CrowdGaussian: Reconstructing High-Fidelity 3D Gaussians for Human Crowd from a Single Image

Supplementary Material

Figure 1. Visualization of the multi-stage occlusion mask generation process. From left to right: original image, with keypoint-based elliptical occlusions applied, further corrupted by irregular Bézier-based erasures, and finally smoothed via morphological operations.

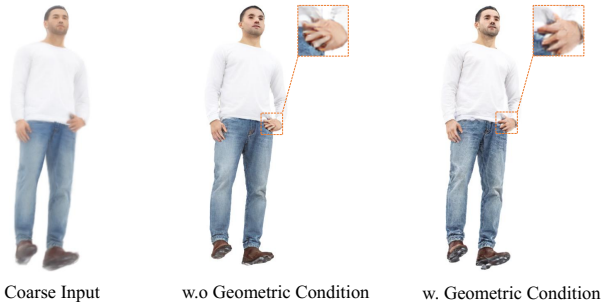

Figure 2. Ablation on geometric conditioning using SMPL normal maps. Left: Coarse input. Middle: Result without normal map conditioning — suffers from hand collapse and facial distortion. Right: Result with normal map — preserves correct hand structure and facial geometry. Zoom-ins show significant improvement in local detail fidelity and structural coherence.

A. Occlusion Simulation for Self-Supervised Training

To ensure robust generalization under real-world occlusions, we design a structured yet diverse masking strategy to synthesize I_{occ} from the full image I_{full}. The occlusion process combines three complementary components: anatomically plausible local erasures, global irregular masks, and structural line cuts—ensuring both realism and diversity in training data.

First, we detect human body keypoints using MediaPipe Pose [8], excluding head landmarks to focus on torso and limb regions. For each selected keypoint $\mathbf{p}_k = (x_k, y_k)$, we apply an elliptical occlusion:

$$\mathcal{M}^{(k)}_{\text{ellipse}} = \left\{ (x, y) \in \mathbb{R}^2 \,\middle|\, \frac{(x - x_k)^2}{a_x^2} + \frac{(y - y_k)^2}{a_y^2} \leq 1 \right\}, \tag{1}$$

where semi-axes $a_x, a_y \sim \mathcal{U}(30, 100)$ and rotation angle $\theta \sim \mathcal{U}(0°, 360°)$ are sampled randomly. Up to $K = 5$ such ellipses are applied, with number and location randomized per instance.

Second, we generate large-scale irregular masks using Bézier curves. A quadratic Bézier path is defined by three control points $\mathbf{C}_0, \mathbf{C}_1, \mathbf{C}_2 \in [0, H] \times [0, W]$, uniformly sampled across the image. The curve is thickened into a filled region via polygon rasterization, forming a continuous occluding band:

$$\mathcal{M}_{\text{bezier}} = \bigcup_{i=1}^{N_b} \text{FillPoly}(\text{Bezier}(\mathbf{C}_0^{(i)}, \mathbf{C}_1^{(i)}, \mathbf{C}_2^{(i)}, t),\ t \in [0, 1]), \tag{2}$$

where $N_b \sim \mathcal{U}\{0, 5\}$ controls the number of such global erasures.

Third, with probability $p = 0.5$, we apply a random straight-line cut that partitions the image into two half-planes. Given two random points (x_1, y_1), (x_2, y_2), the dividing line is:

$$L(x, y) = (y - y_1)(x_2 - x_1) - (x - x_1)(y_2 - y_1) > 0. \tag{3}$$

We select one side for occlusion, ensuring the masked area does not exceed 70% of the total to preserve sufficient visible content.

The final binary mask $\mathcal{M}_{\text{final}}$ is obtained by combining all components:

$$\mathcal{M}_{\text{final}} = \mathcal{M}_{\text{ellipse}} \cup \mathcal{M}_{\text{bezier}} \cup \mathcal{M}_{\text{line}}, \tag{4}$$

followed by morphological closing and dilation (kernel size 5×5, iterations=3) to smooth boundaries and simulate realistic cloth or object overlap.

As illustrated in Figure 1, this multi-step process progressively constructs complex occlusions that mimic real-world scenarios—such as one person being partially blocked by another or obscured by foreground objects.

This multi-level occlusion scheme ensures that the student model (LORM) learns to recover missing geometry under diverse and challenging conditions, while the frozen teacher provides consistent pseudo-ground truths for stable

self-distillation. During training, we render the 3DGS representation from 24 horizontally distributed camera viewpoints uniformly placed on a full 360-degree orbit around the subject, ensuring comprehensive geometric coverage for robust self-supervised learning.

B. Effect of Geometric Conditioning

To evaluate the impact of geometric priors in diffusion-based refinement, we ablate the use of SMPL normal maps as conditioning input. As shown in Figure 2, without geometric guidance (middle), the refiner often produces structural distortions—such as collapsed hands or misaligned limbs—due to ambiguous depth and pose cues in coarse renderings. In contrast, when conditioned on the SMPL normal map (right), our method preserves anatomically plausible hand and facial structures, demonstrating improved 3D consistency.

The zoom-in views highlight that geometric conditioning helps recover fine-scale details that are otherwise hallucinated or distorted. This confirms that the normal map provides effective 3D-aware inductive bias, guiding the diffusion process to respect human body geometry during refinement.

C. Limitations of 2D Inpainting as Preprocessing

A natural alternative to our self-supervised adaptation framework is to first apply 2D image inpainting to occluded person crops and then feed the completed image into a pre-trained 3D generator such as LHM [10]. However, we find this strategy ineffective due to the inherent limitations of 2D inpainting under real-world occlusions.

As shown in Figure 5, while modern diffusion-based inpainters [1] can generate visually plausible textures locally, they often produce boundary artifacts, inconsistent lighting, and semantically implausible content—such as incorrect limb structures or mismatched clothing patterns. These errors are not only visually jarring but also mislead the downstream 3D reconstruction model.

When such inpainted images are passed to LHM, the generated 3DGS exhibits amplified artifacts, including distorted geometry and texture inconsistencies across views. This confirms that 2D inpainting lacks 3D-awareness and cannot reliably recover structurally coherent human shapes under arbitrary occlusion patterns.

D. More Results

We present additional qualitative results to further demonstrate the effectiveness and generalization of our method.

Figure 3 compares our method against recent state-of-the-art approaches—including LHM [10], PSHuman [6], SyncHuman [4], and IDOL [15]—on occluded reconstructions from THuman2.1 [14]. The occlusions are synthesized using either Bézier curves or random rectangular masks.

Figure 4 shows reconstructions on a variety of challenging in-the-wild images containing multiple individuals.

E. Multi-Person Human Mesh Recovery

This section provides background on parametric human modeling and multi-person 3D human recovery from a single image, which forms the foundation of our pipeline.

Human Parametric Model The SMPL [7] and SMPL-X parametric model [7, 9] are widely used for representing 3D human body shape and pose. It represents a deformable human mesh as a function of low-dimensional parameters:

$$\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathcal{W}\left(\mathcal{V}(\boldsymbol{\beta}, \boldsymbol{\theta}), \mathcal{J}(\boldsymbol{\beta}), \mathcal{T}_p(\boldsymbol{\theta}), \mathcal{W}\right), \tag{5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{10}$ denotes the shape parameters (learned from PCA over 3D scans), and $\boldsymbol{\theta} \in \mathbb{R}^{23\times 3}$ represents the joint rotation angles in axis-angle form. The model applies linear blend skinning ($\mathcal{W}$) to a template mesh $\mathcal{V}$, with pose-dependent blend shapes and joint regressor $\mathcal{J}(\boldsymbol{\beta})$ to generate realistic articulations.

Multi-Person Human Mesh Recovery from a Single Image To recover multiple humans from a single image, recent methods [3, 13] estimate per-person SMPL-X parameters along with their camera-space positions. Given an input image I, the goal is to infer a set of parameters for each detected individual $i = 1, \ldots, N$:

$$\{\boldsymbol{\theta}_i, \boldsymbol{\beta}_i, \mathbf{t}_i\}_{i=1}^{N} = f_{\text{pose}}(I), \tag{6}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$ are the pose and shape parameters of person i, and $\mathbf{t}_i \in \mathbb{R}^3$ is the root translation (camera-relative position).

In our pipeline, we adopt PromptHMR [13] to extract these parameters, which provides robust multi-person pose and global positioning, serving as input for subsequent processing stages.

F. Implementation Details

LORM Training. For efficient fine-tuning, we update the transformer block using LoRA with rank 32, while other inherited pre-trained components remain frozen. The model is trained for approximately 2,000 iterations using the AdamW [5] optimizer with a learning rate of 1×10^{-4}.

CrowdRefiner Training. We build CrowdRefiner based on SD-Turbo [11], employing a single-step diffusion process with noise level $\tau = 200$ (maximum diffusion time

Figure 3. Qualitative comparison on occluded human reconstruction (THuman2.1). Input images are occluded via Bézier curves or rectangular masks.

Figure 4. More reconstruction results on diverse in-the-wild multi-person images.

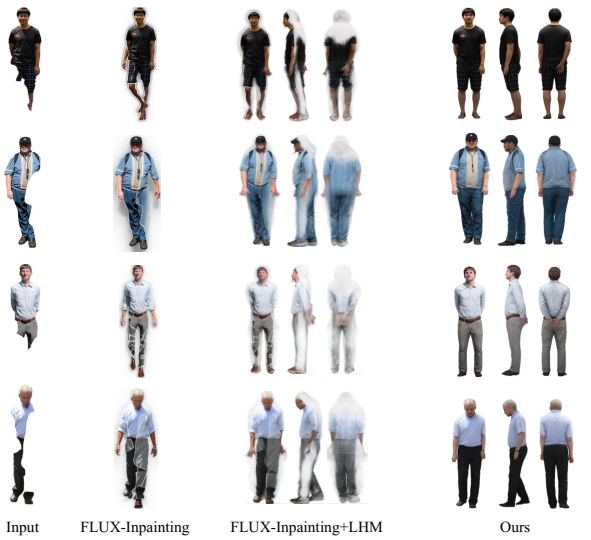

Figure 5. Comparison with the 2D inpainting + LHM pipeline.

$\tau = 1000$), enabling efficient training and inference. We apply LoRA to the VAE decoder with rank 4, while keeping the encoder frozen. The model is trained with the composite loss $\mathcal{L}_{\text{diffuse}}$ for approximately 10,000 steps with a batch size of 2. We apply the SCL strategy by replacing the input with R_{gt} at probability $\rho = 0.2$ during training. The model is optimized using AdamW [5] with a learning rate of 2×10^{-5}.

Evaluation Setup on THuman2.1 For quantitative evaluation on occluded human reconstruction, we select 500 samples from THuman2.1 [14]. Ground truth renderings are generated by rendering the high-fidelity meshes from 24 uniformly distributed viewpoints along a full 360-degree horizontal orbit. These views serve as reference images for computing PSNR, SSIM, and LPIPS to assess reconstruction quality.

G. Evaluation Metrics

We evaluate the performance of CrowdRefiner and compare our method with baselines on occluded human reconstruction using three widely adopted image quality metrics: PSNR, SSIM, and LPIPS. These provide complementary insights into reconstruction fidelity—from pixel accuracy to perceptual similarity.

PSNR Peak Signal-to-Noise Ratio (PSNR) measures the pixel-wise similarity between the predicted rendering I_{pred} and the ground truth I_{gt}. It is computed based on the mean squared error (MSE):

$$\text{MSE} = \frac{1}{N} \sum_{p} (I_{\text{pred}}[p] - I_{\text{gt}}[p])^2, \tag{7}$$

where p indexes all pixels. The PSNR is then:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right), \tag{8}$$

with MAX as the maximum pixel value (e.g., 255). Higher values indicate better photometric accuracy.

SSIM Structural Similarity (SSIM) evaluates how well structural patterns—such as edges and textures—are preserved. Instead of global statistics, it computes similarity in local windows using luminance, contrast, and structure comparisons:

$$\text{SSIM}(I_{\text{pred}}, I_{\text{gt}}) = \frac{(2\mu_p\mu_g + C_1)(2\sigma_{pg} + C_2)}{(\mu_p^2 + \mu_g^2 + C_1)(\sigma_p^2 + \sigma_g^2 + C_2)}, \tag{9}$$

where μ, σ denote local means and standard deviations, and σ_{pg} is the cross-covariance. Constants C_1, C_2 prevent division by zero. The final score is averaged over all patches. Values closer to 1 indicate stronger structural consistency.

LPIPS Learned Perceptual Image Patch Similarity (LPIPS) assesses perceptual quality by comparing deep features from a pre-trained network. Rather than raw pixels, it operates in feature space:

$$\text{LPIPS}(I_{\text{pred}}, I_{\text{gt}}) = \sum_{l} w_l \cdot \|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{gt}})\|_2^2, \tag{10}$$

where ϕ_l extracts features from layer l of a VGG-16 [12] backbone, and w_l are learned weights that emphasize semantically meaningful layers. Lower LPIPS values correspond to higher perceptual similarity, aligning closely with human judgment.

Together, these metrics allow us to assess reconstructions from multiple perspectives: PSNR for pixel-level accuracy, SSIM for structural coherence, and LPIPS for high-level visual realism.

H. Source of In-the-Wild Images

The in-the-wild multi-person images used in our qualitative evaluation are collected from publicly available sources on the web, including search engines such as Google Images [2]. These images are selected solely for visual demonstration and are not used in quantitative benchmarking. We ensure that all displayed examples fall under fair use for academic illustration, with no commercial intent. For reproducibility, we do not claim ownership of these images and encourage readers to refer to the original sources via the search engine.

References

[1] https://huggingface.co/alimama-creative/FLUX.1-dev-Controlnet-Inpainting-Beta. 2

[2] Google Images. https://images.google.com/. 4

[3] Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-hmr: Multi-person whole-body human mesh recovery in a single shot. In *European Conference on Computer Vision*, pages 202–218. Springer, 2024. 2

[4] Wenyue Chen, Peng Li, Wangguandong Zheng, Chengfeng Zhao, Mengfei Li, Yaolong Zhu, Zhiyang Dou, Ronggang Wang, and Yuan Liu. Synchuman: Synchronizing 2d and 3d generative models for single-view human reconstruction. *arXiv preprint arXiv:2510.07723*, 2025. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2, 4

[6] Peng Li, Wangguandong Zheng, Yuan Liu, Tao Yu, Yangguang Li, Xingqun Qi, Xiaowei Chi, Siyu Xia, Yan-Pei Cao, Wei Xue, et al. Pshuman: Photorealistic single-image 3d human reconstruction using cross-scale multiview diffusion and explicit remeshing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16008–16018, 2025. 2

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

[8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019. 2

[10] 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 for single image to 3d in seconds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14184–14194, 2025. 2

[11] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[13] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1148–1159, 2025. 2

[14] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 2, 4

[15] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. Idol: Instant photorealistic 3d human creation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26308–26319, 2025. 2