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

Supplementary materials

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 [7]. 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.

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. Fig. 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.

Table I. **Additional Negative Ablations.** For completeness, we show additional ablations on HuMMan Dataset [2] 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.3. Additional Details for TH21 Experiment

For the TH21 [11] 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 [10] 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 Fig. 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. 3D Pose Estimation from Sparse Views

We train GST on the common 3D pose estimation dataset Human3.6M [3] 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 generalize 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 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 [4].

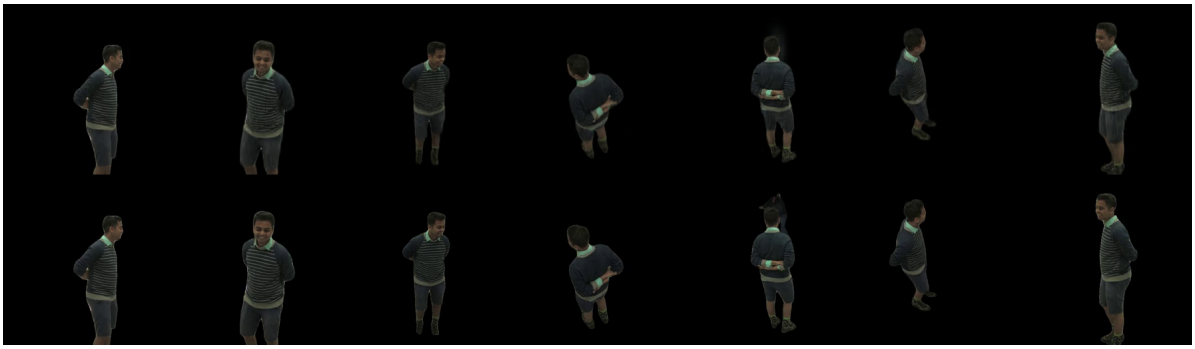

Figure I. **Overfitting to a single sample.** Ground truth (*top*) and renderings (*bottom*) of our model results when overfitting to a single data sample.

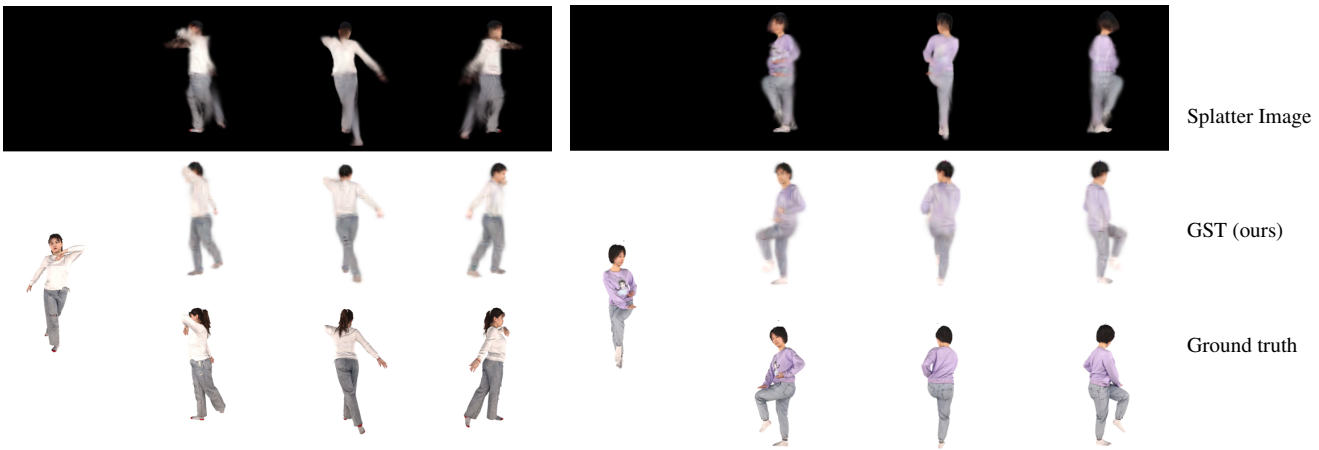

Figure II. **Splatter Image comparison.** Side view comparison with Splatter Image [10] on TH21 [11] for unusual input poses. Input image on the left, Splatter Image rendering in the first row, GST renderings in the second row.

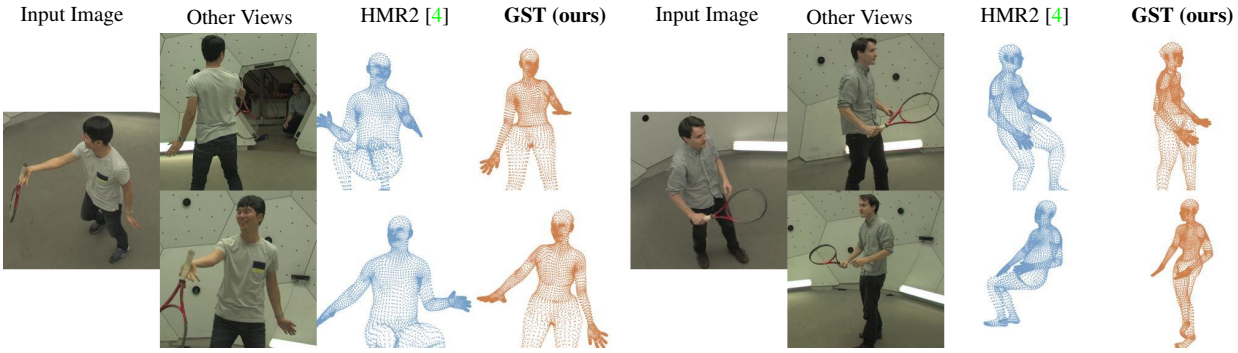

Figure III. **Example of human pose improvements using our method GST.** 3D human body results of our GST and SMPL predictions of HMR2 [4] on a sports sequence from the CMU panoptic dome dataset [6].

B. Additional Visualizations

Fig. III shows additional pose estimation results on the sports sequence of the CMU Panoptic dataset [6]. Fig. IV shows additional examples of novel view synthesis comparisons with SHERF [5]. Fig. V and VI show additional pose comparisons for the RenderPeople [1] dataset. Fig. VII and VIII show examples of novel view synthesis results for the TH21 [11], THuman [12] and RenderPeople [1] datasets.

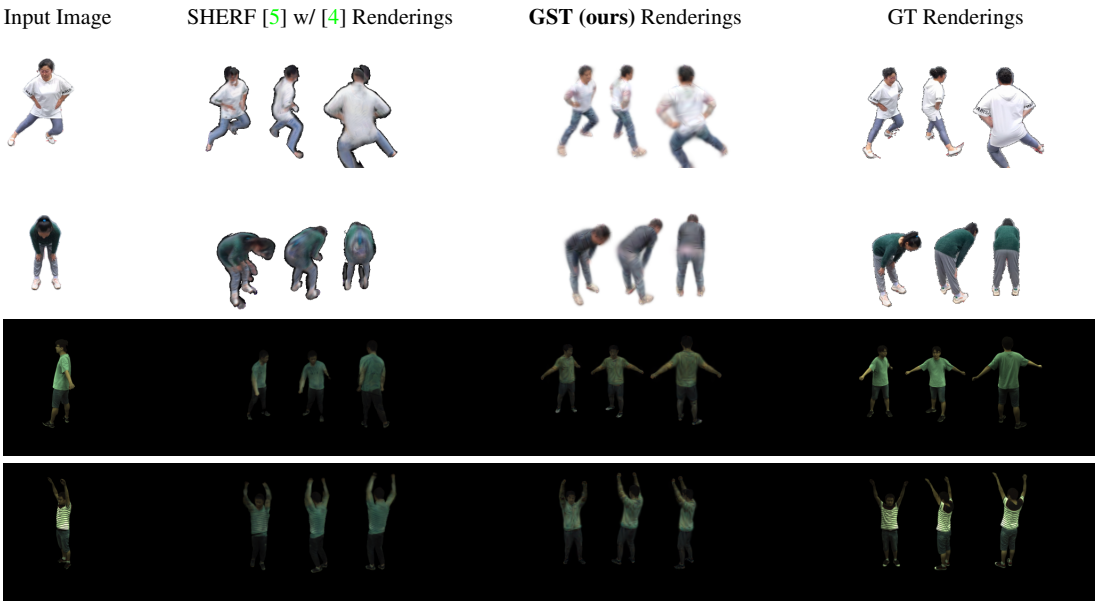

Figure IV. **Single Image NVS** on two subjects of Zju-Mocap [9] and two subjects of HuMMan [2] compared to SHERF [5] (after being adapted with HMR2 to work with single image input only). GST shows improved visual quality, especially when comparing the depicted pose to ground truth.

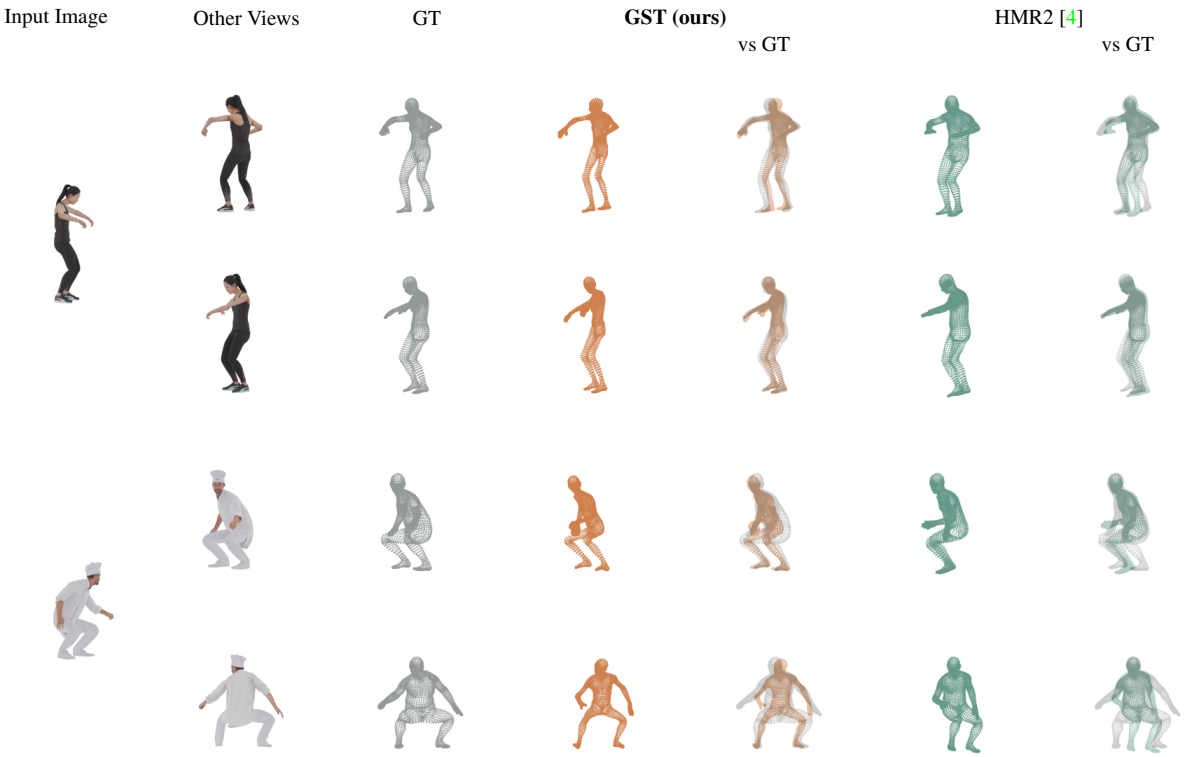

Figure V. **3D Shape Comparison with HMR2.** 3D human body results of our GST on two subjects of RenderPeople [1] dataset compared to Ground Truth SMPL parameters [8], and SMPL predictions of HMR2 [4].

Figure VI. **3D Shape Comparison with HMR2 After Fine-tuning on 2D and 3D Data.** 3D human body results of our GST on five subjects of RenderPeople [1] dataset compared to Ground Truth SMPL parameters [8], and SMPL predictions of HMR2 [4]. We show two versions of HMR2, one finetuned on 2D data only (HMR2-2D), and one finetuned on 3D data (HMR2-3D). Our method is only finetuned on 2D image data, but the results are almost as accurate as HMR2 finetuned on 3D data.

Figure VII. **Results in TH21 [11].** 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 VIII. **Visualization of Single Image Novel View Synthesis Results on THuman and RenderPeople**. We show single image novel view synthesis results on one subject of THuman [12] dataset and one subjects of RenderPeople [1] dataset of our GST (*top row*) compared to Ground Truth renderings (*bottom row*) of each subject.

References

[1] Renderpeople. In *https://renderpeople.com/3d-people*, 2018. 2, 3, 4, 5

[2] 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. 1, 3

[3] 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. 1

[4] 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. 1, 2, 3, 4

[5] 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. 2, 3

[6] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 3334–3342, 2015. 2

[7] 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. 1

[8] 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. 3, 4

[9] 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. 3

[10] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2

[11] 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. 1, 2, 5

[12] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. Deephuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7739–7749, 2019. 2, 5