Human-NGP

Chin-Chia Yang, Ting-Wei Guan

1. Introduction

Neural Radiance Fields (NeRF) [3] have demonstrated remarkable capabilities in synthesizing novel views for static scenes. Recent advancements have extended the applicability of NeRF to dynamic scenes and objects. Notably, HumanNeRF [6] introduced a free-viewpoint rendering approach for complicated human body movements, building upon the NeRF framework. Despite its impressive performance and the efficiency of requiring only a monocular video as input, the training process of HumanNeRF is impeded by the complexity of its neural networks. The model was trained on 4 GeForce RTX 2080 Ti GPUs, necessitating 400K iterations and approximately 3 days to converge. In this project, we incorporated Instant-NGP [4], a state-of-the-art technique for accelerating NeRF, into the HumanNeRF framework. Additionally, we introduced supplementary loss functions to achieve performance comparable to HumanNeRF but with a significantly accelerated convergence time.

2. Related Work

2.1. NeRF

The foundation of computer graphics primitives lies in their representation through mathematical functions that define their visual characteristics. The quality and efficiency of these mathematical representations are pivotal for achieving visual fidelity. NeRF [3] revolutionizes scene representation by employing

multi-layer perceptions (MLPs) to memorize static scenes. It treats the scene as a continuous function, taking 3D coordinates and ray directions as input and producing predicted color and density as output. This technology has significantly advanced both scene representation and view synthesis. Instant-NGP [4] introduces an explicit multi-resolution hash grid representation, enabling the use of smaller networks. As a result, Instant-NGP substantially reduces the training time of NeRF, contributing significantly to the evolution of scene representation techniques.

2.2. Human specific rendering

HumanNeRF [6] operates by taking a single monocular video of a moving person as input and subsequently rendering the resulting volumetric representation from any viewpoint at any frame within the video. Despite its capabilities, the extended training duration underscores the importance of exploring techniques to optimize and enhance the efficiency of the rendering process.

We introduce our method, Human-NGP, which incorporated Instant-NGP [4] into the HumanNeRF framework with supplementary loss functions to achieve performance comparable to HumanNeRF but with a significantly accelerated convergence time.

3. Method

3.1. NeRF

NeRF [3] adopts a neural implicit function to represent a 3D scene, where a neural network $f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ that maps a 3D coordinate $\mathbf{x} = (x, y, z)$ and a unit viewing direction $\mathbf{d} = (\theta, \phi)$ to color $\mathbf{c}$ and density σ. The pixel color of a camera ray $\mathbf{r}$ with N samples is given by,

$$C(r) = \sum_{i=1}^{N} w_i c_i \tag{1}$$

, where

$$w_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)(1 - \exp(-\sigma_i \delta_i)), \tag{2}$$

and δ_i is the distance between the i^{th} point and its adjacent sample.

3.2. HumanNeRF

HumanNeRF represents a moving person with a canonical space F_c warped from an observed pose in an observation space F_o:

$$F_o(\mathbf{x}, \mathbf{p}) = F_c(T(\mathbf{x}, \mathbf{p})), \tag{3}$$

where $F_c : \mathbf{x} \rightarrow (\mathbf{c}, \sigma)$ defines the NeRF model mapping the position $\mathbf{x}$ to color $\mathbf{c}$ and density σ, and $T : (\mathbf{x_o}, \mathbf{p}) \rightarrow \mathbf{x_c}$ defines a motion field mapping points from observed space to canonical space with observed body pose $\mathbf{p} = (J, \Omega)$ as input. The body joints J include 24 standard 3D joint locations, and local joint rotations Ω is a set of rotation vectors in axis-angle representation.

HumanNeRF decomposes the motion filed into two parts:

$$T(\mathbf{x}, \mathbf{p}) = T_{skel}(\mathbf{x}, \mathbf{p}) + T_{NR}(T_{skel}(\mathbf{x}, \mathbf{p}), \mathbf{p}) \tag{4}$$

where T_{skel} represents skeleton-driven deformation, and T_{NR} corresponds to non-rigid body movement. Recognizing potential inaccuracies in pose estimation, the pose correction module $P_{pose}(\mathbf{p})$ is introduced to refine the original pose. Figure 1 provides an overview of HumanNeRF. To streamline the training process, we chose to omit the pose correction module $P_{pose}(\mathbf{p})$ and the non-rigid body movement module T_{NR} as their inclusion resulted in only marginal improvements in the rendered images. Ultimately, MSE loss $\mathcal{L}_{mse}$ and LPIPS loss $\mathcal{L}_{lpips}$ [7] are employed to train HumanNeRF.

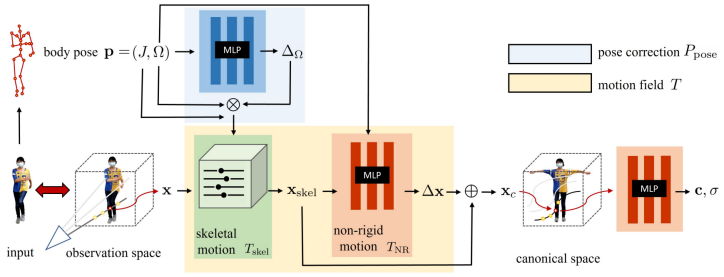

Figure 1: Illustration of HumanNeRF.
[6]

3.3. Instant-NGP

Instant-NGP introduces a novel approach by storing trainable feature vectors in a compact spatial hash table of size T, allowing for a flexible trade-off between reconstruction speed and quality. In contrast to methodologies relying on gradual pruning during training or necessitating prior knowledge of scene geometry, Instant-NGP employs multi-resolution hash tables with L levels, each storing feature vectors of F dimensions. The grid resolution spans from N_{min} to N_{max}. Given an input coordinate, the features stored in these tables are interpolated and concatenated before traversing the MLP. Despite having 20× fewer parameters, Instant-NGP achieves reconstruction quality on par with dense grid encoding. Figure 2 illustrates the multi-resolution hash encoding process, with typical hyperparameter values showcased in Figure 3.

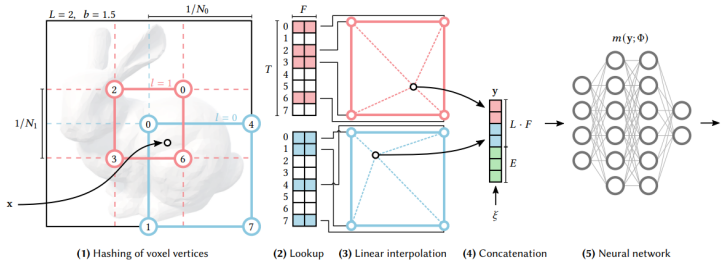

Figure 2: Illustration of the multi-resolution hash encoding.
[4]

Parameter	Symbol	Value
Number of levels	L	16
Max. entries per level (hash table size)	T	2^{14} to 2^{24}
Number of feature dimensions per entry	F	2
Coarsest resolution	N_{min}	16
Finest resolution	N_{max}	512 to 524288

Figure 3: Hash encoding parameters.
[4]

3.4. Loss functions

In our experiment, combining HumanNeRF and Instant-NGP directly leads to a decline in performance. To address this issue, we incorporated the following loss functions to mitigate any performance degradation:

Opacity loss

Our opacity loss is inspired by the hard surface loss and canonical edge loss introduced in NeuMan [2]. This loss function is tailored to alleviate the halo effect around the canonical human. Specifically, we promote the weights of each sample and the accumulated alpha values to be either 1 or 0. The opacity loss function is defined as follows:

$$\mathcal{L}_{opac} = -(\log(\exp^{|w|} + \exp^{|1-w|}) + \log(\exp^{|\alpha|} + \exp^{|1-\alpha|})) \tag{5}$$

where w refers to the weight defined in Equation 2 and $\alpha = \Sigma_{i=1}^{N} w_i$.

Distortion loss

To address the issue of floaters (translucent floating material) and enhance training efficiency, we incorporated the distortion loss proposed in Mip-nerf 360 [1]. This loss is defined in terms of the set of ray distances s and corresponding weights w:

$$\mathcal{L}_{dist}(s, w) = \iint_{-\infty}^{\infty} w_s(u) w_s(v) |u - v| \, du \, dv$$

, where $w_s(u)$ is given by,

$$w_s(u) = \Sigma_i w_i \mathbb{1}_{[s_i, s_{i+1})}(u). \tag{6}$$

The minimization of the distortion loss involves compacting weights into the smallest possible region. This process effectively mitigates the presence of floaters and contributes to an acceleration in training time. The final loss term is given by,

$$\mathcal{L} = \mathcal{L}_{lpips} + 0.5\mathcal{L}_{mse} + 0.1\mathcal{L}_{opac} + \mathcal{L}_{dist}. \tag{7}$$

4. Results

4.1. Dataset

We evaluate our method on the part of the ZJU-MoCap dataset [5]. We select 6 subjects (377, 386, 387, 392, 393, 394) with diverse motions and use images captured by "*camera 1*" for training and sample images from "*camera 2*" to "*camera 23*" for evaluation.

4.2. Quantitative results

Table 1 presents the quantitative results for Neural body [5], HumanNeRF [6], and our proposed method, Human-NGP. The table reveals our superior performance compared to Neural body and a comparable performance to HumanNeRF. Significantly, our method exhibits a convergence speed that is 40 times faster than that of HumanNeRF.

	Subject **377**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	29.11	0.9674	40.95	-
HumanNeRF [6]	30.41	0.9743	24.06	400k
Ours	29.86	0.9672	23.50	10k
	Subject **386**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	30.54	0.9678	46.43	-
HumanNeRF [6]	33.20	0.9752	28.99	400k
Ours	33.41	0.9645	27.70	10k
	Subject **387**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	27.00	0.9518	59.47	-
HumanNeRF [6]	28.18	0.9632	35.58	400k
Ours	28.12	0.9507	35.67	10k
	Subject **392**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	30.10	0.9642	53.27	-
HumanNeRF [6]	31.04	0.9705	32.12	400k
Ours	30.56	0.9559	33.26	10k
	Subject **393**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	28.61	0.9590	59.05	-
HumanNeRF [6]	28.31	0.9603	36.72	400k
Ours	28.30	0.9473	37.12	10k
	Subject **394**			
	PSNR ↑	SSIM ↑	LPIPS* ↓	Iterations
Neural Body [5]	29.10	0.9593	54.55	-
HumanNeRF [6]	30.31	0.9642	32.89	400k
Ours	29.37	0.9478	34.04	10k

Table 1: Quantitative results.

4.3. Qualitative results

Figure 4 presents our qualitative results. The left column showcases our rendered output, the middle column displays the ground truth, and the right column presents the rendered human mask. The results illustrate that our method consistently generates high-quality image masks. For more visualization, please refer to our YouTube video: novel pose and tpose.

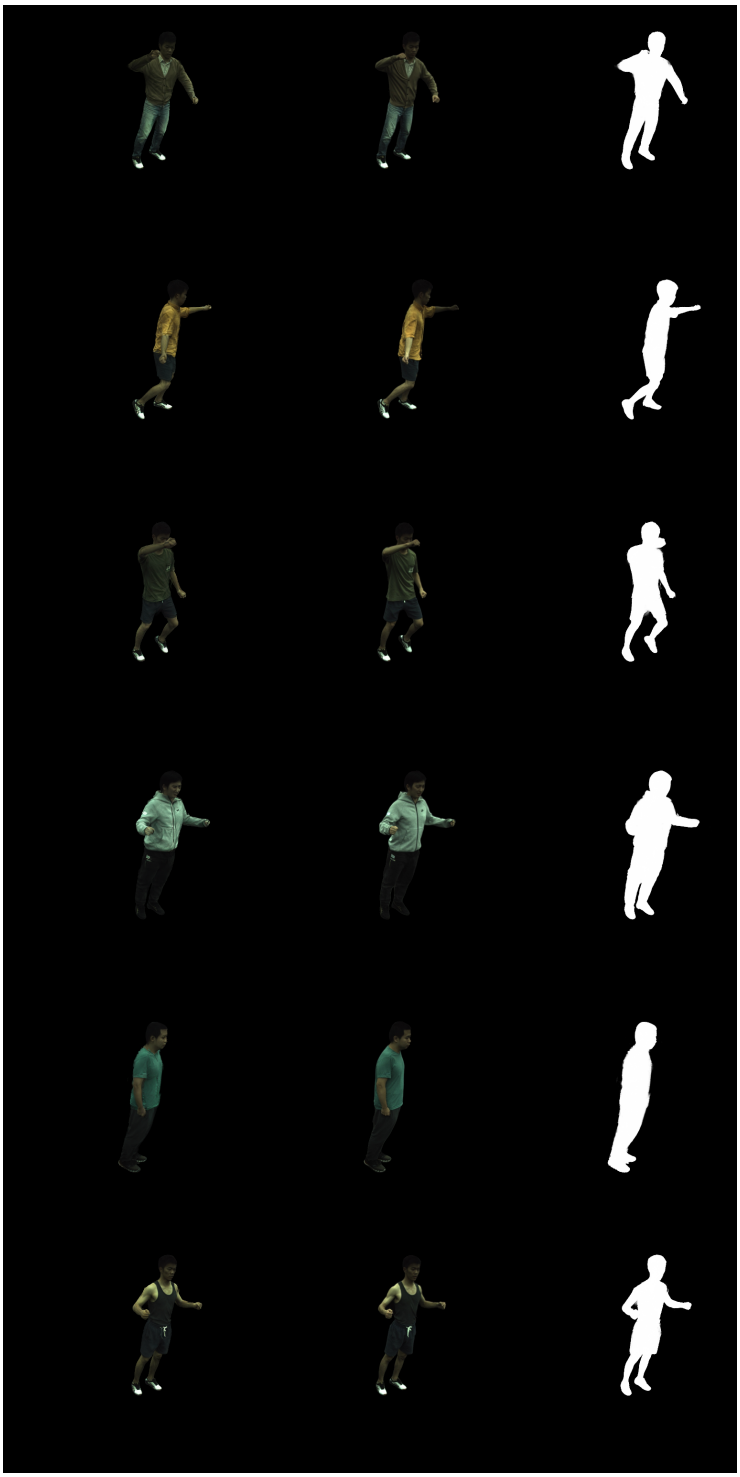

Figure 4: Qualitative results.

5. Conclusion

In this project, we seamlessly integrated HumanNeRF with Instant-NGP, yielding an impressive 40-fold reduction in convergence time. To bolster overall performance, we incorporated supplementary loss functions and demonstrated their effectiveness. Throughout the report-writing process, we initially completed the draft and then utilized the assistance of ChatGPT to refine the entire document, ensuring enhancements in phrasing and grammatical accuracy. It's worth noting that we iteratively refined the report, leveraging our knowledge to ensure fluency and accuracy.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In CVPR, 2022.

[2] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In ECCV, 2022.

[3] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020.

[4] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph., 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, 2021.

[6] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In CVPR, 2022.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, 2018.