Supplementary Material for HumanNeRF

A. Limitations.

In this paper, we propose a generalizable dynamic human neural radiance field method to address issues of the existing approaches. Although very effective, the proposed HumanNeRF still needs hours of fine-tuning and has some limitations. First, we use the regressed parametric human model to handle large pose deformation and complex motions, and it limits our approaches to the single-person setup and fails to handle the multi-person or human-object interaction situations. Also though we have shown the generalization ability of our method, its capability is limited as distributions of human datasets only cover a small portion of the human dynamics and appearances. Moreover, we do not explicitly model lighting conditions, significant brightness or color change between views may cause severe artifacts. For example, due to the switching of the nearest views during our appearance blending, jumping artifacts appear, especially for the significant brightness variance in our sparse input views. Such artifact will be alleviated if the illumination is almost consistent across views, as shown in the FVV results of the supplementary video.

B. Components ablation study.

To better evaluate the components of our pipeline, we also do additional quantitative analysis of different modules of our method, such as without aggregated pixel alignment feature (w/o_F), without pose embedded non-rigid human deformation ($w/o\_\mathbf{MLP}_d$), and without neural appearance blending ($w/o\_\mathbf{MLP}_\mathcal{A}$). Note that our full module achieves the best results.

	PSNR↑	SSIM↑	LPIPS↓	MAE↓
Ours_{wo_F}	18.36	0.8621	0.1503	13.49
$\text{Ours}_{wo\_\mathbf{MLP}_d}$	26.79	0.9704	0.0516	5.251
$\text{Ours}_{wo\_\mathbf{MLP}_\mathcal{A}}$	29.69	0.9620	0.0703	2.016
Ours_{full}	**33.01**	**0.9842**	**0.0334**	**0.9307**

Table 1. **Quantitative evaluation of different Components.**

As shown in Tab. 2, the average error increases rapidly as the camera number decreases.

	two views	four views	six views
PSNR↑	22.44	25.88	**32.59**
SSIM↑	0.9324	0.9552	**0.9817**
LPIPS ↓	0.0887	0.0562	**0.0304**

Table 2. **Quantity evaluation on the different number of input views**. We select six ,four and two camera views for ablation studies in **PSNR**, **SSIM** and **LPIPS** metrics.

C. Discussion about our generalizability .

Despite the requirement of one hour fine-tuning of unseen identities, we would like to point out that our approach serves as a practical and more efficient scheme for dynamic and sparse view setting with significantly less fine-tuning effort than previous methods (see Tab. 3). Our efficient generalizations are many-fold. First, only our generalizable NeRF module already provides meaningful yet blur results in Fig. 1, similar to Neural Human Performer [Kwon *et al.*]. Second, without per-scene fine-tuning, our method provides comparable results to previous general and even per-scene methods. Then, only with efficiently fine-tuning in hours, we can achieve SOTA performance, even for unseen poses.

Figure 1. Results of our generalizable dynamic neural radiance field module on unseen identities.

D. Training time comparison.

We compare our method with other per-scene traning methods in terms of training or fine-tuning time. As shown in Tab. 3, our method is more efficient than other method.

	Ours	Neural Body	Neural Volumes	ST-NeRF
time	1.2h	6.7h	8.4h	9.5h

Table 3. Quantitative comparison against per-scene training methods in terms of **fine-tuning or training time** on the video "batman" with 300 frames of our multi-view dataset.

E. Network Architectures.

We show detailed network architecture specifications of our feature extractor network U (that extracts 2D image features), feature blending network $\mathbf{MLP}_{\mathcal{B}}$, deformation network $\mathbf{MLP}_{d}$, generalizable dynamic neural radiance field Φ and appearance blending network $\mathbf{MLP}_{\mathcal{A}}$.

Layer	**k**	**s**	**d**	**channels**	input
CRB2D Down_0	3	1	1	$4/32$	I
CRB2D Down_1	3	1	1	$32/64$	CRB2D Down_0
CRB2D Down_2	3	1	1	$64/128$	CRB2D Down_1
CRB2D Down_3	3	1	1	$128/256$	CRB2D Down_2
CRB2D Up_1	3	1	1	$256/256$	CRB2D Down_3
CRB2D Up_2	3	1	1	$256/128$	CRB2D Up_1
CRB2D Up_3	3	1	1	$128/64$	CRB2D Up_2
T	3	1	1	$64/32$	CRB2D Up_3

Table 4. Network details of feature extractor network U. **k** is the kernel size, **s** is the stride, **d** is the kernel dilation, and **channels** shows the number of input and output channels for each layer. We denote CRB2D to be ConvBnReLU2D.

Layer	**channels**	input
PE_0	$6/54, 3/27$	view direction d, angle θ
LR_0	$54+27+32*6/256$	PE_0, features f
LR_1	$256/256$	LR_0
LR_2	$256/256$	LR_1
LR_3	$256/256$	LR_2
LR_4	$256/128$	LR_3
LR_5	$128/6$	LR_4

Table 5. Network details of feature blending network $\mathbf{MLP}_{\mathcal{B}}$. PE/LR refers to the positional encoding and LinearRelu layer structure respectively (same as below).

Layer	**channels**	input
PE_0	$24/216$	R_d
LR_0	$216+72+32/256$	PE_0, R_v, F
LR_1	$256/256$	LR_0
LR_2	$256/256$	LR_1
LR_3	$256/256$	LR_2
LR_4	$256/128$	LR_3
LR_5	$128/3$	LR_4

Table 6. Network details of deformation network $\mathbf{MLP}_{d}$. R_d and R_v are the distances and directions between sample point p and the 24 joints of the SMPL skeleton. F is the feature after blending.

Layer	**channels**	input
PE_0	3/63	position x
LR_0	63/256	PE_0
LR_1	256/256	LR_0
LR_2	256/256	LR_1
LR_3	256/256	LR_2
LR_4	27+256/256	PE_0, LR_3
LR_5	256/256	LR_2
LR_6	256/256	LR_2
Density σ	256/1	LR_6
PE_1	3/27	view direction d
LR_7	256+27+32/256	LR_2,PE_1, F
LR_8	256/128	LR_7
Color c	128/3	LR_8

Table 7. Network details of generalizable dynamic neural radiance field Φ.

Layer	**channels**	input
LR_0	$(32+3)*2/256$	$f_r, 0_r, f_l, 0_l$
LR_1	$256/256$	LR_0
LR_2	$256/256$	LR_1
LR_3	$256/256$	LR_2
LR_4	$256/256$	LR_3
LR_5	$256/256$	LR_4
LR_6	$256/128$	LR_5
Blending weights W	$128/3$	LR_4

Table 8. Network details of appearance blending network $\mathbf{MLP}_{\mathcal{A}}$. $f_r, 0_r, f_l, 0_l$ are two adjacent image features and occlusion maps.