NeuHMR: Neural Rendering-Guided Human Motion Reconstruction

Supplementary Material

6. Table of Symbols

For notation simplicity, we adopted alphabetic symbols in this paper to represent essential components in both our designed GHN and NeuHMR. For better symbol-name correspondences, here we justify implications of all symbols used in the paper in Table 5 to help readers comprehend.

Symbols	Implication
Common Symbols	
V	A video of a sequence of frames
T	The number of frames in V
$\mathbf{I}$	An image from a V
$\mathbf{J}$	Human body joints
θ	Angle rotations of $\in \mathbb{R}^{24\times3}$
Generalizable Human NeRF	
$\mathbf{c}$	Radiance of $\mathbf{x}$
σ'	Density of $\mathbf{x}$
$\mathbf{w}'$	Occupancy weight of $\mathbf{x}$
$\mathbf{C}$	Radiance of a pixel
$\mathbf{W}$	Occupancy of a pixel
$\mathcal{T}$	Motion mapping function [48]
$\Phi^{GHN}(\cdot)$	The generalizable human NeRF model
φ	Parameters for $\Phi^{GHN}(\cdot)$
$\mathbf{v}$	SMPL mesh [28]
$\mathcal{F}$	Feature extractor backbone
$\mathbf{F}$	2D feature maps $\mathbf{F} = \mathcal{F}(I)$
π	Camera projection
$\psi(\cdot)$	Rendering network in GHN
n	Number of inputs to GHN $n = 2$
NeuHMR	
N	Number of anchor frames $N = \frac{T}{10}$
$\hat{N}$	Number of anchor frames after FPS
$\mathcal{D}$	Depth map of the rendered avatar
$\mathcal{V}$	Visibility of each joint
$\mathcal{J}$	Registered joint features in 3.3.1
$\Phi^{MLP}(\cdot)$	The motion field model
η	Parameters for $\Phi^{MLP}(\cdot)$
$\mathbf{R}'$	Joint rotation residuals
$\mathbf{T}'$	Joint translation residuals
σ	Standard deviation of the Gaussian

Table 5. Table of symbols used in this paper.

7. Limitations and Future Directions

Optimizing human poses without relying on 2D pseudo ground truth keypoints presents a significant challenge. While NeuHMR mainly follows data-driven optimizations through neural rendering. The performance of NeuHMR is correlated to the rendering quality of the GHN. Employing the most advanced model [31], NeuHMR may learn better correspondences between 2D images and 3D human appearance. However, render human in the presence of changing appearances or incomplete observations, still remains an unresolved problem. NeuHMR is also prone to subpar human renderings due to appearance inconsistencies across videos (*e.g.*, Figure 6). Our method is designed to focus on per-frame optimization without considering valuable temporal information, which may result in unstable estimations across frames. Nevertheless, most existing stabilizers [40, 54] can be integrated as a post-processing measure to enhance our optimized poses.

Figure 6. GHNs are sensible to appearance changing. As shown in the figure, the actor is moving across different scenes with varying lighting conditions, which are different from the anchor frames. This appearance change leads to unmatched rendering.

8. Comparison with More Baselines

In section 4, we present our primary comparison results against two state-of-the-art video-based optimization methods [51, 54], which refine human meshes iteratively per frame and across frames.

Here, we compare NeuHMR with additional baseline methods that focus on per-frame optimization: *(1)* Optimizing human meshes by minimizing the distance between projected 3D mesh key points and pseudo ground-truth 2D key points estimated from ViT-Pose [50]; *(2)* Optimizing human meshes by minimizing the distance between the SMPL UV and the UV-map estimated from DensePose [11]; *(3)* LGD [43] — a method that deforms a canonical SMPL mesh to align with pseudo ground-truth 2D key points; and *(4)*

Methods	MPJPE↓	PVE↓
(1)	101.5	104.2
(2)	116.2	118.0
(3) LGD [43]	102.1	115.3
(4) KAMA [14]	99.8	114.4
Ours	**95.4**	**109.6**

Table 6. Comparison results on more baseline methods.

KAMA [14] — a method that refines inaccurate 3D pose estimations by fitting both 2D and 3D pseudo ground-truth key points.

We conducted the comparison experiments on the EMDB dataset [18] and present the quantitative results in Table 6. Our method, which applies optimization based on low-level visual cues, demonstrates superior robustness and effectiveness compared to all baselines that rely on pseudo ground-truth data, which may introduce unexpected noise. This finding is consistent with our discussion in section 1 and section 4.5.

9. More Extensive Studies

Unless otherwise noted, all extensive studies were performed on the EMDB dataset.

9.1. Optimization of Appearance Model

In NeuHMR (Figure 2), a generalizable Human NeRF is pre-trained, which provides an appearance model for further optimization steps. Human NeRF models need full and clear coverage of every angle of the human for training high quality appearance model. Such coverage should **not** be assumed in any in-the-wild videos. In fact, most videos in EMDB have very sparse views — some body parts are frequently visible while others are hardly shown in the video. Optimizing the appearance model on such extremely unbalanced data hurts its overall performance. As a result, performance drops by *7.0* and *9.9* on MPJPE and MPVPE, respectively, on EMDB with Hyrbik initialization.

9.2. Impact of Appearance Model

To optimize towards low-level visual constraints, a perfect appearance model would be ideal. However, in our experiments, we show that our method is robust even with an average-quality appearance model and already achieves decent results. To further validate our claim, we present additional experiments here that by training the generalizable Human NeRF for only $\frac{1}{3}$ and $\frac{2}{3}$ epochs before using it to optimize poses. According to the results in Table 7, optimization of NeuHMR is robust to different qualities of the appearance model trained by Human NeRF.

Methods	MPJPE↓	PVE↓
$\frac{1}{3}$ epochs	99.2	114.7
$\frac{2}{3}$ epochs	95.6	110.1
All epochs	**95.4**	**109.6**

Table 7. Ablation results on impact of quality of the appearance model.

S	MPJPE↓	PVE↓
$S = 16$	99.7	115.1
$S = 32$	99.0	114.4
$S = 64$	97.5	111.9
$S = 256$	95.6	**109.2**
$S = 128$	**95.4**	109.6

Table 8. Ablation results on number of joint candidates S.

9.3. Study on number of candidate samples S.

In the proposed joint matching constraint, we assume that correct joints should be located near the predictions made by the most advanced HMR estimators. S candidate positions are randomly sampled within a predefined distribution. More samples may lead to a higher probability of being at the "correct" joint position and potentially better results. Here, we conduct extensive experiments by varying the number of S to observe its impact on the final performance of NeuHMR. As reported in Table 8, we observed that NeuHMR is robust to different values of S, outperforming state-of-the-art counterparts for all choices of S. NeuHMR reaches the best MPJPE when $S = 128$ and the best PVE when $S = 256$.

9.4. Generalization ability of the Proposed GHN

NeuHMR relies on a human appearance model that can be generalized from a pre-trained dataset to any given video (section 3.2). Here we visualize the generalized human appearance models for all the actors in the EMDB dataset in Figure 7. According to the rendered results, our GHN shows its capability of generalizing human appearance with acceptable artifacts. Note that our GHN model was trained on ZJU-MoCap sequences only and the design of a monocular GHN is not the focus of this work. By training on larger-scale datasets or adopting the state-of-the-art off-the shelf GHN model, the rendering quality is anticipated to be much better.

10. More Qualitative Results

To further compare the methods qualitatively, we provide SMPL mesh visualizations on 2 outdoor videos and 2 indoor videos by projecting them onto the 2D frames of the videos for different actors. Please see the videos that are submitted along with this supplementary document.

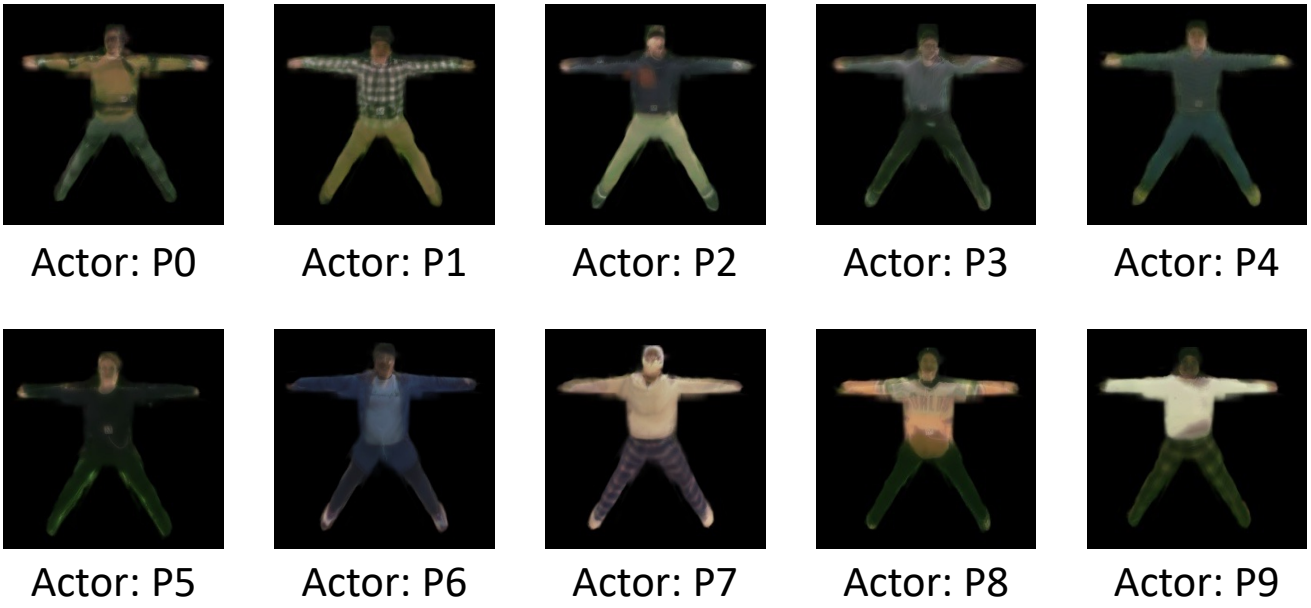

Figure 7. Our generalized human appearance models on all actors in the EMDB dataset. The GHN model was only trained on ZJU-MoCap sequences.

10.1. Comparison against GLAMR [54]

We mainly compare against GLAMR — another state-of-the-art optimization-based HMR method. The video results that have the suffix `_glamr.mp4` show comparisons between GLAMR and NeuHMR. We can easily observe better pose corrections and translation alignments from the results optimized by our methods.

10.2. Comparison against SLAHMR [51]

In section 4, we mentioned that SLAHMR failed on most of long videos in EMDB, which makes it impossible to compare fairly. Here we show visual comparisons in one of the videos named as `compare_slahmr.mp4`. We can see that although SLAHMR seems to be effective at the first few frames, it starts to collapse and fails to optimize the poses for the rest of the video. On the other hand, NeuHMR is able to provide stable refinements to the initially estimated poses on those challenging long sequences.