S^{2}HUMAN: BRIDGING SMPL-X AND STEREO MATCHING FOR POSE-FREE 3D HUMAN RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent years have seen rapid development in 3D human reconstruction, a technology that produces digital human models based on image or video inputs. However, existing methods heavily rely on pre-calibrated camera parameters, thus limiting the applicability in consumer-level scenarios. While intrinsic camera parameters can be readily obtained from device specifications, the pose-free 3D human reconstruction framework remains crucial. In this paper, we propose a pose-free 3D human reconstruction framework named S^{2}Human by bridging SMPL-X and Stereo matching. Stereo matching delivers pose-free capability via extensive 3D reconstruction priors, while SMPL-X supplies human priors to guide the enhancement of geometric accuracy and fine-grained texture details. Specifically, S^{2}Human presents two key innovations. On the one hand, a SMPL-X Estimation module is proposed to acquire a unified SMPL-X model from SMPL-X monocular estimations of different views, providing a robust and consistent human prior. On the other hand, a Posemap Adapter module is designed to inject structured human body information from the unified SMPL-X model into the stereo matching network, significantly enhancing both geometric accuracy and textural details in the reconstructed human model. Experimental results on the THuman2.0 and Twindom datasets demonstrate that S^{2}Human is competitive with the pose-aware methods.

1 INTRODUCTION

3D human reconstruction enables the creation of high-fidelity digital human models, which are critical to a broad range of applications like immersive telepresence Tu et al. (2024), photorealistic digital characters Li et al. (2024), biomechanical analysis Kim et al. (2023), etc. In recent years, it has made remarkable progress, driven by advanced 3D representations, parametric models like SMPL Loper et al. (2023), high-precision sensing and large datasets, overcoming key challenges in detail reconstruction and real-time performance.

Existing 3D human reconstruction methods can be broadly divided into two paradigms. In the first paradigm Peng et al. (2021); Kwon et al. (2024); Xiao et al. (2025), parametric human models such as SMPL Loper et al. (2023) or SMPL-X Pavlakos et al. (2019) are leveraged as templates for providing human prior knowledge of human form. These methods rely on pre-calibrated camera parameters to establish the projection between 2D images and the parametric human model. The second paradigm Hu et al. (2025); Tu et al. (2024); Zheng et al. (2024) utilizes stereo matching networks to predict human point clouds, which are then used for regressing other human 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) parameters. These

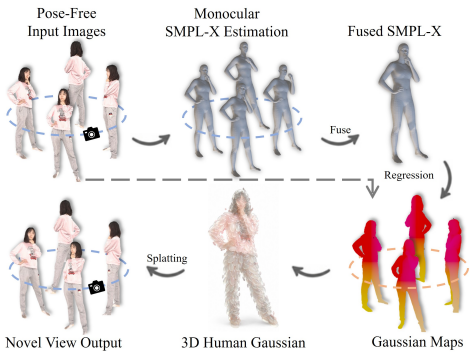

Figure 1: The workflow of S^{2}Human.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

methods are particularly effective for reconstructing humans wearing loose clothing. However, these methods are generally limited to fixed camera setups, requiring a separately trained network for each different camera setup, as their high-fidelity representations exhibit high sensitivity to parameter deviations.

The field of 3D human reconstruction is transitioning from controlled laboratory environments to consumer-level scenarios. However, existing two paradigms both heavily rely on pre-calibrated camera parameters, posing a significant obstacle to the transition. Given that intrinsic camera parameters can be readily obtained from device specifications, most calibration efforts are devoted to extrinsic parameters, namely camera pose. Accordingly, a pose-free framework for 3D human reconstruction remains crucial.

In this paper, we propose a 3D human reconstruction framework named S^2Human, which achieves pose-free capability using extensive 3D reconstruction priors and incorporates human priors to ensure geometric accuracy and fine-grained texture details. As shown in Fig. 1, S^2Human adopts a well-trained VGGT Wang et al. (2025) as the stereo matching network for providing 3D reconstruction priors, and introduces two novel modules for acquiring and injecting human priors. A SMPL-X Estimation module acquires a unified SMPL-X model by fusing SMPL-X monocular estimations from different views under pose-free settings. The fusion is guided by human joint visibility in each view, achieving the correction of errors caused by occlusions and ambiguities in monocular predictions. A Posemap Adapter module is designed to inject structured human body information from the unified SMPL-X model into the stereo matching network. This module represents the body information using a Tri-posemap, and fuses Tri-posemap features with stereo matching features through intra-view as well as inter-view attention mechanisms. Experimental results on the THuman2.0 and Twindom datasets demonstrate that S^2Human achieves competitive performance with pose-aware methods.

In summary, the main contributions of this paper are:

- We bridge SMPL-X and stereo matching to propose S^2Human, which achieves pose-free capability using extensive 3D reconstruction priors and incorporates human priors to ensure geometric accuracy and fine-grained texture details.
- We introduce a SMPL-X Estimation module to acquire a unified SMPL-X model from SMPL-X monocular estimations of different views under pose-free settings, providing a robust and consistent human prior.
- We design a Posemap Adapter module that injects structured human body information from the unified SMPL-X model into the stereo matching network, significantly enhancing both geometric accuracy and textural details in the reconstructed human model.

2 Related Works

2.1 3D Human Reconstruction

3D human reconstruction generates high-fidelity 3D models of the human from images or video sequences. Existing mainstream methods can be divided into two paradigms. The first paradigm is parametric model-based methods, which leverage parametric human models like SMPL Loper et al. (2023) or SMPL-X Pavlakos et al. (2019) as templates to provide anatomical and topological priors for human reconstruction. Early works in this line Peng et al. (2021); Qian et al. (2024); Kocabas et al. (2024); Li et al. (2024) combine the parametric templates with neural radiance fields (NeRF) Mildenhall et al. (2021) or 3D Gaussian Splatting (3DGS) Kerbl et al. (2023) to achieve photorealistic human reconstruction, but often require costly per-subject optimization for each new individual. Subsequent generalizable methods like GHG Kwon et al. (2024) and RoGSplat Xiao et al. (2025) regress 3DGS parameters directly from the templates and the input images, eliminating the need for per-subject tuning. However, during the reconstruction process, these methods require pre-calibrated camera parameters to project sampled points onto images or to unproject image data back into the parametric model. The second category is stereo matching-based methods, which reconstruct humans from calibrated multi-view inputs. The representative work GPS-Gaussian Zheng et al. (2024) and its variants Zhou et al. (2025); Tu et al. (2024); Hu et al. (2025) employ stereo matching networks to predict human point clouds and leverage UNet-style networks Ronneberger

et al. (2015) for 3DGS attributes regression, enabling high-fidelity reconstruction. However, these methods are generally limited to fixed camera setups, as their high-fidelity representations exhibit high sensitivity to parameter deviations, and thus require a separately trained network for each different camera setup. In conclusion, both paradigms advance fidelity but are limited by calibration dependency or fixed camera setups. Recently, the field of 3D human reconstruction is transitioning from controlled laboratory environments to consumer-level scenarios. The reliance of existing methods on pre-calibrated camera parameters severely hinders this ongoing transition. Since intrinsic camera parameters can be readily obtained from device specifications, the pose-free 3D human reconstruction framework remains crucial.

2.2 Pose-Free 3D Reconstruction

Early 3D reconstruction pipelines rely on accurate camera pose calibration, severely restricting their applicability to reconstruction from in-the-wild images with unknown capture parameters. Overcoming this limitation, pose-free 3D reconstruction has emerged as a rapidly advancing research direction in recent years. Recent method DUSt3R Wang et al. (2024), which learns dense cross-view pixel correspondences from massive diverse training scene data, reconstructs globally coherent 3D geometry from in-the-wild images in an end-to-end manner, eliminating the need for camera extrinsic parameters. Building on this framework, subsequent works extend the pose-free paradigm along multiple technical directions: VGGT Wang et al. (2025) and Fast3R Yang et al. (2025) generalize the original pairwise input design to multi-view settings for more robust and complete reconstruction, while LONG3R Chen et al. (2025) and Spann3R Wang & Agapito (2025) adapt the framework to streaming reconstruction for continuous image sequences. Another active research line integrates pose-free principles with 3DGSKerbl et al. (2023), as seen in NoPoSplat Ye et al. (2024) and Splatt3R Smart et al. (2024), to achieve pose-free novel view synthesis. The primary advantage of these approaches is their remarkable flexibility and ease of use, as they require neither a controlled capture setup nor pre-calibrated camera extrinsic parameters, and can be applied to internet data. However, such methods come with some limitations for 3D human reconstruction. First, reconstructions from these methods typically lack real-world metric measurements. Second, these methods are designed as general scene reconstruction systems and therefore lack the priors or architectural designs needed to capture fine-grained details and anatomical plausibility for high-fidelity human reconstruction. Their performance on humans is often coarser compared to methods that incorporate human-specific constraints or priors.

3 Method

3.1 Overview

Given multiple views of a human, S^{2}Human reconstructs the 3D human model without pre-calibrated camera extrinsic parameters. As shown in Fig. 2, S^{2}Human begins with SMPL-X Estimation, which recovers a unified SMPL-X model by fusing monocular predictions of each view. Next, a Human Gaussian Estimation stage uses Posemap Adapter to inject the SMPL-X prior presented by Tri-posemap into the stereo matching network and transforms the stereo features into a detailed 3DGS representation through the 3D Gaussian Regression module. Below we describe SMPL-X Estimation and Human Gaussian Estimation in detail.

3.2 SMPL-X Estimation

SMPL-X Estimation consists of two steps: first, performing independent monocular SMPL-X estimation for each view, and second, fusing these estimations into a unified model based on visibility.

3.2.1 Monocular SMPL-X Estimation.

S^{2}Human operates on a collection of input images, $I = \{I_1, I_2, ..., I_n\}$. In the first step, a monocular SMPL-X estimation model, $\mathcal{F}$, is employed to process each image independently:

$$\Theta_i = (\theta_i, \beta_i, \psi_i, r_i, t_i) = \mathcal{F}(I_i), \tag{1}$$

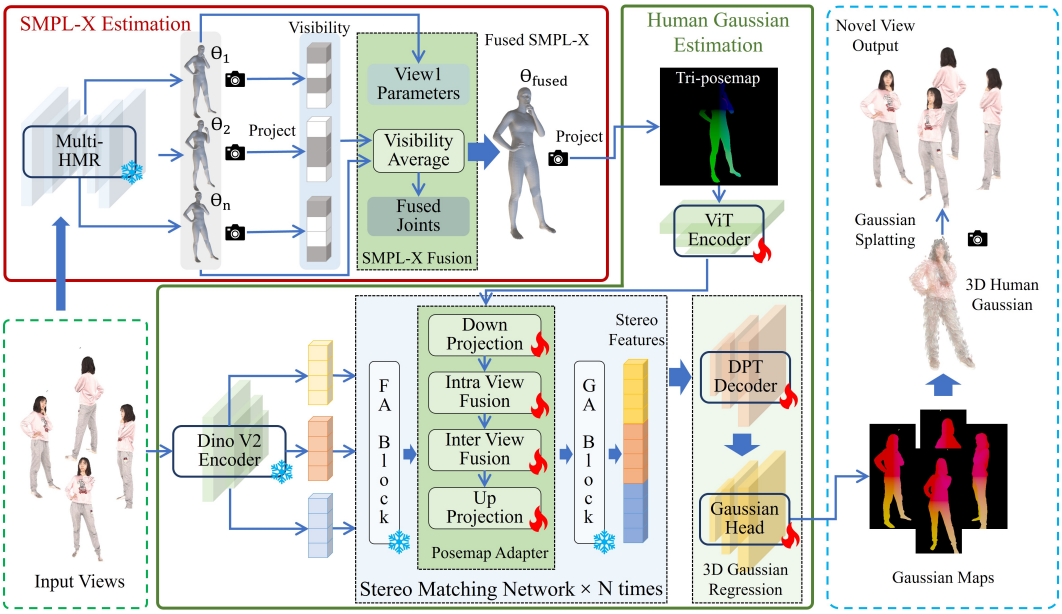

Figure 2: Overview of S^{2}Human. Given multi-view images, S^{2}Human reconstructs the 3D human model without pre-calibrated camera extrinsic parameters. Our pipeline consists of two stages: (1) SMPL-X Estimation, which obtains the SMPL-X model from multi-view images; (2) Human Gaussian Estimation, which reconstructs a detailed, renderable 3D Gaussian representation.

where Θ_i denotes the SMPL-X parameters (pose, shape, expression, global orientation, and global translation) of the i-th view. Specifically, S^{2}Human leverages the pre-trained Multi-HMR Baradel et al. (2024) model to implement the monocular SMPL-X estimation $\mathcal{F}$.

3.2.2 Multi-view SMPL-X Fusion.

S^{2}Human employs a visibility-aware fusion of the SMPL-X parameters to produce a unified SMPL-X model. Specifically, the fused SMPL-X parameters are obtained by performing a visibility-weighted average of the parameters estimated from all individual viewpoints. The weighting coefficients for each view are derived according to its joint visibility, which is computed as the logical AND of 3D and 2D visibility. A joint's 3D visibility is derived according to the number of intersections between the SMPL-X mesh and the ray from the camera origin to the joint. Since these joints lie on or within the mesh, an intersection count greater than two indicates that the joint is invisible. The 2D visibility is obtained by projecting the joints onto the image plane and checking whether the projected 2D points lie within the valid image boundaries. In the pose-free setup where the camera coordinate system of View 1 serves as the world coordinate system, The following parameters from View 1 are directly taken as the fused values including shape parameters, expression parameters, global orientation, global translation, as well as the pose parameters corresponding to the joints that are invisible across all input viewpoints (which cannot be effectively fused via the aforementioned visibility-aware fusion), ensuring coordinate system consistency throughout the fusion process. The fused SMPL-X parameters Θ_{fused} can be represented as:

$$\Theta_{fused} = (\theta_{fused}, \beta_1, \psi_1, r_1, t_1). \tag{2}$$

3.3 Human Gaussian Estimation

This stage bridges the unified SMPL-X model and the stereo matching network for the estimation of 3D Gaussian properties. The stereo matching network takes both the image features and the Tri-posemap features as input. The image features are extracted using DINOv2 Oquab et al. (2024). The Tri-posemap features are extracted from Tri-posemap which is generated from the unified SMPL-X model as a structured, pixel-aligned 2D representation. In the stereo matching network, the image



Figure 3: Qualitative comparison of multi-view reconstruction on THuman2.0.

Table 1: Quantitative comparison of multi-view reconstruction on THuman2.0 and Twindom. For GHG Kwon et al. (2024), results on Twindom could not be obtained due to its requirement for ground-truth SMPL-X parameters (which are unavailable in Twindom).

Method	THuman2.0→THuman2.0			THuman2.0→Twindom		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GPS Gaussian	23.53	**0.9315**	**0.0875**	22.56	**0.9127**	**0.1008**
EVA Gaussian	24.61	0.9261	0.1013	23.50	0.9064	0.1164
GHG	23.39	0.9184	0.0923	/	/	/
S^2Human(ours)	**25.93**	0.9287	0.0888	**24.02**	0.9026	0.1056

features are concatenated with camera and register tokens Wang et al. (2025), and fed into the Frame Attention (FA) block Wang et al. (2025). The Posemap Adapter then fuses the Tri-posemap features with FA features F_{FA}, forming the human-aware FA features F_{FA-H}. Then, F_{FA-H} are fed into the Global Attention (GA) block Wang et al. (2025) to obtain the human-aware GA features F_{GA-H}. Finally, the stereo features F_S which contain F_{FA}, F_{FA-H} and F_{GA-H} are decoded by a 3D Gaussian Regression module into per-pixel 3D Gaussian parameters, thereby accomplishing the 3D reconstruction of the human.

3.3.1 TRI-POSEMAP GENERATION.

The Tri-posemap serves as a structured, pixel-aligned 2D representation of SMPL-X priors. Specifically, the Tri-posemap is generated by projecting the SMPL-X model onto View 1 at the same resolution as the input images. In this map, the RGB channels encode the corresponding X, Y, Z world coordinates of the points on the SMPL-X surface. A fourth channel is added to encode the spatial relationship between the SMPL-X model and the actual clothed human in the image. This channel indicates whether the location contains both the SMPL-X and the actual clothed human,

Table 2: Quantitative comparison of reconstruction on THuman2.0 with different angular separations. Note, GPS-Gaussian Zheng et al. (2024) fails to work effectively when $\Delta = 90°$.

Method	$\Delta = 45°$			$\Delta = 60°$			$\Delta = 90°$		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GPS Gaussian	25.88	**0.9414**	**0.0715**	23.49	**0.9302**	**0.0893**	/	/	/
EVA Gaussian	25.91	0.9326	0.0917	24.45	0.9250	0.1029	21.41	0.9033	0.1314
GHG	23.71	0.9219	0.0918	23.50	0.9194	0.0937	23.25	0.9165	**0.0950**
S^2Human(ours)	**26.07**	0.9303	0.0847	**25.26**	0.9263	0.0934	**23.55**	**0.9176**	0.1103

only one of them, or none of them, corresponding to the encoded values 1, 0.5, or 0, respectively. Subsequently, a Vision Transformer (ViT) is employed to extract features from the Tri-posemap along with the camera and register tokens of View 1. These features will then be integrated via Posemap Adapter to effectively incorporate human priors into the stereo matching network.

3.3.2 POSEMAP ADAPTER.

The Posemap Adapter module injects the Tri-posemap into the stereo matching network. This adapter takes both the FA features F_{FA} and the Tri-posemap features as input. Specifically, Posemap Adapter begins by adding the Tri-posemap features to their corresponding viewpoint's F_{FA}. The fused features are then projected into a lower-dimensional space. This lower-dimensional representation subsequently passes through two specialized fusion blocks: an intra-viewpoint fusion block and an inter-viewpoint fusion block. The intra-viewpoint block applies self-attention to the features of each individual viewpoint, enabling effective comprehension and integration of the Tri-posemap information within local contexts. The inter-viewpoint block then performs self-attention across all viewpoints, allowing the human priors to propagate and fuse globally throughout the entire set of views. The refined features are first up-projected back to the original dimensionality. They are then passed through a zero-initialized convolutional layer and added to F_{FA} as a residual connection forming the human-aware FA features F_{FA-H}.

Table 3: Quantitative results of the SMPL-X Estimation module analysis.

Input view	No.1	No.2	No.3	No.4	All views
Angular Error(rad)↓	0.3092	0.3155	0.3185	0.3171	**0.2853**

3.3.3 3D GAUSSIAN REGRESSION.

The 3D Gaussian Regression module translates the stereo features F_S into explicit Gaussian Maps, pixel-aligned representations where each pixel's channels encode the parameters of a 3D Gaussian. The module begins with the Dense Prediction Transformer (DPT) network Ranftl et al. (2021), which decodes F_S into dense, pixel-aligned feature maps, preserving multi-scale spatial context crucial for high-fidelity reconstruction. Subsequently, a convolution-based Gaussian Parameter Head regresses these feature maps and the input images into the Gaussian Maps. For each input view I_k, the head outputs a corresponding map G_k in which every pixel (u, v) stores the full set of 3D Gaussian parameters:

$$\mathbf{G} = \{\mu,\ \Sigma,\ \mathbf{c},\ \alpha\} = \mathrm{Head}_g(\mathbf{F}_{DPT},\ I). \tag{3}$$

Here, μ, Σ, $\mathbf{c}$, α denote the Gaussian's 3D center, covariance matrix, color, and opacity, respectively. Notably, the color parameter $\mathbf{c}_k(u, v)$ is sampled directly from the input image $\mathrm{I}_k(u, v)$. The complete 3D Gaussian model is obtained by aggregating the parameters from all human pixels across all Gaussian Maps.

3.4 TRAINING AND OPTIMIZATION

S^2Human is trained in two stages using the ground-truth SMPL-X input. The first stage regresses only the Gaussian centers under geometric supervision. In the second stage, the stereo matching network and the 3D Gaussian Regression module for the Gaussian centers are frozen for

the first 3 epochs, and then unfrozen to jointly optimize all Gaussian parameters via a combined photometric-geometric loss.

3.4.1 Loss Function for Stage 1.

The first stage applies point map loss $\mathcal{L}_{pmap}$ and camera parameter loss $\mathcal{L}_{camera}$ to measure the difference between the output point cloud and the ground truth. $\mathcal{L}_{pmap}$ is computed as the L2 distance between the regressed point cloud and the ground truth point cloud, where the ground truth is obtained through back-projection of depth maps. Additionally, a camera head Wang et al. (2025) is employed to estimate camera extrinsics, aiding training and providing a metric prior for evaluation-time pose alignment. $\mathcal{L}_{camera}$ is computed as the L1 loss between the predicted camera extrinsics and the ground truth extrinsics. The loss function for stage 1 can be presented as:

Table 4: Quantitative results of the Posemap Adapter module analysis. Note that the L2 Distance of the full model is not compared with other settings, as the scale of the SMPL-X estimates from the full pipeline may be inconsistent with the ground truth.

Method	THuman2.0→THuman2.0			
	PSNR↑	SSIM↑	LPIPS↓	L2(cm)↓
w/o adapter	25.68	0.9268	0.0893	2.27
full method	25.93	0.9287	0.0888	/
full method (gt SMPL-X)	**26.54**	**0.9331**	**0.0834**	**1.27**

$$\mathcal{L}_{stage1} = \mathcal{L}_{pmap} + \lambda_1 \mathcal{L}_{camera}. \tag{4}$$

3.4.2 Loss Function for Stage 2.

In the second stage, the loss combines a view-consistency term and a per-view photometric term. The view-consistency term measures the error at intermediate viewpoints rendered from Gaussians of all input views, while the per-view term evaluates each input viewpoint rendered from Gaussians of its own Gaussian Map. These two terms are weighted with a 2:1 ratio to form the final L1 loss $\mathcal{L}_{mae}$ and SSIM loss $\mathcal{L}_{SSIM}$. Additionally, during the first few epochs, $\mathcal{L}_{pmap}$ is added to prevent geometry degradation due to the high degrees of freedom in 3D Gaussians. The loss function for stage 2 can be presented as:

$$\mathcal{L}_{stage2} = \lambda_2 \mathcal{L}_{mae} + \lambda_3 \mathcal{L}_{SSIM} + \lambda_4 \mathcal{L}_{pmap}. \tag{5}$$

4 Experiments

4.1 Experiment Setup

4.1.1 Implementation Details.

S^{2}Human is trained on one NVIDIA 4090 GPU with AdamWLoshchilov & Hutter (2017), using 518×518 images as input and 1024×1024 images for rendering supervision. The training comprises two stages: Stage 1 with a 5e-5 learning rate for 10 epochs, and Stage 2 with a 5e-6 learning rate for 16 epochs. Throughout training, the loss weights are empirically set to $\lambda_1 = 5$, $\lambda_2 = 0.8$, and $\lambda_3 = 0.2$. Additionally, λ_4 is empirically set to 1 for the first 6 epochs of Stage 2 and then to 0 for the remaining epochs. The stereo matching network contains 24 layers, with Posemap Adapters inserted only in four of them (the 5-th, 12-th, 18-th, and 24-th); the stereo features from these four layers are all passed to the subsequent DPT for decoding. Notably, the DINOv2 encoder, FA blocks, GA blocks, the camera tokens, the register tokens, the 3D Gaussian Regression module for the Gaussian centers and the camera head are initialized with the pretrained weights from VGGT Wang et al. (2025). The DINOv2 encoder, the camera tokens, the register tokens, FA blocks and GA blocks are kept frozen during training. Additionally, S^{2}Human assumes fixed camera intrinsics corresponding to an FOV of 60° throughout training and inference. For evaluation, each reported result is obtained from a single run of the corresponding method on each test sample.

4.1.2 Evaluation-Time Pose Alignment.

Following previous pose-free reconstruction works Fan et al. (2025); Ye et al. (2024), the camera pose for the target view is optimized during evaluation time to fairly compare with other human

reconstruction works with ground truth camera poses. After obtaining the 3D Gaussians with our method, we freeze the 3D Gaussians and optimize the target camera pose to align the rendered image with the ground truth. It is important to note that this optimization is conducted exclusively for current benchmarking objective and is not required in real-world application scenarios.

4.1.3 DATASETS.

Experiments are conducted on two 3D human reconstruction datasets, THuman2.0 Yu et al. (2021) and Twindom Twindom (2020), which provide high-quality scanned humans with diverse body shapes, poses, and clothing styles for evaluation. THuman2.0 includes 526 instances, with 426 used for training and 100 for evaluation, while 200 Twindom instances are used for cross-domain evaluation.

4.2 COMPARISON STUDY

The proposed method is compared with representative state-of-the-art pose-aware 3D human reconstruction approaches, including GPS-Gaussian Zheng et al. (2024), EVA-Gaussian Hu et al. (2025), and GHG Kwon et al. (2024). Evaluations are conducted under both in-domain and cross-domain settings, demonstrating competitive performance against existing pose-aware methods.

4.2.1 IN-DOMAIN RECONSTRUCTION COMPARISON.

The in-domain reconstruction comparison comprises two components. The first is a multi-view reconstruction experiment: the experiments take images from four viewpoints with a 60-degree angular separation as input, and assess performance using the rendering quality of intermediate viewpoints. The quantitative results presented in Table 1 demonstrate that in the multi-view reconstruction scenario, S^{2}Human achieves the best performance in terms of PSNR and remains competitive with pose-aware methods across other metrics (SSIM and LPIPS). Fig. 3 shows that S^{2}Human achieves competitive results under geometric correctness and fine-grained texture details. The second component involves a reconstruction experiment under varying angular separations. In this setup, two viewpoints are used as input, with angular separations of 45, 60, and 90 degrees between them. The rendering quality of the intermediate viewpoints similarly serves as the evaluation metric. The results presented in Table 2 demonstrate that S^{2}Human maintains robust performance across varying angular baselines, remaining competitive with pose-aware methods on PSNR, SSIM and LPIPS. Additional qualitative results are provided in Appendix A.

Table 5: Quantitative results of reconstruction on THuman2.0 under varying intrinsic settings. Note, to eliminate the confounding factor of varying human footprint caused by different FOVs, the extrinsic parameters are adjusted so that the human's footprint in the images approximately matches across different FOVs.

Field of View	THuman2.0→THuman2.0		
	PSNR↑	SSIM↑	LPIPS↓
50°	27.12	0.9521	0.0638
60°	**27.61**	**0.9537**	**0.0609**
70°	26.92	0.9507	0.0646
60° (original)	25.93	0.9287	0.0888

4.2.2 CROSS-DOMAIN RECONSTRUCTION COMPARISON.

The cross-domain reconstruction comparison is conducted to evaluate the cross-domain generalization capability of S^{2}Human, following a strict zero-shot evaluation protocol: all models in this comparison are trained solely on the source dataset THuman2.0, and directly tested on the target dataset Twindom with no fine-tuning, domain adaptation, or prior exposure to the target domain data. To ensure the fairness and comparability of the experiment, we adopt exactly the same evaluation setup as the in-domain multi-view reconstruction experiment. As demonstrated by the quantitative results in Table 1 , the proposed S^{2}Human attains the best PSNR performance among all compared methods, and remains competitive on other metrics with pose-aware human reconstruction approaches. Additional qualitative results are provided in Appendix A.

4.3 Method Analysis

4.3.1 Analysis of the SMPL-X Estimation Module.

This subsection evaluates the effectiveness of the proposed SMPL-X Estimation module by comparing the fused SMPL-X output with monocular predictions from individual viewpoints. Experiments are conducted on THuman2.0 with four input views separated by 60 degrees. The fused and monocular SMPL-X parameters are quantitatively compared using joint rotation angular error. As shown in Table 3, the proposed module reduces monocular estimation ambiguities and achieves more accurate SMPL-X parameters.

4.3.2 Analysis of the Posemap Adapter Module.

A comprehensive ablation study is conducted on THuman2.0 under the same evaluation setup as the in-domain multi-view experiment to validate the effectiveness of the Posemap Adapter module. The full model is compared against a variant constructed by removing the Posemap Adapter module. The full model is also evaluated using ground-truth SMPL-X parameters. The point clouds from the first stage are evaluated using their distance to the ground truth to further demonstrate the adapter's contribution to 3D human point cloud quality. The results in Table 4 and Fig. 4 confirm that the Posemap Adapter enhances the model's comprehension of human body and fine-grained details, leading to superior point cloud reconstruction and rendering performance. The results also demonstrate that more accurate SMPL-X estimation is positively correlated with higher-quality rendering and reconstruction.

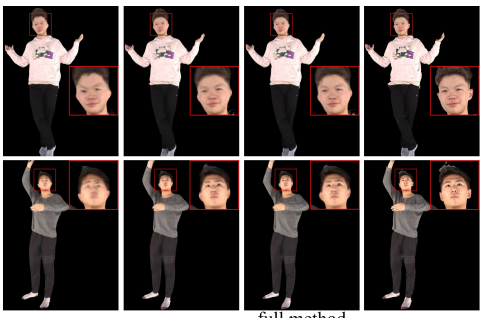

Figure 4: Qualitative results of the Posemap Adapter module analysis.

4.3.3 Analysis of Robustness Under Different Intrinsic Settings

This subsection evaluates the robustness of S^2Human to input images captured with varying intrinsic parameters on THuman2.0 under the same evaluation setup as the in-domain multi-view experiment. Input images with FOV of 50°, 60°, and 70° are used. To eliminate the confounding factor of varying human footprint caused by different FOVs, the extrinsic parameters are adjusted so that the human's footprint in the images approximately matches across different FOVs. This adjustment is achieved by scaling the camera-to-subject distances proportionally to the focal length. Throughout all experiments, S^2Human assumes a fixed intrinsic configuration corresponding to an FOV of 60° during reconstruction. The results in Table 5 demonstrate that S^2Human consistently produces plausible reconstructions across all these settings, exhibiting robustness to variations in camera intrinsics. Additional qualitative results are provided in Appendix A.

5 Conclusion

This paper presents S^2Human, a pose-free 3D human reconstruction framework that bridges SMPL-X and stereo matching for high-fidelity reconstruction. Its core innovations include: a visibility-aware SMPL-X estimation module that fuses monocular predictions of each view to obtain a unified SMPL-X model, providing a robust and consistent human prior, and a Posemap Adapter that injects the Tri-posemap which is a structured, pixel-aligned 2D representation of the SMPL-X priors, into the stereo matching network via intra-view and inter-view attention mechanisms, enhancing geometric accuracy and texture details. Experiments on THuman2.0 and Twindom show that S^2Human achieves competitive performance with state-of-the-art pose-aware methods under both in-domain and cross-domain settings. This work breaks the extrinsic calibration dependency of existing methods and provides a paradigm for integrating human priors into general pose-free stereo matching network offering a practical and scalable solution for pose-free 3D human reconstruction.

AI USE STATEMENT

In this work, we used generative AI tools to aid language polishing and improve the readability of the manuscript. We have reviewed all AI-assisted work and verified the resulting edits for correctness and consistency with the authors' intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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*, pp. 202–218. Springer, 2024.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5273–5284, 2025.

Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, et al. Instantsplat: Sparse-view gaussian splatting in seconds, 2025. URL https://arxiv.org/abs/2403.20309.

Yingdong Hu, Zhening Liu, Jiawei Shao, Zehong Lin, and Jun Zhang. Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse multi-view camera settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2613–2622, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Yu-Hee Kim, In Park, Soo Buem Cho, Seoyon Yang, Il Kim, Kyong-Ha Lee, et al. Three-dimensional virtual reconstructions of shoulder movements using computed tomography images: Model development. *Interactive Journal of Medical Research*, 12(1):e48381, 2023.

Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 505–515, 2024.

Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, et al. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, pp. 451–468. Springer, 2024.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19711–19722, 2024.

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*, pp. 851–866. 2023.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, 2017.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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*, pp. 10975–10985, 2019.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9054–9063, 2021.

Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5020–5030, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12179–12188, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Brandon Smart, Chuanxia Zheng, Iro Laina, and Victor Adrian Prisacariu. Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs, 2024. URL https://arxiv.org/abs/2408.13912.

Hanzhang Tu, Ruizhi Shao, Xue Dong, Shunyuan Zheng, Hao Zhang, Lili Chen, et al. Tele-aloha: A telepresence system with low-budget and high-authenticity using sparse rgb cameras. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Twindom. Twindom dataset. https://web.twindom.com, 2020. Accessed: 2020-07-19.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Junjin Xiao, Qing Zhang, Yonewei Nie, Lei Zhu, and Wei-Shi Zheng. Rogsplat: Learning robust generalizable human gaussian splatting from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5980–5990, 2025.

Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, et al. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21924–21935, 2025.

Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images, 2024. URL https://arxiv.org/abs/2410.24207.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5746–5756, 2021.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19680–19690, 2024.

Boyao Zhou, Shunyuan Zheng, Hanzhang Tu, Ruizhi Shao, Boning Liu, Shengping Zhang, et al. Gps-gaussian+: Generalizable pixel-wise 3d gaussian splatting for real-time human-scene rendering from sparse views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

A SUPPLEMENTARY MATERIAL

A.1 ADDITIONAL QUALITATIVE RESULTS OF CROSS-DOMAIN RECONSTRUCTION

The qualitative results in Fig. 5 of the Supplementary Material correspond to the cross-domain reconstruction comparison described in the main paper. All models are trained solely on THuman2.0 and directly evaluated on Twindom without fine-tuning, domain adaptation, or prior exposure to the target domain data. The figure provides visual comparisons of the reconstruction results on Twindom, complementing the quantitative evaluation reported in Table 1 of the main paper. As shown in Fig. 5 of the Supplementary Material, S^2Human produces plausible reconstruction results on the unseen Twindom dataset and remains competitive with existing pose-aware human reconstruction methods, demonstrating its cross-domain generalization capability.

A.2 ADDITIONAL QUALITATIVE RESULTS UNDER DIFFERENT INTRINSIC SETTINGS

The qualitative results in Fig. 6 of the Supplementary Material correspond to the robustness analysis under different intrinsic settings described in the main paper. The figure presents reconstruction results from input images captured with different camera intrinsic configurations, complementing the quantitative evaluation reported in Table 5 of the main paper. As shown in Fig. 6 of the Supplementary Material, S^2Human produces plausible reconstruction results across different intrinsic configurations. The results show that the method remains robust to variations in camera intrinsics, which is consistent with the quantitative evaluation in Table 5 of the main paper.

A.3 ADDITIONAL QUALITATIVE RESULTS UNDER DIFFERENT ANGULAR SEPARATIONS

The qualitative results in Fig. 7 of the Supplementary Material correspond to the reconstruction experiment with varying angular separations described in the main paper. The figure provides visual comparisons of intermediate-view rendering results under different input view separations, complementing the quantitative evaluation reported in Table 2 of the main paper. As shown in Fig. 7 of the Supplementary Material, S^2Human produces plausible reconstruction results under different angular settings and remains competitive with existing pose-aware methods.

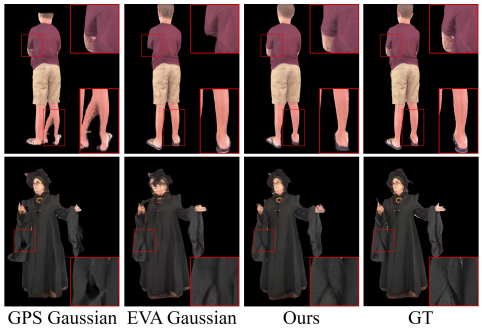

Figure 5: Qualitative comparison on Twindom.

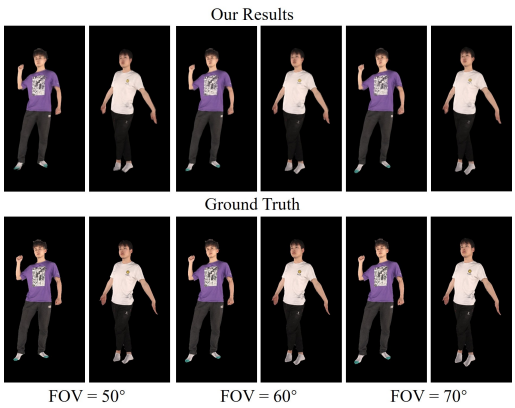

Figure 6: Qualitative comparison of reconstruction on THuman2.0 under varying intrinsic settings.

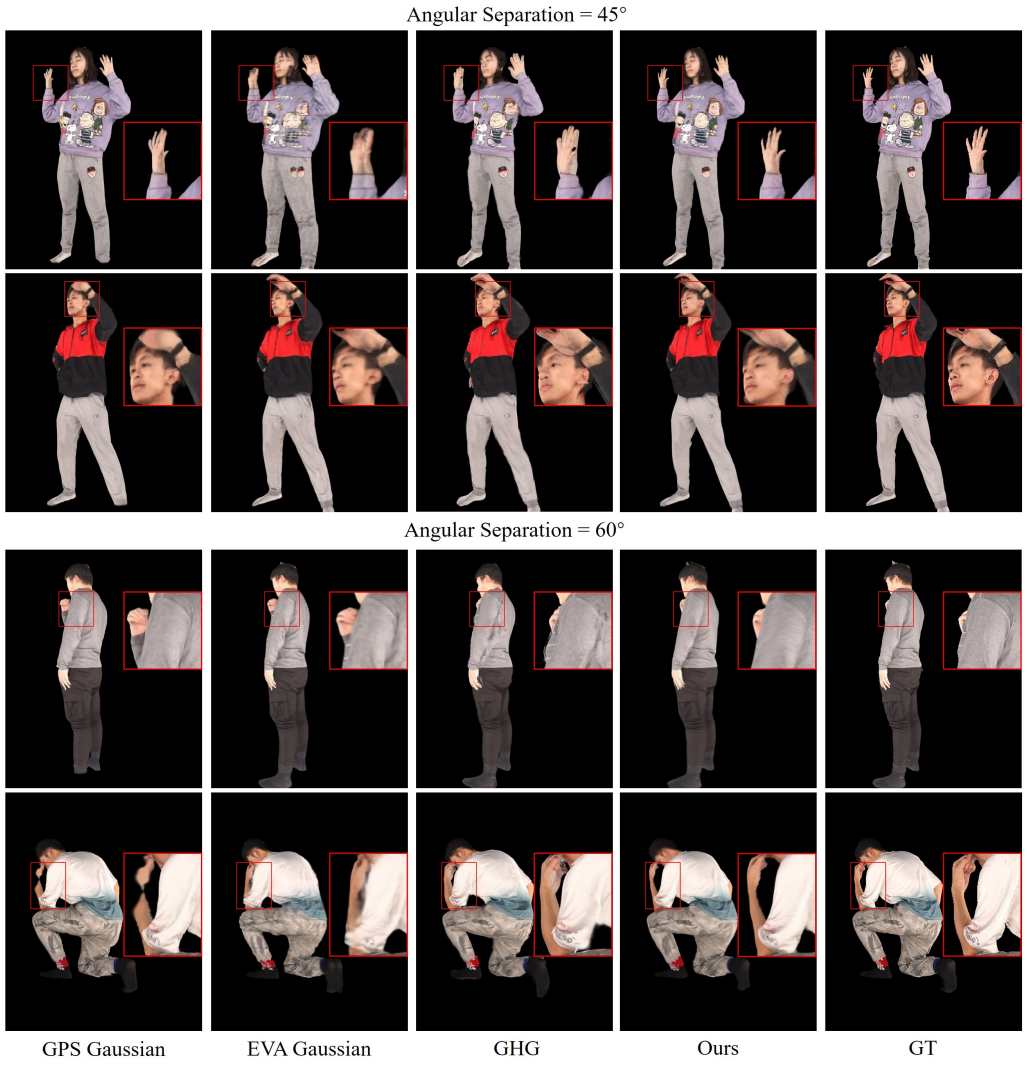

Figure 7: Qualitative comparison of reconstruction on THuman2.0 under varying angular separations.