DiffHuman: Probabilistic Photorealistic 3D Reconstruction of Humans

Supplementary Material

This supplementary material provides additional implementation details, experiments and qualitative results supporting the main manuscript. In particular, Section A gives further details on the losses used to train our DiffHuman model. Section B provides ablation studies investigating: (i) different types of observation sets, (ii) classifier-free diffusion guidance, and (iii) diffusion-via-rendering vs. hybrid diffusion. Moreover, we report mean and standard deviation metrics, to complement the "best of N" metrics reported in the main paper. Finally, Section C provides further qualitative comparisons with competing approaches, as well as some examples of unconditional 3D reconstruction samples.

A. Implementation Details

Our implicit surface diffusion model is trained using 2 different denoising objectives, given by Eqns. 10 and 12 in the main manuscript. In addition, we employ a probabilistic shaded rendering loss to ensure that 3D reconstruction samples are consistent with the 2D conditioning image, as well as several 3D losses on the intermediate implicit surface to stabilise training. These are detailed below.

Shaded rendering loss. We want 3D reconstruction samples, represented by implicit surfaces $\mathcal{S}_\Theta^{(t)}$, to be consistent with the conditioning image $\mathbf{I}$ at every timestep t in the reverse diffusion process. This is achieved by enforcing shaded front renders of $\mathcal{S}_\Theta^{(t)}$ to match $\mathbf{I}$. Shaded renders can be obtained using the albedo and surface normal images that comprise the observation sets we use for diffusion-via-rendering and hybrid diffusion.

Recall that ground-truth observation sets consist of $\boldsymbol{x}_0 = \{\mathbf{A}^F, \mathbf{A}^B, \mathbf{N}^F, \mathbf{N}^B, \mathbf{D}^F, \mathbf{D}^B\}$. The reverse process generates samples by repeatedly estimating denoised observations given noisy observations $\boldsymbol{x}_t$. A denoised estimate obtained using `render` is denoted as $\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}$, while $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$ represents an estimate given by `generate` (see Eqns. 8 and 11 in the main paper). Front albedo $\mathbf{A}^F$ and front normals $\mathbf{N}^F$ may be used in conjunction with the shading neural network $s_\Theta^{(t)}$ to obtain shaded front images $\mathbf{C}^{(t)}$ at each timestep t. We compute separate $\mathbf{C}^{(t)}_{\texttt{render}}$ and $\mathbf{C}^{(t)}_{\texttt{generate}}$ using the elements of $\hat{\boldsymbol{x}}_{0_\Theta}^{(t)}$ and $\bar{\boldsymbol{x}}_{0_\Theta}^{(t)}$, respectively. During training, we apply L_2 losses between shaded renders and the condition $\mathbf{I}$, given by

$$\mathcal{L}_{\text{shaded}}^{\texttt{render}} = \|\mathbf{C}^{(t)}_{\texttt{render}} - \mathbf{I}\|_2^2 \tag{1}$$

$$\mathcal{L}_{\text{shaded}}^{\texttt{generate}} = \|\mathbf{C}^{(t)}_{\texttt{generate}} - \mathbf{I}\|_2^2. \tag{2}$$

Note that shaded render losses $\mathcal{L}_{\text{shaded}}$ have a similar form to the the denoising diffusion objectives $\mathcal{L}_{\text{VLB}}$. However, $\mathcal{L}_{\text{shaded}}$ enforces consistency between an estimated observation set and the *conditioning image* $\mathbf{I}$, while $\mathcal{L}_{\text{VLB}}$ is applied between the estimated and ground-truth observation sets.

Loss	Symbol	Type	Weight
Denoising-via-rendering	$\mathcal{L}_{\text{VLB}}^{\texttt{render}}$	Probabilistic	1.0
Denoising-via-generation	$\mathcal{L}_{\text{VLB}}^{\texttt{generate}}$	Probabilistic	1.0
Shaded rendering	$\mathcal{L}_{\text{shaded}}^{\texttt{render}}$	Probabilistic	1.0
Shaded generation	$\mathcal{L}_{\text{shaded}}^{\texttt{generate}}$	Probabilistic	1.0
On-surface SDF on $d_{\boldsymbol{p}}$	-	Deterministic	1.0
On-surface albedo on $\boldsymbol{a}_{\boldsymbol{p}}$	-	Deterministic	0.2
On-surface normals on $\boldsymbol{n}_{\boldsymbol{p}}$	-	Deterministic	0.2
Near-surface In/Out on $d_{\boldsymbol{p}}$	-	Deterministic	0.2
Near-surface albedo on $a_{\boldsymbol{p}}$	-	Deterministic	0.2
Eikonal	-	Regulariser	0.05

Table 1. Summary of the probabilistic, deterministic and regularisation losses used to train our model. Loss weights are provided.

Deterministic 3D losses. $\mathcal{L}_{\text{VLB}}$ and $\mathcal{L}_{\text{shaded}}$ are probabilistic losses applied within the diffusion framework. In addition, we employ several deterministic 3D losses on surface geometry and albedo, following PIFu [6] and PHORHUM [1], which improves training stability in our experience.

Specifically, we supervise SDF values $d_{\boldsymbol{p}}$, albedo field values $\boldsymbol{a}_{\boldsymbol{p}}$ and per-point normals $\boldsymbol{n}_{\boldsymbol{p}}$ at 3D points $\boldsymbol{p}$ sampled from the ground-truth 3D human surface. $d_{\boldsymbol{p}}$ is enforced to be 0 for these on-surface points. Additionally, we supervise the sign of samples taken around the surface using an inside-outside classification loss implemented using binary cross-entropy. This is applied to the SDF values $d_{\boldsymbol{p}}$ for near-surface points. Moreover, following [6], the albedo colour field $\boldsymbol{a}_{\boldsymbol{p}}$ is also supervised for near-surface points. The ground-truth near-surface albedo at $\boldsymbol{p}$ is approximated using the albedo of the nearest neighbour on the ground-truth surface. Finally, we use an Eikonal geometric regulariser [3] to enforce SDF predictions to have unit norm gradients everywhere.

All losses are summarised in Tab. 1 in this supplementary material, which also provides associated loss weight hyperparameters. Note that deterministic losses are generally weighted lower than probabilistic losses, to encourage sample diversity. Future work may investigate the feasibility of removing deterministic losses altogether.

B. Ablation Studies

This section presents additional ablation studies. We begin with a discussion of "best-of-N" vs. mean metrics, and report means and standard deviations to complement the best-of-N metrics given in the main paper. Then, we provide a

	Render Freq.	Runtime s / sample	3D CD ↓	3D NC ↑	Albedo Front LPIPS ↓	Albedo Front PSNR ↑	Albedo Back LPIPS ↓	Albedo Back PSNR ↑	Normals Front Ang. ↓	Normals Back Ang. ↓	Shaded Front PSNR ↑
Best of $N = 5$	Per step	496	**0.99**	0.85	0.13	22.92	0.25	20.95	21.59	22.88	26.57
	Per 10	86	1.12	**0.87**	0.12	23.24	0.24	**21.07**	**19.05**	**22.46**	27.08
	Per 25	34	1.16	0.86	0.11	**23.26**	0.23	21.06	19.11	22.52	27.09
	Final	**9**	1.16	0.86	**0.11**	23.26	**0.22**	21.05	19.12	22.55	**27.09**
Mean ± Std. $N = 5$	Per step	496	**1.03** ± 0.8	0.83 ± 0.04	0.14 ± 0.03	22.34 ± 2.39	0.27 ± 0.07	20.34 ± 3.35	22.47 ± 3.13	23.42 ± 5.49	26.18 ± 1.90
	Per 10	86	1.33 ± 0.9	**0.85** ± 0.05	0.13 ± 0.04	22.63 ± 2.40	0.24 ± 0.08	**20.46** ± 3.36	**20.14** ± 3.20	**23.15** ± 5.64	26.82 ± 1.91
	Per 25	34	1.37 ± 0.9	0.84 ± 0.05	0.13 ± 0.04	22.64 ± 2.43	0.24 ± 0.08	20.45 ± 3.38	20.20 ± 3.16	23.31 ± 5.56	26.83 ± 1.90
	Final	**9**	1.38 ± 0.9	0.84 ± 0.05	**0.12** ± 0.04	**22.65** ± 2.43	**0.23** ± 0.08	20.46 ± 3.37	20.24 ± 3.14	23.20 ± 5.54	**26.84** ± 1.90

Table 2. Ablation of hybrid implicit surface diffusion. $N = 5$ samples are obtained using 100 DDIM [8] steps. We ablate periodic `render` every 1, 10 and 25 steps, as well as only in the final step. The latter only involves running Marching Cubes [5] for mesh extraction in the final step. All other denoising steps use `generate`. While per step `render` performs best on 3D metrics, predominantly using `generate` results in better perceptual quality in rendered metrics. The **best** and second best results are marked. We report both best-of-N and mean (± std.) metrics for completeness.

	Observations in $\boldsymbol{x}_0$ Type	Observations in $\boldsymbol{x}_0$ View	3D CD ↓	3D NC ↑	Albedo Front LPIPS ↓	Albedo Front PSNR ↑	Albedo Back LPIPS ↓	Albedo Back PSNR ↑	Normals Front Ang. ↓	Normals Back Ang. ↓	Shaded Front PSNR ↑
Best of $N = 5$	$\mathbf{A}, \mathbf{N}$	F, B	1.11	0.84	0.14	21.39	0.26	18.74	20.24	23.86	25.13
	$\mathbf{A}, \mathbf{D}$	F, B	**1.05**	0.78	0.14	21.68	0.27	19.57	19.14	**22.39**	24.94
	$\mathbf{A}, \mathbf{N}, \mathbf{D}$	F	1.18	0.86	0.12	23.04	0.27	19.17	19.13	22.59	25.97
	$\mathbf{A}, \mathbf{N}, \mathbf{D}$	F, B	1.16	**0.86**	**0.11**	**23.26**	**0.22**	**21.05**	**19.12**	22.55	**27.09**
Mean ± Std. $N = 5$	$\mathbf{A}, \mathbf{N}$	F, B	1.31 ± 0.8	0.82 ± 0.06	0.15 ± 0.04	20.96 ± 2.22	0.27 ± 0.08	18.31 ± 2.93	21.69 ± 4.34	24.65 ± 5.63	24.93 ± 2.07
	$\mathbf{A}, \mathbf{D}$	F, B	**1.29** ± 1.0	0.76 ± 0.10	0.15 ± 0.04	21.07 ± 2.33	0.28 ± 0.08	18.83 ± 2.70	20.25 ± 3.62	**22.89** ± 6.17	24.46 ± 2.68
	$\mathbf{A}, \mathbf{N}, \mathbf{D}$	F	1.37 ± 0.8	0.83 ± 0.06	0.13 ± 0.03	21.67 ± 2.30	0.28 ± 0.09	18.70 ± 2.71	**19.75** ± 3.85	23.27 ± 6.21	25.88 ± 2.13
	$\mathbf{A}, \mathbf{N}, \mathbf{D}$	F, B	1.38 ± 0.9	**0.84** ± 0.05	**0.12** ± 0.04	**22.65** ± 2.43	**0.24** ± 0.08	**20.46** ± 3.37	20.24 ± 3.14	23.20 ± 5.54	**26.82** ± 1.90

Table 3. Quantitative comparison between different types observation sets $\boldsymbol{x}_0$ used during implicit surface diffusion. $\mathbf{A}$, $\mathbf{N}$ and $\mathbf{D}$ refer to albedo, surface normal and depth images respectively. F and B designate front and back views. Populating $\boldsymbol{x}_0$ with front and back views of all 3 observation types gives the best all-round performance. Thus, this is the protocol used in the default DiffHuman model presented in the main paper. The **best** and second best results are marked. We report both best-of-N and mean (± std.) metrics for completeness.

more detailed comparison of diffusion-via-rendering vs. our novel hybrid diffusion framework. We investigate different types of observation sets for diffusion, by dropping particular observations from $\boldsymbol{x}_0$. Finally, we implement classifier-free diffusion guidance [4] with an unconditional model and report corresponding metrics.

Mean vs. best-of-N metrics. The main paper reports evaluation metrics using the best-of-$N \in \{1, 5, 10\}$ reconstructions. This is justified for ambiguous metrics that measure performance in ill-posed tasks, where the ground-truth is *but one* plausible solution. Our method is able to yield other solutions that are consistent with the input image but differ from the ground truth. Capturing the ground-truth within the range of solutions modelled by our predicted distributions is sufficient – this is measured by best-of-N.

However, not all metrics correspond to ill-posed tasks. In particular, "shaded front" metrics (*e.g.* PSNR) measure the match between 3D reconstruction samples and the input image. *All* samples should be input-consistent; hence, reporting the mean over N samples is logical. This is arguably also true for LPIPS, which measures perceptual similarity, as noted by [9]. Therefore, Tabs. 2 and 3 in this supplementary material report means and standard deviations, in addition to best-of-N metrics. For completeness, these are provided for both well-posed and ill-posed tasks. Note that standard deviations are generally higher for back albedo and normals than the front. This is desired, and signifies greater diversity in unseen regions. Furthermore, standard deviations are lower for front shading, which should consistently match the conditioning image.

Hybrid implicit surface diffusion. Tab. 1 in the main paper gives a brief ablation of our novel hybrid implicit surface diffusion framework. We provide more detailed results in Tab. 2 in this supplementary material, where we compare denoising via `render`, via `generate`, and using a combination of both. The performances of all these methods are comparable, suggesting that the `generate` neural network has learned to imitate explicit rendering well. However, `generate` has a much reduced runtime – specifically giving a $55\times$ speed-up over the reverse process. A qualitative comparison of these denoising strategies is visualised in Fig. 5 in this supplementary material.

Observations in $\boldsymbol{x}_0$. DiffHuman models a distribution over image-based, pixel-aligned observations of an implicit 3D surface $\mathcal{S}$. The default method utilises three types of observations of the front and back surfaces of $\mathcal{S}$: (i) unshaded albedo colour images $\mathbf{A}^F$ and $\mathbf{A}^B$, (ii) surface normal images $\mathbf{N}^F$ and $\mathbf{N}^B$ and (iii) depth maps $\mathbf{D}^F$ and $\mathbf{D}^B$. In this supplementary material, we investigate the importance of each of these observations. Specifically, we train 3 ablation models by omitting depth, normals and back views in turn from the observation set $\boldsymbol{x}_0$. A quantitative comparison is provided in Tab. 3. Utilising all of aforementioned observation types in $\boldsymbol{x}_0$ gives the best all-round performance on a range of

Train with Drop Cond.	Test with Guidance	3D		Albedo Front		Albedo Back		Normals Front	Normals Back	Shaded Front
		CD ↓	NC ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	Ang. ↓	Ang. ↓	PSNR ↑
Yes	No	1.29	0.84	0.13	22.22	0.25	20.37	20.38	22.90	26.07
Yes	Yes	1.42	0.82	0.14	21.37	0.25	19.98	22.25	24.31	26.99
No	No	**1.16**	**0.86**	**0.11**	**23.26**	**0.23**	**21.06**	**19.11**	**22.46**	**27.09**

Table 4. Quantitative evaluation of classifier-free diffusion guidance [4] applied to DiffHuman. We jointly train an unconditional and conditional implicit surface diffusion model, by randomly dropping the conditioning image $\mathbf{I}$. The effectiveness of guidance with the unconditional model is evaluated in rows 1 and 2. Guidance improves the match between 3D reconstruction samples and the conditioning image, as measured by "Shaded Front" metrics. However, it causes a deterioration of most other metrics. In addition, we report results from the standard DiffHuman model trained without random condition dropping in row 3. This consistently outperforms the model trained with dropping. All metrics are best-of-$N = 5$. We use a guidance weight of 3. The **best** and second best results are marked.

metrics. Dropping back views intuitively worsens metrics computed with back renders. Omitting depth and normals also generally degrades performance - apart from Chamfer distance. However, we note that Chamfer distance is a noisy metric, as evidenced by the large relative standard devitions, and it is difficult to make conclusive judgements from these results.

Classifier-free guidance [4] is an inference-time technique used to trade-off sample quality (including input-consistency) vs. diversity in conditional diffusion models. We experimented with applying guidance to our method, by jointly training a conditional and an unconditional implicit surface diffusion model. In practice, this was achieved by randomly dropping the conditioning image $\mathbf{I}$ as a network input with probability 0.2. Quantitative results are reported in Tab. 4 in this supplementary material. We found that guidance with an unconditional model can indeed improve the match between 3D reconstruction samples and the conditioning image, as measured by metrics corresponding to shaded front renders. However, it caused a deterioration of most other metrics – shown by row 1 vs. row 2 in Tab. 4. Moreover, training with random condition dropping yielded worse performance than a model that always sees a conditioning image. Perhaps a larger and more diverse training dataset is needed to fully realise the benefits of diffusion guidance in this task. Nevertheless, we find it instructive to visualise unconditional generation samples in Fig. 1 in this supplementary material. These exhibit a significant amount of diversity, covering a range of clothing and hair styles, colours and geometries. Moreover, unconditional samples are generated starting from random noise in the silhouette of a particular body shape (see Fig. 1). This is a by-product of the fact that our method applies foreground masking to all neural network inputs. Noise within a silhouette can be considered as a form of implicit conditioning, and allows us to exert control over the body shapes of 3D human samples.

C. Qualitative Results

This section provides further qualitative comparisons with current deterministic approaches to photorealistic 3D human reconstruction. Fig. 3 visualises samples from DiffHuman against reconstructions from PHORHUM [1] and S3F [2] – both of which estimate surface geometry, albedo colour and illumination-dependent shading. Fig. 4 compares DiffHuman with methods that only estimate surface geometry: PIFuHD [7], ICON [10] and ECON [11].

Furthermore, we present qualitative results from additional experiments investigating the feasibility of DiffHuman as a generative model. As mentioned previously, Fig. 1 visualises unconditional 3D human samples generated from random noise in the silhouette of a given body shape. This allows us to loosely control the shape of 3D human samples. We extend this approach, by experimenting with using edge maps as conditioning images – inspired by ControlNet [12]. This allows us to have more fine-grained control over 3D samples, without having to provide a full RGB conditioning image. Qualitative results are given in Fig. 2. These serve as a proof-of-concept for controllable generative applications beyond reconstruction.

Finally, Fig. 5 compares implicit surface diffusion via `render` vs. `generate`, to support the ablation studies presented in Tab. 2.

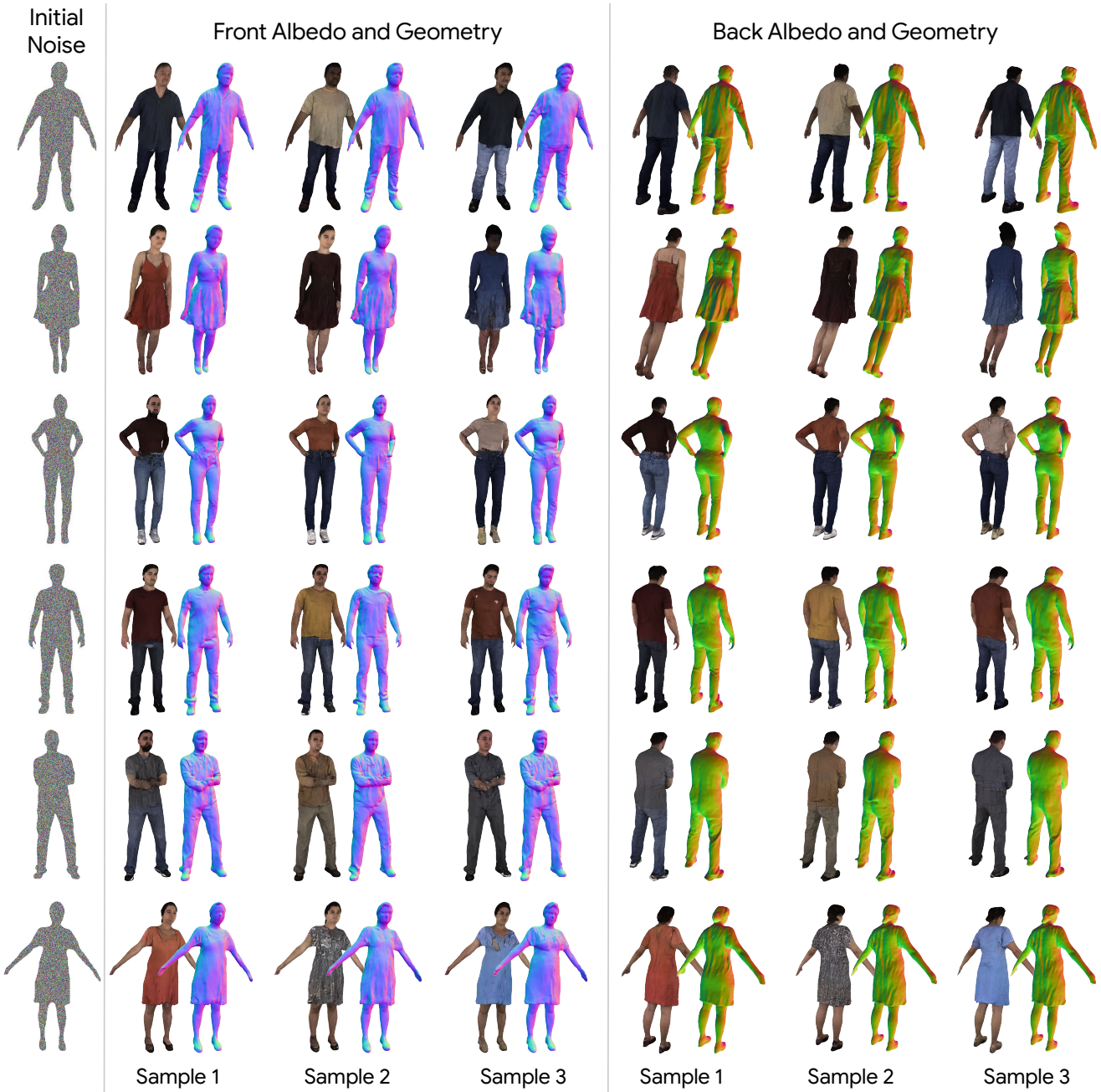

Figure 1. **Unconditional 3D reconstruction samples** generated using an implicit surface diffusion model trained with random condition dropping (following the protocol of classifier-free guidance [4]). These unconditional samples are generated from random noise only, which is masked using a silhouette in the shape of the desired subject. They exhibit significant diversity in terms of clothing styles, colours and geometries, as well as hairstyles, facial features and skin tones. For more ambiguous body shapes, different gendered properties are visible. The silhouette masking can be considered as a form of implicit conditioning, and allows us to exert some control over the 3D samples. Faces and certain body parts are blurrier for these unconditional samples than the conditional samples visualised in other figures. This is somewhat unsurprising, since conditioning images carry a lot of information on these fine features, which unconditional samples are not privy to.

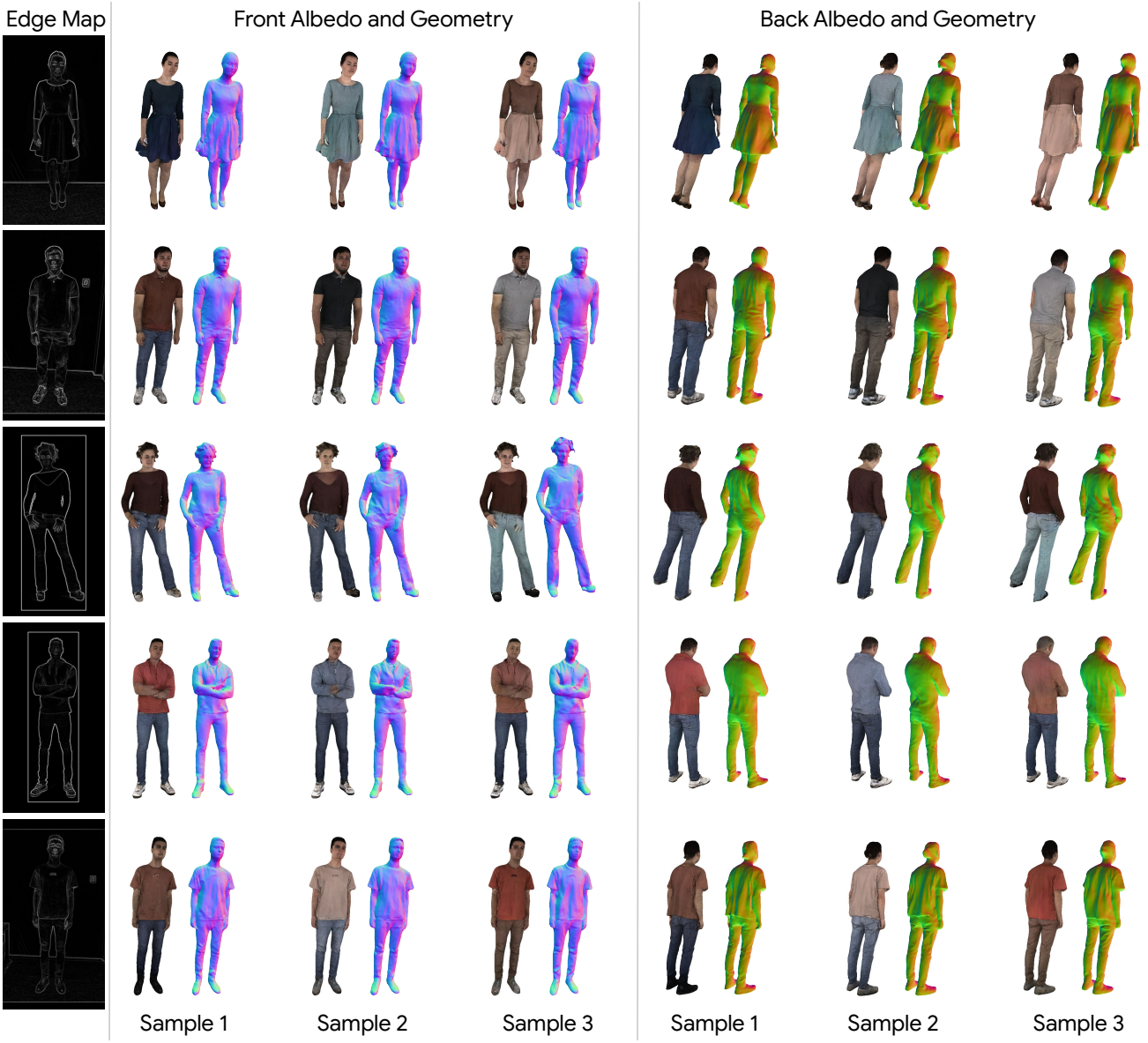

Figure 2. **3D reconstruction samples conditioned on edge map inputs.** These samples are generated using an implicit surface diffusion model that was pre-trained with conditioning RGB images, and then fine-tuned using conditioning edge maps – inspired by ControlNet [12]. Edges are obtained as image gradients using the Sobel operator. The 3D samples exhibit diverse colours, while the surface geometry respects the edge maps. This experiment demonstrates that samples from DiffHuman can be controlled via simpler conditioning inputs than full RGB images, which opens the possibility for generative applications beyond reconstruction from monocular images.

Figure 3. **Qualitative comparison against deterministic monocular 3D human reconstruction methods that predict geometry, surface albedo and shaded colour: PHORHUM [1] and S3F [2].** We show results from the original PHORHUM paper – not our retrained version. Our method, DiffHuman, predicts a distribution over 3D reconstructions from which we can draw multiple samples. We visualise 2 samples from the back and 1 sample from the front for our method. PHORHUM outputs good front predictions, but exhibits flat geometry and blurry colours on the back. S3F [2] yields more detailed geometry, but colours are still often blurry. Moreover, shaded renders of the reconstructions from each of these methods do not consistently match the input image. Our method is able to output multiple samples that are detailed, both in seen and unseen regions. In particular, note the hair geometry in row 1 and diversity of dress styles (from the back) in row 5. Samples from our method exhibit a greater level of input-consistency, as shown by the shaded renders in rows 1, 2 and 4. Furthermore, we can faithfully handle a wider variety of body shapes, such as row 4.

Figure 4. **Qualitative comparison against deterministic monocular 3D human reconstruction methods that predict only surface geometry: PIFuHD [7], ICON [10] and ECON [11].** Our method, DiffHuman, predicts a distribution over 3D reconstructions from which we can draw multiple samples. We visualise 2 samples from the back and 1 sample from the front for our method. Samples from our method exhibiti greater geometric detail, both in seen and unseen and regions. In particular, note the front of the suit jacket in row 1, skirt in row 3, trousers in row 4 and hood in row 5. Moreover, when such details are unlikely – *e.g.* the back of the jacket in row 1, which is typically flat – our method plausibly outputs samples with simpler geometry. Samples differ in hair styles and clothing colours on the back.

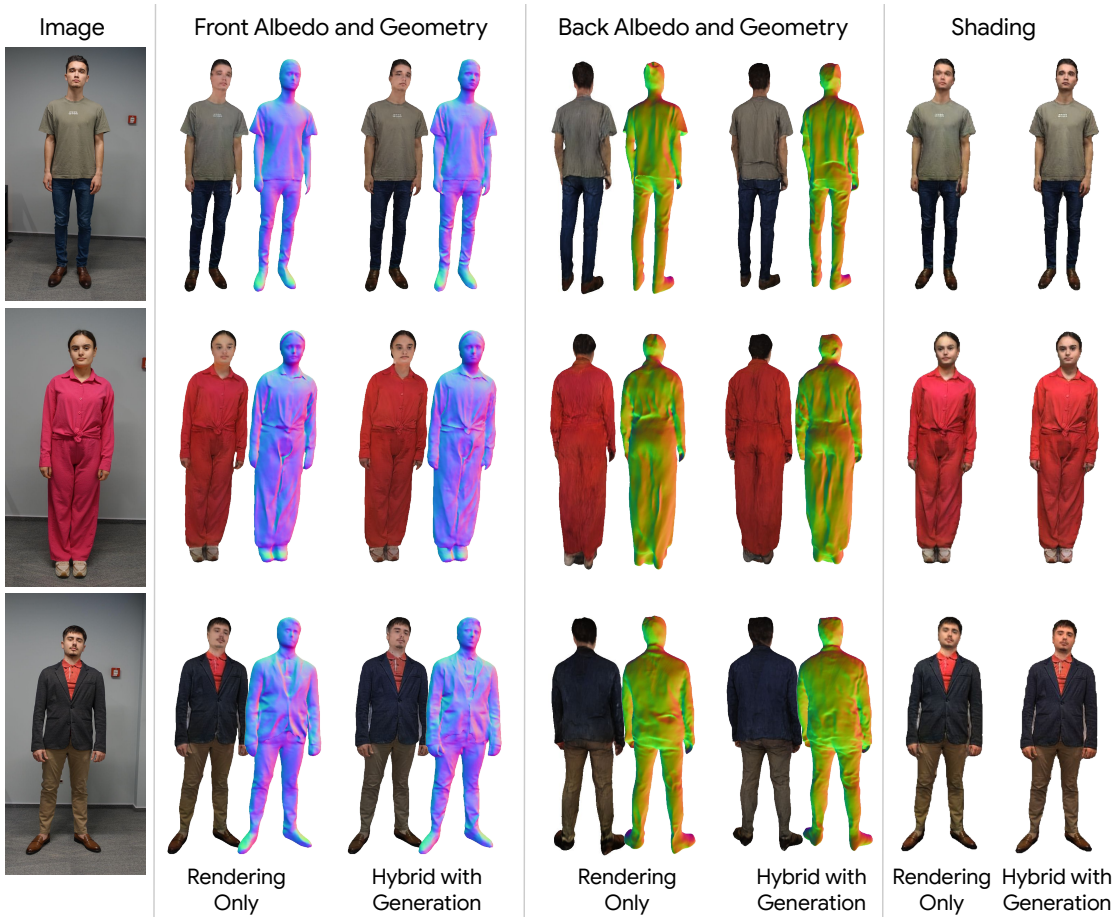

Figure 5. **Qualitative comparison between implicit surface diffusion via rendering, and hybrid diffusion using both rendering and generation.** Diffusion via rendering involves rendering an intermediate 3D representation in each denoising diffusion step to obtain a denoised sample. Hybrid diffusion uses a generator network that imitates rendering during the denoising process, at a much faster runtime. This figure complements Table 1 in the main manuscript and Table 2 in this supplement, by showing that samples from both these denoising processes are similar – quantitatively and qualitatively. This suggests that the generator network learns to imitate explicit rendering sufficiently well. In fact, samples obtained via generation are often perceptually preferable to rendered samples (see the face in row 2). This could be because the generator network focuses solely on synthesising realistic observations, and is not constrained by explicit 3D geometry.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3D reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3, 6

[2] Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanfir, and Cristian Sminchisescu. Structured 3d features for reconstructing relightable and animatable avatars. In *CVPR*, 2023. 3, 6

[3] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *ICML*, 2020. 1

[4] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 2, 3, 4

[5] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3D surface construction algorithm. *SIGGRAPH*, 1987. 2

[6] 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 (ICCV)*, 2019. 1

[7] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 7

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 2

[9] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[10] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 7

[11] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *CVPR*, 2023. 3, 7

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. 2023. 3, 5