HumanNeRF-SE: A Simple yet Effective Approach to Animate HumanNeRF with Diverse Poses

Supplementary Material

7. Network Architexture

7.1. Conv-Filter

We performed channel-by-channel single-layer convolution on the voxel volume we constructed, with convolution kernel weights initialized to one. Our convolution kernel size is 5, padding is 2, and stride is 1 to keep the volume size. Channel-by-channel convolution can preserve the semantic information of high-frequency details.

7.2. Point Refiene

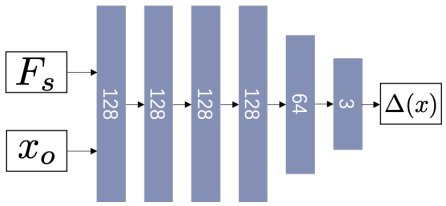

Figure 10. **Visualization of point refine network.** We use F_s and x_o as inputs to the network to obtain offsets $\Delta(x)$ to refine the canonical coordinates after the general rigid deformation. We initialize the bias of the last layer to zero, and the weight is within the range of $(-1e^{-5}, -1e^{-5})$.

7.3. NeRF Network

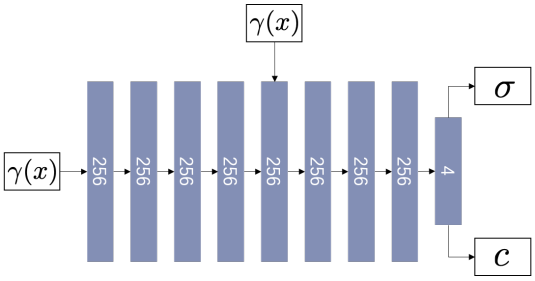

Figure 11. **Visualization of appearance network.** Follow the baseline [56], we use an 8-layer MLP with width=256, taking as input positional encoding γ of position x and producing color c and density σ. A skip connection that concatenates $\gamma(x)$ to the fifth layer is applied. We adopt ReLU activation after each fully connected layer, except for the one generating color c where we use sigmoid.

8. Canonicalization

To map points in the observation space to the canonical space, our baselines utilize neural networks to learn backward warping weight fields. This method is easy to implement but suffers from poor generalization to unseen poses, as the backward weight fields attempt to learn a spatial weight fields that deform with pose variations, necessitating memorization of weight fields for different spatial configurations. Generalizing to unseen poses using such pose-dependent weight fields is difficult. Our method directly queries the nearest SMPL's LBS weights in an explicit voxel, which is highly efficient and generalizable. However, this method suffers from unnatural deformations when the deformation angle is too large (see Figure 8). To address this issue, we use an additional neural network to learn a residual for canonical points, which is specific to the data and empowered by point-level features, considering the different clothing of the performers.

The rotation R_j and translation T_j for the rigid deformation are represented as:

$$\begin{bmatrix} R_j & T_j \\ 0 & 1 \end{bmatrix} = \prod_{i \in p(j)} \begin{bmatrix} R(\omega_i^c) & o_i^c \\ 0 & 1 \end{bmatrix} \left\{ \prod_{i \in p(j)} \begin{bmatrix} R(\omega_i) & o_i \\ 0 & 1 \end{bmatrix} \right\}^{-1} \tag{14}$$

where $p(j)$ is the ordered set of parents of joint j in the kinematic tree, ω_i defines local joint rotations using axis-angle representations, $R(\omega_i) \in \mathbb{R}^{3\times3}$ is the converted rotation matrix of ω_i via the Rodrigues formula, and o_i is the i-th joint center.

9. Ablation Study

Voxel Size. In the Conv-Filter, we voxelized all coordinates, which resulted in some loss of information compared to a dense space. However, this greatly facilitated our subsequent processing and computations. We conducted ablation experiments on the voxel size used in the voxelization process, as shown in Table 4. Voxel size affects mapping granularity to get canonical points. A larger voxel size results in coarser mapping and lower image quality. In contrast, a smaller one requires a bigger convolution kernel to diffuse occupancy and more computing resources, and it did not result in higher accuracy because the prior does not perfectly match the actual human body, and clothing is also outside the prior.

Weight Distribution. In the main text, we discussed the importance of filtering operations. It is worth noting that previous methods did not require similar operations. Our experiments suggest that this is because the learned neural weight field assigns negative weights to irrelevant joints in the data. When we simply use $softmax$ to map the neural weight field to a distribution that is the same as the SMPL

Table 4. Ablation experiment on the voxel size of our method in ZJU-MOCAP. For different subjects, there is a different optimal voxel size. We use 0.02 as voxel size in the experiment section because it has the best overall performance.

	PSNR↑	SSIM↑	†LPIPS↓
voxel=0.01	30.49	0.971	26.386
voxel=0.015	30.69	0.972	25.354
voxel=0.02	31.09	0.974	24.085
voxel=0.025	31.07	0.974	24.705
voxel=0.03	30.72	0.973	25.013

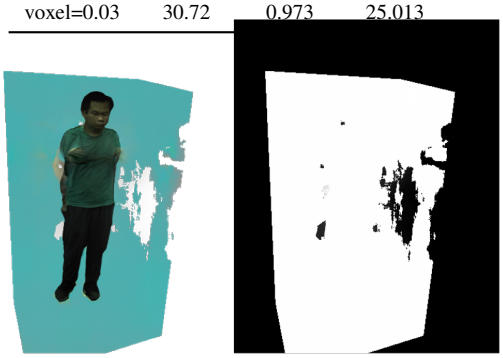

Figure 12. By mapping the learned weight field of HumanNeRF [56] to the same distribution as SMPL weight using softmax, the same phenomenon occurred.

weight, $\sum_{j\in J} w_j = 1, w_j > 0$, similar phenomena occur, see Figure 12.

10. More Results

Quantitative experiments. Due to the space limitation of the main text, the Figure 4 provided in the main text is the few-shot image synthesis result and Table 1 is a overall table.

Table 8, Table 9 and Table 10 are detailed data for Table 1. For ZJU-MOCAP, we use videos from the same six subjects as in previous work. For our IN-THE-WILD DATA, each subject is composed of multiple pose sequences to enable more convincing novel pose experiments. The videos are captured with a single camera, and SMPL are estimated with ROMP [50].

In order to better compare the performance of different methods, we synthesized the results using full input images. Figure 16 shows the synthesis results on ZJU-MOCAP, and Figure 15 shows the image synthesis results on the IN-THE-WILD DATA. Figure 17 directly compares the novel pose image synthesis ability of the three methods under two input conditions.

Compared with Figure 4, the defects of the baselines have been significantly improved on ZJU-MOCAP, but there are still obvious defects in these methods on IN-THE-WILD DATA. This is highly related to the data distribution.

Pose similarity. We found that the novel pose experimental setting in previous studies is unreasonable because the test poses and training poses are very similar, as the characters in ZJU-MOCAP perform similar movements repeatedly. This also encourages us to use custom data and subjective research of results on poses without ground truth images. Novel poses synthesized by HumanNeRF should be sufficiently novel and easy to obtain, rather than being limited to professional laboratories.

Table 5. Pose similarity of test poses and train poses in previous novel pose experiments. The previous setting of novel pose is not novel enough because of the high similarity.

	Min	Max	Average
377	0.876	0.997	0.919
386	0.939	0.995	0.964
387	0.969	0.996	0.985
392	0.845	0.957	0.909
393	0.834	0.998	0.904
394	0.760	0.993	0.842

To quantify the similarity of poses, we calculate the highest cosine similarity with all training poses for each pose in the test data. As shown in Table 5, simply dividing each subject into training and testing data in a 4:1 ratio may result in highly similar poses in the testing data being present in the training data.

In our experiments, high pose similarity does not always result in a decrease in baseline performance, because during the training process, similar poses and viewpoints are limited. For example, the pose similarity of Subject 386 is very high, but the corresponding pose only appears at the beginning and can only see the performer's right side, so when we synthesize this highly similar pose, the performance of the baselines is not good.

Qualitative results. In the qualitative experiments, we drive the model to generate new pose images by using pose sequences of different performers. We make a series of image results into a video to help the human eyes better distinguish the performance of different methods. The video results are included in the supplementary materials in the form of a compressed file.

Visual quality. The rendering results and metrics on the ZJU-MOCAP data are 512p resolution which we followed the popular protocol in [56, 62], while on our collected IN-THE-WILD data the resolution is 1080p. Both results show our superiority over SOTAs. To validate the reliability of these scores calculated with 512p, we compare our method with HumanNeRF on frames with 2K resolution and find that we still surpass HumanNeRF, as shown in Table 6. We provide additional visual comparisons based on higher-resolution frames (Figure 13). The logo and hand show better details than HumanNeRF in both low and high resolution.

More comparisons. To demonstrate the excellent performance of our method in similar tasks, additional compar-

Table 6. Results on subject 392 under different resolution.

2k resolution	PSNR↑	SSIM↑	LPIPS↓	512p resolution	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF	31.36	0.983	0.0318	HumanNeRF	**31.55**	**0.975**	0.0280
Ours	**31.73**	**0.984**	**0.0314**	Ours	31.36	0.973	**0.0276**

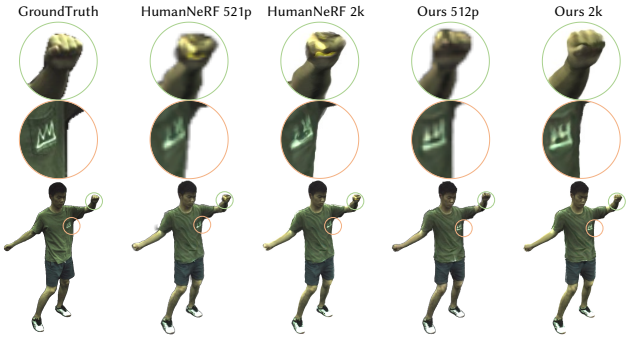

Figure 13. Results on subject 392 under different resolutions.

Table 7. Comparison with more methods. Our method has shown significant advantages in comparison with more methods.

		View	Prior	PSNR↑	SSIM↑	LPIPS↓
THUMAN4.0	TAVA [20]	>1	skeleton	26.607	0.968	0.032
	SLRF [67]	24	nodes	26.152	0.969	0.024
	Posevocab [22]	24	SMPL	30.972	0.977	**0.017**
	Posevocab [22]	1	SMPL	27.820	0.973	0.064
	Ours	1	SMPL	**31.148**	**0.979**	**0.017**
ZJU-MOCAP	SLRF [67]	24	nodes	23.61	0.905	–
	NPC [49]	>1	point clouds	21.88	–	0.134
	SelfRecon [15]	1	SMPL	27.94	0.969	0.043
	Ours	1	SMPL	**29.36**	**0.974**	**0.022**

isons with methods that incorporate explicit information are provided in Table 7.

Dancing visualization. In addition to using cross-subject movements in our qualitative analysis to test the model's novel pose ability, we can also use dance movements from online videos to drive the model as presented in Figure 14.

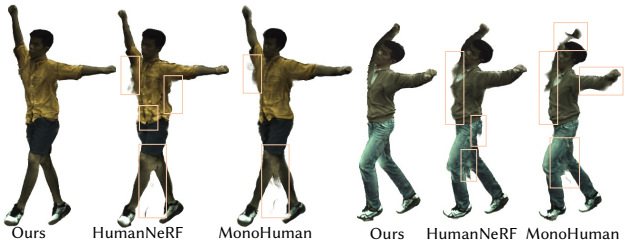

Figure 14. Dancing results. Our results are very clean compared to the blurry and unnatural results of other methods.

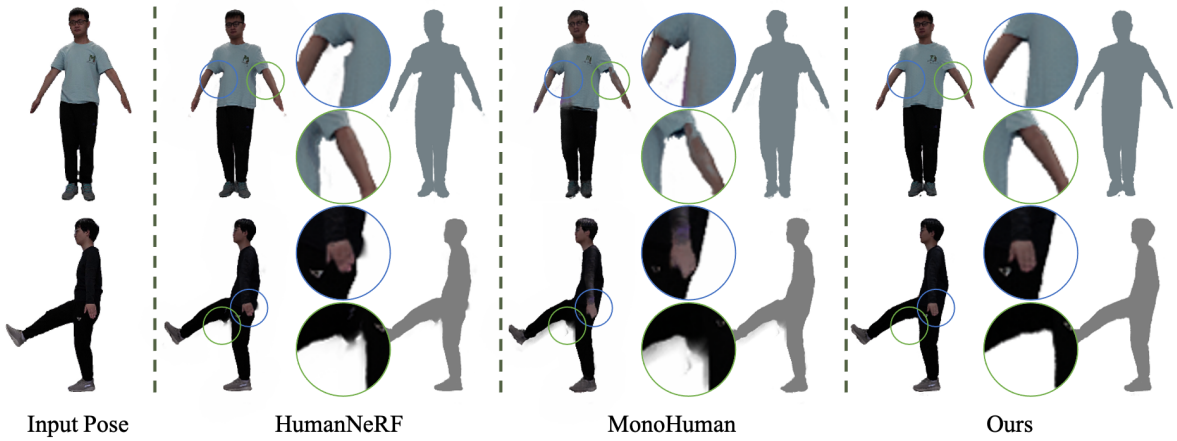

Figure 15. **Image synthesis results with full input on IN-THE-WILD DATA.** Our method maintains good performance at joint junctions, but the results of the baselines are blurry and have unnatural distortions.

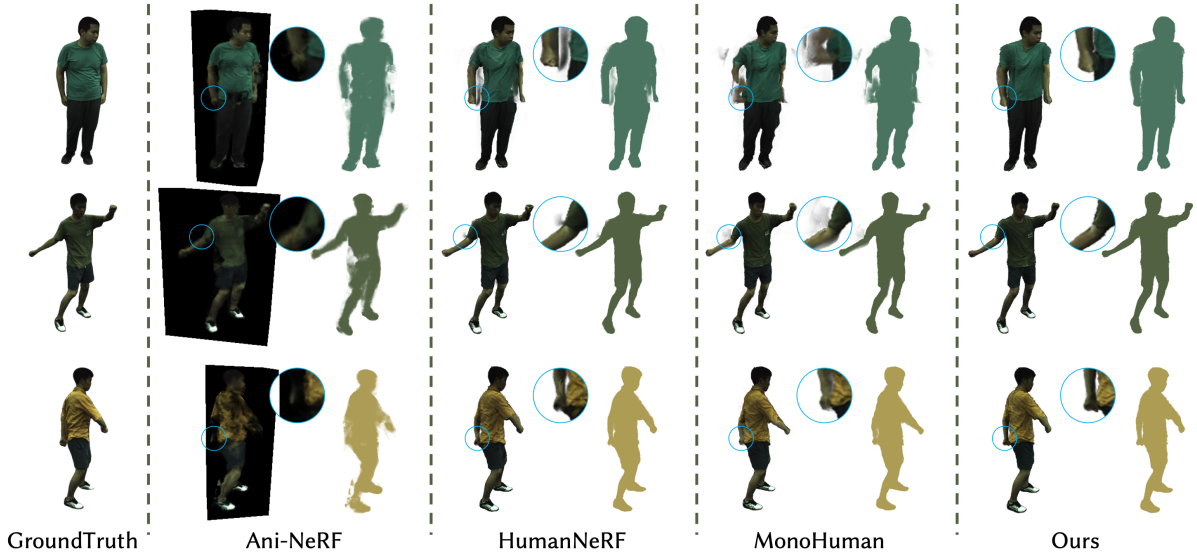

Figure 16. **Image synthesis results with full input on ZJU-MOCAP.** For Subject 386 (line 1), the baselines still have very poor image synthesis results. For Subject 392 and 393, there are still irregular deformations and artifacts in the image synthesis results, whereas our method achieves the best performance in visual comparison.

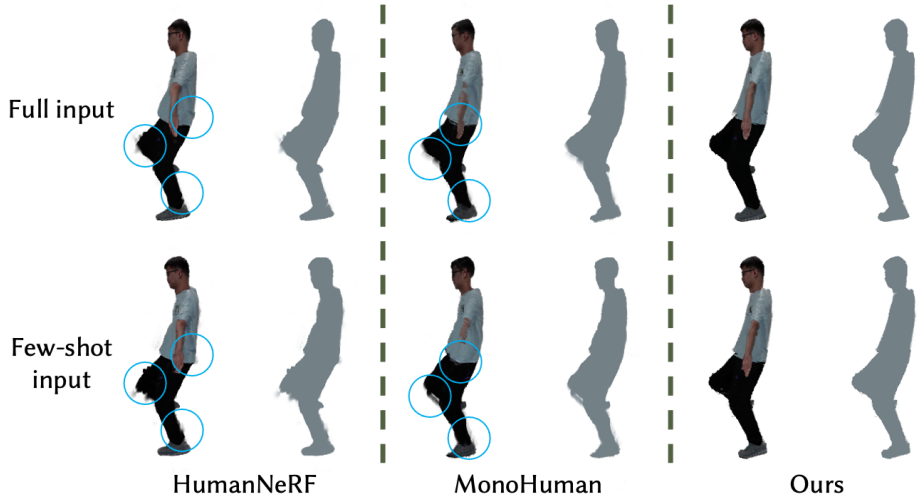

Figure 17. **Comparison between full input and few-shot input.** HumanNeRF [56] and MonoHuman [62] exhibit more artifacts and unnatural deformations with less data. MonoHuman [62] even produces hand missing. Our method maintains almost unchanged performance under the same testing conditions.

Table 8. **Comparison of full input results on ZJU-MOCAP.** Our method shows a leading advantage in LPIPS, which is aligned with human perception. Each data in the table represents the average value of all test frames' metrics in the corresponding video, which is consistent with the previous studies.

Methods	Subject377			Subject386			Subject387		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
Ani-NeRF [40]	22.12	0.8790	51.796	23.92	0.8545	55.697	16.26	0.7850	89.790
HumanNeRF [56]	30.74	0.9795	17.387	33.46	0.9716	36.326	30.30	0.9768	20.010
MonoHuman [62]	31.82	0.9822	17.561	30.10	0.9561	69.107	30.43	0.9755	23.954
Ours	31.34	0.9826	15.129	33.80	0.9741	32.990	29.36	0.9742	21.462

Methods	Subject392			Subject393			Subject394		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
Ani-NeRF [40]	22.78	0.8610	67.744	20.37	0.8417	75.381	22.05	0.8537	69.188
HumanNeRF [56]	31.80	0.9743	26.811	29.80	0.9708	25.615	30.85	0.9702	22.783
MonoHuman [62]	32.06	0.9749	27.043	29.69	0.9701	26.570	31.37	0.9720	23.521
Ours	31.75	0.9740	25.537	29.50	0.9642	25.462	30.81	0.9697	23.929

Table 9. **Comparison of few-shot input results on ZJU-MOCAP.** On the most important metric LPIPS, our method demonstrates the best results. our method exhibits less performance degradation with few-shot input indicates that our method does not overly rely on data to fit the model, unlike previous methods.

Methods	Subject377			Subject386			Subject387		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
Ani-NeRF [40]	21.77	0.8313	66.532	24.48	0.8386	74.342	20.74	0.8279	65.747
HumanNeRF [56]	30.33	0.9799	18.510	31.47	0.9618	48.410	27.92	0.9610	39.941
MonoHuman [62]	31.02	0.9774	22.562	31.26	0.9601	56.916	28.5	0.9619	43.147
Ours	31.07	0.9806	18.484	32.44	0.9644	43.876	28.05	0.9612	39.733

Methods	Subject392			Subject393			Subject394		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
Ani-NeRF [40]	22.49	0.8433	61.917	21.96	0.8384	59.169	21.65	0.8240	61.328
HumanNeRF [56]	31.55	0.9747	28.043	29.45	0.9694	26.930	28.66	0.9629	36.507
MonoHuman [62]	31.48	0.9739	30.090	29.45	0.9691	30.113	28.86	0.9636	36.139
Ours	31.36	0.9734	27.604	29.25	0.9681	27.328	28.51	0.9626	35.479

Table 10. **Comparison of results on the IN-THE-WILD DATA.** Since IN-THE-WILD DATA is not captured in laboratory conditions and only contains monocular information, the accuracy of SMPL estimation is lower compared to ZJU-MOCAP, which leads to a decrease in overall performance metrics. However, this is more in line with real-world application scenarios. Our method demonstrates the best performance, especially in the LPIPS metric, which reflects image quality the most. Even with reduced input data, our method maintains excellent performance, while other methods experience a greater degree of decline.

Full Input	S1			S2			S3		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
HumanNeRF [56]	32.94	0.9737	43.259	26.65	0.9650	41.505	27.32	0.9500	59.620
MonoHuman [62]	32.99	0.9725	46.272	26.74	0.9683	42.198	27.72	0.9509	66.400
Ours	33.72	0.9764	42.343	26.43	0.9709	39.174	27.54	0.9524	57.408

Few-shot Input	S1			S2			S3		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
HumanNeRF [56]	33.37	0.9761	45.012	25.82	0.9599	43.326	27.27	0.9495	62.384
MonoHuman [62]	33.58	0.9744	50.561	26.57	0.9640	45.955	27.48	0.9522	72.143
Ours	33.70	0.9768	41.954	26.45	0.9712	39.894	27.62	0.9525	59.636