connection ensures stable integration without overwriting the original vertex semantics.

Cross-View Refinement. After predicting the contact map C_{ho}, we further refine the initial meshes by focusing on the contact regions. Specifically, we multiply C_{ho} with the vertex features from both views, F_{vA} and F_{vB}, to obtain the masked (contact-related) contact features F_{vA}^{c} and F_{vB}^{c}. We then refine F_{vA}^{c} and F_{vB}^{c} via bidirectional cross-attention (see Eq. (6) (7), and obtain the fused contact feature F_{vAB}^{c} with integrated cross-view interaction cues. Next, we integrate F_{vAB}^{c} with the full-view features F_{vA} and F_{vB}, where F_{vAB}^{c} serves as the query to guide the fusion toward regions requiring fine-grained adjustment, and F_{vA} and F_{vB} serve as keys and values to provide global geometric context. Finally, the fused representation is passed through an MLP to regress per-vertex offsets ΔM^{h}, ΔM^{o} for the human and object meshes, and then obtain the final vertices positions as follows:

$$M_{\text{final}}^{h} = M_{\text{init}}^{h} + \Delta M^{h}, \quad (8)$$

$$M_{\text{final}}^{o} = M_{\text{init}}^{o} + \Delta M^{o}. \quad (9)$$

This cross-view refinement focuses the adjustments on interaction-critical vertices, leading to more accurate contact estimation while maintaining globally consistent geometry.

3.4. Loss Functions for Reconstruction

In this subsection, we describe the loss functions used to train CrossHOI. Following previous works [26, 28, 36] , the overall training objective $\mathcal{L}_{\text{recon}}$ is defined as:

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{est}} + \mathcal{L}_{\text{ref}}, \quad (10)$$

where $\mathcal{L}_{\text{init}}$ optimizes the initial reconstruction and is defined as:

$$\mathcal{L}_{\text{init}} = \mathcal{L}_{\text{param}} + \mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{hbox}}, \quad (11)$$

where $\mathcal{L}_{\text{param}}$ denotes the L1 loss between the predicted and GT SMPL+H parameters (θ_{body}, θ_{hand}) and the 6DoF object parameters (R_{init} and T_{init}). $\mathcal{L}_{\text{coord}}$ denotes the L1 loss between the predicted and GT 3D, 2D human joint coordinates. $\mathcal{L}_{\text{hbox}}$ measures the L1 distance between the predicted and GT hand bounding boxes.

In addition, $\mathcal{L}_{\text{est}}$ denotes the cross-entropy loss between predicted and GT contact map C_{ho} during the contact estimation stage. Finally, $\mathcal{L}_{\text{ref}}$ defines the refinement loss, formulated as:

$$\mathcal{L}_{\text{ref}} = \mathcal{L}_{\text{vertex}} + \mathcal{L}_{\text{edge}}, \quad (12)$$

where $\mathcal{L}_{\text{vertex}}$ is the L1 distance between the predicted human and object meshes (M_{final}^{h}, M_{final}^{o}) and their GT meshes. $\mathcal{L}_{\text{edge}}$ encourages edge-length consistency on the predicted human mesh M_{final}^{h}, promoting locally smooth and physically plausible geometry.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our method on two widely used indoor human-object interaction datasets, BEHAVE [4] and InterCap [14]. BEHAVE captures multi-view recordings of 8 subjects interacting with 20 objects in various environments, and we follow the same split as CHORE [38] and CONTHO [28] for fair comparison. InterCap contains interactions of 10 subjects with 10 objects, and we adopt the official split used in prior works [36, 39].

Evaluation Metrics. We evaluate our method using Chamfer Distance (CD_{human},$\text{CD}_{\text{object}}$) and Contact Quality ($\text{Contact}_{\text{p}}$,$\text{Contact}_{\text{r}}$) to assess reconstruction accuracy and contact prediction. Following prior works [28, 36, 38, 39], we compute the Chamfer Distance between Procrustes-aligned predicted and ground-truth meshes for the human and object (in cm). For Contact Quality, human vertices within 5 cm of the object mesh are considered in contact, and precision and recall are calculated between the predicted and ground-truth contact maps.

4.2. Implementation Details

The experiments are implemented based on the PyTorch framework. For a fair comparison, following previous methods [28, 36], the backbone network is initialized with the pre-trained weights from Hand4Whole [26]. We train the model using the Adam optimizer [16], with a batch size of 16. The initial learning rate is set to 5×10^{-5}, and decayed by a factor of 0.1 after 35 epochs, for a total of 60 epochs. Following prior works [28, 38, 39], the target region for reconstruction (the human-object interaction area) is cropped using the GT box. All experiments are conducted on the GeForce RTX 3090 GPU.

In addition, the cross-view generator is trained separately on BEHAVE and InterCap using image pairs with the

Table 1. Comparison of reconstruction quality (CD_{human}, CD_{object}) and contact estimation ($Contact_p$, $Contact_r$) on the BEHAVE and InterCap datasets. METRO [21] and Graphormer [22] are human mesh recovery methods that are adapted to the HOI reconstruction task as baselines.

Methods	BEHAVE				InterCap			
	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
METRO [21]	46.82	59.13	0.084	0.520	42.83	82.12	0.049	0.641
Graphormer [22]	42.40	36.60	0.047	0.348	38.82	46.18	0.031	0.332
PHOSA [43]	12.17	26.62	0.393	0.266	11.20	20.57	0.228	0.159
Chore [38]	5.58	10.66	0.587	0.472	7.01	12.81	0.339	0.253
CONTHO [28]	4.99	8.42	0.628	0.496	5.96	9.50	0.661	0.432
HOI-TG [36]	4.59	8.00	0.662	0.554	5.43	8.68	0.700	0.473
CrossHOI	**4.27**	**7.68**	**0.687**	**0.576**	**5.17**	**8.38**	**0.724**	**0.491**

largest view difference from the reconstruction dataset. The feature extractor adopts a pre-trained ResNet-50 [13]. The generator is optimized using Adam with an initial learning rate of 1×10^{-4}, a batch size of 32, and trained for 50 epochs, where the learning rate is decayed by a factor of 0.1 after 30 epochs.

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

In this subsection, we report our experimental results on the BEHAVE and InterCap datasets, and compare our method with several state-of-the-art methods, as summarized in Tab. 1. The comparison mainly includes the predefined contact region-based method PHOSA [43], the neural field-based method Chore [38], and the vertex regression-based methods [28, 36]. It can be observed that CrossHOI consistently outperforms all previous methods on both datasets. Specifically, PHOSA [43] shows unsatisfactory performance in both reconstruction and contact estimation, suggesting that relying on predefined contact regions limits its adaptability. On the BEHAVE dataset, compared with CONTHO [28], our method improves human and object reconstruction by 14.4% and 8.7% w.r.t. CD_{human} and CD_{object}, and enhances contact estimation precision and recall by 5.9 pp and 8.0 pp w.r.t. $Contact_p$ and $Contact_r$, respectively. HOI-TG [36] incorporates a graph structure to implicitly encode the topological relationships between 3D vertices. On the InterCap dataset, our method further improves human and object reconstruction by 4.7% and 3.5%, and boosts contact estimation precision and recall by 2.4 pp and 1.8 pp, respectively. These results demonstrate that the introduced cross-view feature prior not only enhances the overall reconstruction quality but also significantly improves contact estimation accuracy.

4.4. Ablation Studies

In this subsection, we conduct ablation experiments on the BEHAVE [4] dataset to evaluate the proposed method.

Evaluation of Cross-view Generator Quality. In our framework, a cross-view generator is pre-trained to infer image features corresponding to another viewpoint. To assess the similarity between the generated and target image features in terms of distribution, we compute the cosine similarity and the Maximum Mean Discrepancy (MMD) for cross-view pairs (I_A, I_B), which measure the consistency from the feature-level and distribution-level perspectives, respectively.

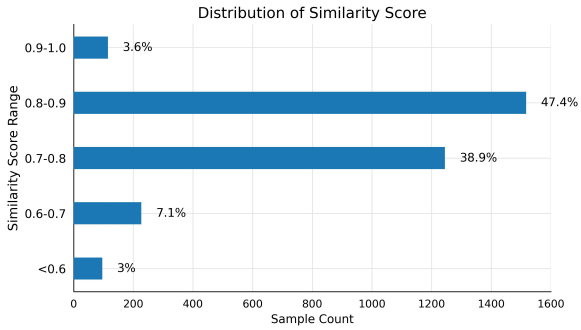

Figure 4. Evaluation of cosine similarity scores for the cross-view pair (I_A, I_B). The average cosine similarity reaches 0.784.

Specifically, for cosine similarity, we calculate the similarity for each cross-view pair, report the mean similarity, and analyze the proportion of pairs across different similarity intervals. For MMD, we randomly select 10 object categories and sample 150 cross-view pairs per category. The results are reported in Fig. 4 and Fig. 5. It can be observed that most sample pairs exhibit similarities between 0.7 and 0.9, while only a small portion (about 3%) fall below 0.6. The average cosine similarity reaches 0.784. Moreover, the MMD values for the selected object categories remain below 0.05, indicating that the generated features are highly consistent with the real feature distributions. This confirms that the pre-trained cross-view generator effectively produces consistent cross-view feature representations suitable for HOI reconstruction.

Effect of Bidirectional Transformer. To verify the effectiveness of the bidirectional Transformer, we compare it with two unidirectional fusion schemes, as reported in Tab. 2. Specifically, A→B denotes using the vertex features from view A (the original view) as queries, while the keys and values come from view B; the opposite applies for B→A. It

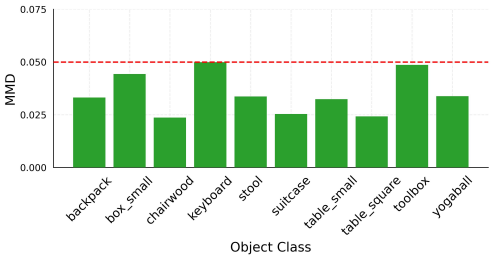

Figure 5. Analysis of distribution-level similarity for cross-view pairs (I_A, I_B).

Table 2. Comparison of different vertex feature fusion strategies on the BEHAVE dataset.

Fusion Direction	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
B→A	4.85	8.46	0.633	0.503
A→B	4.57	8.07	0.665	0.541
A↔B	**4.27**	**7.68**	**0.687**	**0.576**

can be observed that the bidirectional fusion achieves superior performance over both unidirectional settings in terms of reconstruction quality and contact estimation. Moreover, using A as the query outperforms the reverse direction. This indicates that taking the real view as the dominant query while retrieving supplementary features from the generated view leads to more robust feature integration. It also suggests that the generated features contain a certain level of noise. Since the generated and real-view features encode mutually complementary geometric cues, the bidirectional fusion (A↔B) further enhances the completeness of cross-view representations.

Table 3. Comparison of different image feature fusion methods on the BEHAVE dataset.

Method	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
element-wise add	4.47	7.91	0.652	0.530
concat+MLP	4.39	7.82	0.671	0.548
weighted sum	4.34	7.74	0.678	0.552
Ours	**4.27**	**7.68**	**0.687**	**0.576**

In addition, we visualize the reconstructed human-object meshes under the settings (A→B) and (A↔B), as shown in Fig. 6. It can be observed that in scenes with severe human-object occlusions, the unidirectional fusion (A→B) often fails to fully recover the details of the occluded regions, resulting in incomplete or misaligned contact areas. In contrast, the bidirectional fusion (A↔B) enables the model to better integrate complementary information from the two views, thereby reconstructing more accurate and continuous contact regions under complex occlusion conditions. These results further verify the effectiveness of the proposed bidirectional cross-view Transformer, which fully leverages geometric and semantic information across views to achieve more complete and fine-grained 3D human-object reconstruction.

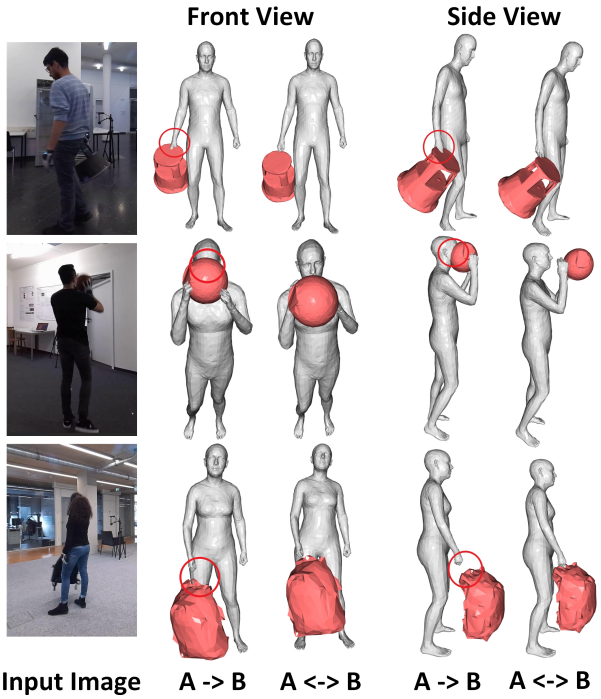

Figure 6. Qualitative comparison of different vertex feature fusion strategies on the BEHAVE dataset. Red circles indicate regions of contact prediction failure.

Table 4. Impact of cross-view features at each stage. Baseline refers to CONTHO [28]. $*$ indicates our reimplementation.

Method	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
Baseline*	5.13	8.51	0.635	0.502
+initial	4.81 (6.2%)	8.27	0.648 (1.3pp)	0.521
+contact	4.36 (9.4%)	7.82	0.679 (3.1pp)	0.560
+refine	4.27 (2.1%)	7.68	0.687 (0.8pp)	0.576
Ours	**4.27** (16.8%)	**7.68**	**0.687** (5.2pp)	**0.576**

Effect of Cross-View Feature. To verify the effectiveness of the proposed cross-view features, we design the following experiments. Specifically, we progressively incorporate cross-view features into the initial reconstruction, contact estimation, and refinement stages, where each variant is built upon the previous one to analyze their cumulative impact. The results are shown in Tab. 4. Baseline denotes CONTHO [28], and each variant represents adding cross-view features at different stages. As shown, the full model achieves the best performance. Notably, incorporating cross-view features during the initial reconstruction and contact estimation stages yields significant gains, indicating that cross-view features effectively improves human-object geometry reconstruction and interaction detail modeling.

Ablation on the Occlusion Subset. To assess the robustness under occlusion, 500 occluded samples are selected from the test set for ablation. As shown in Tab. 5, our

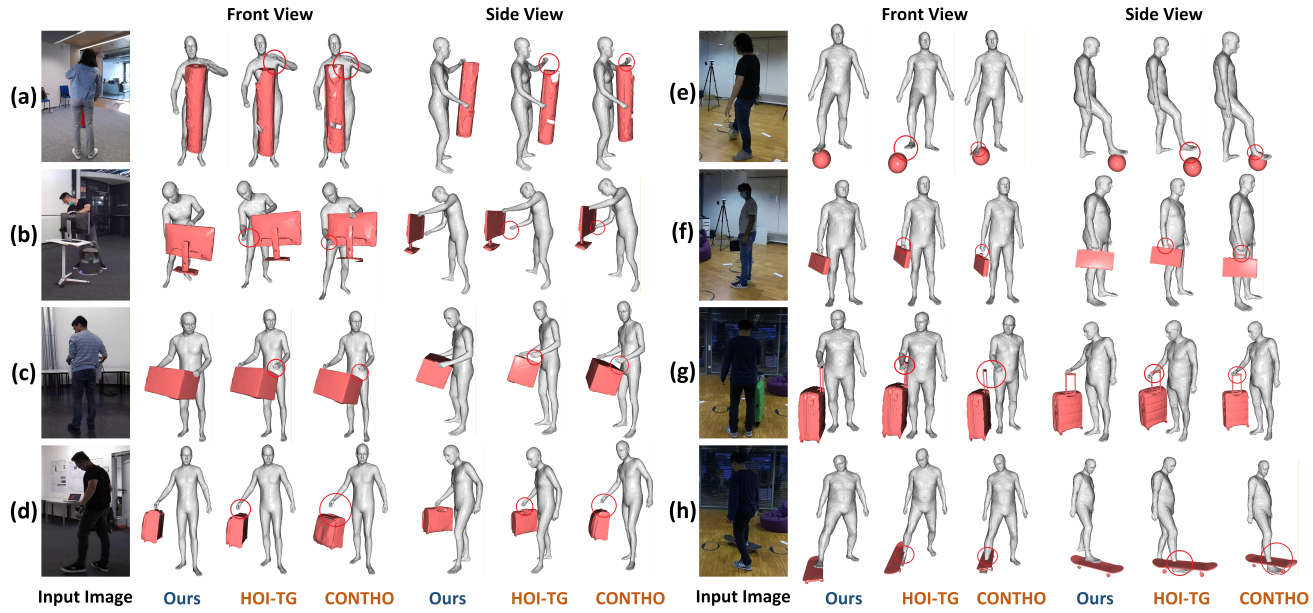

Figure 7. Qualitative comparison of human-object contact estimation with HOI-TG and CONTHO on the BEHAVE dataset (left) and InterCap dataset (right). Red circles indicate regions of contact prediction failure.

Table 5. Comparison with Baseline on occluded samples (500 samples). Baseline refers to CONTHO [28].

Method	$CD_{human}\downarrow$	$CD_{object}\downarrow$	$Contact_p\uparrow$	$Contact_r\uparrow$
Baseline	5.86	10.15	0.573	0.452
Ours	**4.91**	**8.62**	**0.629**	**0.513**

method maintains superior performance even under challenging occlusions, especially in contact estimation, with improvements of 5.6 pp and 6.1 pp w.r.t. precision and recall, respectively.

Comparison of Fusion Strategies. During the initial reconstruction stage, we employ a spatial cross-attention module to fuse image features from two views. We evaluate the impact of feature fusion by comparing several alternative fusion strategies. As shown in Tab. 3, our method achieves the best performance across all metrics. Notably, the performance gains on contact-related metrics are more pronounced than in reconstruction metrics. This is because the overall geometry can be largely recovered from a single view, so fusion brings limited gains. In contrast, local contact regions are occlusion-sensitive and benefit significantly from cross-view information. Compared with other fusion schemes, cross-attention can more effectively model complementary information between the views and adaptively select the most discriminative semantic features from the target view.

4.5. Qualitative Evaluation

In this subsection, we qualitatively demonstrate the performance of our method under complex human-object occlusion scenarios. We select eight representative cases from two datasets to compare with the state-of-the-art methods HOI-TG [36] and CONTHO [28] (see Fig. 7). Our method achieves superior overall reconstruction quality, particularly in accurately recovering human-object contact regions and object poses. Specifically, since HOI-TG relies solely on single-view information: (1) it often fails to reconstruct the occluded parts of humans and objects in many cases, leading to missing contact regions (see (b)(c)(d)(e)); (2) mesh penetration may occur due to inaccurate reasoning under occlusion (see (f)(g)(h)); and (3) the object reconstruction quality is degraded (see (a)). In contrast, our cross-view method effectively alleviates these issues by fusing complementary features from two views. These results demonstrate that our cross-view fusion strategy enhances the geometric consistency and structural completeness of the reconstruction, enabling more robust 3D human-object reconstruction under complex occlusion conditions.

5. Conclusion

In this paper, we present CrossHOI, a novel framework for monocular 3D HOI reconstruction that learns cross-view representations from single-view inputs. Our proposed cross-view generator is pre-trained to learn the mapping from input single view to novel view image features. We then integrate cross-view features to enhance both mesh reconstruction and contact estimation. Extensive experiments on the BEHAVE and InterCap datasets demonstrate that CrossHOI outperforms existing methods in both reconstruction accuracy and contact prediction, particularly in complex occlusion scenarios, validating the effectiveness of our proposed cross-view learning strategy.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62322602), and Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033).

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 2

[2] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. State of the art of virtual reality technology. In *Aerospace Conference*, 2016. 1

[3] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024. 1

[4] 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

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, 2016. 2

[6] Kun-Hung Cheng and Chin-Chung Tsai. Affordances of augmented reality in science learning: Suggestions for future research. *Journal of Science Education and Technology*, 2013. 1

[7] 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. 1

[8] 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. 2

[9] 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. 1, 2

[10] Enric Corona, Gerard Pons-Moll, Guillem Alenya, and Francesc Moreno-Noguer. Learned vertex descent: A new direction for 3d human model fitting. In *ECCV*, 2022. 1

[11] 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. 1

[12] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. *NeurIPS*, 2024. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 6

[14] 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*, 2022. 3, 5

[15] 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, 3

[16] Diederik P Kingma. Adam: A method for stochastic optimization. *ICLR*, 2014. 5

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 2

[18] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 1

[19] 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. 2

[20] 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. 1

[21] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 6

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, 2021. 6

[23] Yongliang Lin, Yongzhi Su, Praveen Nathan, Sandeep Inuganti, Yan Di, Martin Sundermeyer, Fabian Manhardt, Didier Stricker, Jason Rambach, and Yu Zhang. Hipose: Hierarchical binary surface encoding and correspondence pruning for rgb-d 6dof object pose estimation. In *CVPR*, 2024. 2

[24] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *NeurIPS*, 2023. 1

[25] 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*. 2023. 2, 4

[26] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPR*, 2022. 2, 4, 5

[27] 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. 1

[28] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *CVPR*, 2024. 1, 3, 4, 5, 6, 7, 8

[29] 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. 2

[30] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019. 2

[31] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *ICCV*, 2019. 1

[32] Jingnan Shi, Rajat Talak, Harry Zhang, David Jin, and Luca Carlone. Crisp: Object pose and shape estimation with test-time adaptation. In *CVPR*, 2025. 2

[33] 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. 2

[34] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, et al. Aios: All-in-one-stage expressive human pose and shape estimation. In *CVPR*, 2024. 1

[35] 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. 2

[36] Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, and Dongjiang Li. End-to-end hoi reconstruction transformer with graph-based encoding. In *CVPR*, 2025. 1, 3, 4, 5, 6, 8

[37] Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *CVPR*, 2021. 2

[38] 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, 6

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

[40] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *CVPR*, 2024. 1, 3

[41] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *2025 International Conference on 3D Vision (3DV)*, 2025. 3

[42] 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, 2

[43] 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, 2, 6

[44] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 1