H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion

Supplementary Material

Hongyi Xu
Google Research
hongyixu@google.com

Thiemo Alldieck
Google Research
alldieck@google.com

Cristian Sminchisescu
Google Research
sminchisescu@google.com

In this supplementary material we provide additional results, both quantitative and qualitative, as well as comparisons with other methods, ablation studies, and additional background on our model components.

1 Additional Results

Dynamic Human Reconstruction and Rendering. In fig. 1, we provide qualitative results for samples taken from the four datasets we use for experiments. Our method shows very good accuracy for both novel view synthesis and for 3D geometric reconstruction. In addition, our model can generate novel geometry corresponding to poses not in the training set, by altering the input imGHUM pose control $(\boldsymbol{\theta}, \mathbf{T})$, as shown in the fig. 1. We note that differently from novel view synthesis for a fixed pose, the source of error in the image synthesis of a novel pose (not in the training set) could arise from both the NeRF rendering function and the implicit geometric surface skinning. For our RenderPeople sequences where ground-truth images (renderings) and geometries for the novel test poses are available, we quantitatively report both image and geometric errors in tab. 1 of the main paper. We also show, side-by-side, qualitative comparisons in fig. 4 (main paper) and fig. 1 of this material.

Tab. 1 provides additional information on the used datasets. Specifically, the RenderPeople sequences are generated by animating a single rigged scan using motion capture from the CMU [1] and Human3.6M [3] datasets, respectively. The GHS3D dataset is composed of 14 sequences dynamically capturing dressed subjects undertaking freestyle motions (e.g. presentation, dancing or exercising). The PeopleSnapshot [2] dataset consists of monocular videos of a subject rotating in front of a static camera. We select one rotation cycle for training and another (different) cycle for testing. We also demonstrate the capabilities of our method on the Human3.6M dataset which captures the scene using four synchronized and calibrated cameras. There we use the first 320 frames for training (sampled every 8th frame) and the following 160 frames for testing (sampled every 4th frame).

Comparison to Nerfies [4]. For large full-body articulated human motions, as in our typical use case (and illustrated by our sequences), Nerfie overfits to the training images by mixing the human geometry with the background. This results in extremely limited generalisation for both novel view synthesis and 3D geometric reconstruction (tab. 2). We notice that geometry is effectively not reconstructed, leading to zero IoU. Moreover, similarly to D-NeRF [6], the deformation function is conditioned on time (or an embedding of the video frame index), hence the method cannot generalize to novel poses or shapes. To evaluate image (rendering) quality, we crop the image using a 2D bounding box computed from the ground-truth image segmentation (padded with a 20 pixel border), as our region of interest. For H-NeRF, IDR [8] and NeuralBody [5], the evaluation is performed by comparing to the cropped ground-truth image and using the foreground human segmentation. However, for the original NeRF and Nerfies, segmentation is not produced and therefore we compute metrics on the image with background.

Generalization. In addition to pose generalization shown in fig. 1, we further evaluate generalization to novel shapes in fig. 2 for sample sequences taken from the four datasets. The shape extrapolation

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

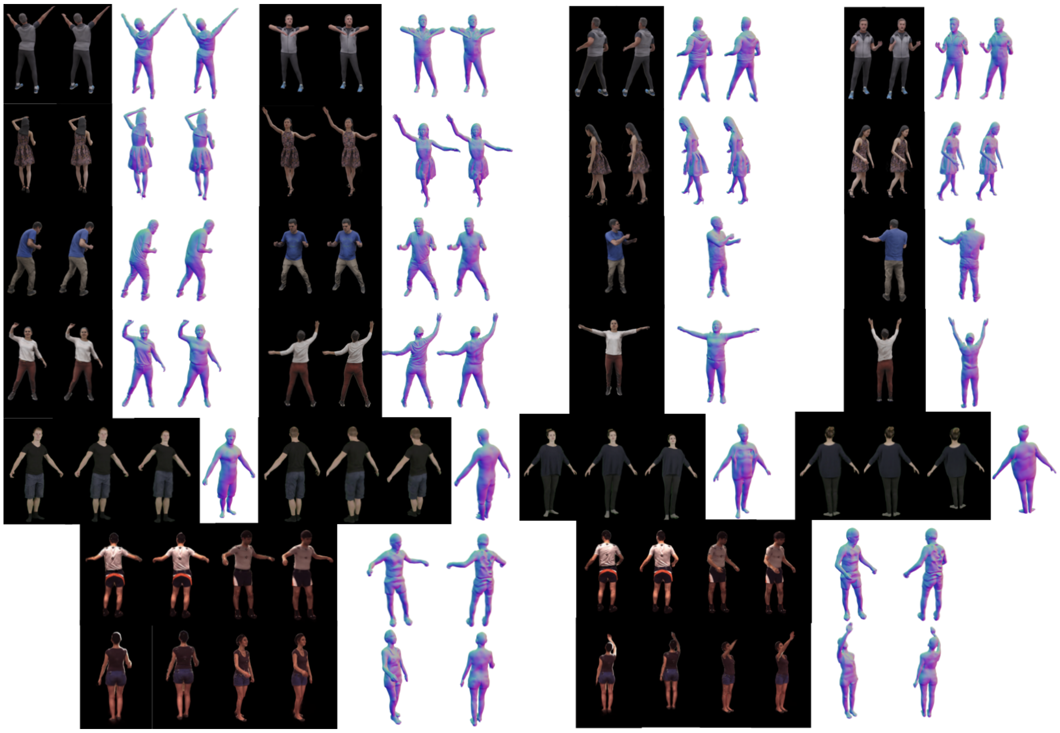

Figure 1: Qualitative results for dynamic sequences. When available, we show ground-truth (left) image and geometry side by side with our results (right). The top 2 rows show single scan animations based on RenderPeople assets, showing novel view synthesis (the left 2) and novel pose generalization (the right 2, under novel views). Similarly, the 3rd-4th row show GHS3D scan sequences, with 2 novel view synthesis and 2 pose generalization illustrations. The 5th row shows the monocular PeopleSnapshot [2] ground-truth image, our rendering of a novel pose under the learned and a novel camera, respectively, as well as our geometric reconstruction from left to the right. The last 2 rows illustrate the pose generalization capability of our models on Human3.6M [3], where we render the novel pose under 2 training camera views and show ground-truth images for side-by-side comparisons.

is achieved by replacing the training imGHUM shape latent code with a different $\boldsymbol{\beta}$, and propagate the shape changes to both image rendering and the geometry. To understand this process, one can consider imGHUM as the inner layer of the human body. The color and signed distance of a spatial point is deforming w.r.t. the body surface. Therefore when we change the underlying body shape, the NeRF radiance field and the signed distance function are updated accordingly. We observe high-quality visual shape generalization results produced by H-NeRF, in fig. 2, even for significant volume changes w.r.t. the ground truth shape.

In fig. 3, we study the image synthesis quality as a function of the differences between testing and training camera views. The view difference (x-axis) is evaluated by computing the angle between the two vectors from the camera position to the 3D center of the person. All methods (H-NeRF, IDR and NeRF) show a degradation of quality as viewpoints deviate significantly from training, but the view synthesis capability of H-NeRF is consistently better than the other two, likely given its capacity to estimate a reasonable 3D surface geometry. The increased error is largely due to the rendering of body parts not visible in the training images. We note that for our RenderPeople and GHS3D (static or dynamic) sequences, the four training cameras are on the four sides of the subject, residing on a sphere with radius 2.4m, oriented towards the center as shown in the last row of fig. 3, where we also show two typical test camera views for our dynamic sequences.

We further evaluate the pose generalization capability of H-NeRF as a function of the pose difference to the training configurations. Fig. 4 shows the image and geometric metrics of two RenderPeople sequences for all of our novel testing poses. We do not observe strong correlation as a function of pose differences w.r.t. training. and suspect that both image and geometric quality are largely affected

Train

Dataset	#seq	#cam	#poses	scene-scale	gt-image	gt-geom	opt
RenderPeople	8	4	39/53/71	2.5	Yes	Yes	No
GHS3D	14	4	31/44/59	2.5	Yes	Yes	Yes
PeopleSnapshot	9	1	19/29/55	2.5	Yes	No	Yes
Human3.6M	5	4	40/40/40	5	Yes	No	Yes

Test

Dataset	#seq	#cam	#poses	gt-image	gt-geom
RenderPeople	8	2	32/47/68	Yes	Yes
GHS3D	14	2	40/47/59	No	No
PeopleSnapshot	9	1	22/31/50	Yes	No
Human3.6M	5	4	40/40/40	Yes	No

Table 1: Dataset statistics. #poses are reported as the minimal/median/maximum number of poses. Scene-scale represents the size of the scene where we divide it to scale into $[-1,-1,-1]$ to $[1,1,1]$. gt-image and gt-geom indicate that ground-truth image and geometry are available, respectively. If opt is yes, we perform fine-tuning of imGHUM during training.

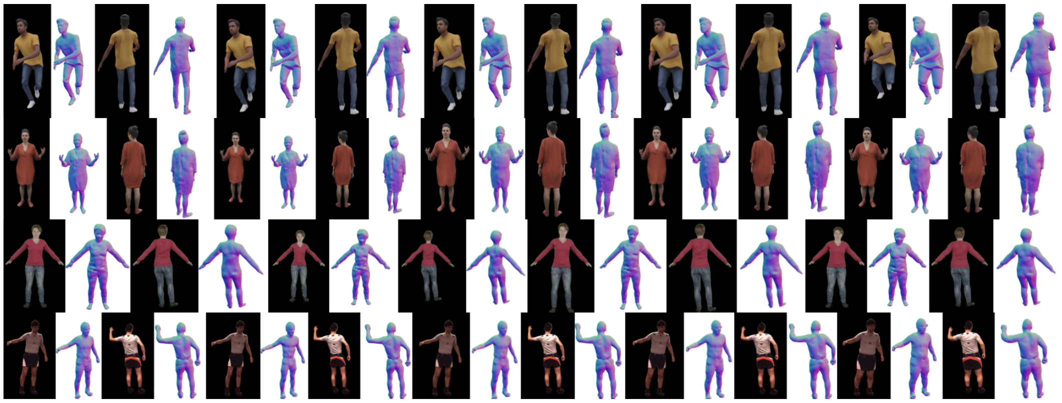

Figure 2: Qualitative evaluation of shape extrapolation for the four datasets. The leftmost is the training shape and we extrapolate both the rendering and the geometric reconstruction to four novel shapes (three for PeopleSnapshot). The shape volume change w.r.t. the learned shape for the 4 rows are: $-5\%, +29\%, +50\%, +87\%$; $-27\%, +48\%, +18\%, +32\%$; $-41\%, +18\%, +9\%$; $-15\%, +4\%, +18\%, +44\%$.

by the accuracy of imGHUM on test poses, and less so by pose differences themselves. However, one still needs to provide sufficient examples for learning pose-dependent geometry and appearance as shown in our frame number ablation experiment (fig.3 in the main paper).

Monocular videos. The PeopleSnapshot dataset consists of monocular videos of people rotating to a full circle in front of the camera. However, the subjects are consistently holding an A-Pose. We therefore additionally evaluate H-NeRF on two RenderPeople and two GHS3D dynamic sequences with large pose variations but trained with only a single camera. Tab. 3 and fig. 5, respectively, show numerical and visual results. Our method still shows strong robustness for novel view synthesis, geometric reconstruction and even pose extrapolation, although at some quality expense compared to models trained using four cameras. We notice that GHS3D-S47 shows a more significant drop in its metrics when trained in the monocular case, which may be caused by the back of the subject not being fully visible in the training images (whereas for the other sequences, the subject is turning in front of the camera). This experiment clearly demonstrates the capability of H-NeRF to integrate temporal image information into consistent neural radiance and SDF representations.

Similarly to NeuralBody [5], we rely on good pose estimation for consistent integration of image observations across multiple dynamic frames. Because of degraded monocular pose estimation, we excluded a small number of side view frames from the PeopleSnapshot dataset aiming to maximize the model quality. Note that for fair comparisons, we trained NeuralBody with the same images and poses as our method. In addition, we have trained a RenderPeople sequence with all frames and have indeed observed degraded novel image synthesis quality caused by the pose estimation inaccuracies (PSNR ↑: 27.5 (full) vs. 28.1 (filtered), SSIM ↑: 0.81 vs. 0.86, LPIPS ↓: 0.23 vs. 0.21). Our process of fine-tuning imGHUM pose and shape (Eq. 11 in the main paper) helps alleviate the problem (see Tab 4) and any other better pose estimation technique would be complementary to our approach.

2 Ablation Studies

Tab. 4 shows ablation studies for each loss for our four datasets. Training H-NeRF with all proposed losses strikes a good balance between image rendering quality and geometric reconstruction accuracy, and achieve the best performance for most metrics. For example, we observe significant drops in image quality when we do not condition the NeRF appearance under the root transformation. We largely benefit from fine-tuning the imGHUM parameters for the PeopleSnapshot dataset where we have the highest geometric fitting error given the use of only monocular videos. The terms $\mathcal{L}_{\text{blend}}$ and $\mathcal{L}_{\text{mask}}$ impact the image metrics significantly. For example, without $\mathcal{L}_{blend}$, the geometric metric is slightly better but leads to significantly worse SSIM for the real-world videos. The terms $\mathcal{L}_{\text{geom}}$, $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{eik}}$ impact the geometric metrics more, and we observe considerably worse reconstruction performance when turning these off. The Eikonal regularization helps smoothing the surface reconstruction and is critical for videos with insufficient views (PeopleSnapshot) or lower image resolution e.g. due to people placed farther away from the camera (Human3.6M). The term $\mathcal{L}_{\text{seg}}$ affects both rendering and geometry and we clearly see performance drops in its absence.

3 Training, Memory Consumption and Timings

We have trained the models for 10k iterations of 4k ray batch size on 8 Nvidia v100 GPUs, which takes about 6-8 hours for each dynamic sequence. For an image of 512×512 resolution, the inference for H-NeRF on a single Nvidia v100 GPU takes about 9.1 sec. The main computation overhead compared to the original NeRF formulation (about 6.5 sec) comes from the imGHUM warping. imGHUM warping also limits the maximum number of query points to be 64^3 due to memory constraints (i.e. 1024 rays consisting of 128 coarse and 128 fine samples). We utilize the coarse scene structuring such that we only query imGHUM for points inside the 3D bounding box $\mathbf{B}$. For exterior points (static background or free space) we use the original position coordinate and a constant distance value (1.0) as input features to the NeRF network. In practice, this significantly reduces the number of imGHUM query points by 90% and largely alleviates the memory constraints.

Parameters. For our training, the weights of $\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{mask}}, \mathcal{L}_{\text{blend}}, \mathcal{L}_{\text{geom}}, \mathcal{L}_{\text{seg}}, \mathcal{L}_{\text{reg}}, \mathcal{L}_{\text{eik}}, \mathcal{L}_{\text{fit}}$ and $\mathcal{L}_{\text{inc}}$ are validated to $1.0, 1.0, 0.1, 0.02, 0.02, 0.0005, 10^{-6}, 1.0$ and 0.02 respectively. We set $\sigma_h = 50$, $\gamma = 200$, and $\eta = 1.0$ for RenderPeople and GHS3D sequences but $\eta = 0.25$ for the other two datasets.

4 imGHUM Fitting

imGHUM shares the same shape and pose latent code with the mesh-based generative human model GHUM [7]. To fit the imGHUM parameters to a given image, we therefore use an off-the-shelf fitting approach for GHUM [9, 10]. We subsequently use the GHUM fit (pose and shape) in order to initialize imGHUM, which shares the same latent code. Specifically, given a monocular image or a set of images collected from multiple views, we use the neural network HUND [10] that takes a cropped human detection, and outputs 137 keypoints with confidences and 15 body-part semantic segmentation masks. HUND further predicts the GHUM shape $\boldsymbol{\beta}$ and pose values $\boldsymbol{\theta}$ based on the various semantic features (keypoints and segmentation masks) extracted from the image. Given HUND predictions, we follow up with a kinematic optimization that better aligns GHUM's predicted keypoints and semantics segmentation with corresponding primitives extracted in the calibrated input cameras. The optimization is formulated as image alignment for which we adopt the self-supervised keypoint and body segmentation loss from [9] (cf. their eq. 11), with L_2 regularization on the latent embeddings of the body shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. The temporal smoothness is applied with a L_2 loss on the pose differences between neighboring frames.

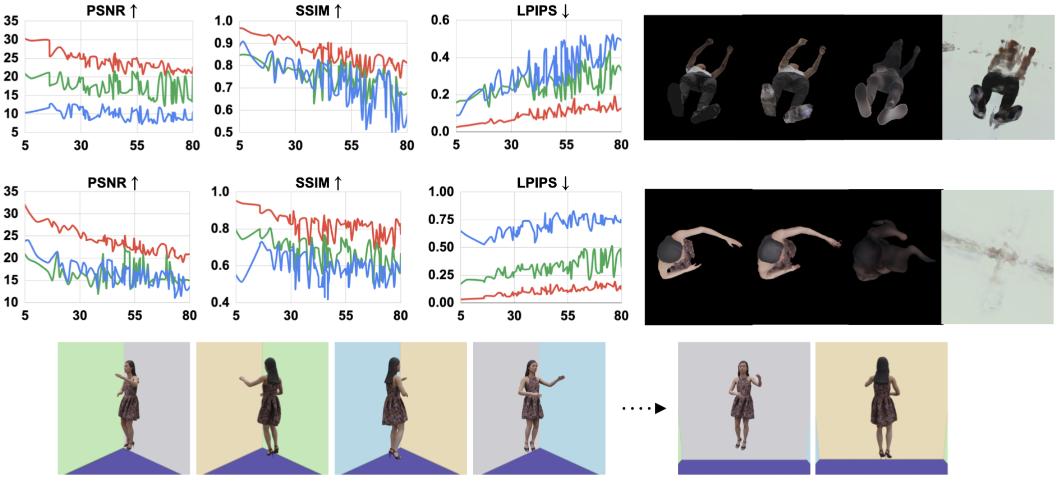

Figure 3: Image evaluation of novel view synthesis w.r.t. the minimal view difference to the training cameras (x-axis, in degrees). We evaluate on two static scenes with a model trained using four cameras. All models degrade in view synthesis quality when the test camera deviates from the training views but H-NeRF (red) significantly outperforms NeRF (blue) and IDR (green). Right: we show the ground truth, H-NeRF, IDR and NeRF from left to right, rendered under significantly different views compared to the training cameras. We note that H-NeRF still produces reasonable rendering, with image metrics degrading mostly due to body parts not visible in the training images (e.g. the elbow pit, bottom of the shoes). The last row shows the four training views on the left, and our two default test views rotated by 45 degrees away from training views.

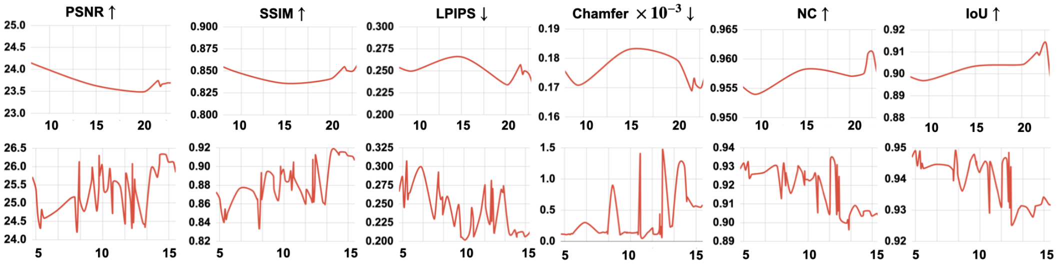

Figure 4: Image and geometric evaluation of H-NeRF on novel poses w.r.t. pose difference to the closest training pose (x-axis, in degrees, per body joint). The plots are based on two of our RenderPeople sequences. We do not observe significant degradation of the reconstruction quality as a function of differences to training poses, as long as imGHUM produces good predictions for the test poses.

Model	Dataset	PSNR ↑	SSIM ↑	LPIPS ↓	Ch $\times 10^{-3}$ ↓	NC ↑	IoU ↑
Nerfie [4]	RenderPeople	10.97	0.7	0.642	57.6	0.487	0
	GHS3D	10.8	0.685	0.635	50.1	0.49	0
H-NeRF (ours)	RenderPeople	**28.78**	**0.913**	**0.125**	**0.217**	**0.950**	**0.917**
	GHS3D	**24.92**	**0.852**	**0.232**	**0.218**	**0.932**	**0.89**

Table 2: Quantitative comparisons to Nerfie [4] on our dynamic sequences. Geometric metrics are only reported when ground-truth is available. We note that for Nerfie, Marching Cubes fails in geometric reconstruction for 5/8 RenderPeople and 13/14 GHS3D sequences. Hence, we only report numbers based on the the scenes where Nerfie produces outputs. Nerfie does not support the rendering of novel poses, therefore all metrics are evaluated on training poses. We do not report metrics for PeopleSnapshot and Human3.6M since no novel camera view with ground-truth images is available.

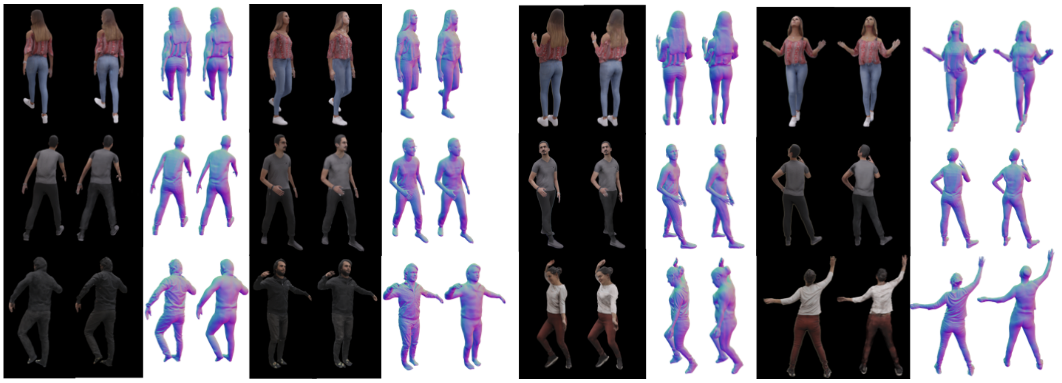

Figure 5: Qualitative evaluation of four dynamic monocular sequences. The first two rows correspond to RenderPeople-Tina and RenderPeople-Nagy, showing two novel view synthesis results for the training poses (on the left) and two novel poses seen from novel viewpoints (ground-truth image and geometry are shown by the left side). The last row shows novel view synthesis for two GHS3D sequences (S36 and S47 respectively).

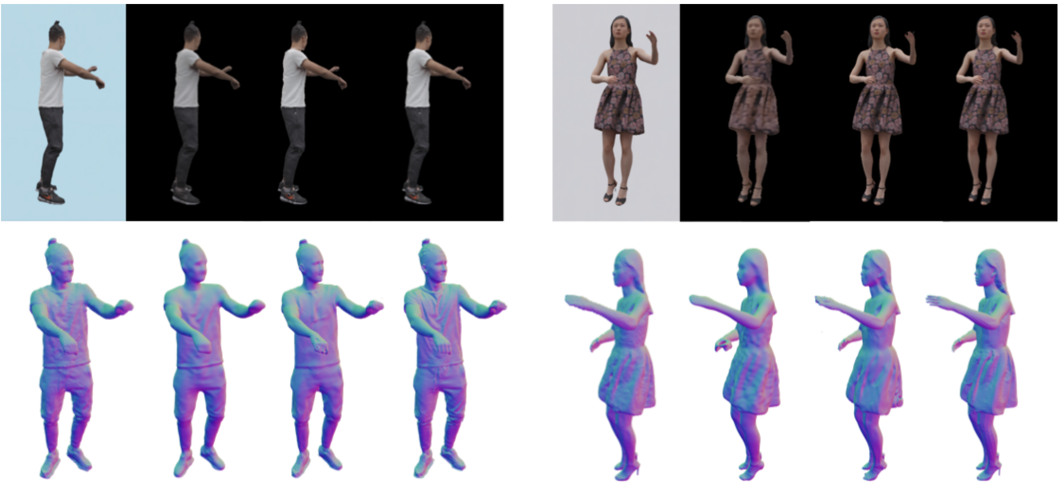

Figure 6: Qualitative evaluation of the static scenes for 32 cameras (from left to right: NeRF, IDR, H-NeRF and ground truth). With growing number of training cameras, the novel view synthesis and geometric reconstruction improve for all methods. NeRF and H-NeRF starts to converge whereas the IDR shows some differences due to the different image formation process.

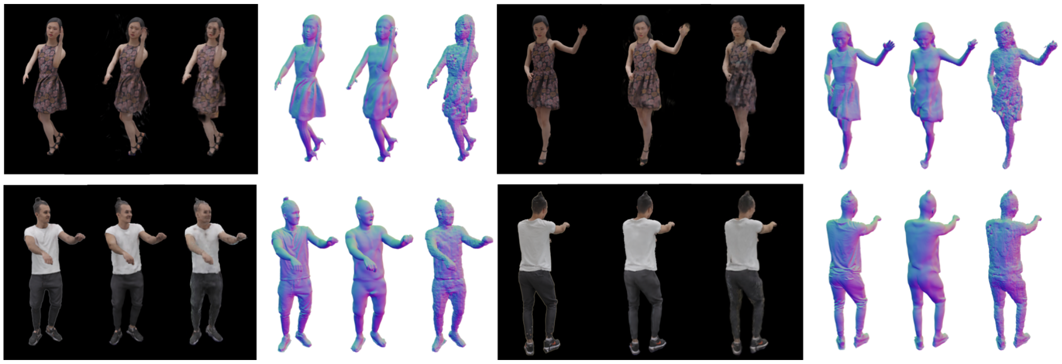

Figure 7: Qualitative comparison of H-NeRF and NeuralBody trained with sparse video frames (from left to right: ground truth, H-NeRF and NeuralBody). We illustrate novel view synthesis and geometric reconstruction for a RenderPeople sequence (trained with 10 frames) and a GHS3D sequence (trained with 20 frames). Our qualitative results (H-NeRF) consistently outperform NeuralBody.

Sequence	#cam	PSNR ↑	SSIM ↑	LPIPS ↓
RenderPeople-Tina	1	24.74/23.48/21.79	0.863/0.850/0.819	0.142/0.178/0.224
	4	27.23/23.56/23.53	0.9013/0.855/0.854	0.146/0.185/0.219
RenderPeople-Nagy	1	26.41/23.81/23.58	0.882/0.843/0.839	0.128/0.263/0.312
	4	30.44/23.73/24.16	0.936/0.838/0.849	0.112/0.286/0.317
GHS3D-S36	1	28.83	0.874	0.141
	4	29.41	0.896	0.134
GHS3D-S47	1	23.9	0.852	0.153
	4	24.2	0.871	0.117

Sequence	#cam	Ch $\times 10^{-3}$ ↓	NC ↑	IoU ↑
RenderPeople-Tina	1	0.327/0.419	0.94/0.922	0.898/0.86
	4	0.274/0.3	0.95/0.935	0.918/0.886
RenderPeople-Nagy	1	0.339/0.477	0.96/0.938	0.944/0.93
	4	0.119/0.143	0.97/0.95	0.957/0.95
GHS3D-S36	1	0.124	0.935	0.8853
	4	0.079	0.947	0.916
GHS3D-S47	1	0.434	0.907	0.8
	4	0.118	0.948	0.905

Table 3: Quantitative evaluation on monocular RenderPeople and GHS3D sequences, compared to models trained using four cameras. For RenderPeople, the three image metrics are reported corresponding to rendering of the training poses under novel cameras, novel poses under the training camera, and novel poses under novel cameras, respectively. Geometric metrics are reported as training poses/novel poses. For GHS3D, we only have ground-truth images and scans for the training poses and therefore we report the numbers for training poses under novel camera views.

Metric	Dataset	Full	$-\mathcal{L}_{\text{seg}}$	$-\mathcal{L}_{\text{mask}}$	$-\mathcal{L}_{\text{blend}}$	$-\mathcal{L}_{\text{geom}}$	$-\mathcal{L}_{\text{reg}}$	$-\mathcal{L}_{\text{eik}}$	-opt	-$\mathbf{T}$	-noise
PSNR ↑	RenderPeople	**28.35**	26.46	23.96	27.85	26.67	26.55	26.68	**28.35**	24.43	26.67
	GHS3D	**27.26**	26.74	25.63	26.88	26.79	25.68	26.72	26.70	25.72	26.90
	PeopleSn.	**28.12**	27.04	27.83	27.48	**28.12**	28.07	28.02	25.69	28.1	28.04
	Human3.6M	**24.4**	24.3	15.7	24.1	24.34	24.3	24.3	24.3	24.2	24.2
SSIM ↑	RenderPeople	**0.908**	0.874	0.819	0.872	0.882	0.881	0.878	**0.908**	0.867	0.879
	GHS3D	0.887	0.887	0.876	0.885	0.887	0.877	0.886	0.888	0.875	**0.891**
	PeopleSn.	**0.86**	0.84	0.835	0.478	0.858	0.858	0.858	0.329	0.832	0.857
	Human3.6M	0.874	0.875	0.69	0.651	0.871	0.87	0.874	**0.877**	0.865	0.869
LPSIS ↓	RenderPeople	**0.101**	0.115	0.152	0.106	0.113	0.115	0.115	**0.101**	0.125	0.119
	GHS3D	**0.203**	0.208	0.248	0.205	0.205	0.218	0.209	0.211	0.212	0.210
	PeopleSn.	0.206	0.228	0.205	0.211	0.196	0.196	**0.194**	0.235	0.206	0.196
	Human3.6M	**0.129**	0.133	0.671	0.138	0.13	0.132	0.134	0.135	0.134	0.132
Ch ↓	RenderPeople	0.121	0.129	0.107	**0.082**	0.128	0.136	0.103	0.121	0.1	0.107
	GHS3D	0.113	0.13	0.136	0.107	0.185	0.151	**0.104**	0.235	0.123	0.128
NC ↑	RenderPeople	**0.961**	0.958	0.958	**0.961**	0.947	0.958	0.959	**0.961**	0.96	0.96
	GHS3D	**0.947**	0.946	0.944	0.946	0.928	0.943	0.946	0.943	0.944	**0.947**
IoU ↑	RenderPeople	**0.943**	0.939	0.923	0.937	0.845	0.941	0.939	**0.943**	0.941	0.939
	GHS3D	0.915	0.913	0.914	**0.923**	0.851	0.911	0.915	0.91	0.91	0.909

Table 4: Ablation study on losses (Chamfer $\times 10^{-3}$). The last three columns (-opt, -$\mathbf{T}$, -noise) show results without imGHUM fine tuning, not conditioning the NeRF appearance using a root transformation $\mathbf{T}$, and not applying Gaussian noise to NeRF's condition code $(\boldsymbol{\theta}, \mathbf{T}, \mathbf{v})$, respectively. We do not apply imGHUM parameter fine-tuning on the RenderPeople sequence.

References

[1] CMU graphics lab motion capture database. 2009. `http://mocap.cs.cmu.edu/`.

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8387–8397. IEEE, 2018.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014.

[4] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[6] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, jun 2021.

[7] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2021.

[8] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Adv. Neural Inform. Process. Syst.*, 33, 2020.

[9] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Eur. Conf. Comput. Vis.*, 2020.

[10] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI*, pages 12749–12756, 2020.