GST: Precise 3D Human Body from a Single Image with Gaussian Splatting Transformers

Lorenza Prospero Abdullah Hamdi Joao F. Henriques Christian Rupprecht
Visual Geometry Group, University of Oxford
{lorenza,abdullah,joao,chrisr}@robots.ox.ac.uk

Abstract

Reconstructing realistic 3D human models from monocular images has significant applications in creative industries, human-computer interfaces, and healthcare. We base our work on 3D Gaussian Splatting (3DGS), a scene representation composed of a mixture of Gaussians. Predicting such mixtures for a human from a single input image is challenging, as it is a non-uniform density (with a many-to-one relationship with input pixels) with strict physical constraints. At the same time, it needs to be flexible to accommodate a variety of clothes and poses. Our key observation is that the vertices of standardized human meshes (such as SMPL) can provide an adequate density and approximate initial position for Gaussians. We can then train a transformer model to jointly predict comparatively small adjustments to these positions, as well as the other Gaussians' attributes and the SMPL parameters. We show empirically that this combination (using only multi-view supervision) can achieve fast inference of 3D human models from a single image without test-time optimization, expensive diffusion models, or 3D points supervision. We also show that it can improve 3D pose estimation by better fitting human models that account for clothes and other variations. The code is available on the project website https://abdullahamdi.com/gst/.

1. Introduction

Reconstructing realistic 3D human models from monocular images is crucial for virtual reality and creative industries, as well as possibly improving human pose estimation for human-computer interfaces and health applications. It is also an integral component of "3D spatial computing" for mainstream consumer products incorporating 3D vision in VR and augmented reality (AR). To be practical in homes, offices, and workplaces, these products require precise 3D rendering, speed, compactness, flexibility, and realism. However, reconstructing 3D models from a single image remains challenging. Previous approaches that have shown success in this problem often utilize 2D priors in a multi-view setup

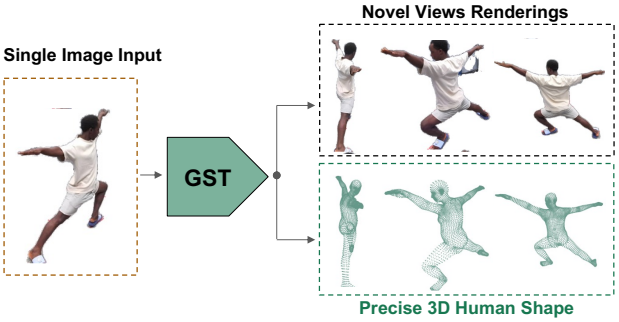

Figure 1. **Gaussian Splatting Transformers**. We propose a method that, given a *single* input image, predicts an accurate 3D human pose and shape, along with a color model that enables novel view rendering, including clothing. GST relies solely on multi-view supervision (no direct 3D supervision).

[8, 14, 20, 32, 36, 38, 40, 45–47, 57]. Specifically, in human modeling, issues arise due to intricate 3D details such as facial features, clothing, and joints, which present challenges for deep learning methods. Early methods addressed these challenges using a learned Signed Distance Function (SDF) on a human template to predict detailed 3D meshes [53, 70]. Later works incorporated Neural Radiance Fields (NeRFs) to capture texture details [24, 61], or leveraged pre-trained diffusion models to generate dense views from a single frontal image, reducing prediction ambiguity [7, 16, 22, 61, 65, 71]. However, they typically suffer from low speed, hindering real-time deployment.

In this work, we present GST (Gaussian Splatting Transformer), illustrated in Fig. 1, a direct method that learns to predict 3D Gaussian Splatting [28] for 3D representation, allowing for fast rendering and flexible editing abilities compared to others. Our method does not rely on diffusion priors and is, therefore, capable of near real-time predictions. This is essential for downstream applications and ensures that the inference can be easily incorporated with prior-based approaches. GST leverages multi-view supervision instead of precise (and expensive) 3D point clouds. Despite this, it

Table 1. **Single Image 3D Human Reconstruction Methods**. Comparison of various 3D representation models, highlighting key attributes such as speed, method of obtaining the model, type of 3D representation, usage of diffusion prior, and supervision technique.

Method	Speed	Obtained by	3D Representation	Diffusion Prior	Supervision
PIFU [53]	10 seconds	Inference	SDF	✗	Direct 3D
HumanLRM [61]	7 seconds	Inference	NeRF (Triplane)	✓	Direct 3D + MV
SiTH [22]	2 minutes	Inference	SDF	✓	Direct 3D
SIFU [71]	6 minutes	Optimization	SDF	✓	Direct 3D
GTA [70]	0.55 seconds	Optimization	SDF	✗	Direct 3D
SHERT [65]	23 seconds	Inference	Mesh	✓	Direct 3D
R2Human [13]	0.04 seconds	Inference	NeRF (MLP)	✗	Direct 3D
ConTex-Human [16]	60 minutes	Optimization	NeRF+Mesh	✓	Direct 3D
Ultraman [7]	20 minutes	Optimization	Mesh	✓	✗
ANIM [44]	few seconds	Inference	SDF	✗	Direct 3D
SHERF [24]	0.75 seconds	Inference	NeRF (MLP)	✗	Multi-View
GST (ours)	**0.02s seconds**	Inference	Gaussian Splatting	✗	Multi-View

predicts accurate 3D joint and body poses while maintaining the perceptual quality of renderings from novel views. Table 1 summarizes the characteristics of prior work.

GST is inspired by recent works on single view 3D reconstruction [56]. However, the complexity of human pose in 3D space poses significant challenges to the direct applications of methods that associate one 3D point (or Gaussian) to each pixel. Therefore, we augment our model also to predict the pose parameters of the SMPL [37] model. The SMPL model is used as the *scaffolding* on which the Gaussians are positioned and rendered. Each Gaussian is loosely tied to a vertex on the SMPL model by an offset. This has two advantages. First, it provides a good initialization of the density and pose of the Gaussians, including back faces, which are notoriously difficult for single-view methods. Second, we find that the joint optimization of pose and appearance also improves the SMPL pose prediction.

To the best of our knowledge, GST is the first work that efficiently combines accurate 3D human prediction with improved visual quality, utilizing only multi-view supervision and without relying on diffusion priors. In summary, our contributions are the following:

1) We propose GST, a 3D human model prediction method that does not rely on diffusion priors and performs novel view synthesis from a single image input. This makes it particularly amenable to real-time modeling tasks, where multiple views are uneconomical or impractical.

2) We evaluate our method and compare it to other state-of-the-art models. Although prior methods only solve for 3D pose estimation or 3D reconstruction, our method still performs equally or better on perceptual *and* 3D pose estimation metrics *without 3D supervision*.

2. Related work

3D Reconstruction using Image Priors. In the area of prior-based 3D reconstruction, contemporary zero-shot text-to-image generators [2, 14, 48, 49, 51, 52] have shown significant improvement by leveraging enhanced synthesis priors [6, 9, 39, 45, 60]. DreamFusion [45] stands out as a pioneering work that distilled a pre-existing diffusion model [52] into a NeRF framework [3, 41] using text prompts. This innovation spurred further research in both text-to-3D synthesis [8, 32] and image-to-3D reconstruction [36, 38, 46, 54]. The latter approach leverages supplementary reconstruction losses focused on frontal camera perspectives [36] and subject-specific diffusion guidance [46, 47]. In addition, task-specific priors have been explored [23, 26, 50], as well as additional control mechanisms [40]. More recently, Gaussian-Splatting approaches [21, 28] have improved the efficiency of 3D generation optimization through rapid Gaussian Splatting rasterization [57, 58, 68]. In contrast, our method GST does not rely on diffusion priors, providing a simpler and more cost-effective framework specialized for human 3D reconstruction and jointly predicts the precise internal *3D human body joints*.

3D Human Pose Estimation. Many approaches in the literature focus on predicting 3D human pose and shape from a single image [10, 18, 27, 30, 33, 34]. Relevant for our work are the approaches that directly regress the body shape and pose from a single image. The first work to introduce this approach was HMR [27], which uses a CNN to regress SMPL [37] parameters. Dedicated designs have been proposed for the HMR architecture; HoloPose [19] suggests a pooling strategy based on the 2D locations of body joints, while HKMR [17] relies on SMPL hierarchical structure to make predictions. PARE [29] introduces a body-part-guided attention mechanism to handle occlusions better, and PyMAF [66, 67] incorporates a mesh alignment module for

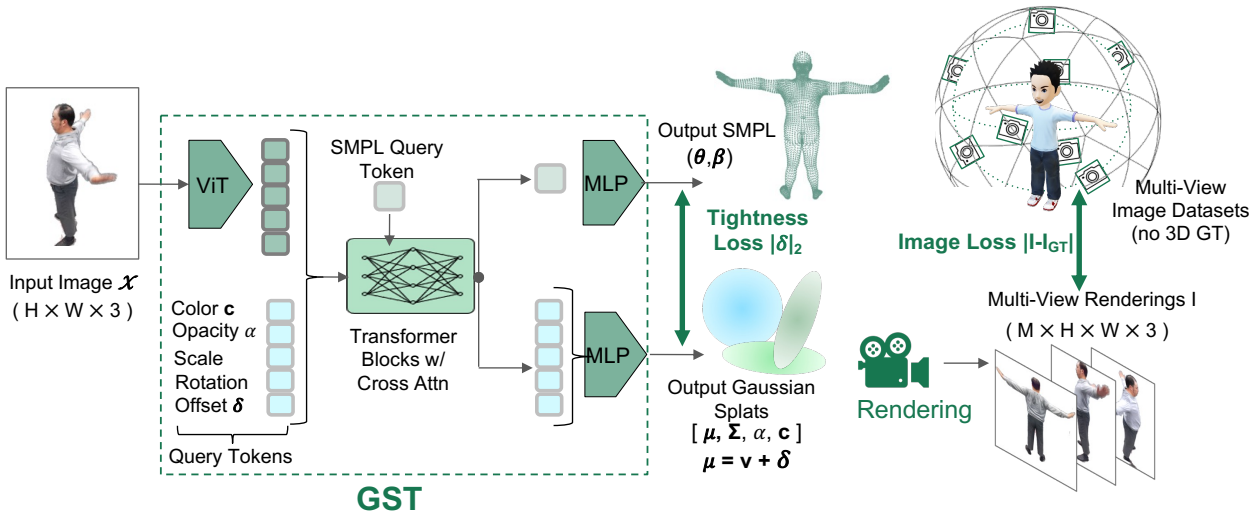

Figure 2. **Overview of the pipeline of Gaussian Splatting Transformer (GST)**. Given a single input image, GST uses a Vision Transformer (ViT) to predict both a 3D human pose (in the form of SMPL parameters) and a refined full-color 3D model (in the form of 3D Gaussian Splats). Additional input tokens facilitate the output of individual Gaussians' color $\mathbf{c}$, opacity α, scale, rotation, and position offset $\boldsymbol{\delta}$. Each Gaussian's position $\boldsymbol{\mu}$ is relative to one vertex of the SMPL model $\mathbf{v}$ by the offset $\boldsymbol{\delta}$, and so this model can be considered a refinement or residual over the interpretable SMPL mesh, facilitating multi-view rendering with higher visual fidelity.

SMPL parameter regression. More recently HMR2 [18] utilizes a transformer to predict the SMPL parameters and train on a large pool of 3D data and unprotected 2D joint labels, while TokenHMR [12] improved HMR2 by leveraging tokenized encoding and reduced the 2D basis in training HMR2. In contrast to all of these methods, our GST does not rely on 3D supervision and utilizes novel view synthesis training of a transformer to predict the Gaussian splats that are grounded on predicted SMPL parameters.

Monocular 3D Human Reconstruction. Recent advancements in 3D human reconstruction from single images have resulted in diverse methods, each employing different data sources, 3D representations, and supervision strategies [7, 13, 16, 22, 24, 25, 44, 61, 64, 65, 70, 71]. PIFU [53] is one of the earlier works that successfully uses the learned Sign Distance Function (SDF) representation with direct 3D supervision to reconstruct a detailed 3D mesh of humans from a single image. ANIM [44] incorporates sparse voxel depth features with the input image features and uses direct 3D supervision to train on the RGBD input (depth is needed). SHERF [24] developed on PIFU's 3D representation and adopted a NeRF representation for decoding the 3D human, training with multi-view supervision. While GST follows SHERF in the multi-view supervision, we utilize the more explicit Gaussian Splatting representation [28] for 3D, allowing for more flexible control and better 3D alignment. Similar to our method, A-NeRF [55] jointly optimizes human pose and 3D reconstruction, however it takes videos as input and does not generalise to unseen subjects.

With the recent wave of success of generative image and text models [42, 48, 49, 51, 52], several methods try to leverage these foundation models to improve the performance of 3D human reconstruction [7, 13, 16, 22, 61, 71]. For example, SIFU [71] integrates GPT-predicted captions with diffusion models for back-view generation and texture refinement and builds on GTA [70], its predecessor, which learns a triplane SDF decoder. Similarly, SiTH [22] employees Diffusion-prior to generate back views and decode the SDF and texture colours, while HumanLRM [61] generate multi-views with pre-trained diffusion and then train a tri-plane NeRF Large Reconstruction Model (LRM) for decoding. SHERT [65] utilizes semantic mesh and whole texture inpainting with the help of diffusion priors to create detailed 3D Mesh of humans. Most of these methods use pre-trained diffusions to improve the texturing with optimization, which slows down the process and prevents scaling for long videos, unlike our GST, which does not use any diffusion priors and runs at almost real-time inference, allowing for flexibility and potential integration with priors.

3. Gaussian Splatting Transformers (GST)

This section presents our methodology for reconstructing 3D human models from a single image using Gaussian Splatting Transformers (GST), as illustrated in Fig. 2.

3.1. Architecture

Our model predicts 3D Gaussian splatting parameters from a single input image using a transformer architecture, including tokenization, processing through transformer blocks, and

decoding into Gaussian parameters. We detail the model architecture (3.1) and the loss functions (3.2).

Image Encoder Architecture. Our backbone follows HMR2 [18] and uses a ViT [11] to map an image to a series of visual tokens. The input is an RGB image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, which is divided into non-overlapping patches $\mathbf{p}_j \in \mathbb{R}^{p \times p \times 3}$, with $j \in \{1, \dots, HW/p^2\}$. The patches are vectorized and affinely transformed into patch tokens $\mathbf{x}_j \in \mathbb{R}^d$.

The patch tokens are processed through a series of Transformer blocks [59]. The final output is a set of tokens $\mathbf{y}_j \in \mathbb{R}^d$ encapsulating the transformed image information.

Human Shape Representation. The SMPL model [37] represents the 3D human mesh shape as a mesh. SMPL is a low-dimensional parametric model defined by pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{24 \times 3 \times 3}$ and shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, outputting mesh vertices' 3D positions $\mathbf{v} = \text{SMPL}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{6890 \times 3}$.

Decoder Architecture. We build on HMR2 [18], which predicts the SMPL representation $(\boldsymbol{\theta}, \boldsymbol{\beta})$ from the image representation $\mathbf{y}_j$ through a cross-attention mechanism. Specifically, a single (fixed) token t_{SMPL} attends to all image tokens $\mathbf{y}_j$ through a series of cross-attention layers. An MLP decodes the token into the pose parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

This representation could be learned with image-pose pairs $(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\beta})$. However, here we focus on multi-view supervision, as 3D supervision is costly and scarce.

To train with multi-view supervision, the model needs to generate an image. We use recent advances in fast neural rendering: Gaussian Splatting [28]. This scene representation is defined by a set of 3D Gaussians, each characterized by a mean position $\boldsymbol{\mu} \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3 \times 3}$, the opacity $\alpha \in \mathbb{R}$ and a colour $\mathbf{c} \in \mathbb{R}^3$.

We link the 3D body shape and pose with the Gaussian scene representation, such that each vertex $\mathbf{v}_n$ in the mesh is assigned a Gaussian $G_n = (\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \alpha_n, \mathbf{c}_n)$. We allow the Gaussians to move away from the original vertex positions by a learned offset $\boldsymbol{\delta}_n$ to model clothes and other visual shape features that the SMPL model cannot capture.

$$\boldsymbol{\mu}_n = \mathbf{v}_n + \boldsymbol{\delta}_n, \tag{1}$$

This combination ensures the 3D model captures both shape and appearance, allowing more realistic reconstructions.

Similar to prior work [56], we factorize and simplify the covariance into the product of a rotation matrix and a diagonal matrix, enforcing a reduced number of degrees of freedom from 9 to 6: $G_n \in \mathbb{R}^{14}$.

It is theoretically possible to assign five tokens per Gaussian, one for each parameter: rotation, offset, scale, color, and opacity. However, this would result in over 34k tokens, which is computationally infeasible to decode with a standard Transformer. We thus group vertices into K groups, reducing the number of tokens to $5K + 1$ (in practice, we set $K = 26$). As discussed before, the additional token is used to predict the SMPL shape parameters.

This representation allows initialization with the pretrained weights of HMR2 [18] since we only introduce additional (fixed but learned) tokens in the decoder architecture. A set of Gaussians can be assembled from the predictions, which can be rendered into an image from any viewpoint.

3.2. Loss Functions

We use a combination of losses to train our model to ensure accurate and visually realistic 3D reconstructions.

Image Reconstruction Loss. We use a combination of Mean Squared Error (MSE) to measure the difference between the M multi-view ground truth images $\hat{\mathbf{I}}_i$ and rendered images $\mathbf{I}_i$, a perceptual loss to capture high-level features and textures with LPIPS metric [69], and a masked loss on the rendered opacity $\mathbf{I}_i^\alpha$ to remove background splats [68]:

$$\mathcal{L}_{\text{img}} = \frac{1}{M} \sum_{i=1}^{M} \Big(\left\| \hat{\mathbf{I}}_i - \mathbf{I}_i \right\|_2^2 + \lambda_{\text{perceptual}} \cdot \text{LPIPS}(\hat{\mathbf{I}}_i, \mathbf{I}_i) + \lambda_\alpha \left\| \hat{\mathbf{M}}_i - \mathbf{I}_i^\alpha \right\|_2^2 \Big), \tag{2}$$

where $\hat{\mathbf{M}}_i$ is the background mask of the ground truth images $\hat{\mathbf{I}}_i$, and $\lambda_{\text{perceptual}}$ and λ_α are weighting hyperparameters for the perceptual and transparency losses respectively. The transparency loss is necessary to reduce floating Gaussians that do not contribute to the foreground object.

Gaussian Tightness Regularization. To ensure that the predicted Gaussian Splats in Sec. 3.1 follow the SMPL parameters closely, we introduce a Gaussian tightness regularization that ensures the generated Gaussian splats [28] do not diverge and remain faithful to the underlying SMPL parameters as follows:

$$\mathcal{L}_{\text{tight}} = \frac{1}{V} \sum_{n=1}^{V} \left\| \boldsymbol{\delta}_n \right\|_2, \tag{3}$$

where $\boldsymbol{\delta}_n$ is defined in (1) and $V = 6890$ is the number of Gaussian splats (number of vertices in SMPL).

The total loss function is a weighted sum of the image losses (MSE, perceptual, and alpha) and tightness:

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{tight}} \mathcal{L}_{\text{tight}}, \tag{4}$$

where λ_{tight} is the weighting hyperparameter for the tightness regularization. As we show later in Sec. 5.3, this regularization plays an important role in the precision of the 3D human body predicted by GST. By minimizing this combined loss, our GST model learns to generate accurate and visually pleasing 3D reconstructions from a single image.

4. Experiments

In this section we describe our evaluation setup and the baselines for our comparisons.

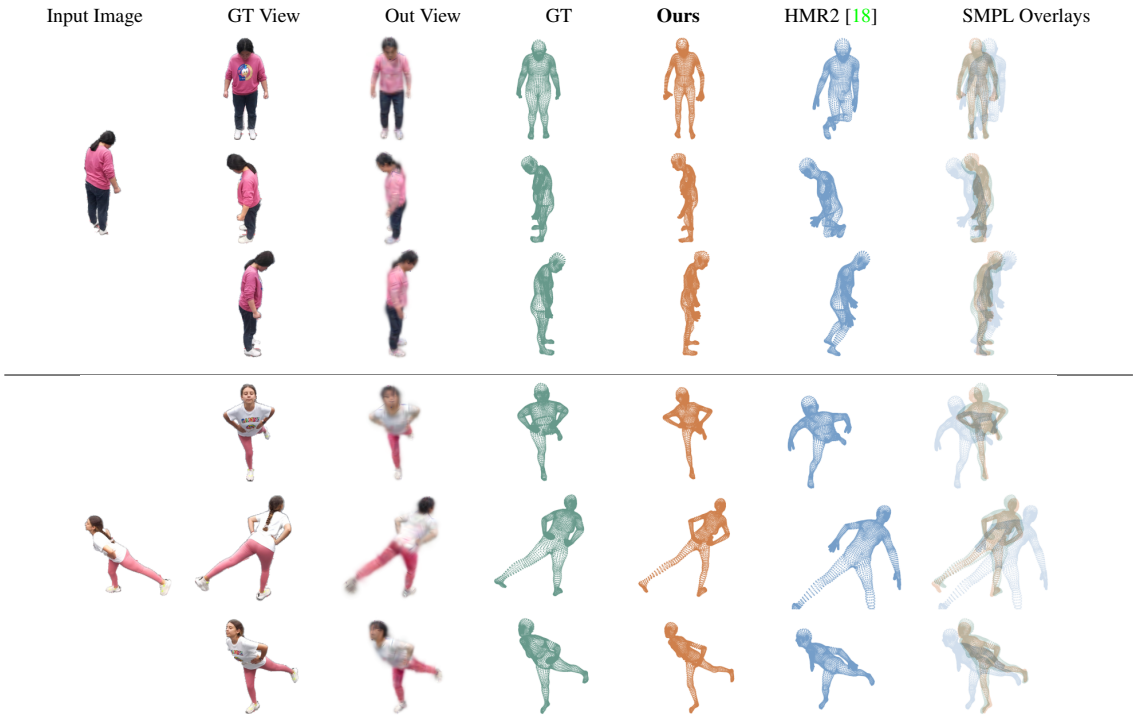

Figure 3. **3D SMPL Shape.** 3D human body results of our GST on two subjects of HuMMan [4] dataset compared to Ground Truth renderings, Ground Truth SMPL parameters [37], and SMPL predictions of HMR2 [18]. The overlay of the 3 SMPL bodies (ours, HMR2, and GT) shows that our predicted SMPL is more precise, while our predicted Gaussian splats maintain visual quality from novel views.

4.1. Datasets and Metrics

Datasets. Similar to previous works [24], we utilize four comprehensive human datasets for evaluation: THuman [72], RenderPeople [1], ZJU MoCap [43], and HuMMan [4]. For ZJU MoCap, the dataset is divided following the SHERF setup [24]. Similarly, for HuMMan, we adhere to the official split (HuMMan-Recon), using 317 sequences for training and 22 for testing, with 17 frames sampled per sequence. For THuman, we select 90 subjects for training and 10 for testing, and for the RenderPeople dataset, we randomly sample 450 subjects for training and 30 for testing. Those four datasets used for evaluation above are all small in terms of subject diversity. To showcase the capabilities of GST on a large dataset, we train our GST also on the TH21 dataset [63], which contains 2,500 3D scans, with diverse subject diversity. We randomly select 200 scans for evaluation.

Evaluation Metrics. When the Ground Truth 3D SMPL parameters are available as in RenderPeople [1] and HuMMan [4], we adopt 3D Human Joints precision MPJPE as a metric [27]. MPJPE refers to Mean Per Joint Position Error: the average L2 error across all joints after aligning with the root node. To quantitatively assess the quality of rendered novel view and novel pose images, we report peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS)[69]. Consistently with prior works[15, 24, 72], we project the 3D human bounding box onto each camera plane to derive the bounding box mask, subsequently reporting these metrics based on the masked regions.

Baselines. In addition to earlier works on Human NeRF with multi-view setting, NHP [31] and MPS-NeRF [15], we compare to recent single-image methods SHERF [24] for novel view synthesis and HMR2 [18] and TokenHMR [12] for 3D Human reconstruction precision. Different from SHERF, our method does not take as input ground truth SMPL parameters, therefore we adapt SHERF to use HMR2 [18] or TokenHMR [12] SMPL predictions for a fair comparison to our method. We also include a fast and salable Splatter Image [56], a state-of-the-art single image 3D reconstruction method for novel view synthesis tables.

4.2. Implementation Details of GST

Our model follows the implementation of HMR2 [18] for the predictions of SMPL parameters. We extend the HMR2 decoder implementation to process some additional learnable tokens for the predictions of the Gaussian parameters. The Gaussian parameters (color, rotation, scale, opacity, offset) are predicted for $K = 26$ groups of 265 Gaussians. The token output is passed through a linear layer to obtain the final parameters. We use pre-trained weights from HMR2

Table 2. **Human Novel View Synthesis and 3D Keypoints Evaluation Performance Comparison.** We compare GST on the RenderPeople [1] and HuMMan [4] datasets. For each dataset, we report PSNR, SSIM, and LPIPS for novel view synthesis, as well as MPJPE (in mm) for 3D keypoints evaluation. The top section methods use the Ground Truth input SMPL parameters, which are shown for reference, while the bottom section methods only use the single image input (our setup). ↑ means the larger is better; ↓ means the smaller is better.

Method	GT 3D input	*RenderPeople*				*HuMMan*			
		Novel View			3D Shape	Novel View			3D Shape
		PSNR↑	SSIM↑	LPIPS↓	MPJPE (mm)↓	PSNR↑	SSIM↑	LPIPS↓	MPJPE (mm)↓
NHP [31]	✓	20.59	0.81	0.22	0.000	18.99	0.84	0.18	0.000
MPS-NeRF [15]	✓	20.72	0.81	0.24	0.000	17.44	0.82	0.19	0.000
SHERF [24] w/ GT	✓	22.88	0.88	0.14	0.000	20.83	0.89	0.12	0.000
HMR2 [18]	✗	-	-	-	101.0	-	-	-	133.4
TokenHMR [12]	✗	-	-	-	77.9	-	-	-	91.4
SHERF [24] w/ [18]	✗	13.55	0.62	0.37	101.0	18.00	0.85	0.18	133.4
SHERF [24] w/ [12]	✗	15.24	0.70	0.33	77.9	16.41	0.84	0.17	91.4
GST (Ours)	✗	**17.80**	**0.81**	**0.25**	**67.6**	**18.40**	**0.87**	**0.14**	**64.6**

Table 3. **Novel View Synthesis Performance Comparison.** We compare GST on the ZJU_MoCap [43] and THuman [72] datasets on novel view synthesis. The top section methods use the Ground Truth input SMPL parameters (allowing for changing the pose) and are shown for reference, while the bottom section methods only use the single image input (our setup).

Method	*ZJU MoCap*			*THuman*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
PixelNeRF [62]	-	-	-	16.51	0.65	0.35
NHP [31]	21.66	0.87	0.17	22.53	0.88	0.17
MPS-NeRF [15]	21.86	0.87	0.17	21.72	0.87	0.18
SHERF [24] /w GT	22.87	0.89	0.12	24.66	0.91	0.10
Splatter Img [56]	19.50	0.80	0.28	19.20	0.80	0.20
SHERF [24] /w [18]	19.11	0.81	0.21	17.27	**0.85**	**0.16**
SHERF [24] /w [12]	20.72	0.85	0.16	**19.29**	0.84	0.18
GST (Ours)	**21.26**	**0.85**	**0.16**	16.34	0.84	0.20

for the ViT and the decoder and freeze the ViT weights during training. For our experiments, we use the loss weights $\mathcal{L}_{\text{perceptual}} = 0.01$, $\mathcal{L}_{\alpha} = 0.1$, $\mathcal{L}_{\text{tight}} = 0.1$, and we train on square image crops of size 256. We train on a single A6000 GPU with a batch size of 32 for 3 days. At test time, GST can simultaneously perform 3D human pose estimation and 3D reconstruction in a single forward pass at 47fps.

5. Results

In this section, we discuss the results obtained on four datasets in various evaluation settings.

5.1. 3D Human Shape Results

The primary focus of this work is the ability to infer a precise 3D human body from a single image without explicit 3D supervision. We show quantitative results in Table 2 on RenderPeople [1] and HuMMan datasets [4]. We compute the MPJPE error with respect to the ground truth SMPL joints before and after training. The results show that without explicit 3D supervision, our training improves the quality of the pose estimation from the pretrained HMR2 [18] and TokenHMR [12]. Furthermore, Fig. 3 shows some examples of our predictions in comparison with the HMR2 initialization and the ground truth SMPL pose. Our poses visually appear better aligned to the ground truth, emphasizing the results in Table 2. During training, the shape of the SMPL models also changes, with the 3D human shape becoming thinner. SMPL models human body shape without clothing. Our method decouples the body shape from additional layers, such as clothing. We hypothesize that this leads the model to estimate the underlying body shape of the human, effectively using the offsets to model clothes and other deformations.

5.2. Novel View Synthesis Results

We evaluate our method in the task of novel view synthesis across 4 datasets and compare the results with SHERF [24]. For a fair comparison with our method, which does not assume ground truth SMPL parameters are available, we evaluate SHERF using the estimated HMR2/TokenHMR pose and shape parameters instead of the ground truth ones. Our results are in Tables 2 and 3. Visual results are shown in Figures 4, 5 and 6. Note that the underlying 3D body is consistent, despite a slight blurriness of the Gaussians, and follows precise 3D geometry.

To showcase the capabilities of GST on a large dataset, we train our GST on multi-view images rendered from the TH21 dataset [63], which contains 2,500 3D scans and shows the results on 200 randomly sampled test scans in Table 4 and Fig.7. It clearly shows less blurriness than the other datasets. For reference, we include Splatter Image [56] in Table 4, where our GST predicts precise 3D body pose and shape in addition to the renderable representation unlike Splatter Image.

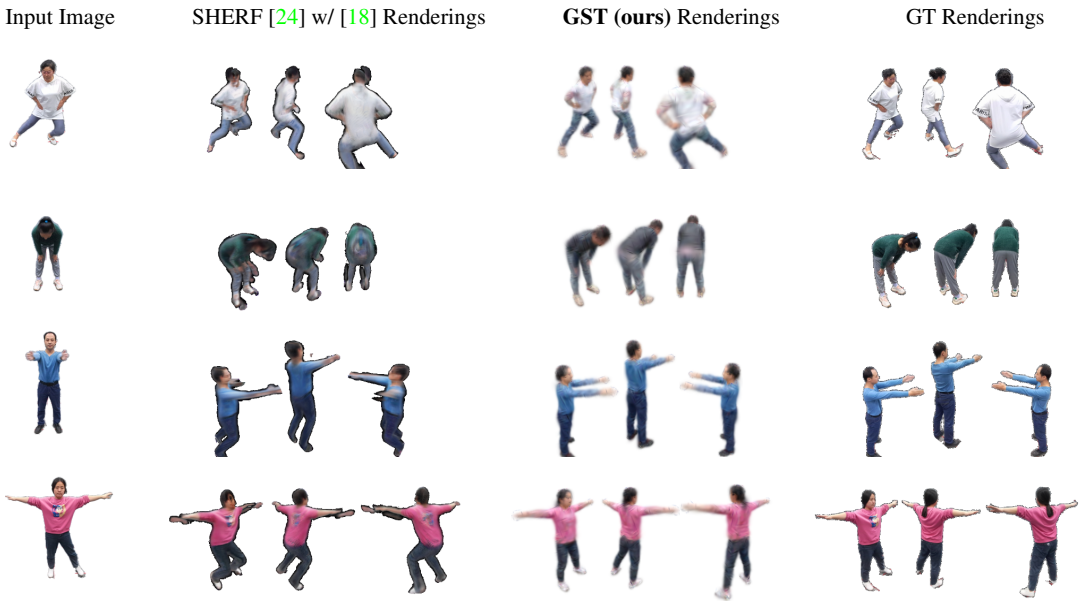

Figure 4. **Single Image NVS**. GST on 4 subjects of HuMMan [4] dataset compared to Ground Truth renderings, and SHERF [24] (after being adapted with HMR2 to work with single image input only). GST depicts the *correct* human pose (compared with ground truth).

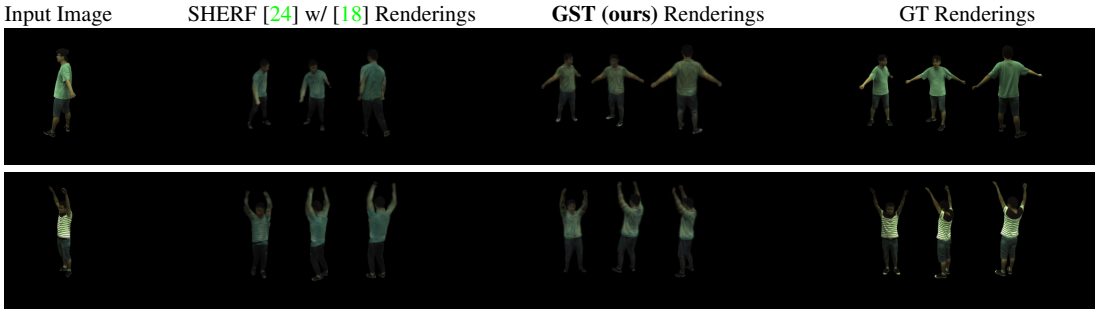

Figure 5. **Single Image NVS** on 2 subjects of Zju-Mocap [43] compared to SHERF [24] (after being adapted with HMR2 to work with single image input only). GST shows improved visual quality, especially when comparing the depicted pose to ground truth.

Table 4. **Novel View Synthesis on Large-Scale TH21**. We compare GST to fast and large-scale multi-view baseline [56] that do not need 3D annotations on the 2,500 examples from TH21 [63]. Unlike Splatter Image [56], our GST predicts precise 3D body pose and shape in addition to the renderable representation.

Method	Output 3D Body	Novel View PSNR↑	SSIM↑	LPIPS↓
Splatter Img [56]	✗	**23.74**	**0.91**	0.10
GST (Ours)	✓	22.20	0.90	**0.09**

Table 5. **Ablation study**. We show an ablation study on HuMMan Dataset [4] where the left shows the design choices and the right part shows the results. For each setup, we report PSNR, SSIM, and LPIPS for novel view synthesis, as well as MPJPE (in mm) for 3D keypoints evaluation.

LPIPS loss	Tightness loss	Transparency loss	Novel View PSNR↑	SSIM↑	LPIPS↓	3D Shape MPJPE (mm)↓
✓		✓	21.77	0.87	0.12	82.3
	✓	✓	21.80	0.86	0.15	53.6
✓	✓		21.77	0.87	0.12	52.3
✓	✓	✓	21.79	0.87	0.12	**50.8**

5.3. Ablation and analysis

Ablation Study. We present an ablation study of different design choices and key elements in our architectures and losses and their effect on the 2D and 3D results of a single image to 3D of humans on the HuMMan dataset [4] in Table 5. For these experiments, we report PSNR, SSIM and LPIPS metrics computed on the entire image. It shows the importance of combining the LPIPS, tightness, and transparency loss on the final 3D precision while maintaining the

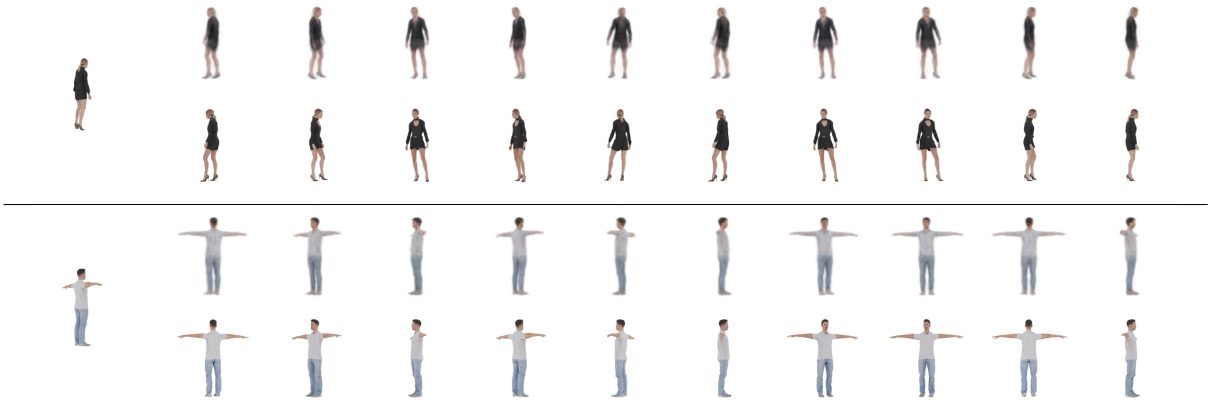

Figure 6. **Visualization of Single Image Novel View Synthesis Results on RenderPeople**. We show single image novel view synthesis results on two subjects of RenderPeople [1] dataset of our GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject.

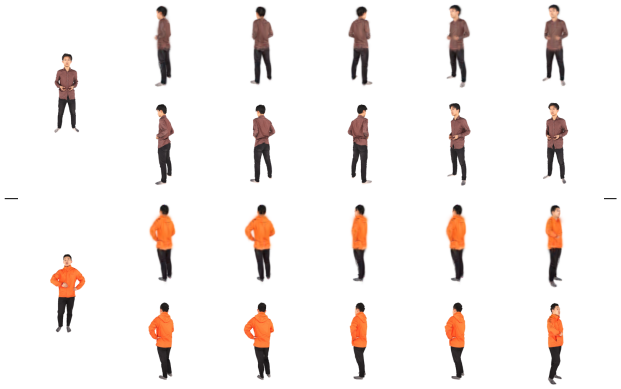

Figure 7. **Scaling Up Training of GST on TH21.** We show rendering results for GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject. The training of 2,500 subjects in TH21 [63] reduces the blurriness observed in other datasets and demonstrates the scalability merit of our GST Transformer training.

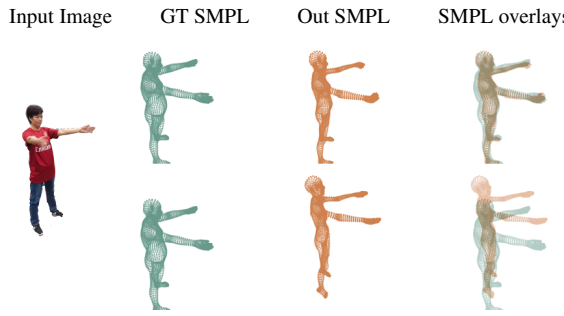

Figure 8. **Tightness Regularization.** Renderings and SMPL models with (*top*) and without (*bottom*) tightness regularization. The regularization maintains a precise body shape and pose.

visual fidelity intact. The tightness regularization of (3) has the highest impact on 3D precision, as it favors solutions in which the majority of the pose corrections are obtained with the SMPL parameters, and the Gaussians are only used for small refinements. In contrast, removing the tightness regularization encourages unrealistic and less precise poses, with much larger adjustments obtained with the Gaussian offsets. We also visualize this effect in Fig. 8. We conduct additional ablations in *the Appendix*.

3D Pose Estimation from Sparse Views. We train GST on the common 3D pose estimation dataset Human3.6M [5] using the default split for train and test subjects (subjects 9 and 11 are used for testing). This dataset is not ideal for our method as it only has 4 views and very few subjects, therefore it's difficult to generalise to unseen poses and subjects. Additionally, the human masks provided with the dataset are not always precise and our method tends to model parts of the background together with the human. This affects the quality of both the visual results and the 3D pose estimation. The visual metrics for our GST are evaluated on a squared crop of size 256x256 around the human with a PSNR of 18.68 and a 3D error of MPJPE $\downarrow$ = 63.7 mm compared to 50.0 mm for HMR2 [18].

6. Conclusions and Discussion

In this paper, we introduced *GST*, a novel approach for human 3D representation that predicts 3D Gaussian Splatting [28], enabling fast rendering with accurate poses. GST leverages multi-view supervision to accurately predict 3D joint and body poses while preserving the perceptual quality of novel view renderings. This dual capability combines precise pose estimation with high-quality rendering, bridging two research paradigms and showcasing the benefits of our approach (See Fig. 3).

Limitations. The main limitation in our method is the requirement of multi-view datasets to train. Another issue is the slight blurriness that appears on some of the renderings

as a result of the generalization limitation of the transformer trained on very small datasets we employ (in terms of subject diversity). A possible solution to this is to use multiple datasets or bigger datasets.

References

[1] Renderpeople. In https://renderpeople.com/3d-people, 2018. 5, 6, 8, 12, 15, 19

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 2

[4] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, Fangzhou Hong, Mingyuan Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. HuMMan: Multi-modal 4d human dataset for versatile sensing and modeling. In *17th European Conference on Computer Vision, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 557–577. Springer, 2022. 5, 6, 7, 12, 16, 17

[5] Cristian Sminchisescu Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2011. 8

[6] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2

[7] Mingjin Chen, Junhao Chen, Xiaojun Ye, Huan ang Gao, Xiaoxue Chen, Zhaoxin Fan, and Hao Zhao. Ultraman: Single image 3d human reconstruction with ultra speed and detail. 2024. 1, 2, 3

[8] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22246–22256, 2023. 1, 2

[9] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[10] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4

[12] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1323–1333, 2024. 3, 5, 6

[13] Qiao Feng, Yuanwang Yang, Yu-Kun Lai, and Kun Li. R²human: Real-time 3d human appearance rendering from a single image. *arXiv preprint arXiv:2312.05826*, 2023. 2, 3

[14] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 89–106, 2022. 1, 2

[15] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *arXiv preprint arXiv:2203.16875*, 2022. 5, 6

[16] Xiangjun Gao, Xiaoyu Li, Chaopeng Zhang, Qi Zhang, Yanpei Cao, Ying Shan, and Long Quan. Contex-human: Free-view rendering of human from a single image with texture-consistent synthesis, 2023. 1, 2, 3

[17] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020. 2

[18] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa*, and Jitendra Malik*. Humans in 4D: Reconstructing and tracking humans with transformers. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 5, 6, 7, 8, 12, 15, 16

[19] Riza Alp Guler and Iasonas Kokkinos. Holopose: Holistic 3d human reconstruction in-the-wild. In *CVPR*, 2019. 2

[20] Abdullah Hamdi, Bernard Ghanem, and Matthias Nießner. Sparf: Large-scale learning of 3d sparse radiance fields from few input images. 2023. 1

[21] Abdullah Hamdi, Luke Melas-Kyriazi, Jinjie Mai, Guocheng Qian, Ruoshi Liu, Carl Vondrick, Bernard Ghanem, and Andrea Vedaldi. Ges: Generalized exponential splatting for efficient radiance field rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[22] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[23] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7909–7920, 2023. 2

[24] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 5, 6, 7

[25] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 20418–20431, 2024. 3

[26] Tomas Jakab, Ruining Li, Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Farm3D: Learning articulated 3d animals by distilling 2d diffusion. *arXiv preprint arXiv:2304.10535*, 2023. 2

[27] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 2, 5

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 2, 3, 4, 8

[29] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *ICCV*, 2021. 2

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[31] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 5, 6

[32] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[34] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[35] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 12

[36] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 1, 2

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2, 4, 5, 15, 16

[38] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360{\deg} reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[39] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 2

[40] Aryan Mikaeili, Or Perel, Mehdi Safaee, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Sked: Sketch-guided text-based 3d editing. *ICCV*, 2023. 1, 2

[41] 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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 2

[42] OpenAI. Gpt-4 technical report, 2023. 3

[43] 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*, pages 9054–9063, 2021. 5, 6, 7

[44] Marco Pesavento, Yuanlu Xu, Nikolaos Sarafianos, Robert Maier, Ziyan Wang, Chun-Han Yao, Marco Volino, Edmond Boyer, Adrian Hilton, and Tony Tung. Anim: Accurate neural implicit model for human reconstruction from a single rgb-d image, 2024. 2, 3

[45] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *International Conference on Learning Representations (ICLR)*, 2022. 1, 2

[46] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 2

[47] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Ben Mildenhall, Nataniel Ruiz, Shiran Zada, Kfir Aberman, Michael Rubenstein, Jonathan Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. *ICCV*, 2023. 1, 2

[48] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 2, 3

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[50] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 2

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language

understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2, 3

[53] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2304–2314, 2019. 1, 2, 3

[54] Hoigi Seo, Hayeon Kim, Gwanghyun Kim, and Se Young Chun. Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. *arXiv preprint arXiv:2304.02827*, 2023. 2

[55] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[56] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4, 5, 6, 7, 12, 13

[57] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 1, 2

[58] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *eccv*, 2024. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 4

[60] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[61] Zhenzhen Weng, Jingyuan Liu, Hao Tan, Zhan Xu, Yang Zhou, Serena Yeung-Levy, and Jimei Yang. Template-free single-view 3d human digitalization with diffusion-guided lrm. *Preprint*, 2023. 1, 2, 3

[62] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 6

[63] 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 (CVPR)*, 2021. 5, 6, 7, 8, 12, 13, 14

[64] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 896–905, 2024. 3

[65] Xiaoyu Zhan, Jianxin Yang, Yuanqi Li, Jie Guo, Yanwen Guo, and Wenping Wang. Semantic human mesh reconstruction with textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3

[66] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, 2021. 2

[67] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu. Pymaf-x: Towards well-aligned full-body model regression from monocular images. *PAMI*, 2023. 2

[68] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. *arXiv*, 2024. 2, 4

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 5

[70] Zechuan Zhang, Li Sun, Zongxin Yang, Ling Chen, and Yi Yang. Global-correlated 3d-decoupling transformer for clothed avatar reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3

[71] Zechuan Zhang, Zongxin Yang, and Yi Yang. Sifu: Side-view conditioned implicit function for real-world usable clothed human reconstruction. 2024. 1, 2, 3

[72] 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. 5, 6, 18

A. Additional Results and Analysis

A.1. Additional Ablations

In addition to the ablations described in Table 5 in the main paper, we report here three variations to the GST model that did not result in a performance improvement. The ablations are provided in Table I.

More Gaussians. The first design change we tested is an increase in the number of Gaussians per vertex. We increase the number of splats by predicting two or three independent offsets per vertex. Because random initialization breaks the symmetry, the model can learn to move each splat independently even though all two/three are anchored to the same vertex. Contrary to our assumption, an increase in the number of splats did not result in a increased visual quality of the renderings.

Setting Opacity to 1. Predicting opacity is not strictly necessary to render humans, therefore we tried simplifying the model by removing this parameter. We removed the opacity prediction during training and manually set the opacity to 1 for all the Gaussians.

Single-view + Multi-view Images. Next, to increase the subject diversity in the small datasets we use, we tried including some single view images in our training pipeline. For this experiment, we use crops of images containing humans from the MSCOCO dataset [35]. The single view images are used for training together with the multi-view images from the original dataset. For the single view images, the model predictions are supervised using the same input image. The results do not show any notable improvement.

Table I. **Additional Negative Ablations.** For completeness, we show additional ablations on HuMMan Dataset [4] that did not give positive improvements to our best setup of Table 5 in the main paper. For each setup, we report PSNR, SSIM, and LPIPS for novel view synthesis, as well as MPJPE (in mm) for 3D keypoints evaluation.

Ablation setup	Novel View			3D Shape
	PSNR↑	SSIM↑	LPIPS↓	MPJPE (mm)↓
our best model	21.79	0.87	0.12	50.8
2 Gaussians per vertex	21.25	0.87	0.12	50.1
3 Gaussians per vertex	21.18	0.87	0.12	53.2
setting opacity to 1	21.17	0.87	0.11	58.4
single-view + multi-view	21.47	0.87	0.12	53.4

A.2. Overfitting Example

To test that the number of Gaussians is sufficient to produce sharp details, we train our model to overfit a single data sample. We obtain an almost perfect reconstruction with PSNR of 41. Image I shows examples of the renderings we obtained. This result confirms our assumption that with a large enough dataset, our model would be able to learn sharper details than it currently learns on the small scale datasets.

A.3. Additional Details for TH21 Experiment

For the TH21 [63] experiment in Table 4 in the main report, we use 72 views rendered in a loop around the subject. We train both our method and Splatter Image [56] using 256x256 images. Despite our model performing worse than Splatter Image in terms of visual metrics, our model also predicts the SMPL paramters for 3D pose estimation. This is both useful for downstream tasks, but also ensures that the underlying 3D shape is plausible for a human. This difference can be noticed in the examples in Figure II, where GST can reconstruct a plausible human shape despite the uncommon input pose, while Splatter Image fails to reconstruct arms and legs.

A.4. HMR2 Finetuning Comparison

As we discussed in the main paper, we train GST starting from a pretrained version of HMR2 [18] and the resulting model almost halves the MPJPE error on the two datasets, compared to the original pretrained version. We now instead compare GST with a *finetuned* version of HMR2 to test the quality of the 3D pose estimation of our method.

Table II. **HMR2 finetuning comparison** MPJPE (in mm) for 3D keypoints evaluation.

Method	3D annotations	*RenderPeople* MPJPE (mm)↓	*HuMMan* MPJPE (mm)↓
GST	✗	<u>64.6</u>	<u>67.6</u>
HMR2 [18] pretrained	✗	101.0	133.4
HMR2 [18] finetune w/ 2D data	✗	127.40	163.77
HMR2 [18] finetune w/ 2D + 3D data	✓	**57.33**	**61.20**

We finetune HMR2 on the two datasets in Table 2 in the main paper: HuMMan [4] and RenderPeople [1]. The results are reported in Table II. To reproduce a similar training setup to our method (that does not require any 3D ground truth annotations), we finetune HMR2 using only 2D keypoints annotations. We use images from all views in the dataset, but restrict the supervision to only use the 2D keypoints loss.

The results show that the 2D information alone is not enough for HMR2 to improve the quality of the 3D pose estimation on the two datasets, and the finetuned model MPJPE error is worse than the pretrained one.

For completeness, we also report the errors when finetuning HMR2 with additional 3D annotations: 3D keypoints and ground truth SMPL parameters. We would like to emphasize that we think this is an unfair comparison to our method, since our method does not use ground truth SMPL parameters or 3D keypoints for training. The MPJPE of the HMR2 version finetuned with 3D data is only 7mm better than ours on RenderPeople and 6mm better than ours on HuMMan.

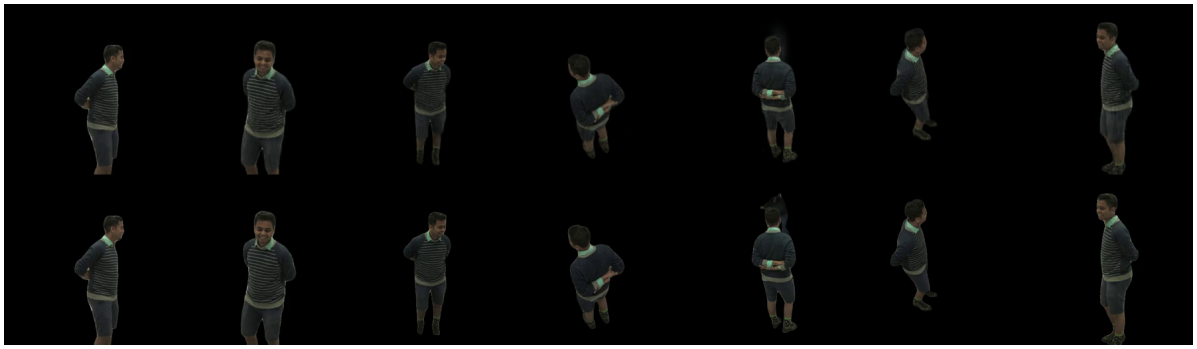

Figure I. **Overfitting to a single sample.** Ground truth (*top*) and renderings (*bottom*) of our model results when overfitting to a single data sample.

Figure II. **Splatter Image comparison.** Side view comparison with Splatter Image [56] on TH21 [63] for unusual input poses. Input image on the left, Splatter Image rendering in the first row, GST renderings in the second row.

B. Additional Visualizations

We provide some additional examples of novel view synthesis and 3D pose estimation results.

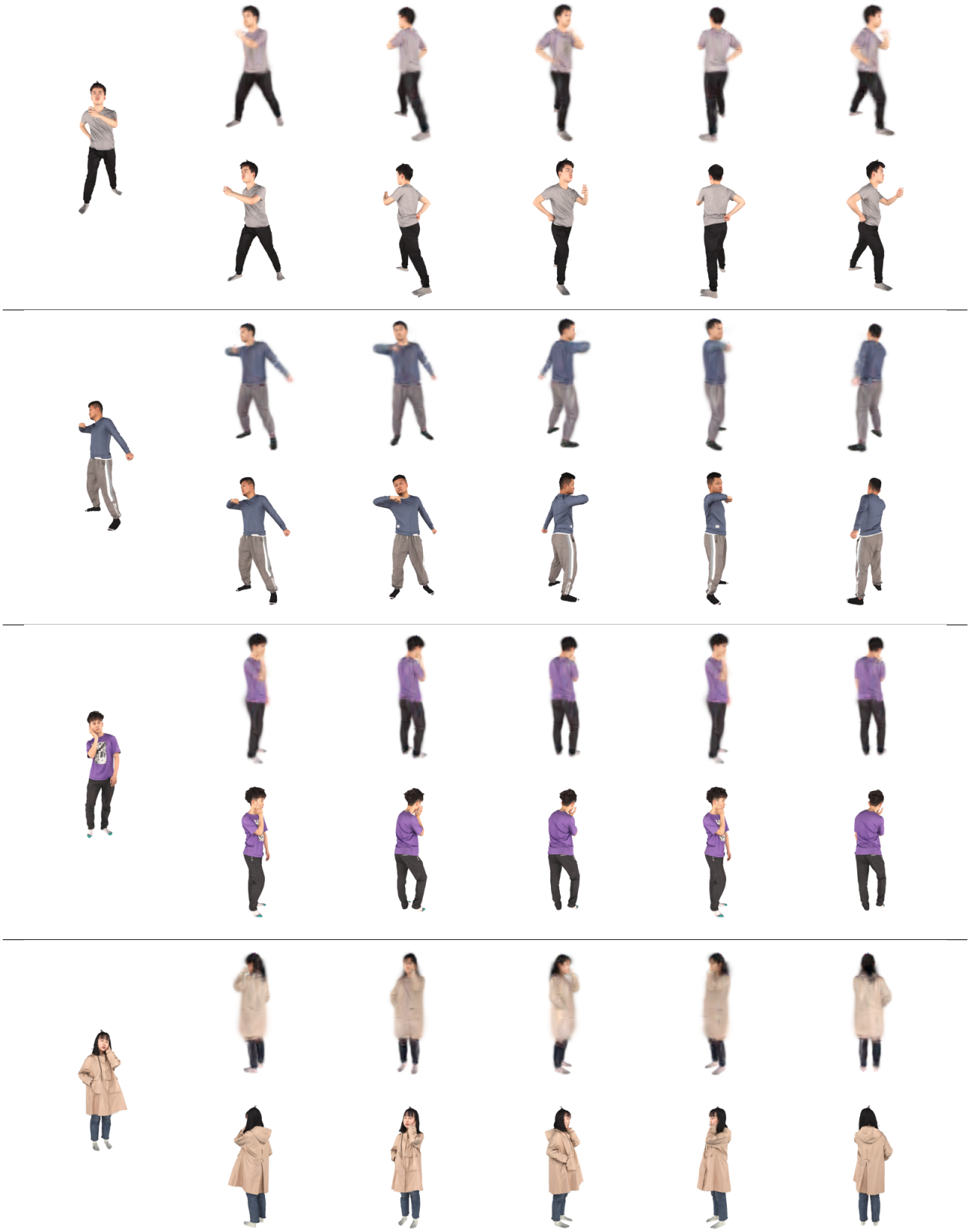

Figure III. **Results in TH21 [63].** Rendering results for GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject. An example of loose clothes is in the last row.

Figure IV. **Additional 3D SMPL Shape Results for the RenderPeople dataset [1]**. We show 3D human body results of our GST on three subjects of RenderPeople [1] dataset compared to Ground Truth renderings, Ground Truth SMPL parameters [37], and SMPL predictions of HMR2 [18]. The overlay of the 3 SMPL bodies (ours, HMR2, and GT) shows that our predicted SMPL is more precise while our predicted Gaussian splats maintain visual quality from novel views.

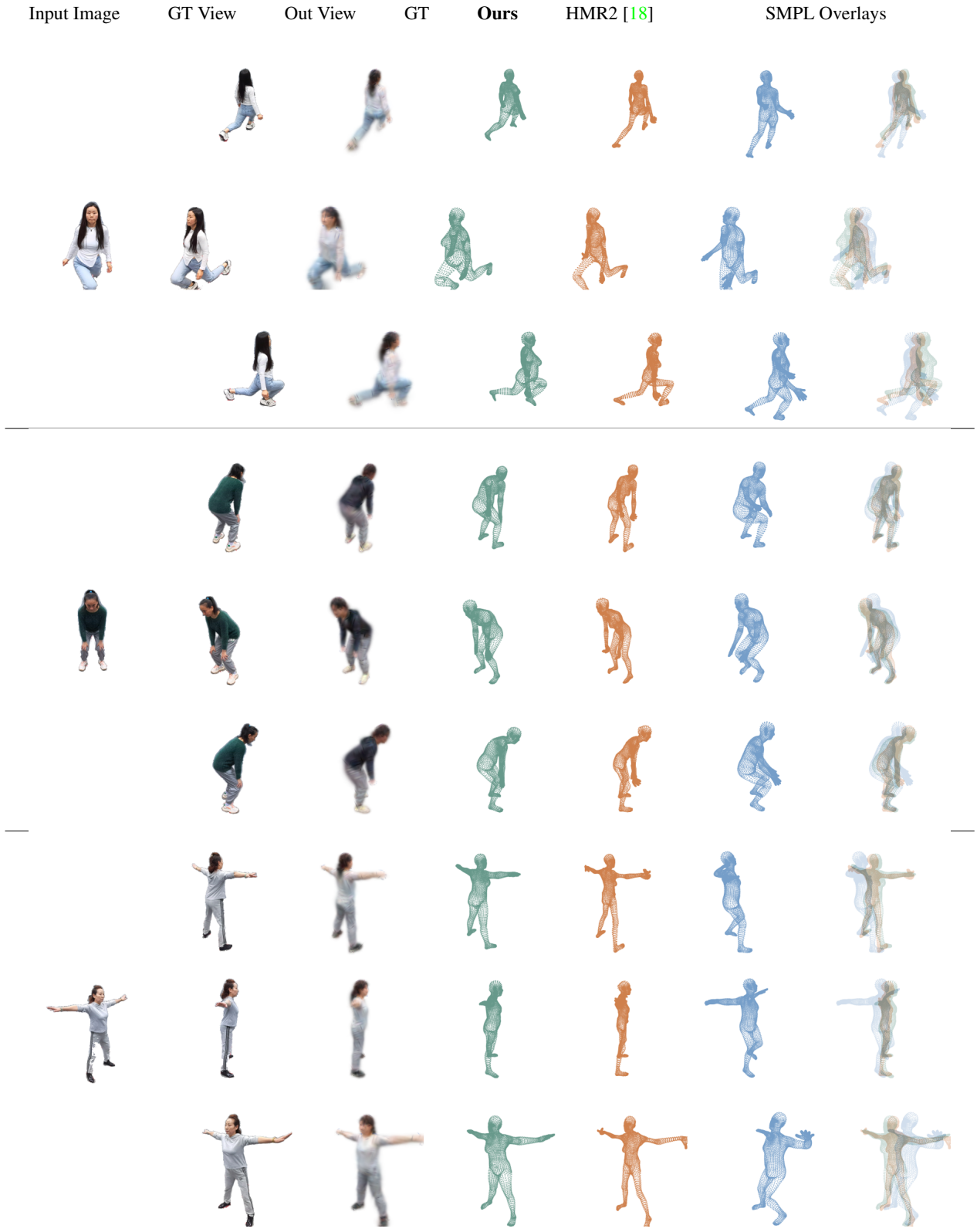

Figure V. **Additional 3D SMPL Shape Results for the HuMMan dataset [4].** We show 3D human body results of our GST on three subjects of HuMMan [4] dataset compared to Ground Truth renderings, Ground Truth SMPL parameters [37], and SMPL predictions of HMR2 [18]. The overlay of the 3 SMPL bodies (ours, HMR2, and GT) shows that our predicted SMPL is more precise while our predicted Gaussian splats maintain visual quality from novel views.

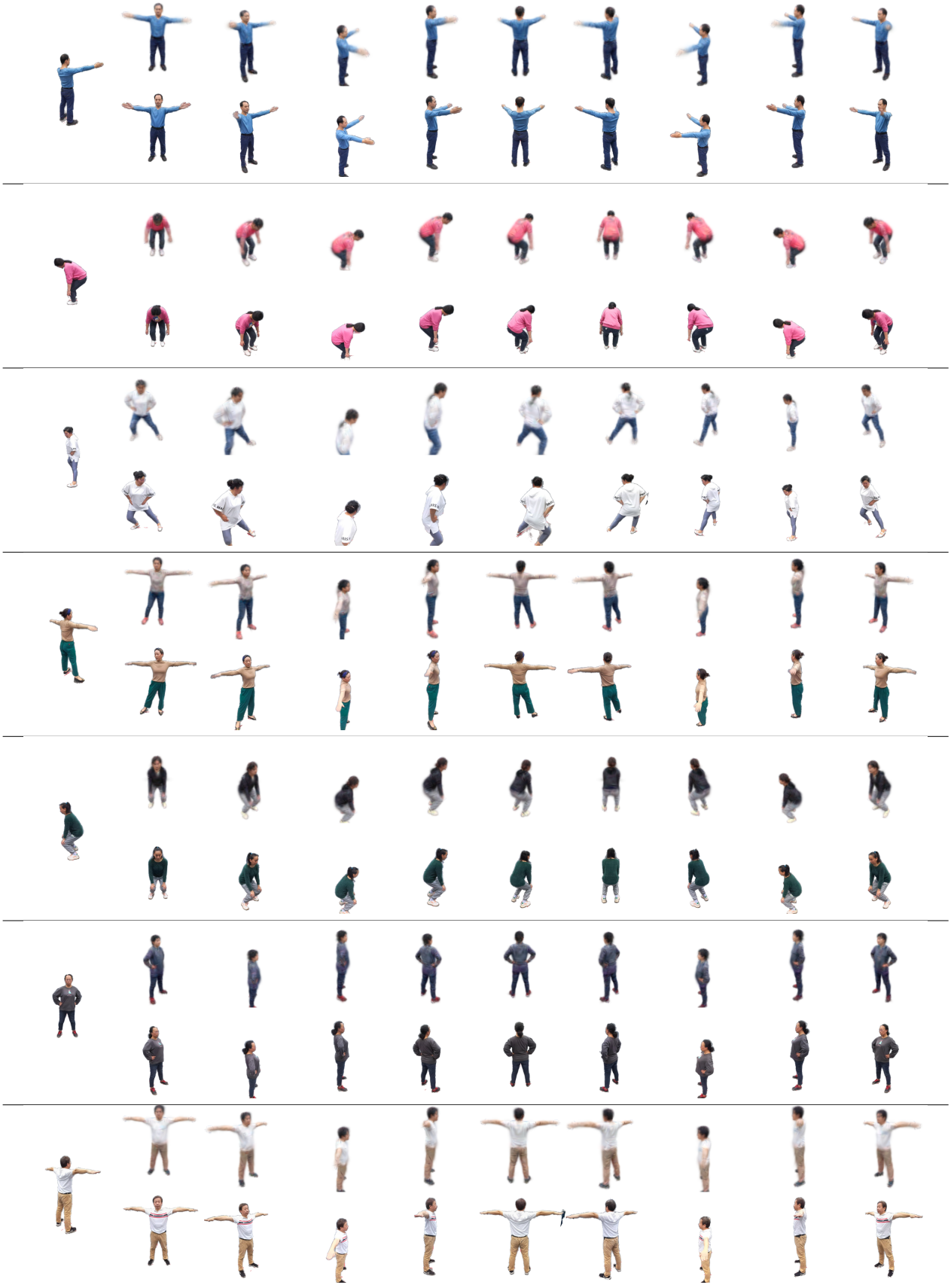

Figure VI. **Visualization of Single Image Novel View Synthesis Results on HuMMan**. We show single image novel view synthesis results on 7 subjects of HuMMan [4] dataset of our GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject.

Figure VII. **Visualization of Single Image Novel View Synthesis Results on THuman**. We show single image novel view synthesis results on 5 subjects of THuman [72] dataset of our GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject.

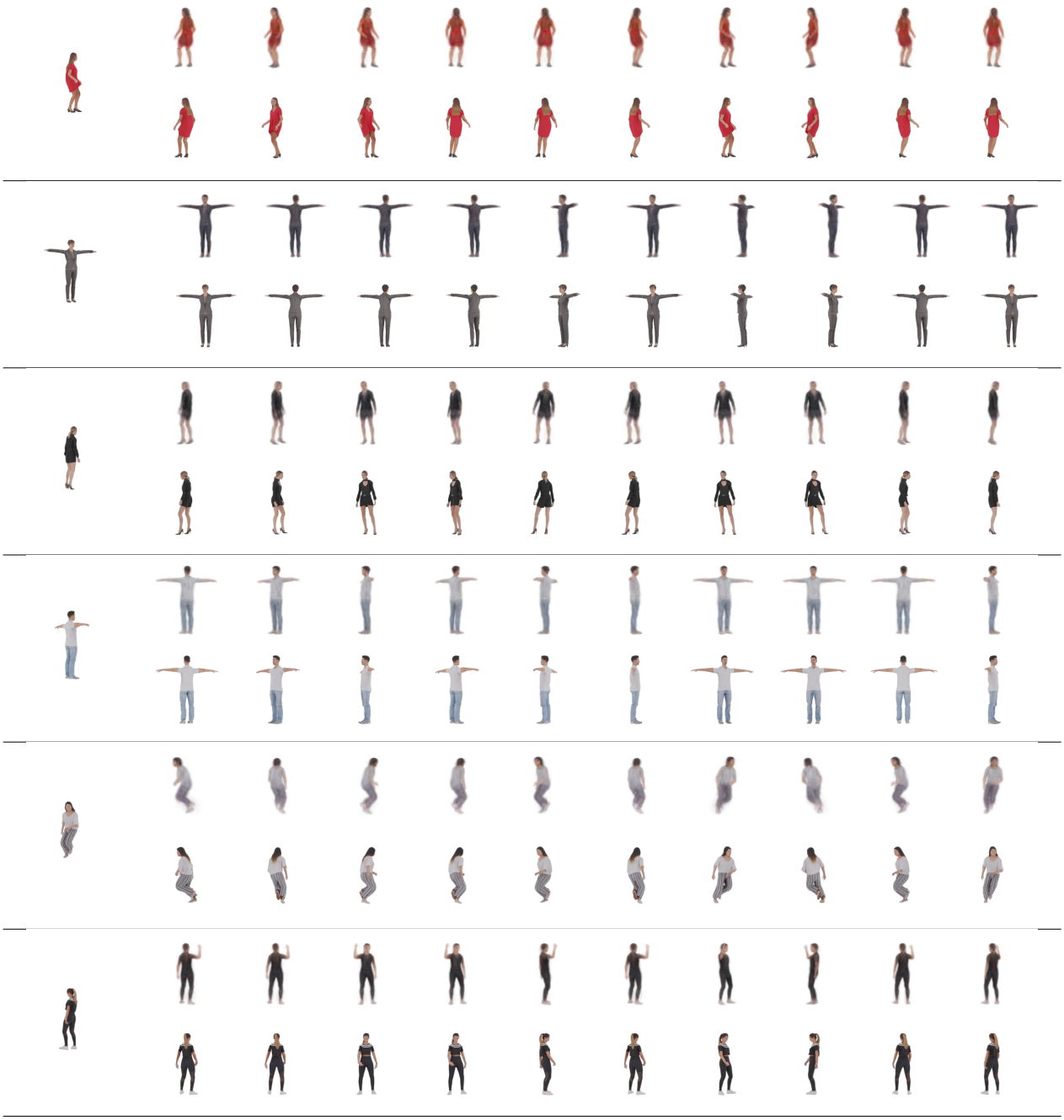

Figure VIII. **Visualization of Single Image Novel View Synthesis Results on RenderPeople**. We show single image novel view synthesis results on 6 subjects of RenderPeople [1] dataset of our GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject.