Reconstructing 3D Human-Object Interaction via a Unified Triplane Space

Yuhang Chen[1,2] and Chenxing Wang[1,2⋆]

[1] School of Automation, Southeast University, Nanjing, China
[2] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing, China
{yh-chen,cxwang}@seu.edu.cn

Abstract. Reconstructing 3D Human-Object Interaction (HOI) from a single image is challenging due to the diversity of the reconstructed targets. Ordinarily, the human body follows a fixed parametric template, whereas objects exhibit diverse and irregular shapes. Such differences lead to inconsistent geometric priors and scale variations, making joint modeling within a unified framework highly difficult. Existing vertex-level methods struggle to effectively handle this structural discrepancy. We propose to bridge this gap by introducing a unified triplane feature map that models both human and object meshes within a shared 3D latent space. Our framework comprises three components: (1) a HOI Triplane Variational AutoEncoder (HOIT-VAE) that learns a compact and stable triplane latent space from large-scale human-object mesh data; (2) a HOI Triplane Transformer (HOIT-T) that predicts triplane features from a single image, regularized by the learned priors of HOIT-VAE; (3) a pose-fitting method for estimating the human pose from the reconstructed human mesh. Experiments on the BEHAVE and InterCap datasets show that our approach achieves competitive reconstruction performance, validating the effectiveness of the unified triplane representation in capturing the mesh structures and interactions between humans and objects.

Keywords: 3D human-object interaction · 3D reconstruction · Triplane representation

1 Introduction

Reconstructing 3D human-object interaction (HOI) from a single image plays a fundamental role in robotics [6] and AR/VR [1]. While human pose estimation [12, 16] and object pose estimation [41, 50] have each achieved remarkable progress, jointly reconstructing their interactions introduces additional complexity. The human body follows a fixed template (Skinned Multi-Person Linear Model, SMPL [29]), whereas object templates vary widely across categories in both shape and topology. This structural discrepancy makes joint reconstruction

⋆ Corresponding author.

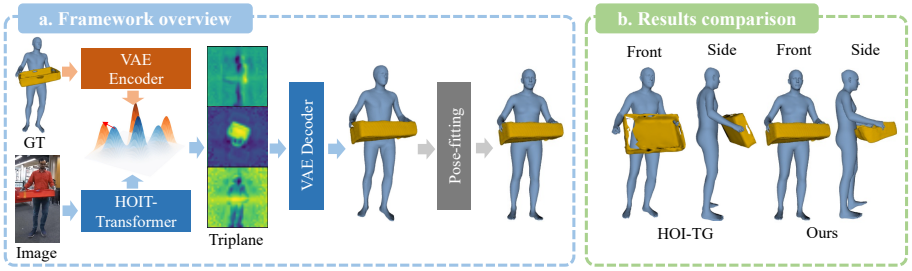

Fig. 1: Framework overview and comparison with the state-of-the-art (SOTA) methods. (a) Our framework aligns the triplane latent distributions from VAE-encoded ground-truth meshes with those predicted from a single image by HOIT-T, enabling geometry-consistent supervision for unified human–object reconstruction. (b) Compared with HOI-TG [48], our approach produces more accurate and spatially aligned human and object meshes.

particularly challenging and often leads to misaligned contact or inconsistent spatial relationships between the human and the object.

Recent studies have shifted toward joint reconstruction frameworks that explicitly model HOI, where interaction constraints are typically imposed at the vertex-level. For instance, StackFLOW [21] models interactions through relative offsets between human and object vertices, while CONTHO [38] leverages contact maps to prevent implausible interactions. Although such strategies effectively capture local contact patterns, they often overfit to the local geometry and neglect the global mesh structure. To maintain a balance between local interaction details and global structural consistency, HOI-TG [48] explicitly supervises vertex positions and integrates graph convolution [24] with self-attention [47] to implicitly model local interactions. However, since the human body template is unique while object templates are highly diverse, applying uniform positional constraints introduces an imbalance. The model becomes disproportionately sensitive to errors on large or elongated objects while under-constraining small or compact ones, leading to inconsistent optimization across categories. As a result, the learned representation is biased toward the human template, causing persistently high reconstruction errors for objects. This discrepancy suggests that the model lacks a unified feature space to jointly encode humans and objects.

In contrast, Xie *et al.* [55] represent HOI as joint point clouds and reconstruct them holistically via diffusion models [35]. Although promising, the unordered nature of point clouds makes one-stage reconstruction unstable, often requiring refinement steps that increase computational costs and limiting downstream usability.

Motivated by these limitations, we propose using a triplane feature map as a unified 3D representation for HOI. As illustrated in Fig. 1, the triplane encodes both human and object occupancy fields into three orthogonal feature planes [4], providing a compact yet spatially structured representation of 3D geometry. This

formulation preserves distinct geometric characteristics for each entity while embedding them within a shared latent space that supports joint reasoning about their relative spatial relationships. Consequently, monocular 3D HOI reconstruction is reformulated as an image-to-triplane transfer problem, which provides a tractable formulation for modeling joint human-object geometry.

To learn a structured triplane representation directly from 3D geometry, we propose the **H**uman-**O**bject **I**nteraction **T**riplane **V**ariational **A**uto**E**ncoder (**HOIT-VAE**). The main challenge lies in accurately reconstructing geometrically delicate regions, such as thin structures. To address this, HOIT-VAE replaces the binary occupancy function with a semi-continuous probability representation derived from signed distances and encodes high-resolution 3D geometry into a compact triplane latent space using a transformer-based architecture. The decoder then reconstructs human and object occupancy grids within this shared latent space. Trained on the synthetic ProciGen dataset [55], HOIT-VAE transfers to BEHAVE [2], InterCap [20] and HODome [62], showing that the triplane representation can model diverse human-object geometry without relying on object templates.

We further propose a hybrid pipeline that includes an image-to-triplane reconstruction stage for predicting human-object geometry from a single RGB image, and a model-based pose-fitting stage for estimating the final human pose parameters. For the image-to-triplane stage, we develop the **H**uman-**O**bject **I**nteraction **T**riplane **T**ransformer (**HOIT-T**). HOIT-T maps image features extracted by a DINO-based backbone [40] into the triplane latent space through a transformer decoder conditioned on 2D visual cues. To mitigate spatial ambiguity caused by image cropping, we introduce a camera-aware embedding that incorporates geometric cues and camera features for explicit 3D grounding. Supervised by triplane features reconstructed from the pre-trained HOIT-VAE, HOIT-T achieves competitive performance on BEHAVE [2] and InterCap [20], particularly on object and combined metrics. Finally, a pose-fitting method is employed to estimate the human pose under the guidance of the mesh reconstructed by HOIT-T. By leveraging the robust structural priors of SMPL-H [42] and the strong pose priors of DPoser-X [32], we recover the interaction details that are compromised by the quantization or smoothing effects inherent in the compressed triplane feature map.

In summary, our key contributions are as follows:

- We present a unified triplane representation for 3D HOI that jointly encodes the human mesh and the object mesh within a shared 3D space. This design bridges the structural gap between parametric human templates and diverse object geometries, enabling joint modeling of global structure and relative spatial relationships.
- We propose HOIT-VAE, a variational autoencoder that learns a continuous and structured triplane latent space for human and object meshes without requiring object templates, facilitating high-quality reconstruction and geometrically consistent representation learning.

- We develop a hybrid reconstruction pipeline consisting of image-to-triplane prediction by HOIT-T and a pose-fitting method for final human pose refinement. Our method achieves leading performance on object and combined metrics across BEHAVE and InterCap, offering a more balanced reconstruction of both the human and the object compared to existing methods.

2 Related Work

Single-view 3D human-object interaction reconstruction. The rapid advancement of deep learning has driven remarkable progress in single-view 3D human pose estimation [12, 16] and 6D object pose estimation [5, 17, 25, 28]. Building on these successes, recent works attempt to reconstruct full 3D human-object interactions (HOI) by enforcing geometric and physical constraints between template-based meshes.

Early approaches, constrained by the lack of 3D-annotated HOI data, relied on coarse cues such as 2D mask alignment [61], 3D bounding boxes [51], or even inferring object geometry solely from human poses [44], often producing physically implausible results. The release of large-scale datasets like BEHAVE [2] and InterCap [20] has enabled data-driven methods that directly supervise interactions. Open3DHOI [49] studies in-the-wild open-vocabulary 3D HOI reconstruction with a new dataset and a Gaussian-based optimizer. Most existing approaches employ vertex-level constraints, typically formulated as (1) relative distances (*e.g.*, STACKFLOW [21], CHORE [56], PICO [7]) or (2) explicit contact maps based on distance thresholds (*e.g.*, CONTHO [38], Xie *et al.* [57]). To further leverage temporal and vision-language priors, recent methods incorporate tracking-based consistency [58] or reasoning via 2D foundational models (InteractVLM [11]). However, such explicit or sparse constraints often bias the model towards local contact regions while weakening global spatial consistency.

To address this, HOI-TG [48] supervises only vertex positions, allowing the model to learn implicit interaction priors. While effective for the human body with a fixed SMPL template, it generalizes poorly to objects with diverse shapes. More recently, HDM [55] and TriDI [45] introduce diffusion-based models for template-free or joint reconstruction. Although unifying humans and objects within a single representation, the unordered and iterative nature of these point-cloud or diffusion-based models requires multiple inference steps, which limits efficiency and practical usability.

Learning 3D latent representations of humans and objects. Single-view 3D reconstruction is ill-posed. Earlier works on human [16, 23, 26, 27] and object reconstruction [8, 52, 59] typically adopt encoder-decoder frameworks to directly regress 3D geometry from images. These methods rely heavily on large-scale paired 2D-3D data and remain vulnerable to monocular depth ambiguity, which limits reconstruction fidelity and generalization.

Recent advances in generative modeling [33, 43, 46] and the availability of large-scale 3D datasets [3, 9, 10, 34, 54] have shifted research toward latent-space learning. Instead of predicting dense 3D outputs directly, these methods first

encode complex geometry into compact latent representations and then learn to infer these latents from images. Reformulating reconstruction as inference in a structured latent space improves training stability and output realism.

In human mesh recovery, where the human body is represented by the SMPL model [29], VAE [22] and VQ-VAE-based [39] frameworks have been widely used to encode pose parameters [12, 42] or mesh vertices [14, 15] into latent spaces. However, such methods inherently depend on the uniform topology of SMPL, making them unsuitable for jointly modeling humans and objects with diverse and irregular structures.

In 3D object reconstruction, the lack of a unified representation format has led to diverse latent formulations. Vector-set methods (*e.g.*, 3DShape2VecSet [60], Michelangelo [64], Hunyuan3D [63]) achieve compact representations but often lose geometric consistency due to purely data-driven learning. Sparse volumetric representations [36, 65] preserve spatial priors but are computationally expensive. Recently, triplane-based methods such as LRM [19] and Direct3D [53] have shown a favorable trade-off between efficiency and geometric fidelity by maintaining structured 3D priors.

Motivated by these observations, we represent both human and object meshes using a shared triplane feature space, enabling unified HOI reconstruction.

3 Method

We propose a unified framework for reconstructing 3D human-object interactions from a single RGB image via a shared triplane representation. As illustrated in Fig. 2, our pipeline consists of three main stages: latent space learning via HOIT-VAE, image-to-triplane prediction via HOIT-T, and geometry-guided human pose fitting.

3.1 Preliminaries

We aim to jointly reconstruct a human mesh $\mathbf{M}_{\text{human}}$ and an object mesh $\mathbf{M}_{\text{object}}$ from a single RGB image $\mathbf{I}$ depicting their interaction. Following previous works [38, 48, 55], we assume the availability of 2D human and object masks. Each mesh is expressed via an occupancy function, and these occupancy fields are encoded using a triplane feature map that serves as our unified 3D representation.

Semi-continuous occupancy function. Each mesh $\mathbf{M}_i$ ($i \in \{\text{human}, \text{object}\}$) is represented by an occupancy function $o_i(\mathbf{x}) : \mathbb{R}^3 \rightarrow \{0, 1\}$ that indicates whether a point $\mathbf{x}$ lies inside or outside the surface. To better model near-surface continuity, we approximate this function using a signed-distance-based semi-continuous formulation:

$$\tilde{o}_i(\mathbf{x}) = \begin{cases} 1, & \text{if } \text{sdf}_i(\mathbf{x}) < -s \\ 0.5 - \frac{0.5 \cdot \text{sdf}_i(\mathbf{x})}{s}, & \text{if } -s \leq \text{sdf}_i(\mathbf{x}) \leq s \\ 0, & \text{if } \text{sdf}_i(\mathbf{x}) > s \end{cases} \tag{1}$$

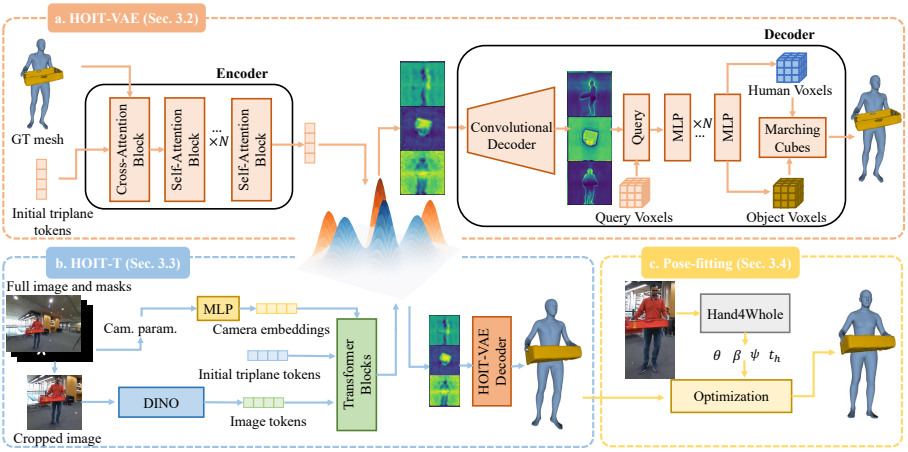

Fig. 2: Overview of the framework. (a) HOIT-VAE learns a low-resolution triplane latent space from ground-truth meshes via attention blocks, decoding them into high-resolution occupancy grids for human and object mesh extraction via marching cubes [30]. **(b) HOIT-T** predicts triplane tokens from a single image by combining DINO [40] features, mask-derived camera embeddings, and transformer blocks with adaptive Layer Normalization (adaLN) [43]. The HOIT-VAE decoder then reconstructs the final 3D meshes from these features. **(c) Pose-fitting** estimates SMPL-H [42] parameters $(\theta, \beta, \psi, t_{\mathrm{h}})$ initialized via Hand4Whole [37].

where $s = \frac{1}{512}$ controls the smooth transition width, and $\mathrm{sdf}(\cdot)$ denotes the signed distance function. This semi-continuous representation not only produces smoother mesh surfaces but also better retains structural fidelity in thin or fine regions, which are easily lost under binary occupancy modeling, as analyzed in Sec. 4.5. The goal of reconstruction is thus to estimate $\tilde{o}_i$ from $\mathbf{I}$.

Triplane representation. Rather than directly predicting $\tilde{o}_{\mathrm{human}}$ and $\tilde{o}_{\mathrm{object}}$, we introduce a triplane feature map $\mathbf{T} \in \mathbb{R}^{W_T \times H_T \times 3 \times C_T}$ as an intermediate representation. Here, C_T denotes the number of feature channels and $W_T \times H_T$ represents the resolution of the map. This representation consists of three orthogonal feature planes: $\mathbf{T}_{xy}$, $\mathbf{T}_{xz}$, and $\mathbf{T}_{yz}$. For any 3D point $\mathbf{x}$, the corresponding triplane feature is queried via trilinear interpolation:

$$\mathbf{f}(\mathbf{x}) = g(\mathbf{T}_{xy}, \mathbf{p}_{xy}) \oplus g(\mathbf{T}_{xz}, \mathbf{p}_{xz}) \oplus g(\mathbf{T}_{yz}, \mathbf{p}_{yz}), \tag{2}$$

where $\mathbf{p}_{xy}, \mathbf{p}_{xz}, \mathbf{p}_{yz}$ are the projections of $\mathbf{x}$, and $\oplus$ denotes concatenation. The occupancy probability is then predicted as:

$$\tilde{o}_i(\mathbf{x}) = \mathrm{MLP}_{\mathbf{T} \to o}\big[\mathbf{f}(\mathbf{x})\big]. \tag{3}$$

Consequently, monocular 3D HOI reconstruction is formulated as learning to predict the triplane feature map $\mathbf{T}$ from the input image $\mathbf{I}$.

Human pose. In our pose-fitting stage, we represent the 3D human body using the SMPL-H model [42], extending the standard SMPL to include expressive

hand poses. The model maps pose $\theta \in \mathbb{R}^{51\times 3}$, shape $\beta \in \mathbb{R}^{10}$, global orientation $\psi \in \mathbb{R}^3$, and translation $t_{\mathrm{h}} \in \mathbb{R}^3$ to mesh vertices $\mathbf{V}_{\mathrm{h}} \in \mathbb{R}^{6890\times 3}$.

3.2 HOIT-VAE

Directly learning high-resolution triplane feature maps is computationally expensive. Following latent 3D encoding [53], we design a variational autoencoder, HOIT-VAE, to learn a compact triplane latent representation $\mathbf{T}_{\mathrm{vae}} \in \mathbb{R}^{W'_T\times H'_T\times 3\times C'_T}$ from paired human-object meshes, where C'_T denotes the channel number, $W'_T \times H'_T$ denotes the resolution of the encoded triplane map. These latents are decoded into occupancy fields. We sample $\mathbf{T}_{\mathrm{vae}}$ from a Gaussian distribution $p_\phi(\mathbf{T}_{\mathrm{vae}}|\mathbf{M}_{\mathrm{human}}, \mathbf{M}_{\mathrm{object}})$ parameterized by the encoder ϕ.

Input. To accurately capture the geometric structure of human-object meshes, we apply Farthest Point Sampling (FPS) to sample a large number ($N_{\mathrm{human}} + N_{\mathrm{object}}$) of points $\mathbf{P} \in \mathbb{R}^{(N_{\mathrm{human}}+N_{\mathrm{object}})\times(3+3)}$ on the surfaces of the human mesh and the object mesh. The six channels consist of the normals and normalized positions of the points. We enrich the point cloud representation by augmenting $\mathbf{P}$ with Fourier Positional Encoding, which explicitly encodes spatial structure, resulting in a positional feature $\mathbf{f_P} \in \mathbb{R}^{(N_{\mathrm{human}}+N_{\mathrm{object}})\times C_P}$, where C_P denotes the channel number. The encoder takes as input the concatenation of $\mathbf{P}$ and $\mathbf{f_P}$: $\mathbf{f'_P} = \mathbf{P} \oplus \mathbf{f_P}$.

Encoder. The encoder models $p_\phi(\mathbf{T}_{\mathrm{vae}}|\mathbf{f'_P})$ by predicting the mean $\mathrm{E}(\mathbf{T}_{\mathrm{vae}}) \in \mathbb{R}^{W'_T\times H'_T\times 3\times C'_T}$ and variance $\mathrm{Var}(\mathbf{T}_{\mathrm{vae}}) \in \mathbb{R}^{W'_T\times H'_T\times 3\times C'_T}$. We start with learnable tokens $\mathbf{e}_0 \in \mathbb{R}^{(W'_T\times H'_T\times 3)\times C'_T}$. This initial representation is then refined to $\mathbf{e}_1$ by a vanilla cross-attention layer that incorporates the structural information from the point cloud features:

$$\mathbf{e}'_0 = \mathrm{CrossAttn}(\mathrm{LN}(\mathbf{e}_0); \mathbf{f'_P}) + \mathbf{e}_0, \tag{4}$$

$$\mathbf{e}_1 = \mathrm{MLP}(\mathrm{LN}(\mathbf{e}'_0)) + \mathbf{e}'_0, \tag{5}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization. Then, several self-attention layers are utilized to further capture the intra-plane and inter-plane dependencies across the triplane tokens:

$$\mathbf{e}'_i = \mathrm{SelfAttn}(\mathrm{LN}(\mathbf{e}_i); \mathrm{LN}(\mathbf{e}_i)) + \mathbf{e}_i, \tag{6}$$

$$\mathbf{e}_{i+1} = \mathrm{MLP}(\mathrm{LN}(\mathbf{e}'_i)) + \mathbf{e}'_i, \tag{7}$$

where i denotes the index of each self-attention layer. Taking the final output tokens as the input, we apply linear projections to predict the $\mathrm{E}(\mathbf{T}_{\mathrm{vae}})$ and $\mathrm{Var}(\mathbf{T}_{\mathrm{vae}})$, from which the latent $\mathbf{T}_{\mathrm{vae}}$ is sampled.

Decoder. The decoder upsamples $\mathbf{T}_{\mathrm{vae}}$ to the full triplane feature map $\mathbf{T}$ via convolutional layers and decodes it into semi-continuous occupancy fields $\tilde{o}_i$ for both human and object meshes using MLPs, as defined in Eq. 3. Finally, two

voxel grids with a resolution of r^3 are used to query the human and object occupancy fields, respectively, and the final meshes are reconstructed using the marching cubes algorithm [30]. The voxel resolution r greatly influences reconstruction fidelity, as analyzed in Tab. 3 and Fig. 6.

Learning objective. To supervise occupancy reconstruction, we sample a combined query set $\mathcal{Q} = \mathcal{Q}_s \cup \mathcal{Q}_u$ composed of near-surface points $\mathcal{Q}_s$ to capture fine details and uniformly sampled points $\mathcal{Q}_u$ for global coverage. Training is guided by two losses: (1) a Binary Cross Entropy loss between predicted and ground-truth occupancy values $\hat{o}_i(\mathbf{q})$ and $\tilde{o}_i(\mathbf{q})$ for $i \in \{\text{human}, \text{object}\}$, and (2) a KL-divergence term $\mathcal{L}_{\text{KL}}$ that regularizes the latent distribution toward a standard normal prior. The full objective is:

$$\mathcal{L}_{\text{VAE}} = \sum_i \mathbb{E}_{\mathbf{q}\in\mathcal{Q}}[\text{BCE}(\tilde{o}_i(\mathbf{q}), \hat{o}_i(\mathbf{q}))] + \lambda\mathcal{L}_{\text{KL}}, \tag{8}$$

with $\lambda = 1\text{e}{-6}$.

3.3 HOIT-T

The HOIT-VAE maps human and object meshes into a triplane latent space, providing a compact geometric prior for reconstruction. Based on this, the HOIT-T learns to predict the corresponding latent distribution $p_\epsilon(\mathbf{T}_{\text{img}}|\mathbf{I}_{\text{crop}})$ directly from a cropped RGB image $\mathbf{I}_{\text{crop}}$, where ϵ denotes the network parameters. The training objective minimizes the divergence between this predicted distribution and the HOIT-VAE posterior $p_\phi(\mathbf{T}_{\text{vae}}|\mathbf{f}'_{\mathbf{P}})$.

Image encoder. Given an input image, HOIT-T first extracts visual tokens $\mathbf{h} \in \mathbb{R}^{N_e \times d_e}$ using DINO [40], a vision transformer pre-trained with self-distillation that captures both structure and texture with useful geometry-aware representations [13].

Camera features and conditioning. We follow HDM [55] to derive camera features from the fixed full-image intrinsics and the crop bounding box computed by the masks of humans and objects. The crop-adjusted intrinsics $\mathbf{K}_{\text{crop}}$ are obtained by normalizing focal lengths and principal points, and the camera translation $\mathbf{t}_c$ is computed using HDM's closed-form two-constraint solver. Importantly, $\mathbf{t}_c$ is obtained analytically rather than learned, ensuring reproducibility and fair comparison. Further details appear in the supplementary material.

The adjusted intrinsic matrix $\mathbf{K}_{\text{crop}}$ and translation $\mathbf{t}_c$ are flattened and passed through an MLP to produce a camera embedding $\tilde{\mathbf{c}} \in \mathbb{R}^{128}$. This embedding is injected into each transformer block through adaptive LayerNorm (adaLN) [43]. Let $\mathbf{f}_j$ denote the triplane token features at the j-th transformer block, representing the intermediate 3D feature state being progressively refined. The modulation is defined as:

$$\gamma, \eta = \text{MLP}(\tilde{\mathbf{c}}), \tag{9}$$

$$\text{adaLN}_c(\mathbf{f}_j) = \text{LN}(\mathbf{f}_j) \cdot (1 + \gamma) + \eta. \tag{10}$$

This design enables camera-aware feature modulation, effectively aligning image features with the triplane features.

Transformer architecture. Each transformer block consists of a cross-attention module, a self-attention module, and an MLP, taking as input the image tokens $\mathbf{h}$ and the camera embedding $\tilde{\mathbf{c}}$. We initialize learnable triplane tokens $\mathbf{f}_0 \in \mathbb{R}^{(W'_T \times H'_T \times 3) \times C'_T}$ and denote the input of the j-th block as $\mathbf{f}_j^{\text{in}}$. The cross-attention module first attends from $\mathbf{f}_j^{\text{in}}$ to the image tokens $\mathbf{h}$, injecting visual cues into the triplane representation:

$$\mathbf{f}_j^{\text{cross}} = \text{CrossAttn}(\text{adaLN}_c(\mathbf{f}_j^{\text{in}}), \mathbf{h}) + \mathbf{f}_j^{\text{in}}. \tag{11}$$

The updated tokens are then refined through a self-attention layer to capture inter-token dependencies, followed by an MLP for feature transformation:

$$\mathbf{f}_j^{\text{self}} = \text{SelfAttn}(\text{adaLN}_c(\mathbf{f}_j^{\text{cross}})) + \mathbf{f}_j^{\text{cross}}, \tag{12}$$

$$\mathbf{f}_j^{\text{out}} = \text{MLP}(\text{adaLN}_c(\mathbf{f}_j^{\text{self}})) + \mathbf{f}_j^{\text{self}}. \tag{13}$$

The output $\mathbf{f}_j^{\text{out}}$ is passed to the next block. After all blocks, the final triplane tokens $\mathbf{f}^{\text{out}}$ are linearly projected to predict the mean and variance of the triplane distribution, $\text{E}(\mathbf{T}_{\text{img}})$ and $\text{Var}(\mathbf{T}_{\text{img}})$.

Learning objective. We employ three losses to supervise the model training: (1) the MSE loss $\mathcal{L}_{\text{MSE}}$ between $\mathbf{T}_{\text{img}}$ and $\mathbf{T}_{\text{vae}}$; (2) the cosine similarity loss $\mathcal{L}_{\text{SIM}}$ between $\mathbf{T}_{\text{img}}$ and $\mathbf{T}_{\text{vae}}$, which encourages feature alignment in the triplane latent space; (3) the KL-divergence loss $\mathcal{L}_{\text{KL}}$ for aligning the predicted distribution with the HOIT-VAE's posterior. The loss function is

$$\mathcal{L}_{\text{all}} = \lambda_1 \mathcal{L}_{\text{MSE}} + \lambda_2 \mathcal{L}_{\text{SIM}} + \lambda_3 \mathcal{L}_{\text{KL}}, \tag{14}$$

where $\lambda_1 = \lambda_2 = 1.0, \lambda_3 = 1e-6$ are the loss weights. Ablation studies on three terms are shown in Tab. 4 and Fig. 6.

3.4 Pose-fitting

To address the loss of hand details during the triplane compression and reconstruction process, we employ a model-based pose-fitting stage to fit the triplane-derived geometry to an SMPL-H [42] model, as illustrated in Fig. 2(c).

We initialize the human parameters $\{\theta, \beta, \psi, t_{\text{h}}\}$ using the Hand4Whole [37] framework. We then optimize these parameters to minimize a composite loss function $\mathcal{L}_{\text{human}}$:

$$\min_{\theta,\beta,\psi,t_{\text{h}},s} \mathcal{L}_{\text{human}} = \lambda_{\text{dist}} \mathcal{L}_{\text{dist}}(\mathbf{V}_{\text{h}}(\theta, \beta, \psi, t_{\text{h}}), s\mathbf{P}_{\text{h}}) + \lambda_\theta \mathcal{L}_{\text{prior}}(\theta) + \lambda_\beta |\beta|^2, \tag{15}$$

where $\mathbf{V}_{\text{h}} \in \mathbb{R}^{6890 \times 3}$ denotes the vertices of the SMPL-H [42] model determined by body pose θ, shape β, global orientation ψ, and translation t_{h}. s denotes the scale factor. $\mathcal{L}_{\text{dist}}$ represents the Chamfer Distance between $\mathbf{V}_{\text{h}}$ and the human

point cloud $\mathbf{P}_{\mathrm{h}} \in \mathbb{R}^{6890\times 3}$ sampled from the predicted human mesh decoded from the triplane features predicted by HOIT-T. To ensure the optimized pose remains within the manifold of natural human poses, we incorporate a pose prior term $\mathcal{L}_{\mathrm{prior}}$ from DPoser-X [32]. More details of the optimization process are provided in the supplementary material.

4 Experiments

4.1 Implementation details

Normalized camera coordinates. Following the setup in HDM [55], we apply sphere normalization to unify all HOI meshes and point clouds in a normalized camera coordinate system, where the XYZ coordinates are all bounded within $[-0.5m, 0.5m]$. All comparative evaluations with other methods are consistently conducted in this normalized camera coordinate system, ensuring a fair comparison in alignment with HDM.

Model structures and training details. Our HOIT-VAE takes as input $N_{\mathrm{human}} = 40960$ human points and $N_{\mathrm{object}} = 40960$ object points with normals uniformly sampled from the meshes. The encoder maps the input point clouds to $32 \times 32 \times 3$ triplane tokens with latent dimension 16. The encoder contains 1 cross-attention block and 8 self-attention blocks, with each attention layer comprising 12 heads of dimension 64. The decoder comprises 4 ResNet [18] blocks to up-sample the triplane tokens to $256 \times 256 \times 3$, with latent dimension 32. We apply two voxel grids with resolution 512^3 to query the triplane feature map and use 5 MLPs to decode the queried triplane features into occupancy probabilities. Finally, we apply marching cubes [30] to reconstruct the meshes from the voxel grids. During training, we sample $N_s = 20480$ near-surface points (10240 for the human mesh and 10240 for the object mesh) and $N_u = 20480$ uniform points for supervision. We employ the AdamW [31] optimizer with a learning rate of $5e-5$. We train the model with a batch size of 8 on 4 NVIDIA RTX A6000 (48GB) GPUs for 130K steps.

For our HOIT-T, we apply the frozen DINO-ViT-B/14 [40] as the image encoder, which takes a 448×448 RGB image with the background removed as input and generates 1024 feature tokens of dimension $d_e = 768$. The image-to-triplane transformer consists of 12 blocks, with each attention layer comprising 12 heads of dimension 64. During training, we employ the AdamW [31] optimizer with a learning rate of $1e-4$. We train the model with a batch size of 12 on 4 NVIDIA RTX A6000 (48GB) GPUs for 100K steps.

The details of our pose-fitting method are provided in the supplementary material.

4.2 Datasets

We train HOIT-VAE on the synthetic dataset ProciGen [55] and evaluate it on the test sets of the real datasets BEHAVE [2], InterCap [20] and HODome [62].

Table 1: Reconstruction results on BEHAVE [2] and InterCap [20]. The symbol '-' indicates that Contact_p and Contact_r cannot be calculated without a human template. 'w.': with; 'w.o.': without; 'Fit.': pose-fitting. 'Hum., Obj., Comb.': F-score@0.01m of human, object and combined meshes.

	Methods	Hum. ↑	Obj. ↑	Comb. ↑	Contact_p ↑	Contact_r ↑
BEHAVE	CHORE [56]	0.3454	0.4258	0.3966	0.587	0.472
	CONTHO [38]	0.5690	0.3177	0.5173	0.628	0.496
	HOI-TG [48]	**0.5913**	0.3278	0.5354	**0.662**	0.554
	HDM [55]	0.3925	0.5049	0.4604	-	-
	Ours (w.o. Fit.)	0.5560	**0.5391**	**0.5723**	-	-
	Ours (w. Fit.)	0.5199	0.5356	0.5497	0.619	**0.684**
InterCap	CHORE [56]	0.4064	0.5135	0.4687	0.339	0.253
	CONTHO [38]	0.4261	0.2305	0.3840	0.661	0.432
	HOI-TG [48]	0.4465	0.2118	0.3953	0.700	0.473
	HDM [55]	0.4399	0.6072	0.5344	-	-
	Ours (w.o. Fit.)	**0.6323**	**0.6192**	**0.6432**	-	-
	Ours (w. Fit.)	0.5882	0.6107	0.6298	**0.726**	**0.560**

HOIT-T is trained separately on the training splits of BEHAVE and InterCap and evaluated on their respective test sets, following the same configurations as previous works [38, 55, 56] for a fair comparison. Since our method requires accurate occupancy supervision, we apply a unified watertight preprocessing pipeline to all meshes, as detailed in the supplementary material.

4.3 Metrics and baselines

Metrics. Following HDM [55], we evaluate the reconstruction performance using the F-score based on the Chamfer Distance between point clouds sampled on the mesh surfaces. The Chamfer Distance is computed in the normalized camera coordinates. We compute the F-score with a threshold of 0.01m [35, 55] and report the scores for human, object and combined meshes separately. Following previous methods [26, 38], interaction quality is assessed by generating contact maps from human vertices within 5cm of the object mesh, reporting precision and recall against ground truth.

Baselines. We compare our method with CHORE [56], CONTHO [38], HOI-TG [48] and HDM [55] on BEHAVE and InterCap test sets in Tab. 1. As CONTHO and HOI-TG are template-based methods, we sample the same number of points from the mesh surfaces and compute the F-score using the same evaluation protocol as HDM for a fair comparison.

4.4 Comparison with state-of-the-art methods

As shown in Tab. 1, our method achieves competitive performance on BEHAVE and InterCap, with relatively strong results on object and combined metrics.

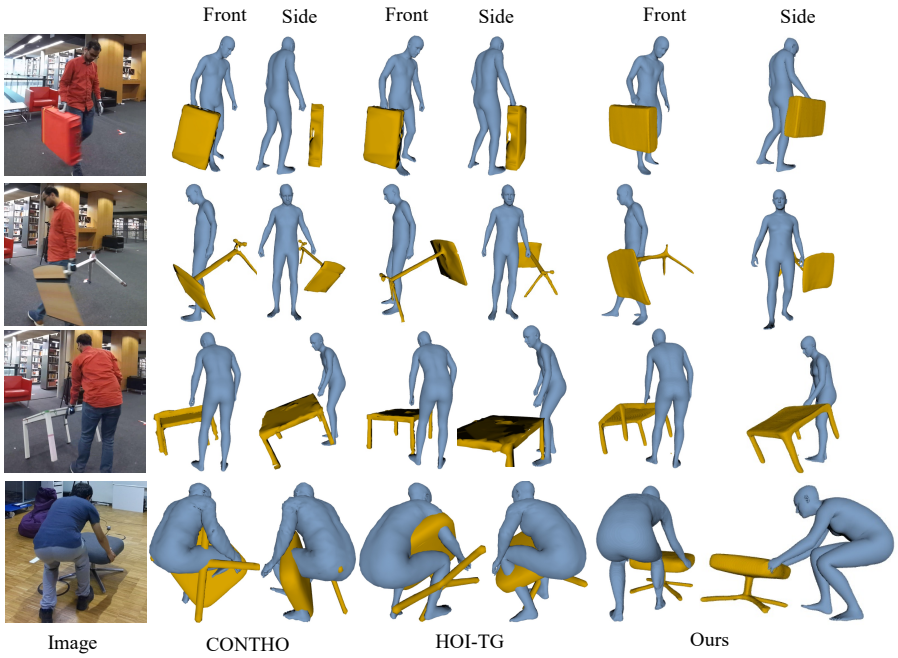

Fig. 3: Qualitative comparison of 3D human and object reconstruction with CONTHO [38] and HOI-TG [48] on BEHAVE [2] and InterCap [20] datasets.

For the human metric, template-based methods such as HOI-TG [48] and CONTHO [38] remain higher on BEHAVE [2], reflecting the strength of parametric body templates; nevertheless, our unified triplane representation achieves the best balance between human and object and yields the top combined performance. On InterCap [20], our framework consistently outperforms all baselines across human, object, and combined metrics.

As shown in Fig. 3, the qualitative comparison shows that our method reconstructs both human and object meshes with more plausible relative placement than previous methods, especially for symmetric objects such as suitcases and tables, where previous methods frequently produce flipped or misoriented poses.

The pose-fitting stage slightly decreases the surface-level F-scores, because it maps the decoded template-free human surface to the fixed-topology SMPL-H body and may lose some surface details. However, it estimates the final SMPL-H human pose parameters and improves the plausibility of the human body structure, as shown in Fig. 4. In particular, while the raw triplane output may contain blurry hand regions, the fitting procedure produces more plausible hand postures. It is also worth noting that this stage optimizes only the SMPL-H human parameters and does not refine the object mesh. The slight changes in object metrics before and after fitting are caused by the evaluation protocol, where the human-object pair is normalized as a whole before metric computation.

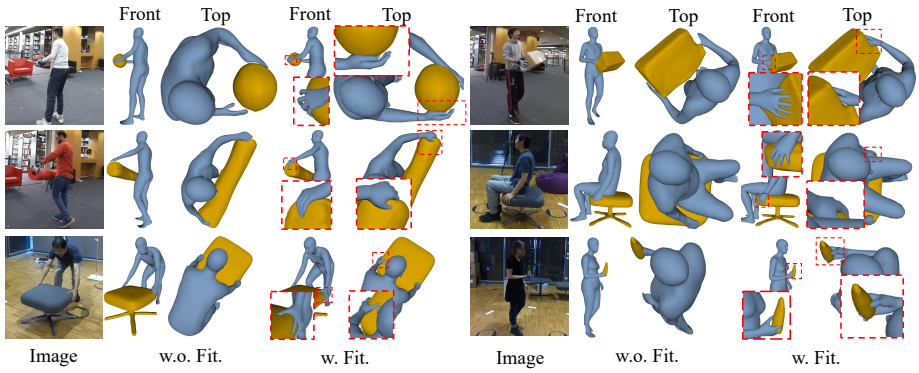

Fig. 4: Effectiveness of the pose-fitting method.

Table 2: Quantitative results of HOIT-VAE on BEHAVE [2], InterCap [20] and HODome [62] datasets (F-score@0.01m).

Datasets	Hum. ↑	Obj. ↑	Comb. ↑
BEHAVE	0.9858	0.9912	0.9913
InterCap	0.9797	0.9912	0.9872
HODome	0.9309	0.9596	0.9594

Table 3: Ablation study of HOIT-VAE on BEHAVE [2] (F-score@0.01m).

Res.	Semi.	Hum. ↑	Obj. ↑	Comb. ↑
128	✗	0.9505	0.9625	0.9611
128	✓	0.9647	0.9746	0.9740
256	✗	0.9696	0.9731	0.9738
256	✓	0.9848	0.9900	0.9905
512	✗	0.9745	0.9814	0.9804
512	✓	**0.9858**	**0.9912**	**0.9913**

Table 4: Ablation study of the HOIT-T loss terms on BEHAVE [2] (F-score@0.01m).

$\mathcal{L}_{\text{MSE}}$	$\mathcal{L}_{\text{SIM}}$	$\mathcal{L}_{\text{KL}}$	Hum. ↑	Obj. ↑	Comb. ↑
✓	✗	✗	**0.5922**	0.3885	0.5602
✓	✓	✗	0.4838	0.4434	0.4934
✓	✓	✓	0.5560	**0.5391**	**0.5723**

Table 5: Ablation study of the HOIT-T camera embeddings on BEHAVE [2] (F-score@0.01m).

Cam. Embed.	Hum. ↑	Obj. ↑	Comb. ↑
✗	0.5002	0.4794	0.5144
✓	**0.5560**	**0.5391**	**0.5723**

Thus, small changes in the fitted human mesh can slightly affect the overall normalization and lead to minor variations in the measured object scores.

4.5 Ablation studies

Results of HOIT-VAE. As demonstrated in Tab. 2, HOIT-VAE generalizes to several real-world datasets, including BEHAVE [2], InterCap [20], and HODome [62], despite being trained exclusively on the synthetic dataset ProciGen [55]. Notably, the HODome dataset contains unseen and out-of-domain object categories, further highlighting the versatility of our model. Qualitative

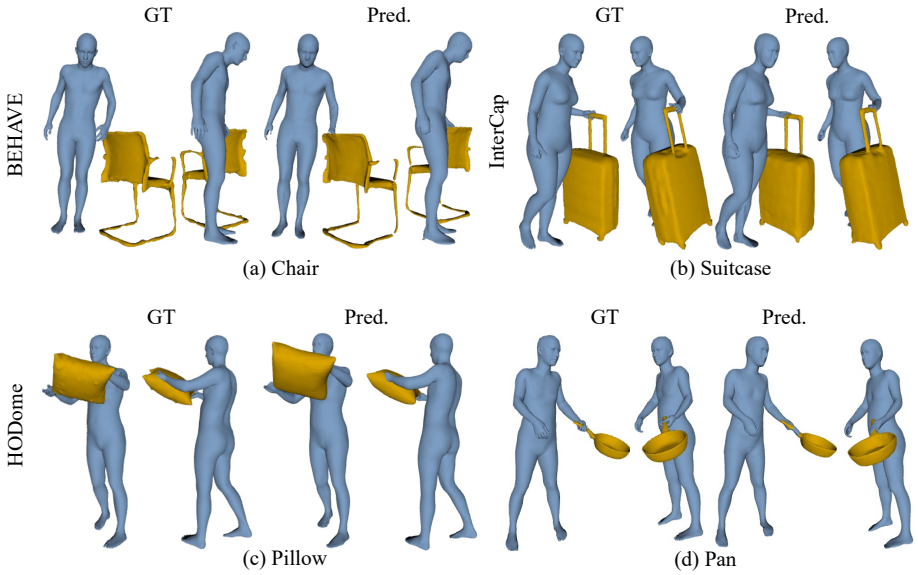

Fig. 5: Qualitative results of HOIT-VAE. (i) and (ii) show in-category reconstructions for *Chair* (BEHAVE) and *Suitcase* (InterCap), respectively. (iii) and (iv) demonstrate generalization to out-of-category objects *Pillow* and *Pan* from HODome. Compared to ground truth (GT), our predictions (Pred.) maintain complete topological structures and plausible relative positioning.

results across various objects and datasets are presented in Fig. 5, where the reconstructed meshes are generally consistent with the ground truth.

Ablation study of HOIT-VAE. Table 3 evaluates grid resolution and semi-continuous occupancy in HOIT-VAE. Increasing voxel resolution from 128^3 to 512^3 steadily improves accuracy. Moreover, semi-continuous representation enhances performance across all metrics. Figure 6 visualizes these configurations; without semi-continuous occupancy or high resolution, reconstructed surfaces lose smoothness and details.

Ablation study of HOIT-T. We analyze loss design and camera-aware embedding in HOIT-T. As Tab. 4 shows, training solely with $\mathcal{L}_{\text{MSE}}$ yields the highest human score but poor object accuracy due to human template overfitting. Introducing cosine similarity loss $\mathcal{L}_{\text{SIM}}$ improves object smoothness and completeness despite a slight human accuracy drop. Combining $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{SIM}}$, and $\mathcal{L}_{\text{KL}}$ yields the best overall performance. Table 5 confirms that removing camera-aware embedding causes a performance drop, indicating that explicit camera geometry is helpful for spatial grounding. Qualitative results in Fig. 6 align with these findings.

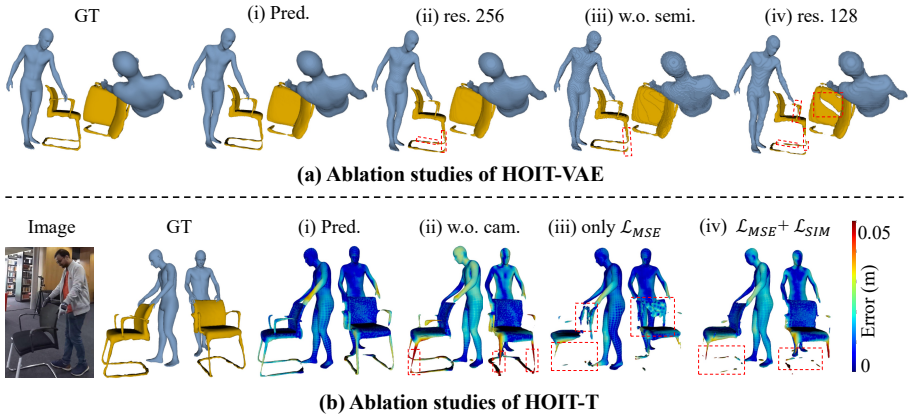

Fig. 6: Qualitative results of ablation studies. (a) HOIT-VAE: (i) corresponds to our full model using 512^3 voxel grids with the semi-continuous occupancy representation; (ii) and (iv) reduce the voxel resolution while keeping the semi-continuous formulation unchanged; (iii) denotes the variant using binary occupancy at 256^3 resolution. **(b) HOIT-T**: (i) Full model with camera embedding and all three loss terms ($\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SIM}} + \mathcal{L}_{\text{KL}}$); (ii) The same configuration as (i) but without camera embedding; (iii-iv) Models with camera embedding retained but trained using different loss strategies: (iii) only $\mathcal{L}_{\text{MSE}}$, and (iv) $\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SIM}}$.

5 Conclusion

In this paper, we present a framework for reconstructing 3D human-object interactions from a single image through a shared triplane feature space. By representing both human and object geometries within a common triplane latent space, our method bridges the structural gap between parametric human models and diverse object shapes. The proposed HOIT-VAE learns a compact, geometry-consistent triplane representation from 3D meshes, and the HOIT-T predicts these triplane features directly from a single image via camera-aware transformers. Furthermore, the pose-fitting method estimates SMPL-H human pose parameters from the reconstructed human mesh. Experiments on the BEHAVE and InterCap datasets show that the proposed hybrid pipeline achieves competitive reconstruction performance, especially on object and combined metrics.

Limitations and discussion. Our HOIT-T has limited generalization to unseen object categories and may fail in heavily occluded cases where large object regions are hidden by the human body. Our HOIT-VAE may fail to capture the fine-grained surfaces of specific objects with high-frequency textures. Failure cases, additional experimental details and discussions are provided in the supplementary material. Nevertheless, this work offers valuable insights into unified 3D HOI representation learning and paves the way toward more complex multi-human, multi-object interaction reconstruction.

References

1. Anthes, C., García-Hernández, R.J., Wiedemann, M., Kranzlmüller, D.: State of the art of virtual reality technology. In: 2016 IEEE aerospace conference. pp. 1–19. IEEE (2016)
2. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15935–15946 (June 2022)
3. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737 (June 2023)
4. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (ECCV). pp. 333–350. Springer (2022)
5. Chen, J., Sun, M., Zheng, Y., Bao, T., He, Z., Li, D., Jin, G., Rui, Z., Wu, L., Jiang, X.: Geo6d: Geometric-constraints-guided direct object 6d pose estimation network. IEEE Transactions on Multimedia (TMM) **27**, 5770–5783 (2025)
6. Chen, J., Li, X., Cao, J., Zhu, Z., Dong, W., Liu, M., Wen, Y., Yu, Y., Zhang, L., Zhang, W.: Rhino: Learning real-time humanoid-human-object interaction from human demonstrations. arXiv preprint arXiv:2502.13134 (2025)
7. Cseke, A., Tripathi, S., Dwivedi, S.K., Lakshmipathy, A.S., Chatterjee, A., Black, M.J., Tzionas, D.: Pico: Reconstructing 3d people in contact with objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1783–1794 (June 2025)
8. Dai, A., Ruizhongtai Qi, C., Niessner, M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5868–5877 (July 2017)
9. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 35799–35813 (2023)
10. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13142–13153 (June 2023)
11. Dwivedi, S.K., Antić, D., Tripathi, S., Taheri, O., Schmid, C., Black, M.J., Tzionas, D.: Interactvlm: 3d interaction reasoning from 2d foundational models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22605–22615 (June 2025)
12. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1323–1333 (June 2024)
13. El Banani, M., Raj, A., Maninis, K.K., Kar, A., Li, Y., Rubinstein, M., Sun, D., Guibas, L., Johnson, J., Jampani, V.: Probing the 3d awareness of visual foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21795–21806 (June 2024)

14. Fiche, G., Leglaive, S., Alameda-Pineda, X., Agudo, A., Moreno-Noguer, F.: Vq-hps: Human pose and shape estimation in a vector-quantized latent space. In: European Conference on Computer Vision (ECCV). pp. 471–490. Springer (2024)
15. Fiche, G., Leglaive, S., Alameda-Pineda, X., Moreno-Noguer, F.: Mega: Masked generative autoencoder for human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5366–5378 (June 2025)
16. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa, A., Malik, J.: Humans in 4d: Reconstructing and tracking humans with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14783–14794 (October 2023)
17. Hai, Y., Song, R., Li, J., Hu, Y.: Shape-constraint recurrent flow for 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4831–4840 (June 2023)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (June 2016)
19. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: Lrm: Large reconstruction model for single image to 3d. In: International Conference on Learning Representations (ICLR). vol. 2024, pp. 50678–50702 (2024)
20. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: Intercap: Joint markerless 3d tracking of humans and objects in interaction. In: DAGM German Conference on Pattern Recognition. pp. 281–299. Springer (2022)
21. Huo, C., Shi, Y., Ma, Y., Xu, L., Yu, J., Wang, J.: Stackflow: monocular human-object reconstruction by stacked normalizing flow with offset. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI). pp. 902–910 (2023)
22. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
23. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11127–11137 (October 2021)
24. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4501–4510 (June 2019)
25. Li, F., Vutukur, S.R., Yu, H., Shugurov, I., Busam, B., Yang, S., Ilic, S.: Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 2123–2133 (October 2023)
26. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1954–1963 (June 2021)
27. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12939–12948 (October 2021)
28. Lin, X., Peng, Y., Wang, L., Zhong, X., Zhu, M., Feng, Y., Yang, J., Liu, C., Chen, Q.: Cleanpose: Category-level object pose estimation via causal learning and knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5990–6000 (October 2025)

29. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 851–866 (2015)
30. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm, p. 347–353. Association for Computing Machinery, New York, NY, USA (1998), https://doi.org/10.1145/280811.281026
31. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
32. Lu, J., Lin, J., Dou, H., Zeng, A., Deng, Y., Liu, X., Cai, Z., Yang, L., Zhang, Y., Wang, H., Liu, Z.: Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9988–9997 (October 2025)
33. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision (ECCV). pp. 23–40. Springer (2024)
34. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5442–5451 (October 2019)
35. Melas-Kyriazi, L., Rupprecht, C., Vedaldi, A.: Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12923–12932 (June 2023)
36. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: Autosdf: Shape priors for 3d completion, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 306–315 (June 2022)
37. Moon, G., Choi, H., Lee, K.M.: 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 (CVPR) Workshops. pp. 2308–2317 (June 2022)
38. Nam, H., Jung, D.S., Moon, G., Lee, K.M.: 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 (CVPR). pp. 10218–10227 (June 2024)
39. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
41. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: European Conference on Computer Vision (ECCV). pp. 163–182. Springer (2024)
42. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10975–10985 (June 2019)

43. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (October 2023)
44. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Object pop-up: Can we infer 3d objects and their poses from human interactions alone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4726–4736 (June 2023)
45. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5523–5535 (October 2025)
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
48. Wang, Z., Zheng, Q., Ma, S., Ye, M., Zhan, Y., Li, D.: End-to-end hoi reconstruction transformer with graph-based encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27706–27715 (June 2025)
49. Wen, B., Huang, D., Zhang, Z., Zhou, J., Deng, J., Gong, J., Chen, Y., Ma, L., Li, Y.L.: Reconstructing in-the-wild open-vocabulary human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17426–17436 (June 2025)
50. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17868–17879 (June 2024)
51. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 334–343 (June 2021)
52. Wu, R., Zhuang, Y., Xu, K., Zhang, H., Chen, B.: Pq-net: A generative part seq2seq network for 3d shapes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 829–838 (June 2020)
53. Wu, S., Lin, Y., Zhang, F., Zeng, Y., Xu, J., Torr, P., Cao, X., Yao, Y.: Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 121859–121881 (2024)
54. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 803–814 (June 2023)
55. Xie, X., Bhatnagar, B.L., Lenssen, J.E., Pons-Moll, G.: Template free reconstruction of human-object interaction with procedural interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10003–10015 (June 2024)
56. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Chore: Contact, human and object reconstruction from a single rgb image. In: European Conference on Computer Vision (ECCV). pp. 125–145. Springer (2022)

57. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Visibility aware human-object interaction tracking from single rgb camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4757–4768 (June 2023)
58. Xie, X., Lenssen, J.E., Pons-Moll, G.: Intertrack: Tracking human object interaction without object templates. In: 2025 International Conference on 3D Vision (3DV). pp. 1427–1439 (2025)
59. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32 (2019)
60. Zhang, B., Tang, J., Niessner, M., Wonka, P.: 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. ACM Transactions On Graphics (TOG) **42**(4), 1–16 (2023)
61. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3d human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV). pp. 34–51. Springer (2020)
62. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: 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). pp. 8834–8845 (June 2023)
63. Zhao, Z., Lai, Z., Lin, Q., Zhao, Y., Liu, H., Yang, S., Feng, Y., Yang, M., Zhang, S., Yang, X., et al.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. arXiv preprint arXiv:2501.12202 (2025)
64. Zhao, Z., Liu, W., Chen, X., Zeng, X., Wang, R., Cheng, P., FU, B., Chen, T., Yu, G., Gao, S.: Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 73969–73982 (2023)
65. Zheng, X.Y., Pan, H., Wang, P.S., Tong, X., Liu, Y., Shum, H.Y.: Locally attentional sdf diffusion for controllable 3d shape generation. ACM Transactions on Graphics (TOG) **42**(4), 1–13 (2023)