This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Human Interaction-Aware 3D Reconstruction from a Single Image

Gwanghyun Kim[1*] Junghun James Kim[2*] Suh Yoon Jeon[1*] Jason Park[1] Se Young Chun[1,2,3†]
[1]Dept. of Electrical and Computer Engineering, [2]IPAI, [3]INMC
Seoul National University, Republic of Korea
{gwang.kim, jonghean12, euniejeon, jsp11235, sychun}@snu.ac.kr

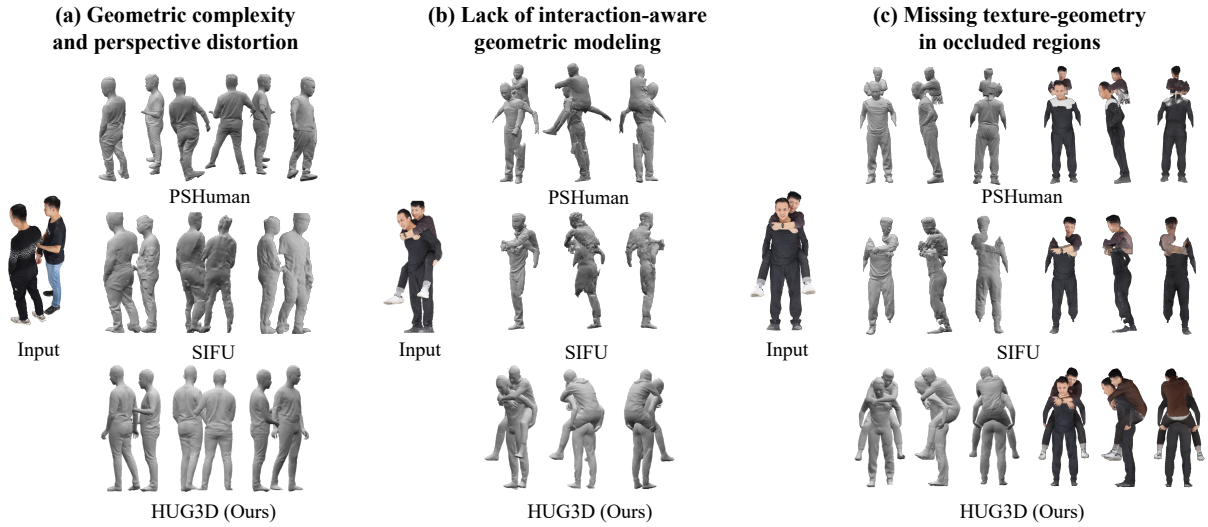


Figure 1. Core challenges in monocular multi-human 3D reconstruction: (a) geometric complexity and perspective distortion, (b) lack of interaction-aware geometric modeling, and (c) missing texture and geometry in occluded regions. HUG3D addresses all three challenges.

Abstract

Reconstructing textured 3D human models from a single image is fundamental for AR/VR and digital human applications. However, existing methods mostly focus on single individuals and thus fail in multi-human scenes, where naive composition of individual reconstructions often leads to artifacts such as unrealistic overlaps, missing geometry in occluded regions, and distorted interactions. These limitations highlight the need for approaches that incorporate group-level context and interaction priors. We introduce a holistic method that explicitly models both group- and instance-level information. To mitigate perspective-induced geometric distortions, we first transform the input into a canonical orthographic space. Our primary component, Human Group-Instance Multi-View Diffusion (HUG-MVD), then generates complete multi-view normals and images by jointly modeling individuals and group context to resolve occlusions and proximity. Subsequently, the Human Group-Instance Geometric Reconstruction (HUG-GR) module optimizes the geometry by leveraging explicit, physics-based interaction priors to enforce physical plausibility and accurately model inter-human contact. Finally, the multi-view images are fused into a high-fidelity texture. Together, these components form our complete framework, HUG3D. Extensive experiments show that HUG3D significantly outperforms both single-human and existing multi-human methods, producing physically plausible, high-fidelity 3D reconstructions of interacting people from a single image. Project page: jongheean11.github.io/HUG3D_project

1. Introduction

Reconstructing detailed 3D human models from visual input [10, 11, 15, 22, 31, 32, 35, 40, 41] is a fundamental task in computer vision, supporting applications in augmented and virtual reality (AR/VR) [24, 27], digital humans, and social behavior understanding. While monocular 3D reconstruction from a single RGB image [10, 22, 40] has made substantial progress, most existing methods are limited to isolated individuals in controlled environments. However, these approaches often fail to generalize to real-world scenes involving multiple interacting people, where occlusion, perspective distortion, and spatial entanglement

*Authors contributed equally. †Corresponding author.

introduce significant ambiguity and modeling challenges.

In particular, we identify three core challenges in monocular multi-human 3D reconstruction: **(1) Geometric complexity and perspective distortion.** Multi-human scenes often involve substantial depth variation and complex spatial layouts, which can lead to perspective distortions. While most methods assume orthographic views leading to distortions on real-world perspective inputs (Fig. 1(a)), a few perspective-aware models [21, 34] remain limited to single-object cases and struggle with multi-human complexity. The scarcity of annotated multi-human data [37, 41] further hinders generalization across camera poses and interaction patterns. **(2) Lack of interaction-aware geometric modeling.** Most methods reconstruct individuals independently, overlooking contextual cues like contact, occlusion, and spatial proximity (Fig. 1(b)). This often results in unrealistic outputs such as overlapping limbs or unnatural distances. Although group-wise SMPL-X approaches [1, 25, 33] offer early signs of interaction modeling, full-surface, textured reconstruction remains underexplored. **(3) Missing geometry and texture in occluded regions.** Occlusions between people obscure critical body parts, causing incomplete geometry and textures (Fig. 1(c)). While generative inpainting [18, 30] can hallucinate plausible content, only a few approaches [2, 4] jointly address geometry and appearance, and even fewer try handling multi-view consistency under occlusion.

To address these challenges, we propose a holistic method for human interaction-aware 3D reconstruction from a single image. Our method incorporates both group- and instance-level information with three main components: (1) *Canonical Perspective-to-Orthographic View Transform (Pers2Ortho).* To make multi-view diffusion tractable under severe geometric distortion, we transform the input perspective image into a canonical orthographic space. From the image, we estimate a partial 3D textured geometry and reproject it to obtain consistent multi-view RGB and normal representations. (2) *Human Group-Instance Multi-View Diffusion (HUG-MVD).* This diffusion model jointly completes missing geometry and textures while serving as an implicit prior that enforces physically plausible human interactions. (3) *Textured Mesh Reconstruction.* Our *Human Group-Instance Geometry Reconstruction (HUG-GR)* module first optimizes the mesh using group-level, instance-level, and physics-based supervision. Subsequently, the multi-view images are fused into a final, high-fidelity texture using occlusion-aware blending. Collectively, these components form our holistic framework, **HUG3D**.

Experiments show that HUG3D outperforms baselines, producing physically plausible, high-fidelity textured reconstructions of interacting people from a single image.

2. Related Works

Single-Human 3D Reconstruction. Significant advances in single-human 3D reconstruction have been driven by parametric models such as SMPL [23, 28]. Early methods fit models to 2D cues [3], while later works regress parameters directly from images using deep networks [13]. Implicit and volumetric approaches [7, 26, 31, 32] further improve geometric detail. Recent efforts enhance view consistency and texture. SIFU [40] optimizes UVs via SDF, and SiTH [10] uses back-view synthesis. PSHuman [22] generates multi-view RGBs via diffusion, and LHM [29] uses transformers over image and SMPL-X tokens. Though effective for isolated individuals, these methods struggle in multi-human settings. Naïve application per instance leads to artifacts like overlapping meshes and inconsistent scale, underscoring the need for interaction-aware models.

Multi-Human 3D Reconstruction. Reconstructing multiple humans in 3D remains challenging due to occlusions, inter-person interactions, and depth ambiguities. Early approaches reconstructed each individual independently [20, 33, 41], but this often led to physically implausible results. Later methods improved spatial coherence by introducing global constraints [8] or jointly regressing shapes and poses in a shared coordinate system [26]. Multi-view and video-based approaches [11, 17, 19] further enhanced reconstruction accuracy, though they require video or multi-view inputs, which can be costly to acquire and process. More recently, learning-based methods have sought to tackle multi-human reconstruction from a single image. While interaction-aware networks [5–7] and model-free architectures [26] improve plausibility, their performance remains limited in single-view, single-frame scenarios.

Scene-Level Human Reconstruction and Interaction Modeling. Several methods incorporate human-scene or human-human interaction to improve physical plausibility. POSA [8] enforces realistic body-ground contact but focuses on single-person scenarios. Group-level priors [25] enhance pose coherence, while BUDDI [25] learns a diffusion-based prior for plausible two-person interactions. Other approaches rely on explicit contact labels [12], but typically focus on coarse geometry or pose estimation. However, these methods often yield results with low texture fidelity or limited to SMPL-X mesh predictions, falling short of producing fully textured, detailed 3D reconstructions of multi-human interacting.

3. Human Interaction-Aware 3D Reconstruction from a Single Image

Our HUG3D framework operates in three main stages, as illustrated in Fig. 2. First, the Canonical Perspective-to-

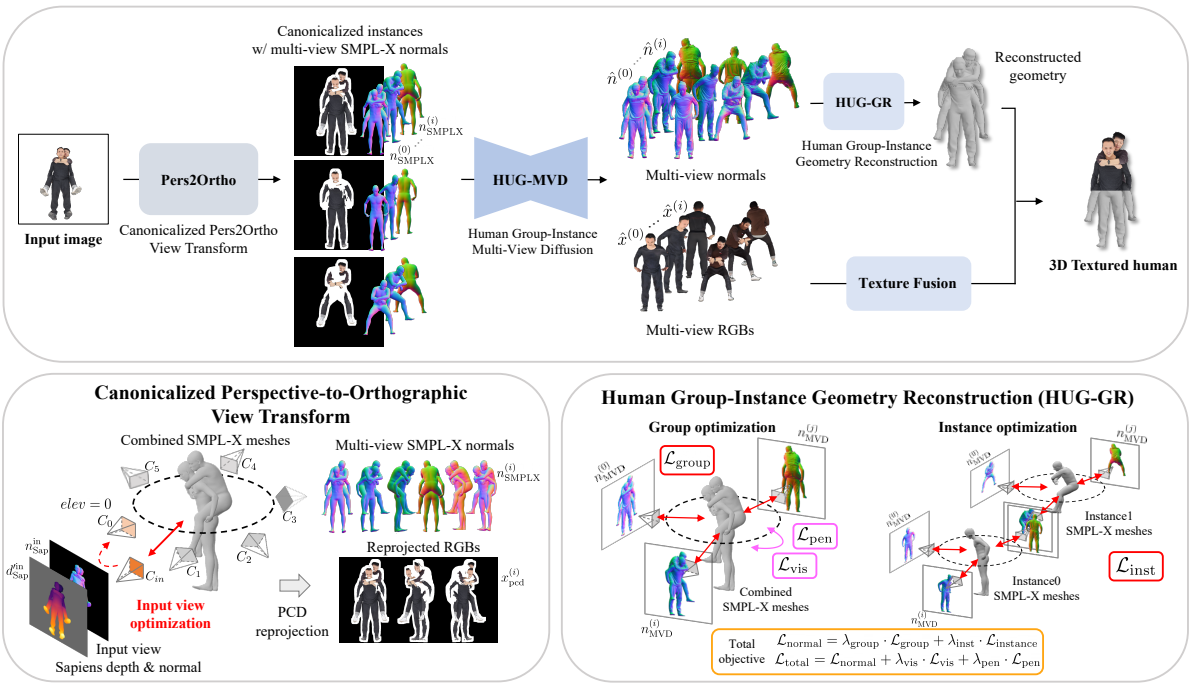
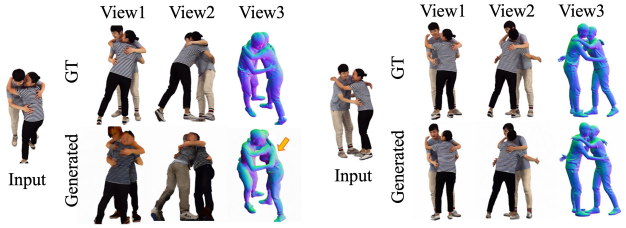


Figure 2. Overview of our HUG3D framework. Given a single perspective image, (1) the *Canonical Perspective-to-Orthographic View Transform (Pers2Ortho)* converts it into a canonical multi-view orthographic representation to resolve scale ambiguity and enable consistent multi-view reasoning. (2) The *Human Group-Instance Multi-View Diffusion (HUG-MVD)* model completes occluded geometry and texture while maintaining plausible interactions. (3) The *Textured Mesh Reconstruction* stage refines the mesh and generates high-fidelity textures with our physics-based *Human Group-Instance Geometry Reconstruction (HUG-GR)*, which enforces physical consistency via optimization with multi-view normal cues and interaction constraints.



(a) Trained on **perspective** images (b) Trained on **orthographic** images

Figure 3. Comparison of results from multi-view diffusion trained on perspective vs. orthographic images.

Orthographic View Transform (*Pers2Ortho*) module converts the input perspective image into a consistent multi-view orthographic representation. Next, our Human Group-Instance Multi-View Diffusion (*HUG-MVD*) model completes occluded geometry and texture while ensuring plausible interactions. Finally, the Textured Mesh Reconstruction stage refines the mesh with our physics-based *HUG-GR* module and synthesizes a high-fidelity texture.

3.1. Canonical Perspective-to-Orthographic View Transform

Multi-human scenes exhibit severe depth variation and occlusion, which violate the orthographic assumptions typically adopted in multi-view diffusion models. As a result, directly learning interactions from a single perspective-view image is challenging due to extreme geometric complexity (see Fig. 3). To overcome this limitation, we introduce Pers2Ortho module, a canonical view transformation module that estimates partial 3D geometry from the input perspective view and constructs consistent multi-view representations robust to geometric distortion.

Initial Geometry, Segmentation, and Camera Setup. Given a single RGB input, we first estimate an initial SMPL-X mesh, instance segmentation masks, and the perspective camera parameters $(\mathbf{K}_{\text{in}}, \mathbf{R}_{\text{in}}, \mathbf{T}_{\text{in}})$ of $\mathcal{C}_{\text{in}}$. We employ RoBUDDI, our more robust variant of BUDDI [25], as detailed in supplementary Sec. A.1. To define a canonical processing space, the SMPL-X mesh is first normalized to a tight bounding box. Around this normalized geometry, we place six orthographic cameras $\{\mathcal{C}_0, \ldots, \mathcal{C}_5\}$ at fixed azimuths $\{0°, 45°, 90°, 180°, 270°, 315°\}$ with zero elevation. Each camera has extrinsic parameters $(\mathbf{R}_i, \mathbf{T}_i)$. This canonical camera rig ensures spatial alignment across all instances, providing a stable basis for downstream multi-view generation.

Partial 3D Construction. To enhance the fidelity of the canonical representation, we initialize a partial 3D mesh $\mathcal{M}$ with the initial SMPL-X mesh and refine it using geometric supervision from Sapiens [14]. Specifically, Sapiens predicts affine-invariant depth ($d'^{\text{in}}_{\text{Sap}}$) and surface normals ($n^{\text{in}}_{\text{Sap}}$)

from the perspective input view $\mathcal{C}_{\text{in}}$. The mesh vertices are optimized to align with these predictions by minimizing the geometry loss:

$$\mathcal{L}_{\text{geo}} = \left\| d'^{\text{in}}_{\text{Sap}} - d'^{\text{in}}_{\mathcal{M}} \right\|_2^2 + 1 - \left\langle n^{\text{in}}_{\text{Sap}}, n^{\text{in}}_{\mathcal{M}} \right\rangle, \tag{1}$$

where $d'^{\text{in}}_{\mathcal{M}}$ and $n^{\text{in}}_{\mathcal{M}}$ denote the depth and normal maps rendered from the mesh. The first term enforces depth consistency via L2 distance, while the second term enforces orientation consistency via cosine similarity. To handle possible topology mismatches between the initial mesh and Sapiens predictions, we adopt a remeshing strategy [22].

Multi-View Input Generation via PCD Reprojection. For the multi-view diffusion stage, we first render a complete set of normal maps $\{n^{(i)}_{\text{SMPLX}}\}_{i=0}^{5}$ from all six canonical views $\{\mathcal{C}_i\}$ from the initial SMPL-X mesh for the pose guidance.

Second, we generate partial RGB inputs $\{x^{(0)}_{\text{pcd}}, x^{(1)}_{\text{pcd}}, x^{(5)}_{\text{pcd}}\}$ by reprojecting the input image onto the refined partial 3D mesh $\mathcal{M}$ from three key canonical views (0°, 45°, and 315°). Concretely, we render a depth map $d'^{\text{in}}_{\mathcal{M}}$ from $\mathcal{M}$ and build a dense point cloud (PCD) $\mathcal{P}$ in the coordinate system of $\mathcal{C}_{\text{in}}$. This PCD is reprojected into each orthographic view $\mathcal{C}_i$ as:

$$\mathcal{P}_i = \Pi^i_{\text{ortho}}(\mathcal{P}) = \mathbf{R}_i \cdot \mathcal{P} + \mathbf{T}_i, \tag{2}$$

where Π^i_{ortho} denotes the orthographic projection. RGB values from $\mathcal{C}_{\text{in}}$ are then transferred onto $\mathcal{P}_i$, yielding partial RGB maps $x^{(i)}_{\text{pcd}}$. Unlike mesh vertex coloring, which often produces sparse and low-quality textures, our PCD reprojection preserves dense appearance details in visible regions while maintaining spatial consistency across canonical views. These inputs serve as robust conditioning signals for multi-view diffusion. Further details can be found in Sec. A.2 of the supplementary.

3.2. Human Group-Instance Multi-View Diffusion

We introduce Human Group-Instance Multi-View Diffusion (HUG-MVD), a diffusion model that leverages both group-level and instance-level priors to resolve occlusions and inter-person interactions in multi-human scenes.

Interaction- and Occlusion-Aware Multi-View Diffusion. Given reprojected partial RGB inputs and canonical-view SMPL-X normal maps, HUG-MVD reconstructs geometry and appearance missing due to occlusion. Unlike standard single-human diffusion models, our formulation incorporates group-level priors by jointly training on two complementary sources. First, diverse single-human datasets with full supervision [9, 38] provide coverage over a wide range of body shapes and identities. Second, multi-human datasets with partial geometry [37] capture realistic inter-person occlusions and interactions, though with limited identity variation. This combination enables the model to generalize across diverse individuals while remaining robust to complex group-level occlusions.

To simulate realistic occlusions for single-human training data, we generate masked inputs $x^{(i)}_{\text{mask}}$ from the reprojected point clouds $x^{(i)}_{\text{pcd}}$ and visibility masks M_{vis}, as described in Sec. A.3.3. These masks indicate unobserved regions for each canonical view $\mathcal{C}_i$. The model is conditioned on SMPL-X normal maps $\{n^{(i)}_{\text{SMPLX}}\}_{i=0}^{5}$, which provide geometric guidance via ControlNet [39], steering the denoising process toward consistent, human-like reconstructions.

Multimodal Training Objective. Given 6 canonical views, the model predicts complete RGB images $\{\hat{x}^{(i)}\}$ and surface normals $\{\hat{n}^{(i)}\}$ simultaneously using a shared denoising diffusion process. To train the model, we minimize the following objective $\mathcal{L}_{\text{diff}}$ as below:

$$\begin{aligned}\sum_{i=0}^{5} \Bigg(& \mathbb{E}_{t,\epsilon}\left[\left\| \epsilon - \epsilon_\theta(z^{(i)}_{t,\text{rgb}}, t, x^{(i)}_{\text{mask}}, n^{(i)}_{\text{SMPLX}}) \right\|_2^2 \right] \\ & + \mathbb{E}_{t,\epsilon}\left[\left\| \epsilon - \epsilon_\theta(z^{(i)}_{t,\text{normal}}, t, x^{(i)}_{\text{mask}}, n^{(i)}_{\text{SMPLX}}) \right\|_2^2 \right] \Bigg),\end{aligned} \tag{3}$$

where $z^{(i)}_{t,\text{rgb}}$ and $z^{(i)}_{t,\text{normal}}$ are the noisy latents for the RGB and normal map at timestep t, respectively. The denoising network ϵ_θ learns to estimate the clean signals for both modalities, conditioned on the masked RGB inputs $x^{(i)}_{\text{mask}}$ and SMPL-X normal maps $n^{(i)}_{\text{SMPLX}}$. By jointly optimizing RGB and normal prediction, the model enforces consistency between appearance and geometry.

Joint Group-Instance Inference. At inference, a single HUG-MVD model jointly performs group- and instance-level reconstruction, predicting complete normals $\{\hat{n}^{(i)}\}$ and RGB images $\{\hat{x}^{(i)}\}$ from partial RGB inputs and SMPL-X guidance.

To maintain consistency between group- and instance-level latent representations, instance-specific latents $z^{(i)}_{t,\text{inst}(k)}$ are injected into the group latent $z^{(i)}_{t,\text{group}}$ at their corresponding spatial regions. Formally, at diffusion timestep t:

$$\begin{aligned} z^{(i)}_{t,\text{group}} \leftarrow & \sum_{k=1}^{K} \Big[\alpha \cdot \mathbb{1}^{(i,k)}_{\text{inst}} \cdot z^{(i)}_{t,\text{inst}(k)} + (1-\alpha) \cdot \mathbb{1}^{(i,k)}_{\text{inst}} \\ & \cdot z^{(i)}_{t,\text{group}} \Big] + \left(1 - \sum_{k=1}^{K} \mathbb{1}^{(i,k)}_{\text{inst}} \right) \cdot z^{(i)}_{t,\text{group}}, \end{aligned} \tag{4}$$

where $\mathbb{1}^{(i,k)}_{\text{inst}}$ indicates the spatial extent of instance k in view $\mathcal{C}_i$, and $\alpha \in [0,1]$ controls blending strength. This composition allows fine-grained instance details to inform the global group representation, improving geometric consistency across overlapping surfaces.

Further details with detailed illustrations of HUG-MVD are provided in Sec. A.3 of the supplementary.

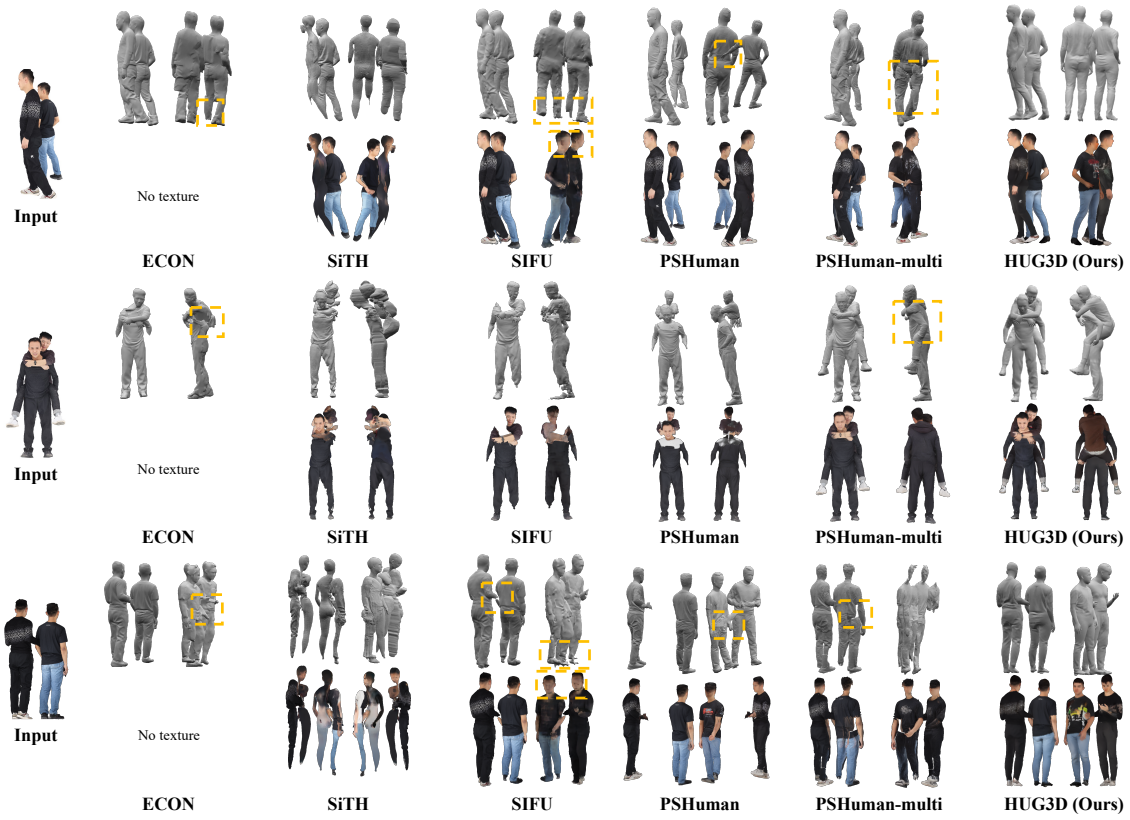


Figure 4. Qualitative comparison on multi-human 3D reconstruction from a single image. HUG3D outperforms baselines by correcting perspective distortion, preserving inter-human contact, and hallucinating plausible textures under heavy occlusion.

3.3. Textured Mesh Reconstruction

3.3.1. Human Group-Instance Geometry Reconstruction

We refine the initial SMPL-X mesh to produce physically plausible geometry for multi-human scenes, leveraging both group- and instance-level supervision and physics-inspired constraints.

Group-Instance Normal Supervision. The mesh $\mathcal{M}$ is optimized to match predicted normals from HUG-MVD at both group and instance levels. Let $n^{(i)}_{\text{MVD,group}}$ and $n^{(i)}_{\mathcal{M},\text{group}}$ denote predicted and rendered normals for the group, and $n^{(i,k)}_{\text{MVD,inst}}$ and $n^{(i,k)}_{\mathcal{M},\text{inst}}$ for instance k. We define:

$$\mathcal{L}_{\text{normal}} = \lambda_{\text{group}} \cdot \mathcal{L}_{\text{group}} + \lambda_{\text{inst}} \cdot \mathcal{L}_{\text{instance}}, \tag{5}$$

$$\begin{aligned} \mathcal{L}_{\text{group}} &= \sum_{i=0}^{5} \left(1 - \langle n^{(i)}_{\text{MVD,group}}, n^{(i)}_{\mathcal{M},\text{group}} \rangle\right), \\ \mathcal{L}_{\text{instance}} &= \sum_{i=0}^{5} \sum_{k=1}^{K} \left(1 - \langle n^{(i,k)}_{\text{MVD,inst}}, n^{(i,k)}_{\mathcal{M},\text{inst}} \rangle\right), \end{aligned} \tag{6}$$

where $n_{\mathcal{M}}$ is rendered from the currently optimized mesh via a differentiable renderer at each step.

Interpenetration Loss. To prevent implausible overlaps between body parts, we define an interpenetration loss over tolerance pairs $(i, j) \in \text{V}$, where V denotes the set of part pairs derived from contact regions in the initial SMPL-X meshes. For each pair, let $s_1^{i,j}$ and $s_2^{i,j}$ be the closest points on the surfaces of parts i and j. The loss penalizes distances below a threshold tol, with ξ acting as a smoothing temperature that controls the softness of the penalty:

$$\mathcal{L}_{\text{pen}} = \text{mean}_V \left[\xi \, \ln\!\left(1 + e^{(\text{tol} - |s_1^{i,j} - s_2^{i,j}|)/\xi}\right) \right]. \tag{7}$$

Visibility Loss. We enforce consistency between rendered visibility and ground-truth masks:

$$\mathcal{L}_{\text{vis}} = \frac{1}{2B} \sum_{k=1}^{K} \sum_{b=1}^{B} \frac{E_b^k}{M_b^k + \epsilon}, \tag{8}$$

where E_b^k and M_b^k denote incorrectly occluded and total visible pixels for body part b in instance k. This loss encourages each body part in the rendered mesh to match its expected visible region, ensuring accurate silhouettes and occlusion boundaries, especially in complex multi-human interactions.

Total Optimization Objective. The final mesh is optimized using:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{normal}} + \lambda_{\text{vis}} \cdot \mathcal{L}_{\text{vis}} + \lambda_{\text{pen}} \cdot \mathcal{L}_{\text{pen}}. \tag{9}$$

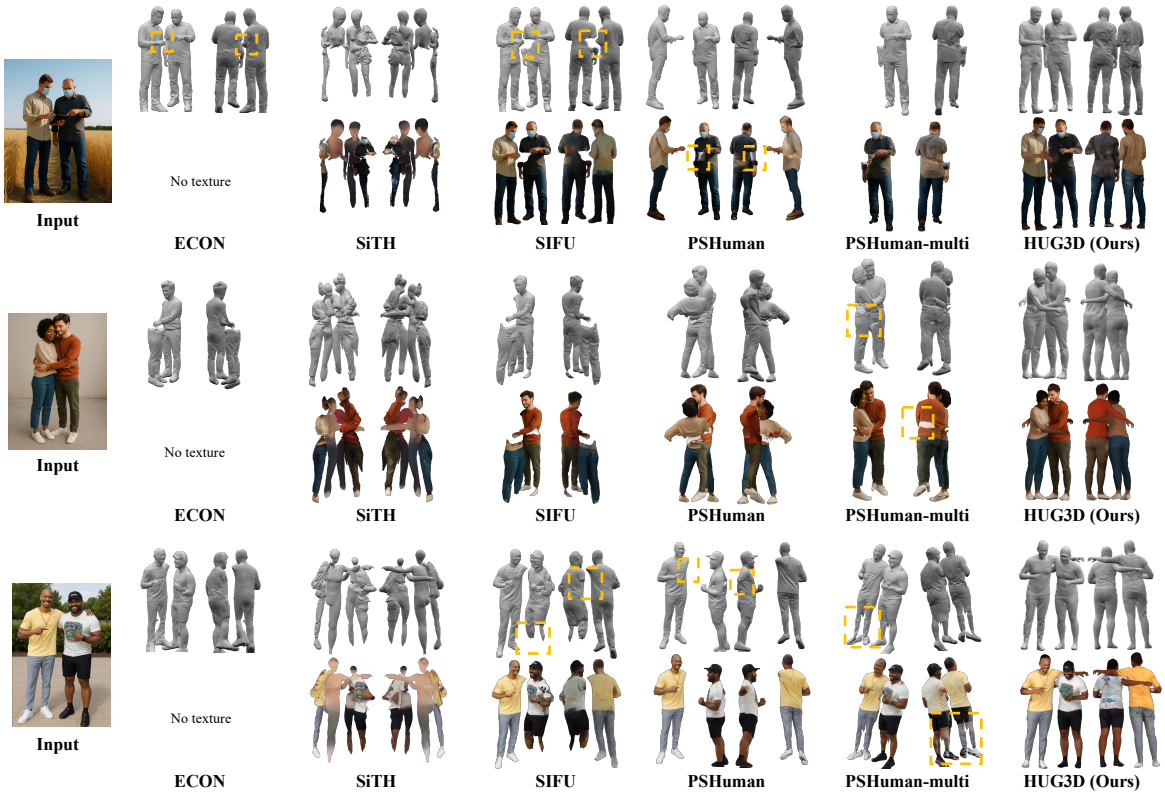


Figure 5. Qualitative results on in-the-wild images. Our method demonstrates robust multi-human reconstructions, outperforming baseline approaches and highlighting its practical applicability.

We apply finer learning rates to high-frequency semantic regions (e.g., hands, face) to improve local accuracy while maintaining overall shape stability.

Additional details for HUG-GR are provided in Sec. A.4 of the supplementary.

3.3.2. Texture Construction

Full-body vertex texture is generated by projecting multi-view RGBs onto the optimized mesh. To improve fidelity, occluded and low-confidence regions are blended using view-aware confidence masks. High-fidelity face restoration is applied for oblique or occluded views. Please refer to Sec. A.5 of the supplementary for additional details.

4. Experiments

4.1. Experimental Setup

Implementation Details. The *Pers2Ortho* module is based on BUDDI [25] for SMPL-X fitting and camera parameter estimation. For partial 3D construction, the initial SMPL-X mesh is optimized for 200 iterations using Adam [16] with a learning rate of 0.02. *HUG-MVD* is initialized from PSHuman [22], based on Stable Diffusion 2.1 [30], and integrates a frozen normal-map ControlNet [39]. Training is performed on a single NVIDIA A100 (80GB) GPU with a batch size of 16 and gradient accumulation of 8 steps, using Adam (lr=5×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$). A two-stage curriculum is employed: (1) 1,000 steps without inpainting masks, and (2) 1,000 steps with inpainting masks to simulate occlusion. Training takes approximately two days. A DDPM scheduler with 1,000 diffusion steps is used during training, while inference uses a DDIM scheduler ($\eta = 1.0$) with 40 denoising steps. A blending factor of $\alpha = 0.8$ balances diversity and fidelity. *HUG-GR* optimizes the mesh over 200 iterations with a learning rate of 0.01 and loss weights: $\lambda_{\text{group}} = 1.0$, $\lambda_{\text{inst}} = 0.2$, $\lambda_{\text{pen}} = 2.0$, and $\lambda_{\text{vis}} = 1.0$. Further implementation details are provided in Sec. A of the supplementary.

Training Data. Training samples of HUG-MVD are rendered from raw scans which are composed of textured mesh sequences. We train our model using the Hi4D dataset [37] for multi-human supervision, along with THuman2.0 [38] and CustomHumans [9] for diverse single-human poses and appearances, enabling robust learning across varied interactions and body configurations. Additional details are provided in Sec. A.3.1 of the supplementary.

Evaluation Dataset. We conduct experiments on the MultiHuman [41] dataset, which provides multi-person, multi-view sequences with full 3D mesh supervision. For quan-

Table 1. Quantitative comparison of geometric metrics for multi-human 3D reconstruction. HUG3D achieves the best overall scores in all metrics including CD, P2S, and NC, and also outperforms other baselines in CP, indicating better interaction-aware reconstruction.

Method	CD ↓	P2S ↓	NC ↑	F-score ↑	bbox-IoU ↑	Norm $L2$ ↓	CP ↑
SIFU	5.644	2.284	0.754	29.244	0.778	0.028	0.089
SiTH	9.251	3.185	0.709	21.037	0.708	0.040	0.135
PSHuman	15.579	6.088	0.617	9.749	0.659	0.069	0.027
DeepMultiCap	13.719	2.555	0.749	18.125	0.513	0.049	0.083
Ours	**3.631**	**1.752**	**0.811**	**41.504**	**0.847**	**0.019**	**0.240**

Table 2. Quantitative evaluation on texture quality of multi-human 3D reconstruction.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
SIFU	15.202	0.793	0.202
SiTH	13.798	0.789	0.233
PSHuman	11.418	0.742	0.304
Ours	**16.456**	**0.809**	**0.168**

Table 3. Quantitive evaluation on geometry and texture quality within occluded region.

Method	Norm $L2$ ↓	PSNR ↑	SSIM ↑
SIFU	0.197	6.157	0.559
SiTH	0.197	6.355	0.539
PSHuman	0.252	4.621	0.510
DeepMultiCap	0.217	-	-
Ours	**0.140**	**8.388**	**0.602**

Table 4. Ablation study of key components in HUG3D. We report geometry (CD, P2S, Norm $L2$) and texture (PSNR, LPIPS) metrics under various configurations.

Module	Method	CD ↓	P2S ↓	Occ.Norm $L2$ ↓	PSNR ↑	LPIPS ↓	Occ.PSNR ↑
HUG-MVD	Trained on group-only data	4.564	2.203	0.157	15.641	0.191	7.423
	Trained on instance-only data	4.645	2.245	0.156	15.840	0.185	7.726
	w/o Instance-to-group latent composition	4.646	2.249	0.159	16.198	0.183	7.916
HUG-GR	Instance-only normal supervision	4.642	2.250	0.156	16.180	0.183	7.902
	Group-only normal supervision	4.620	2.230	0.159	16.169	0.183	7.678
	Ours (Full)	**4.316**	**2.122**	**0.153**	**16.454**	**0.179**	**8.082**

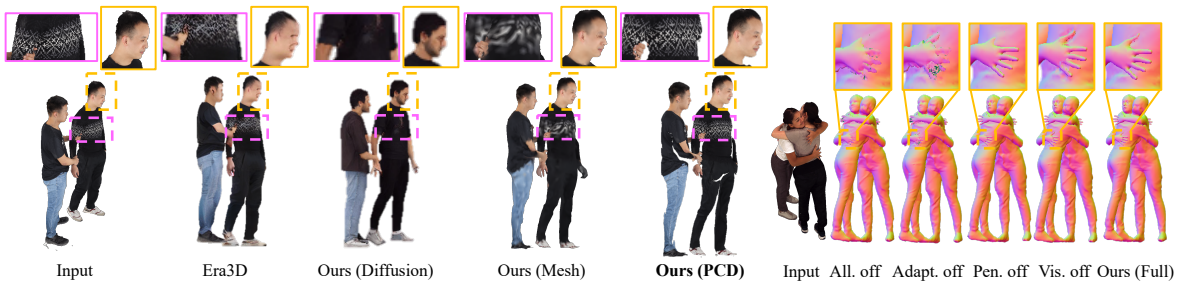
titative evaluation, we focused on two-person interacting scenes, including 20 sequences - six of these feature closely interacting pairs with heavy occlusions and complex spatial entanglements, while the remaining sequences capture more natural interactions. For each scene, we select a random initial viewpoint and render four perspective views, to evaluate under natural camera distortion, at azimuth offsets of $\{0°, 90°, 180°, 270°\}$, yielding 80 images in total. See Sec. B.2 of the supplementary for the details.

Baselines. To the best of our knowledge, we are the *first* to tackle *multi-human 3D reconstruction with both geometry and texture*, and no existing public baselines directly address this task (see Tab. S5 in supplementary). We therefore compare our method against two categories of prior works: (i) single-human reconstruction from a single image, and (ii) multi-human reconstruction from multi-view images or videos. To ensure a fair comparison, we follow the evaluation protocol of [5]. For single-human methods—ECON [36], SIFU [40], SiTH [10], and PSHuman [22]—we crop each person using the dataset's ground-truth instance masks, reconstruct them independently, and then align the results into a shared coordinate frame and composite to form the complete scene. We also report PSHuman's performance when applied directly to uncropped multi-person images. For multi-human methods, we evaluate DeepMultiCap [41], designed for multi-view images, and Multiply [11], designed for videos, under single image setting. All SMPL-based methods use ground-truth SMPL-X poses to isolate reconstruction quality from pose estimation errors. Further details are provided in supplementary Sec. B.1 and results on multi-human baselines can be found in Sec. C.1 of the supplementary material.

Evaluation Metrics. We evaluate geometry with L_1 Chamfer distance (CD) [cm] and 1-directional point-to-surface distance (P2S) [cm], normal consistency (NC), F-score, and bbox-IoU between reconstructed and ground-truth meshes. To assess surface detail consistency, we compute L_2 normal error between predicted and ground-truth normal renders - across four rotated views at $\{0°, 90°, 180°, 270°\}$ relative to the input view. This is also separately computed specifically within occluded regions to evaluate robustness under occlusions. Physical realism is quantified through contact precision score (CP) defined by the overlap between the estimated and ground-truth inter-body contact map. Texture fidelity is assessed using PSNR, SSIM, and LPIPS, computed across same views as L_2 normal error. Also separately computed specifically within occluded regions. More details are provided in Sec. B.3 of the supplementary.

4.2. Results

Quantitative Results. As shown in Tab. 1, our method outperforms on all geometric metrics, with the lowest CD, P2S



(a) Ablation on Pers2Ortho **(b) Ablation on HUG-GR optimization loss**

Figure 6. Ablation study on key components. (a) Pers2Ortho improves projection sharpness.(b) Each HUG-GR loss boosts geometric plausibility.

and highest NC. CP is markedly higher than other baselines, indicating superior physical realism and fidelity of inter-instance contacts. Tab. 2 shows our method achieved the highest PSNR and SSIM scores along with the lowest LPIPS. Moreover, in the occluded-region evaluation, Tab. 3, we significantly outperform the baselines in both Normal $L2$ error for geometry and PSNR/SSIM for texture. This indicates better perceptual quality of hallucinated textures where baselines typically fall.

Qualitative Results. Fig. 4 illustrates that our approach surpasses all baselines in multi-human 3D reconstruction from a single image. Pers2Ortho enables accurate shape recovery in challenging viewpoints such as elevated scenes, while baseline models produce distorted results. Through HUG-MVD and HUG-GR, we model inter-human interactions, producing meshes that faithfully align with contact regions in the input image. In contrast, baselines either suffer from interpenetration or fail to preserve contact. Visual inspections confirm that HUG3D successfully reconstructs complete human meshes even under substantial occlusion. Hallucinated textures plausibly infer clothing and shape details not visible in the input. Competing methods, by contrast, exhibit broken surfaces, floating limbs, and missing textures. Fig. 5 shows results on in-the-wild images. Our approach consistently reconstructs multiple humans with high fidelity, surpassing baseline methods and demonstrating real-world applicability.

Additional results can be found in Secs. C and D of the supplementary.

4.3. Ablation Study

Pers2Ortho. As shown in Fig. 6(a), we evaluate our PCD-based Pers2Ortho against several baselines: the generative method Era3D [21], and our custom diffusion and mesh reprojection variants. These competing methods often produce blurry results with lost facial details, whereas our strategy consistently preserves high-fidelity features from the original view.

HUG-MVD. Tab. 4 shows that training HUG-MVD on group-only data [37] suffers from limited geometry and texture diversity, while instance-only data [9, 38] fails to capture inter-person interactions. Our full setting, combining both sources, significantly outperforms either alone, highlighting their complementary benefits. Also, disabling instance-to-group latent composition leads to inconsistent surfaces and breaks continuity across instances.

HUG-GR. We analyze the impact of each component in HUG-GR. As shown in Fig. 6(b), omitting adaptive-region specific optimization degrades quality in high-frequency areas, such as hands. Removing interpenetration loss leads to mesh interpenetration and unrealistic overlaps, demonstrating its importance for physically grounded reconstruction. Excluding the visibility loss impairs alignment and produces unnatural surface at contact regions. In Tab. 4, by comparing variants using only group-level or only instance-level losses, we found that incorporating both losses simultaneously yields the best performance.

More ablations and analysis can be found in Sec. E of the supplementary.

5. Discussion

Limitations. HUG3D focuses on inter-human occlusion and does not handle occlusions from external objects, which we plan to address in future work. Although robust in the wild, our method can fail under severe depth ambiguity or heavy occlusion due to the inherent limitations of the single-image setting. Further discussion appears in Sec. G of the supplementary.

Conclusion. We presented HUG3D, a three-stage framework for high-fidelity 3D reconstruction of interacting people from a single RGB image. HUG3D addresses key challenges such as occlusions, complex interactions, and geometric distortions by combining a canonical view transform (Pers2Ortho), a group-instance multi-view diffusion model (HUG-MVD), and a physics-based reconstruction stage (HUG-GR) for refined geometry and texture. Extensive experiments show that HUG3D significantly outperforms single-human and prior multi-human methods, producing visually accurate, physically plausible reconstructions. Our framework enables reliable applications in AR/VR, telepresence, and digital human modeling, advancing realistic multi-human 3D reconstruction from a single image.

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea(NRF) grants funded by the Korea government(MSIT) (Nos. RS-2022-NR067592, RS-2025-02263628), the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University and AI-Bio Research Grant through Seoul National University.

References

[1] 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. *ECCV*, 2024. 2

[2] Amir Barda, Matheus Gadelha, Vladimir G Kim, Noam Aigerman, Amit H Bermano, and Thibault Groueix. Instant3dit: Multiview inpainting for fast editing of 3d objects. *arXiv preprint arXiv:2412.00518*, 2024. 2

[3] 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. *ECCV*, 2016. 2

[4] Chenjie Cao, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Mvinpainter: Learning multi-view consistent inpainting to bridge 2d and 3d editing. *arXiv preprint arXiv:2408.08000*, 2024. 2

[5] Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jaeshin Yoon, and Seungryul Baek. 3d reconstruction of interacting multi-person in clothing from a single image. *WACV*, 2024. 2, 7

[6] Kennard Chan, Fayao Liu, Guosheng Lin, Chuan-Sheng Foo, and Weisi Lin. Robust-pifu: Robust pixel-aligned implicit function for 3d human digitalization from a single image. *ICLR*, 2024.

[7] Mihai Fieraru, Mihai Zanfir, Teodor Szente, Eduard Bazavan, Vlad Olaru, and Cristian Sminchisescu. Remips: Physically consistent 3d reconstruction of multiple interacting people under weak supervision. *NeurIPS*, 2021. 2

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

[9] Hsuan-I Ho, Lixin Xue, Jie Song, and Otmar Hilliges. Learning locally editable virtual humans. *CVPR*, 2023. 4, 6, 8

[10] I Ho, Jie Song, Otmar Hilliges, et al. Sith: Single-view textured human reconstruction with image-conditioned diffusion. *CVPR*, 2024. 1, 2, 7

[11] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. *CVPR*, 2024. 1, 2, 7

[12] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CVPR*, 2018. 2

[13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. *CVPR*, 2018. 2

[14] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. *ECCV*, 2024. 3

[15] Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, and Hanbyul Joo. Chupa: Carving 3d clothed humans from skinned shape priors using 2d diffusion probabilistic models. *ICCV*, 2023. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014. 6

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

[18] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2

[19] Inhee Lee, Byungjun Kim, and Hanbyul Joo. Guess the unseen: Dynamic 3d scene reconstruction from partial 2d glimpses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1062–1071, 2024. 2

[20] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *CVPR*, 2019. 2

[21] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, et al. Era3d: high-resolution multiview diffusion using efficient row-wise attention. *NeurIPS*, 2024. 2, 8

[22] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Mengfei Li, Xiaowei Chi, Siyu Xia, Wei Xue, et al. Pshuman: Photorealistic single-view human reconstruction using cross-scale diffusion. *arXiv preprint arXiv:2409.10141*, 2024. 1, 2, 4, 6, 7

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 2015. 2

[24] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. *CVPR*, 2021. 1

[25] Lea Müller, Vickie Ye, Georgios Pavlakos, Michael Black, and Angjoo Kanazawa. Generative proxemics: A prior for 3d social interaction from images. *CVPR*, 2024. 2, 3, 6

[26] Armin Mustafa, Akin Caliskan, Lourdes Agapito, and Adrian Hilton. Multi-person implicit reconstruction from a single image. *CVPR*, 2021. 2

[27] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. Holoportation: Virtual 3d teleportation in real-time. *ACM Symposium on User Interface Software and Technology*, pages 741–754, 2016. 1

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. *CVPR*, 2019. 2

[29] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025. 2

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022. 2, 6

[31] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, 2019. 1, 2

[32] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. *CVPR*, 2020. 1, 2

[33] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. *CVPR*, 2022. 2

[34] Yuhan Wang, Fangzhou Hong, Shuai Yang, Liming Jiang, Wayne Wu, and Chen Change Loy. Meat: Multiview diffusion model for human generation on megapixels with mesh attention. *CVPR*, 2025. 2

[35] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. *CVPR*, 2022. 1

[36] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. *CVPR*, 2023. 7

[37] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. *CVPR*, 2023. License: Non-commercial academic use only. See https://hi4d.ait.ethz.ch/. 2, 4, 6, 8

[38] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. *CVPR*, 2021. 4, 6, 8

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *ICCV*, 2023. 4, 6

[40] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. *CVPR*, 2024. 1, 2, 7

[41] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *ICCV*, 2021. 1, 2, 6, 7