DECON: Reconstruction of Clothed-Geometric Multiple Humans from a Single Image via Geometry-Guided Decoupling

Yiming Jiang[1], **Wenfeng Song**[2*], **Shuai Li**[1,3*], **Aimin Hao**[1,4]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University
[2]College of Computer Science, Beijing Information Science and Technology University
[3]Zhongguancun Laboratory, China
[4]Research Unit of Virtual Human and Virtual Surgery (2019RU004), Chinese Academy of Medical Sciences
jiangyimingjym@buaa.edu.cn, songwenfenga@gmail.com, {lishuai, ham}@buaa.edu.cn,

Abstract

3D multi-human reconstruction from single images holds significant potential for advancing AR/VR applications. While remarkable progress has been made in single-human reconstruction, existing methods face challenges when reconstructing multiple humans. These challenges include: (1) severe inter-occlusion that disrupts individual body structures, and (2) the absence of physically plausible relative positioning among subjects. We present **DECON**, a novel **DE**couple-and-re **CON**struct framework that systematically addresses these limitations through two technical innovations: (1) a decouple-and-reconstruct framework with multi-view synthesis. It separates individuals and reconstructs detailed 3D bodies from a single image. (2) a Perspective-Aware Position Optimization (PAPO) approach. It ensures realistic positioning by fixing overlaps and gaps between subjects. Extensive experiments demonstrate our method's capability to reconstruct fully separated, anatomically complete 3D humans with clothed-geometric details and plausible interactions. Quantitative evaluations show a 54% reduction in Chamfer Distance and 35% in Point-to-Surface Distance compared to state-of-the-art methods.

Code — https://github.com/IridescentJiang/DECON

1 Introduction

Despite recent progress in single-image 3D reconstruction of individual humans (He et al. 2025; Li et al. 2024; Zhuang et al. 2024b; Jiang et al. 2025), clothed-geometric single-image 3D reconstruction of multi-human with close interactions remains an open challenge. The research gap between the two is noteworthy, as real-world scenarios encompass numerous instances of human-human interactions. While previous research has explored multi-human reconstruction, existing approaches require multi-view image inputs (Zheng et al. 2021), multi-frame video sequences (Jiang et al. 2024), or exhibit insufficient personal characteristics (Cha et al. 2024). Single-image multi-human 3D reconstruction holds application potential in AR/VR scenarios, enabling the creation of personalized stereoscopic photo albums, or 3D printing reconstructed group photos for home display.

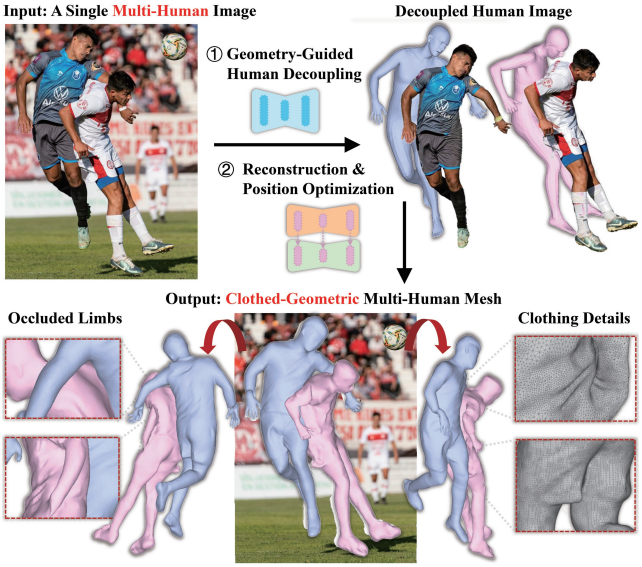

Figure 1: **Given a single RGB image, DECON can reconstruct clothed-geometric multi-human interacting 3D results.** The reconstructed models exhibit three critical properties: (1) clearly separated individuals, (2) anatomically complete bodies with preserved clothing details, and (3) spatial relationships between humans faithfully match the input.

This task is challenging as it requires achieving pixel-level restoration without depth information or other geometric information. Moreover, compared to single-person cases, multi-human scenarios inherently involve mutual occlusions and require careful consideration of interpersonal interactions - all of which must be inferred solely from a single 2D RGB image.

We propose a novel framework, DECON, for single-image multi-human 3D reconstruction. As shown in Fig. 1, our method requires only a single RGB image as input and outputs 3D human models that preserve clothing details, maintain separated individual entities, and accurately restore interacting behaviors. Recent methods using diffusion model (Ho, Jain, and Abbeel 2020) have achieved remarkable progress in human geometry reconstruction: some employing SDS loss for distillation (Wang et al. 2024), oth-

*Corresponding author.

ers combining Large Language Model (LLM) or ControlNet (Zhang, Rao, and Agrawala 2023) to extract additional constraints from images (Zhuang et al. 2024b; Weng et al. 2024).

However, they have fundamental limitations in multi-human scenarios arising from two inherent challenges: i) severe inter-person occlusion and overlapping that compromises individual body completeness, and ii) complicated spatial relationships that connect both relative positioning and interaction patterns among subjects.

To overcome these challenges, we propose a novel 'decouple-and-reconstruct' paradigm. It first separates entangled individuals and reconstructs the complete human models by generative multi-view synthesis. Then, it restores the correct spatial positioning of inter-human spatial relationships. Specifically, we adopt the following approaches: i) we design a geometry-guided human decoupling strategy. This strategy achieves a fully isolated human image by utilizing parameterized geometric body information as a topological prior. We integrate the strategy with a generative multi-view synthesis approach. This approach tackles the issue of information incompleteness due to occlusions by employing the parametric human model-guided novel view images generation, thereby ensuring consistency across the synthesized multi-view images. ii) We propose a Perspective-Aware Position Optimization (PAPO) method that utilizes single-image perspective scaling effects to refine interacting spatial relationships between individuals.

In experiments, we compare our method with current state-of-the-art single-image and monocular video-based human reconstruction methods, demonstrating that our method can reconstruct multiple 3D human models with separate bodies, correct limb structures, and rich geometric details of clothing, while also exhibiting appropriate interactivity within the scene. Comprehensive ablation studies validate the effectiveness of our key components. Our source code will be publicly released.

Our contributions can be summarized as follows:

- We propose **DECON**, a novel decouple-and-reconstruct framework for clothed-geometric single-image multi-human 3D reconstruction, integrating multi-human decoupling and a generative multi-view synthesis module to reconstruct clothed full-body geometries.
- We introduce a simple yet powerful Perspective-Aware Position Optimization method to recover plausible relative positioning between individuals, addressing erroneous inter-person overlaps and implausible separations.
- We provide comprehensive evaluations of our method using the multi-human datasets MultiHuman (Zheng et al. 2021) and Hi4D (Huang et al. 2024a), demonstrating the state-of-the-art geometric fidelity and relative positional validity for multi-human reconstruction.

2 Related Work

Conditionally controlled image generation. Diffusion-based inpainting frameworks (Brack et al. 2024; Ju et al. 2024; Zhuang et al. 2024a; Zuo et al. 2023; Suvorov et al. 2022) enable high-fidelity editing through conditional score prediction. Key innovations include: (1) Efficient inversion mechanisms (LEDITS++ (Brack et al. 2024)); (2) Structural control modules (BrushNet (Ju et al. 2024), PowerPaint (Zhuang et al. 2024a)); (3) Regularization techniques (SCAT (Zuo et al. 2023), LaMa (Suvorov et al. 2022)). Based on mask-aware consistency, these methods achieve pixel-level image completion. Novel-view synthesis has evolved from 3D-aware GANs to hybrid architectures combining diffusion models with geometric priors. Wonder3D (Long et al. 2024) synchronizes normal/RGB generation via cross-domain diffusion for sub-2-minute mesh reconstruction. In the human 3D reconstruction field, MagicMan (He et al. 2025) combines SMPL-X parametric models with hybrid multi-view attention for pose-texture co-optimization. These approaches effectively bridge incomplete 2D inputs with full 3D geometry recovery, enabling the reconstruction of complete 3D human bodies from partial observations through conditionally guided diffusion models.

Monocular image-based clothed single-human reconstruction. Implicit functions (Chen and Zhang 2019; Mescheder et al. 2019; Park et al. 2019; Peng et al. 2020) enabled continuous clothed human reconstruction (He et al. 2020; Huang et al. 2020; Saito et al. 2019; Alldieck, Zanfir, and Sminchisescu 2022; Liao et al. 2023; Yang et al. 2023; Song et al. 2025). However, single-view inputs suffer from depth ambiguity and self-occlusion. Recent solutions include: (1) Parametric priors (PaMIR (Zheng et al. 2022), ICON/ECON (Xiu et al. 2022, 2023), HFHuman (Jiang et al. 2025)) that enhance geometry but rely on model accuracy; (2) Tri-plane transformers (GTA (Zhang et al. 2023), SIFU (Zhang, Yang, and Yang 2024)) tackle with view ambiguity; (3) Multi-view Score Distillation Sampling (Poole et al. 2022) (TeCH (Huang et al. 2024b), GeneMAN (Wang et al. 2024)) and (4) image conditioned diffusion models (SiTH (Ho, Song, and Hilliges 2023), Human-LRM (Weng et al. 2024), DiffHuman (Sengupta et al. 2024)) generating novel-view details. While effective, these methods critically depend on parametric model quality for novel-view consistency. Recently, MagicMan (He et al. 2025) and PSHuman (Li et al. 2024) leverages multi-view image generation to simultaneously optimize parametric model poses while synthesizing novel views, ensuring tighter alignment between the parametric skeleton and observed human posture in images, thereby yielding more accurate reconstructed models.

Monocular multi-human reconstruction. Current research in clothed-geometric multi-person 3D reconstruction from a single image remains in its nascent stage. Existing methods mainly focus on the estimation of parameterized models or multi-frame/multi-image input. SAT-HMR (Su et al. 2024), Multi-HMR (Baradel et al. 2024), and CloseInt (Huang et al. 2024a) enable efficient monocular multi-human estimation, yet their outputs remain limited to textureless parametric human models lacking clothing details. DeepMultiCap (Zheng et al. 2021) and ContactField (Lee et al. 2025) achieve high-fidelity through sparse multi-view inputs. Multiply (Jiang et al. 2024) addresses dynamic multi-human scenarios via hierarchical neural rendering and hybrid instance segmentation. In single-image multi-human

3D reconstruction, Mustafa et al. (Mustafa et al. 2021) pioneered monocular clothed multi-human reconstruction. Cha et al. (Cha et al. 2024) achieved interacting-body recovery relying on latent code-based identity representations. ECON (Xiu et al. 2023) provides a multi-human reconstruction function based on single-person reconstruction. Although they successfully recover the overall shapes and the relative position, the fine details of clothing remain imprecise. Inspired by prior solutions, our approach reconstructs geometric fidelity multiple-interacting 3D humans from a single image through a decouple-and-reconstruct framework and Perspective-Aware Position Optimization.

3 Method

As shown in Fig. 2, we propose a decouple-and-reconstruct framework for single-image multi-human 3D human reconstruction (Sec.3.2). In the decoupling stage, we employ estimated human geometry as guidance to extract individual human images from the original input. During reconstruction, these decoupled human images are processed by our multi-view synthesizer to generate complete 3D human models with detailed clothing geometry. Finally, Perspective-Aware Position Optimization (PAPO) restores each individual's spatial positioning in 3D space according to the original image composition (Sec.3.3).

3.1 Preliminary

Skinned multi-person linear model. The SMPL model (Loper et al. 2015) is a parametric model for human body representation that maps pose $\theta \in \mathbb{R}^{3\times24}$ and shape $\beta \in \mathbb{R}^{10}$ to a mesh with 6890 vertices. The map function M_{SMPL} can be represented as:

$$M_{SMPL}(\theta,\beta) : \mathbb{R}^{3\times24} \times \mathbb{R}^{10} \mapsto \mathbb{R}^{3\times6890}, \tag{1}$$

where θ affects joint positions and orientations. β controls body size. We also use SMPL-X (Pavlakos et al. 2019) as the guidance of multi-view human image synthesis.

3.2 Decouple-and-Reconstruct Framework

Geometry-guided human decoupling. This stage aims to decouple complete individual humans from multi-human images. Due to mutual occlusions between subjects, each person's visual information in the image is inherently incomplete. We address this by leveraging parametric human models (SMPL (Loper et al. 2015)) as geometric priors to ensure anatomical completeness. While existing methods can estimate parametric models from multi-human images, we specifically adopt SAT-HMR (Su et al. 2024) as our Geometry Estimator. SAM 2 (Ravi et al. 2024), a foundation model capable of segmenting various image categories, serves as our image segmentation method. Since the segmented human images remain incomplete, we employ PowerPaint (Zhuang et al. 2024a) as our image inpainting method to achieve pixel-level completion for subsequent reconstruction. Indeed, our proposed method allows for the flexibility of replacing any existing multi-human pose estimation, segmentation, and image inpainting method within our workflow. Throughout the decoupling pipeline, the parametric human model (geometry prior) estimated by the Geometry Estimator provides guidance through three critical interactions: i) During image segmentation, the geometry prior's corresponding human bounding boxes provide segmentation scope guidance to the image segmentation method. ii) We implement Geometry-Guided Decoupling Optimization (GGDO) through differentiable rasterization to enhance alignment between the geometry prior and segmented human images. iii) In the inpainting phase, the geometry prior M_{SMPL} defines completion regions m_{geo} to guide the image inpainting method in generating complete human images from partial segments.

Specifically for the GGDO process, the differentiable renderer from PyTorch3D (Ravi et al. 2020), denoted as DR, is utilized to generate the geometric mask m_{geo} of the estimated SMPL model M_{SMPL}, which can be represented as:

$$DR(M_{SMPL}) \rightarrow m_{geo}. \tag{2}$$

The optimization has two stages. In the first stage f_{G-I}, we use the segmented image mask m_s to optimize the geometry prior's translation parameters $t \in \mathbb{R}^3$, ensuring that the geometry projection is close to the image, which can be represented as:

$$\begin{aligned} f_{G-I} &= \min_{t}(\mathscr{L}_{m\_diff}), \\ \mathscr{L}_{m\_diff} &= |m_{geo} - m_s|, \end{aligned} \tag{3}$$

where $\mathscr{L}_{m\_diff}$ is the MSE loss between m_{geo} and m_s. After f_{G-I}, the optimized translation t' is obtained. In the second stage f_{G-II}, we optimize the pose and shape parameters θ and β:

$$f_{G-II} = \min_{\theta,\beta,t'}(\mathscr{L}_{m\_diff}) \tag{4}$$

Ablation studies for the GGDO are presented in Sec. 4.3.

Generative multi-view synthesis human reconstruction. Single-view-to-multi-view generation followed by synthesized 3D reconstruction effectively ensures full circumferential completeness of human bodies. Inspired by MagicMan (He et al. 2025), we integrate generative multi-view synthesis-based 3D reconstruction while utilizing estimated parametric human models (Pavlakos et al. 2019) as geometry priors. Through ControlNet-style (Zhang, Rao, and Agrawala 2023) constraints, we use decoupled human images as color and style references for the multi-view rendering of the parametric human model. A Denoising UNet is then employed to generate novel-view imagery from these renderings. During training, we also iteratively optimize the pose parameters of the parametric human model to enhance its alignment with the decoupled human image. Guided by geometry priors, geometrically consistent human multi-view images are generated for 3D reconstruction. Following the Wonder3D (Long et al. 2024) paradigm, 3D human meshes are ultimately reconstructed using NeuS (Wang et al. 2021).

3.3 Perspective-Aware Position Optimization

For multi-human scene reconstruction tasks, it is crucial to recover the relative positions and interactions between subjects. Multi-human reconstruction requires precise inter-mesh positioning but faces two challenges: 1) decoupled images lose original spatial contexts, and 2) inaccurate SMPL

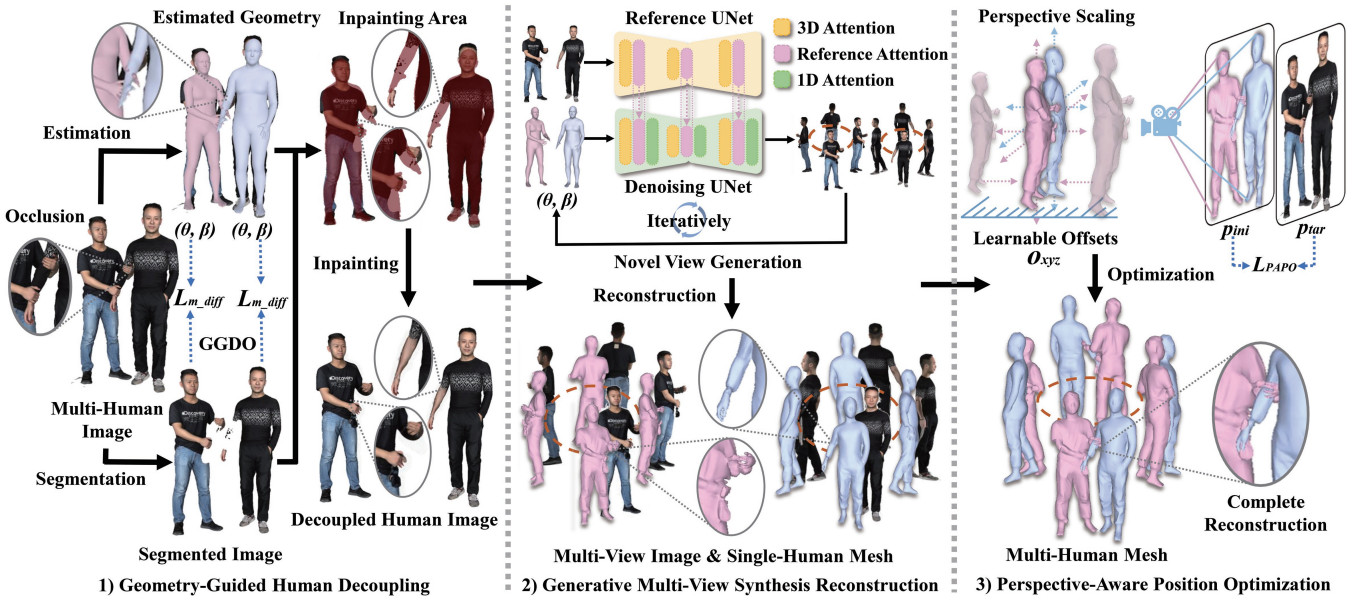

Figure 2: **Pipeline.** Our decouple-and-reconstruct framework operates in three stages: First, we leverage geometry-guided human decoupling to separate individual humans. Then, we employ generative multi-view synthesis for complete human model reconstruction. Finally, we assemble these models through Perspective-Aware Position Optimization.

position estimates. To this end, we propose PAPO. The key motivation of PAPO lies in using perspective scaling effects to optimize inter-mesh positioning. Considering that each estimated subject M_{SMPL} from the Geometry-Guided Human Decoupling stage already has a coarse position, we use the coarse position as the initial positioning. Specifically, we obtain initial XYZ-axis offsets $o_{xyz} \in \mathbb{R}^3$ for each subject and parameterize the offsets as learnable parameters. We then get the projection p_{ini} through the differentiable renderer DR from all subjects with their initial offsets from the camera viewpoint. Simultaneously, we perform pixel-wise summation of the segmented images of all subjects to create the target p_{tar}. Notably, all segmented images maintain the original input resolution - their pixel-wise summation essentially generates a background-free multi-human image. The PAPO f_P can be represented as:

$$\begin{aligned} f_{PAPO} &= \min_{o_{xyz}}(\mathscr{L}_{PAPO}), \\ \mathscr{L}_{PAPO} &= |p_{ini} - p_{tar}|, \end{aligned} \tag{5}$$

where $\mathscr{L}_{PAPO}$ is the MSE loss between projections p_{ini} and targets p_{tar} to individually optimize each subject's offsets. Ablation studies for the PAPO are presented in Sec. 4.3.

4 Experiments

4.1 Implementation Details

Datasets and baselines. The novel view generator is trained on the THuman2.1 dataset (Yu et al. 2021), which contains 2,445 human scans. The training data includes RGB images and normal maps captured at 72 uniformly spaced viewpoints from 0° to 360°. For the geometry estimator and the geometry-guided image inpainting module, we directly use the pre-trained models from SAT-HMR (Su et al. 2024) and PowerPaint (Zhuang et al. 2024a), respectively.

We compare our method with three existing approaches: PIFu (Saito et al. 2019), ECON (Xiu et al. 2023), and Multiply (Jiang et al. 2024). Our benchmark selection is based on the following reason: While methods like SIFU (Zhang, Yang, and Yang 2024) and MagicMan (He et al. 2025) achieve state-of-the-art single-human reconstruction, their architectures inherently lack multi-human decoupling modules. This critical functionality gap necessitates our comparison with PIFu and ECON, which, despite being primarily optimized for single-human reconstruction, retain essential capabilities for multi-human processing. Our choice specifically excludes newer single-human methods to ensure fair evaluation under multi-person interaction scenarios. All experiments are conducted on a single NVIDIA RTX 4090 GPU. More details are provided in the Supplementary Material.

We chose the MultiHuman dataset (Zheng et al. 2021) for comparison with PIFu and ECON. The test set consists of 16 groups of multi-human models, categorized into four types: two-person with minimal contact (4 groups), two-person with natural contact (4 groups), two-person with close contact (4 groups), and three-person (4 groups).

We chose the Hi4D dataset (Huang et al. 2024a) for comparison with Multiply, a method designed for multi-human 3D reconstruction from videos. This dataset contains challenging human interaction videos between pairs of people with ground truth meshes. We chose one video as the test set. Specifically, we extract a single frame f from the video as input to our method and sample varying numbers of frames (30, 70) before and after f to test Multiply.

Metrics. We utilize two quantitative metrics: Point-to-Surface Distance (P2S) and Chamfer Distance (CD). P2S quantifies the proximity of the reconstructed shape points to the surface of the ground-truth shape. CD measures the shape similarity between two point sets by computing average bidirectional point-to-surface distances.

4.2 Evaluation

Comparison with single-image-based human reconstruction methods. We compare our method with existing high-performance methods PIFu (Saito et al. 2019) and ECON (Xiu et al. 2023). In creating visualization results, ECON appends the predicted mesh with the predicted SMPL-X face and hand parts. For fairness, our method's results also automatically stitch the predicted mesh with the predicted SMPL-X parts. We use Poisson Surface Reconstruction (Kazhdan, Bolitho, and Hoppe 2006) for stitching.

The visualization results in Fig. 3 demonstrate our method's superior performance in both limb completeness and plausible relative positioning between subjects. For challenging multi-occlusion cases involving three individuals (Rows 1&3), our approach reconstructs physically reasonable poses with ground-truth-aligned positioning. Notably, existing methods demonstrate fundamental limitations in spatial relationship prediction. Besides, existing methods would show an obvious segmentation border (Row 1 front view) or fail to achieve effective segmentation under severe overlapping conditions (Row 3). In scenarios with close two-person contact (Rows 2), our framework successfully reconstructs two distinct bodies while the baseline methods fail to maintain the proper separation between subjects. More results are provided in the supplementary material.

We quantitatively compare our method against existing high-performance methods, as shown in Table. 1. We evaluate our method in four distinct categories: two-person with minimal contact, two-person with natural contact, two-person with close contact, and three-person scenarios, measuring both CD and P2S metrics. The metrics are computed for all the people in the scene after translation optimization. Our approach achieves the best results across most metrics, particularly excelling in three-person cases, where our CD and P2S are 7.534×10^{-3} and 4.271×10^{-3} lower than ECON (Xiu et al. 2023), respectively, demonstrating significant advantages in multi-human reconstruction. Competing methods cannot accurately estimate the relative spatial relationships between individuals and struggle to distinguish overlapping human bodies in the image. In general, our method outperforms PIFu (Saito et al. 2019) by 4.919×10^{-3} and 1.682×10^{-3} in CD and P2S and outperforms ECON (Xiu et al. 2023) by 3.948×10^{-3} and 2.476×10^{-3}.

Comparison with monocular video-based multi-human reconstruction method. We compare our method with state-of-the-art methods MultiPly (Jiang et al. 2024). Since MultiPly is a video-based method, we provide MultiPly with 1 frame, 30 frames, and 70 frames of video as input. We select the 1-frame video as the input for our method. In Fig. 4, we present a comparison of test results on a set of images from the Hi4D (Huang et al. 2024a) dataset (row 1) and a set of in-the-wild images (row 2). With 30-frame input, MultiPly struggles to fully separate the two human subjects from the background. With 30-frame input, MultiPly produces relatively complete results. With 70-frame input, although there is slight redundancy in the clothing around the legs and arms, the human figures are fully reconstructed with rich surface details. Using only a single image as input, our method outperforms MultiPly with a 30-frame input in terms of geometric accuracy and achieves results comparable to MultiPly with a 70-frame input.

In Table 2, we present the quantitative results for this set of images from the Hi4D (Huang et al. 2024a) dataset. The time refers to the inference time required for each person using a single NVIDIA 4090 GPU. As shown, our method achieves the best performance in the CD metric, while P2S is not far behind. Furthermore, the inference time of our method is only 1/22 of MultiPly with a 30-frame input and 1/45 of MultiPly with a 70-frame input. Despite presenting more surface redundancy, MultiPly with a 30-frame input offers a more precise bending angle estimation for a bent posture, leading to better metrics than the 70-frame input version.

In-the-wild images evaluation. We conduct experiments across a diverse set of in-the-wild scenarios, encompassing various complex occlusions, different crowd sizes, and outdoor lighting conditions. As shown in Fig 5, our approach effectively captures intricate details of clothing and accurately reconstructs the relative positions between individuals. More results are provided in the Supplementary Material.

4.3 Ablation Studies

Ablation on PAPO. As shown in Table. 1, our method significantly reduces CD and P2S in all categories after PAPO, with overall improvements of 1.955×10^{-3} and 1.684×10^{-3}, respectively. Fig. 6 presents the frontal and side views of the partial qualitative results. For the side view, we annotate the contact regions between two individuals in red and green. Before optimization, these regions exhibit a noticeable gap, while after optimization, they achieve tight physical contact. In the red-circled regions, the frontal view of the reconstructed model demonstrates improved alignment with the input after optimization.

Ablation on GGDO. Fig. 7 presents partial qualitative ablation results on GGDO. The inpainted images (right in Rows (c) and (d)) exhibit a reduction in erroneous redundant regions after optimization. Row (b) shows the segmented images from the input image. Rows (c) and (d) depict the parametric human model mask, which defines the inpainting range, and the resulting inpainted images. The yellow contour on the mask corresponds to the segmented human silhouette in row (b). Through optimization, the mask exhibits improved alignment with the yellow contour. The optimization reduces redundant inpainting artifacts in specific regions: the elbow and wrist for the left individual, the toes for the middle individual, and the fingers for the right.

5 Conclusion

We present **DECON**, a novel framework addressing clothed-geometric single-image multi-human 3D recon-

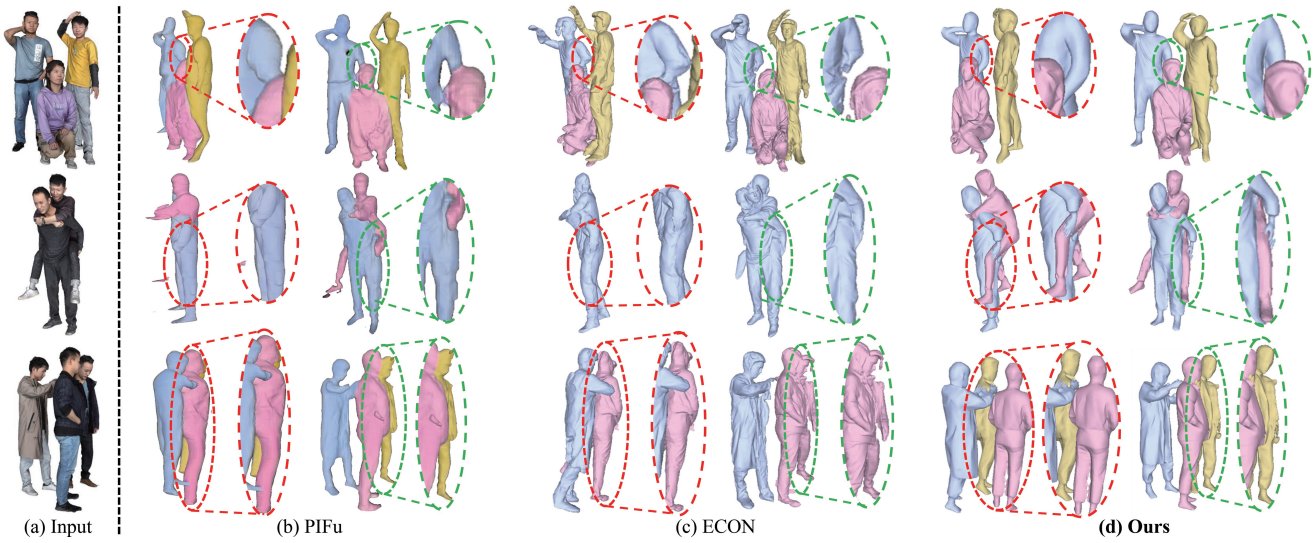

Figure 3: **Qualitative evaluation against single-image-based human reconstruction methods, including (b) PIFu (Saito et al. 2019) and (c) ECON (Xiu et al. 2023).** Our approach achieves superior results on both the front and sides of the model, ensuring better independence for each subject and avoiding the fusion artifacts observed in other methods. The key areas of the side results are circled in red, and the key areas of the front/back results are circled in green. In order to distinguish different subjects, we manually colored the models. Please **zoom in** to see details.

Categories	Two minimal contact		Two natural contact		Two close contact		Three-person		All	
Methods	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓
PIFu (Saito et al. 2019)	5.998	4.083	5.743	3.494	3.892	**2.076**	17.455	9.553	8.272	4.801
ECON (Xiu et al. 2023)	4.333	3.874	4.023	3.840	7.504	5.177	13.346	9.487	7.301	5.595
Ours(w/o PAPO)	4.266	3.895	2.371	1.986	7.246	7.786	7.348	5.545	5.308	4.803
Ours	**3.099**	**3.114**	**1.570**	**1.328**	**2.931**	2.817	**5.812**	**5.216**	**3.353**	**3.119**

Table 1: **Quantitative evaluation on MultiHuman (Zheng et al. 2021).** We compare our method with PIFu (Saito et al. 2019), ECON (Xiu et al. 2023), and our method without PAPO. Our method achieves the lowest values on nearly all metrics. The best and the second-best results are highlighted in gray. All CD and P2S values in the table are scaled by 10^{-3}.

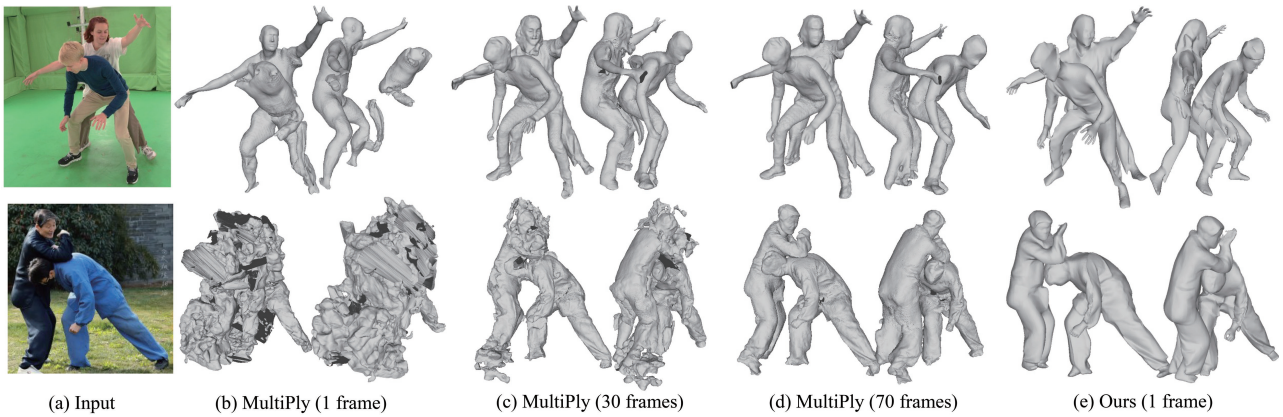

Figure 4: **Qualitative evaluation against MultiPly (Jiang et al. 2024), the SOTA monocular video-based multi-human reconstruction method.** Our method outperforms Multiply when provided with 30 frames as input and achieves comparable results when given 70 frames as input. However, in contrast to Multiply with a 70-frame input, our approach requires only a single image as input, and the inference time is merely 1/45 of that. Please **zoom in** to see details.

Methods	Input	CD↓	P2S↓	Time (h/p)
MultiPly (Jiang et al. 2024)	1 frame	12.062	8.034	1.75
	30 frames	7.621	**3.351**	7.5
	70 frames	9.628	3.356	15
Ours	1 frame	**5.519**	5.385	**0.33**

Table 2: **Quantitative evaluation on Hi4D (Huang et al. 2024a) with MultiPly (Jiang et al. 2024).** Our method requires only a single image as input and surpasses all input modes of MultiPly in terms of the CD metric, while the inference time is merely 1/45 of MultiPly with a 70-frame input. All CD and P2S values in the table are scaled by 10^{-3}.

Figure 5: **In-the-wild images evaluation.** Our method performs well across a diverse set of scenarios, encompassing various complex occlusions, different crowd sizes, and outdoor lighting conditions. Please 🔍 **zoom in** for details.

struction with interacting relationships. Our decouple-and-reconstruct paradigm has two contributions: 1) decoupling of entangled individuals through parametric priors and then reconstructing the complete body by generative multi-view synthesis, and 2) Perspective-Aware Position Optimization of multi-person spatial arrangements. Extensive experiments demonstrate that DECON outperforms existing single-image and video-based approaches in reconstructing mutually separated, geometrically detailed 3D humans with plausible interactions. This work fills a critical gap in reconstructing interacting human groups from monocular imagery, enabling applications in VR and 3D photography.

Limitation and future work. DECON fails to reconstruct the human with severe (>50%) occlusion. In the future, we plan to address this by using the large human model.

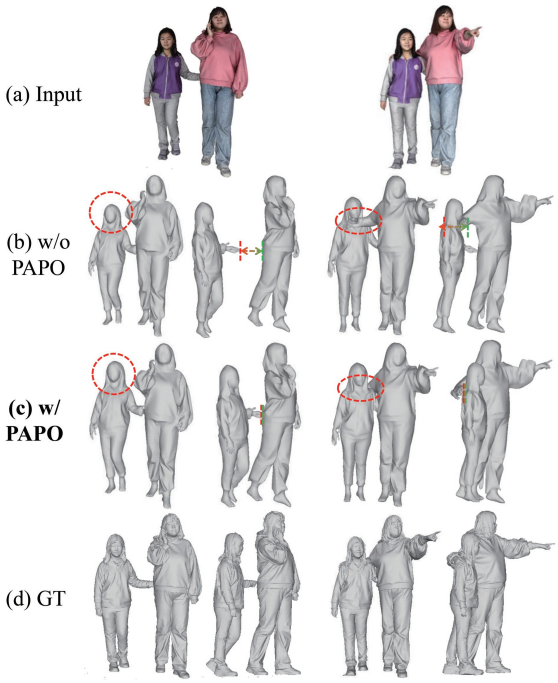

Figure 6: **Visualization results of ablation on the PAPO.** After optimization, the relative positioning of the individuals exhibits improved spatial coherence.

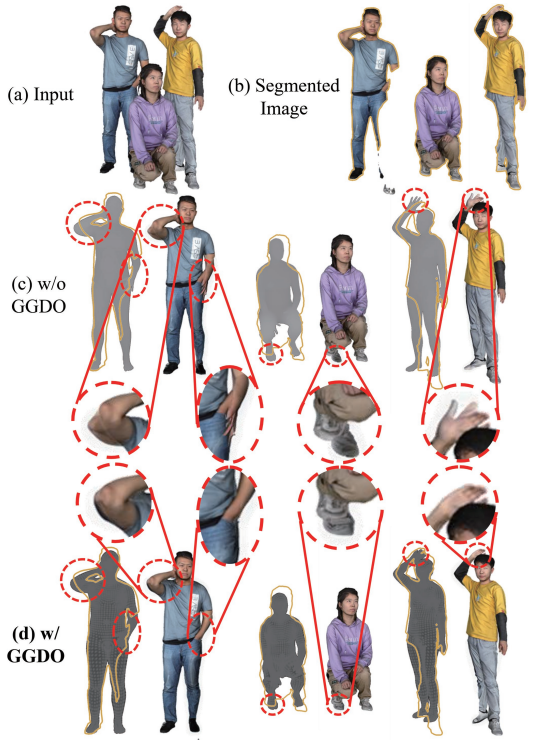

Figure 7: **Visualization results of ablation on the GGDO.** The inpainted image exhibits a reduction in erroneous redundant regions after GGDO.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (62525204, 62572062, 62441201, 62272021), Beijing Natural Science Foundation (L232102), the Fundamental Research Funds for the Central Universities, and Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A2406), Zhejiang University.

References

Alldieck, T.; Zanfir, M.; and Sminchisescu, C. 2022. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, 1506–1515.

Baradel, F.; Armando, M.; Galaaoui, S.; Brégier, R.; Weinzaepfel, P.; Rogez, G.; and Lucas, T. 2024. Multi-HMR: Multi-person Whole-Body Human Mesh Recovery in a Single Shot. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV*, volume 15081, 202–218.

Brack, M.; Friedrich, F.; Kornmeier, K.; Tsaban, L.; Schramowski, P.; Kersting, K.; and Passos, A. 2024. LEDITS++: Limitless Image Editing Using Text-to-Image Models. In *CVPR*, 8861–8870.

Cha, J.; Lee, H.; Kim, J.; Truong, N. N. B.; Yoon, J. S.; and Baek, S. 2024. 3D Reconstruction of Interacting Multi-Person in Clothing from a Single Image. In *WACV*, 5291–5300.

Chen, Z.; and Zhang, H. 2019. Learning implicit fields for generative shape modeling. In *CVPR*, 5939–5948.

He, T.; Collomosse, J.; Jin, H.; and Soatto, S. 2020. Geo-pifu: Geometry and pixel aligned implicit functions for single-view human reconstruction. *NeurIPS*, 33: 9276–9287.

He, X.; Wu, Z.; Li, X.; Kang, D.; Zhang, C.; Ye, J.; Chen, L.; Gao, X.; Zhang, H.; and Zhuang, H. 2025. MagicMan: Generative Novel View Synthesis of Humans with 3D-Aware Diffusion and Iterative Refinement. In *AAAI*, 3437–3445.

Ho, H.-I.; Song, J.; and Hilliges, O. 2023. SiTH: Single-view Textured Human Reconstruction with Image-Conditioned Diffusion. *arXiv preprint arXiv:2311.15855*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.

Huang, B.; Li, C.; Xu, C.; Pan, L.; Wang, Y.; and Lee, G. H. 2024a. Closely Interactive Human Reconstruction with Proxemics and Physics-Guided Adaption. In *CVPR*, 1011–1021.

Huang, Y.; Yi, H.; Xiu, Y.; Liao, T.; Tang, J.; Cai, D.; and Thies, J. 2024b. TeCH: Text-Guided Reconstruction of Lifelike Clothed Humans. In *3DV*, 1531–1542.

Huang, Z.; Xu, Y.; Lassner, C.; Li, H.; and Tung, T. 2020. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 3093–3102.

Jiang, Y.; Song, W.; Li, S.; and Hao, A. 2025. HFHuman: High-Fidelity Human Reconstruction from Single Image with Multi-Modality Fusion. *IEEE Transactions on Visualization and Computer Graphics*, 1–12.

Jiang, Z.; Guo, C.; Kaufmann, M.; Jiang, T.; Valentin, J.; Hilliges, O.; and Song, J. 2024. MultiPly: Reconstruction of Multiple People from Monocular Video in the Wild. In *CVPR*, 109–118.

Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. BrushNet: A Plug-and-Play Image Inpainting Model with Decomposed Dual-Branch Diffusion. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV*, volume 15078 of *Lecture Notes in Computer Science*, 150–168.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *SGP*, volume 7.

Lee, H.; You, T.; Park, H.; Shim, W.; Kim, S.; and Lim, H. 2025. ContactField: Implicit Field Representation for Multi-Person Interaction Geometry. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *NeurIPS*, volume 37, 38079–38104.

Li, P.; Zheng, W.; Liu, Y.; Yu, T.; Li, Y.; Qi, X.; Li, M.; Chi, X.; Xia, S.; Xue, W.; Luo, W.; Liu, Q.; and Guo, Y. 2024. PSHuman: Photorealistic Single-view Human Reconstruction using Cross-Scale Diffusion. *CoRR*, abs/2409.10141.

Liao, T.; Zhang, X.; Xiu, Y.; Yi, H.; Liu, X.; Qi, G.-J.; Zhang, Y.; Wang, X.; Zhu, X.; and Lei, Z. 2023. High-fidelity clothed avatar reconstruction from a single image. In *CVPR*, 8662–8672.

Long, X.; Guo, Y.; Lin, C.; Liu, Y.; Dou, Z.; Liu, L.; Ma, Y.; Zhang, S.; Habermann, M.; Theobalt, C.; and Wang, W. 2024. Wonder3D: Single Image to 3D Using Cross-Domain Diffusion. In *CVPR*, 9970–9980.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics*, 34(6): 248:1–248:16.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 4460–4470.

Mustafa, A.; Caliskan, A.; Agapito, L.; and Hilton, A. 2021. Multi-Person Implicit Reconstruction From a Single Image. In *CVPR*, 14474–14483.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 165–174.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *CVPR*, 10975–10985.

Peng, S.; Niemeyer, M.; Mescheder, L.; Pollefeys, M.; and Geiger, A. 2020. Convolutional Occupancy Networks. In *ECCV*, 523–540.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2:

Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Ravi, N.; Reizenstein, J.; Novotný, D.; Gordon, T.; Lo, W.; Johnson, J.; and Gkioxari, G. 2020. Accelerating 3D Deep Learning with PyTorch3D. *CoRR*, abs/2007.08501.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Li, H.; and Kanazawa, A. 2019. PIFu: Pixel-Aligned Implicit Function for High-Resolution Clothed Human Digitization. In *ICCV*, 2304–2314.

Sengupta, A.; Alldieck, T.; Kolotouros, N.; Corona, E.; Zanfir, A.; and Sminchisescu, C. 2024. DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans. In *CVPR*, 1439–1449.

Song, W.; Ding, Y.; Hou, F.; Li, S.; Hao, A.; and Hou, X. 2025. CtrlAvatar: Controllable Avatars Generation via Disentangled Invertible Networks. In *AAAI*, 6959–6967.

Su, C.; Ma, X.; Su, J.; and Wang, Y. 2024. SAT-HMR: Real-Time Multi-Person 3D Mesh Estimation via Scale-Adaptive Tokens. *CoRR*, abs/2411.19824.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust Large Mask Inpainting with Fourier Convolutions. In *WACV*, 3172–3182.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, W.; Ye, H.; Hong, F.; Yang, X.; Zhang, J.; Wang, Y.; Liu, Z.; and Pan, L. 2024. GeneMAN: Generalizable Single-Image 3D Human Reconstruction from Multi-Source Human Data. *CoRR*, abs/2411.18624.

Weng, Z.; Liu, J.; Tan, H.; Xu, Z.; Zhou, Y.; Yeung-Levy, S.; and Yang, J. 2024. Template-free single-view 3d human digitalization with diffusion-guided lrm. *arXiv preprint arXiv:2401.12175*.

Xiu, Y.; Yang, J.; Cao, X.; Tzionas, D.; and Black, M. J. 2023. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 512–523.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 13296–13306.

Yang, X.; Luo, Y.; Xiu, Y.; Wang, W.; Xu, H.; and Fan, Z. 2023. D-if: Uncertainty-aware human digitization via implicit distribution field. In *ICCV*, 9122–9132.

Yu, T.; Zheng, Z.; Guo, K.; Liu, P.; Dai, Q.; and Liu, Y. 2021. Function4D: Real-Time Human Volumetric Capture From Very Sparse Consumer RGBD Sensors. In *CVPR*, 5746–5756.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 3836–3847.

Zhang, Z.; Sun, L.; Yang, Z.; Chen, L.; and Yang, Y. 2023. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. *NeurIPS*, 36: 7818–7830.

Zhang, Z.; Yang, Z.; and Yang, Y. 2024. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. In *CVPR*, 9936–9947.

Zheng, Y.; Shao, R.; Zhang, Y.; Yu, T.; Zheng, Z.; Dai, Q.; and Liu, Y. 2021. DeepMultiCap: Performance Capture of Multiple Characters Using Sparse Multiview Cameras. In *ICCV*, 6219–6229.

Zheng, Z.; Yu, T.; Liu, Y.; and Dai, Q. 2022. PaMIR: Parametric Model-Conditioned Implicit Representation for Image-Based Human Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3170–3184.

Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2024a. A Task Is Worth One Word: Learning with Task Prompts for High-Quality Versatile Image Inpainting. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *ECCV*, volume 15116 of *Lecture Notes in Computer Science*, 195–211.

Zhuang, Y.; Lv, J.; Wen, H.; Shuai, Q.; Zeng, A.; Zhu, H.; Chen, S.; Yang, Y.; Cao, X.; and Liu, W. 2024b. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. *arXiv preprint arXiv:2412.14963*.

Zuo, Z.; Zhao, L.; Li, A.; Wang, Z.; Zhang, Z.; Chen, J.; Xing, W.; and Lu, D. 2023. Generative Image Inpainting with Segmentation Confusion Adversarial Training and Contrastive Learning. In Williams, B.; Chen, Y.; and Neville, J., eds., *AAAI*, 3888–3896.