Supplementary Material: TotalSelfScan: Learning Full-body Avatars from Self-Portrait Videos of Faces, Hands, and Bodies

Anonymous Author(s)
Affiliation
Address
email

In this supplementary material, we describe the implementation details and provide the detailed quantitative results of each person. Moreover, we provide a video to show the qualitative results of our method. The code and dataset will be released upon the publication of the paper.

1 Implementation details

Network architecture and hyper-parameters. For each part p, the signed distance field F_s^p, color field $F_{\mathbf{c}}^p$, and non-rigid displacement field T_{nr}^p are all represented as MLP networks whose dimension of hidden layers is 256. The signed distance field F_s^p has eight layers with the softplus activation and has a skip connection to the middle layer. The input of F_s^p is the positional encoding of point position $\gamma_{\mathbf{x}}(\mathbf{x}) \in R^{39}$ and the outputs are signed distance $s(\mathbf{x}) \in R$ and geometry feature $\mathbf{z}(\mathbf{x}) \in R^{256}$.

The color field F_c^p has four layers with the ReLU activation and has a skip connection to the middle layer. The inputs of F_c^p are the positional encoding of point position $\gamma_{\mathbf{x}}(\mathbf{x}) \in R^{39}$, the positional encoding of view direction $\gamma_{\mathbf{d}}(\mathbf{d}) \in R^{27}$, normal of point position $\mathbf{n}(\mathbf{x}) \in R^3$, and latent code $\boldsymbol{\ell}^p \in R^{128}$. The output of F_c^p is the color $\mathbf{c}(\mathbf{x}) \in R^3$.

The non-rigid displacement field T_{nr}^p has eight layers with the ReLU activation and has a skip connection to the middle layer. The inputs of T_{nr}^p are the positional encoding of point position $\gamma_{\mathbf{x}}(\mathbf{x}) \in R^{39}$ and latent code $\boldsymbol{\phi}^p \in R^{128}$ and the output is the displacement $\Delta\mathbf{x} \in R^3$.

The λ_1 in Equation (10) is set to 0.1. The λ_2^ps in Equation (10) are set to 0.01 for the body and face and 0.001 for the hands.

Training details. The Adam optimizer [1] is adopted for training and the learning rate is set as $5e^{-4}$ which decays exponentially to $5e^{-5}$ during the learning procedure. The training is done on a single Nvidia 2080-Ti GPU. Each part model is trained for $100k$ iterations, which takes approximately 6 hours. In addition, the appearance latent code optimization adopts the same optimizer and the learning rate as the training but conducts approximately $15k$ iterations for each part model except the body.

Inverse LBS. Given a sample point $\mathbf{x}$ in the observation space and the human pose $\mathbf{p}$, the inverse linear blend skinning [2, 7, 6] can be written as follows:

$$T_{ilbs}(\mathbf{x}, \mathbf{p}) = \left(\sum_{k=1}^{K} w^k(\bar{\mathbf{x}})\mathbf{B}(\mathbf{p})^k\right)^{-1} \bar{\mathbf{x}}, \tag{1}$$

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Quantitative results of 3D reconstruction of each part for each subject on *SynTotalHuman* dataset.

	m1		m2		f1		f2	
Head	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
NeuralBody [5]	1.64	1.38	1.64	1.41	1.68	1.81	1.23	1.26
AniNeRF [3]	1.55	1.34	1.73	1.54	2.24	2.08	1.68	1.82
AniSDF [4]	0.45	0.59	1.05	0.99	0.90	1.11	0.64	0.95
Ours	0.40	0.53	0.67	0.74	0.69	1.04	0.61	0.95
Left hand	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
NeuralBody [5]	1.54	1.37	1.00	0.93	1.97	1.80	1.38	1.27
AniNeRF [3]	1.03	0.89	1.06	0.96	1.37	1.17	1.00	0.93
AniSDF [4]	0.52	0.52	0.71	0.74	1.11	1.03	0.90	0.90
Ours	0.44	0.49	0.37	0.37	0.58	0.52	0.87	0.88
Right hand	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
NeuralBody [5]	1.25	1.10	1.02	1.02	1.26	1.14	0.92	0.88
AniNeRF [3]	1.16	1.01	1.17	1.06	1.19	1.05	0.98	0.92
AniSDF [4]	0.47	0.48	0.66	0.64	1.05	0.90	0.88	0.77
Ours	0.33	0.34	0.33	0.35	0.79	0.67	0.67	0.56
Total	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓	P2S↓	CD↓
NeuralBody [5]	2.13	1.86	2.29	2.15	2.12	1.97	2.10	1.93
AniNeRF [3]	2.49	2.15	2.28	2.07	2.72	2.29	2.72	2.35
AniSDF [4]	2.04	2.14	2.12	2.18	1.74	1.81	1.69	1.75
Ours	1.98	2.07	1.97	2.03	1.73	1.76	1.68	1.71

Table 2: Quantitative results of image synthesis under novel human pose of each part for each subject on *SynTotalHuman* dataset.

	m1			m2			f1			f2		
Head	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody[5]	19.20	0.918	0.174	20.44	0.915	0.188	19.19	0.922	0.175	16.92	0.901	0.198
AniNeRF[3]	19.74	0.926	0.167	21.07	0.930	0.173	21.53	0.932	0.130	21.33	0.925	0.141
AniSDF[4]	22.50	0.939	0.093	20.98	0.925	0.113	22.52	0.940	0.103	21.62	0.933	0.106
Ours	22.64	0.944	0.071	21.32	0.926	0.088	21.32	0.930	0.102	23.66	0.939	0.076
Left hand	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody[5]	18.75	0.889	0.245	18.92	0.885	0.235	17.71	0.895	0.243	18.26	0.890	0.228
AniNeRF[3]	19.23	0.902	0.177	21.27	0.904	0.193	22.16	0.923	0.148	20.39	0.911	0.144
AniSDF[4]	22.05	0.935	0.105	21.99	0.934	0.107	19.93	0.922	0.122	21.64	0.934	0.080
Ours	21.78	0.933	0.096	23.11	0.942	0.065	21.25	0.938	0.083	22.45	0.945	0.062
Right hand	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody[5]	19.23	0.894	0.243	19.08	0.894	0.243	18.01	0.901	0.252	18.12	0.893	0.233
AniNeRF[3]	19.48	0.905	0.200	20.58	0.902	0.220	21.69	0.910	0.178	19.99	0.897	0.172
AniSDF[4]	21.86	0.930	0.117	21.52	0.927	0.121	19.62	0.920	0.154	21.86	0.938	0.084
Ours	22.80	0.936	0.090	23.38	0.941	0.073	22.61	0.949	0.070	22.56	0.945	0.066
Total	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeuralBody[5]	19.19	0.831	0.231	20.27	0.854	0.248	20.22	0.869	0.241	20.22	0.866	0.237
AniNeRF [3]	21.46	0.881	0.199	24.31	0.894	0.197	24.91	0.884	0.173	24.14	0.896	0.171
AniSDF[4]	25.54	0.907	0.144	26.51	0.916	0.136	23.34	0.914	0.129	26.87	0.928	0.107
Ours	26.78	0.919	0.127	26.94	0.919	0.123	23.36	0.915	0.112	27.52	0.932	0.095

where $\bar{\mathbf{x}}$ denotes the homogeneous coordinate of $\mathbf{x}$ and $\mathbf{B}(\mathbf{p})^k \in SE(3)$ denotes the transformation matrix of bone k. $w^k(\bar{\mathbf{x}})$ is the blend weight of bone k, which is acquired by retrieving the blend weight of the closest vertex on the template mesh. K is the number of bones.

2 Detailed quantitative results

In this section, we present the detailed quantitative results of 3D reconstruction and image synthesis on *SynTotalHuman* dataset in Table 1 and 2.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[2] 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] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14314–14323, 2021.

[4] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022.

[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 *CVPR*, pages 9054–9063, 2021.

[6] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems*, 34, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020.