HOSNeRF: Dynamic Human-Object-Scene Neural Radiance Fields from a Single Video
Supplementary Material

Jia-Wei Liu[1]*, Yan-Pei Cao[2]†, Tianyuan Yang[1], Zhongcong Xu[1], Jussi Keppo[4,5], Ying Shan[2], Xiaohu Qie[3], Mike Zheng Shou[1]†

[1] Show Lab, National University of Singapore [2] ARC Lab, [3] Tencent PCG
[4] Business School, [5] Institute of Operations Research and Analytics, National University of Singapore

The supplementary material is structured as follows:

- Sec. 1 provides further implementation details of the proposed HOSNeRF.
- Sec. 2 presents additional details on the network designs of our HOSNeRF.
- Sec. 3 summarizes additional comparisons of our HOSNeRF against state-of-the-art (SOTA) approaches.

Furthermore, we also provide a **supplementary video** showcasing per-scene 360° free-viewpoint renderings from our HOSNeRF on all six scenes of our HOSNeRF dataset.

1. Implementation Details

We conducted all our experiments on 4 Tesla V100 GPUs, using the PyTorch [12] deep learning framework.

Optical Flow Supervision. We first map the deformed points $\mathbf{x}_{\mathrm{d}}$ from the deformed space at timestep t to canonical points $\mathbf{x}_{\mathrm{c}}$ in the canonical space. Then we compute their corresponding deformed points at timestep $t-1$, denoted as $\hat{\mathbf{x}}_{\mathrm{d}_{t-1}}$, through forward deformation:

$$\hat{\mathbf{x}}_{\mathrm{d}_{t-1}} = \Psi^{\mathrm{coarse}}_{\mathrm{c}\to\mathrm{d}_{t-1}}\left(\mathbf{x}_{\mathrm{c}},\, \mathcal{J},\, \mathcal{R}\right) + \Delta\mathbf{x}_{\mathrm{c}\to\mathrm{d}_{t-1}}. \tag{1}$$

We project $\hat{\mathcal{X}}_{\mathrm{d}_{t-1}} = \left\{\hat{\mathbf{x}}^i_{\mathrm{d}_{t-1}}\right\}$ onto the reference camera at timestep $t-1$ and to obtain their corresponding pixel locations $\hat{\mathcal{P}}_{\mathrm{d}_{t-1}} = \{\hat{\mathbf{P}}^i_{\mathrm{d}_{t-1}}\}$. We then compute the optical flow induced by these points with respect to the pixel locations $\mathcal{P}_{\mathrm{d}_t} = \{\mathbf{P}^i_{\mathrm{d}_t}\}$, from which the rays of $\mathcal{X}_{\mathrm{d}} = \left\{\mathbf{x}^i_{\mathrm{d}}\right\}$ are cast. Finally, we minimize the error between the induced flow and the estimated flow:

$$\mathcal{L}_{\mathrm{Flow}} = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}\sum_{i=1}^{N} w^{\mathbf{r},i}\left|\left|\left(\hat{\mathbf{P}}^{\mathbf{r},i}_{\mathrm{d}_{t-1}} - \mathbf{P}^{\mathbf{r},i}_{\mathrm{d}_t}\right) - \mathbf{f}_{\mathbf{P}^{\mathbf{r},i}_{\mathrm{d}_t}}\right|\right|, \tag{2}$$

where $w^{\mathbf{r},i} = T_i\left(1-\exp\left(-\sigma_i\delta_i\right)\right)$ is the ray termination weights from the volume rendering equation, and $\mathbf{f}_{\mathbf{P}^{\mathbf{r},i}_{\mathrm{d}_t}}$ is the estimated 2D backward optical flow using RAFT [16] at $\mathbf{P}^t_{\mathbf{r},i}$.

Coordinate System Alignment. To integrate the state-conditional scene model and dynamic human-object model, we initially synchronize their coordinate systems during preprocessing, as they are originally processed and defined in separate coordinate systems. To achieve this, we utilize the SMPL [8] parameters acquired from the pre-trained human pose estimation model ROMP [15] and adopt the scene-SMPL alignment approach from NeuMan [4]. This technique requires that the human subject always stands on the ground. Subsequently, we align the two coordinate systems through the Perspective-n-Point (PnP) [6] method and resolve any scale ambiguities by restricting the feet meshes of the SMPL model to touch the ground plane [4]. In this context, the near and far parameters for the scene model are set to 0.1 and 10^6, respectively, while those for the dynamic human-object model are determined by the coarse bounding box calculated from the human-object poses.

Human and Object Masks. To estimate human and object masks, we utilize the pre-trained Mask-RCNN [3] model. Consequently, the majority of object classes in our dataset come from the COCO [7] dataset. During preprocessing, we successfully segment all humans and most objects in our dataset. However, for objects that are not detected due to occlusions or out-of-domain classes, we manually segment them. To ensure complete separation of the foreground from the background, we then dilate the human and object masks by 5%. The proposed three-stage training

*Work is partially done during internship at ARC Lab, Tencent PCG.
†Corresponding Authors.

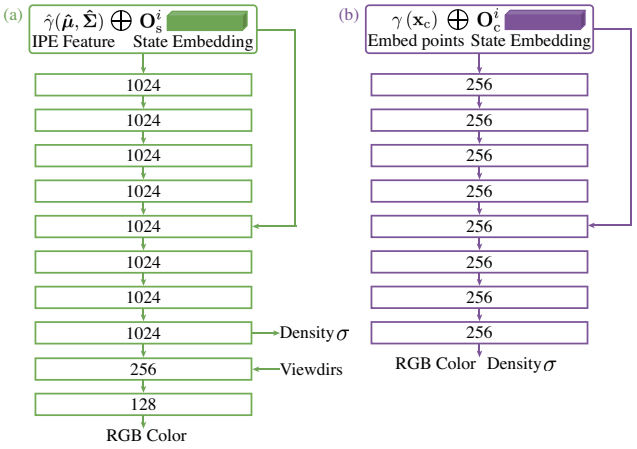

Figure 1: State-conditional network designs for the scene base model (a) and the canonical space model (b).

pipeline of our HOSNeRF method is beneficial, especially the third stage, which involves fine-tuning for foreground-background merging. This enables training with *coarse* human and object masks. In contrast, HumanNeRF [17] depends on manual intervention to correct coarse segmentation errors.

Optimization Parameters. We optimize our HOSNeRF using Adam optimizer [5]. We set the base learning rates for our training process as follows: 0.002 for the first stage to train the background, 0.0006 for the second stage to train the dynamic human-object model, and 0.00006 for the third stage to fine-tune the complete HOSNeRF model. For most of the scenes, we balance the loss terms using the following weighting factors: $\omega_{\mathrm{MSE}} = 0.2$, $\omega_{\mathrm{LPIPS}} = 1.0$, $\omega_{\mathrm{Cycle}} = 0.01$, $\omega_{\mathrm{Flow}} = 0.01$. The three stages are trained for 500k, 400k, and 200k iterations, respectively.

2. Network Details

Object State Embeddings. To address the issue of humans interacting with different objects at different times, we introduce two new learnable object state embeddings that serve as conditions for learning our human-object representation and scene representation, respectively. In a dynamic scene with N object states, we define N learnable state embeddings $\mathcal{O}_{\mathrm{s}} = \left\{\mathbf{O}_{\mathrm{s}}^{i}\right\}$ $(i = 1, 2, \cdots, N)$ to represent object states in the scene model, and N learnable state embeddings $\mathcal{O}_{\mathrm{c}} = \left\{\mathbf{O}_{\mathrm{c}}^{i}\right\}$ $(i = 1, 2, \cdots, N)$ to represent object states in the canonical space. The feature dimension of $\mathcal{O}_{\mathrm{s}}$ and $\mathcal{O}_{\mathrm{c}}$ are both set to 64 in our model. To obtain the number of object states, we manually label the transition timesteps for each video when the human picks up or puts down objects. Alternatively, we could use pretrained affordance detection methods to detect these transition timesteps. In our newly collected dataset, we provide the ground-truth transition timesteps for all the scenes.

State-Conditional Scene Network. As shown in Fig. 1(a), we employ a 10-layer multilayer perceptron (MLP) as our state-conditional scene base network, following the approach outlined in Mip-NeRF 360 [1]. Specifically, at state i, we utilize a concatenation of the IPE features $\hat{\gamma}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ of ray intervals with the scene state embedding $\mathbf{O}_{\mathrm{s}}^{i}$ as input to the scene MLP. To achieve this, we employ a skip connection that concatenates the input to the fifth layer. For the activation functions, we use ReLU after each fully connected layer, except for predicting density, for which we use Softplus, and for predicting color, for which we use Sigmoid.

State-Conditional Canonical Space Network. As illustrated in Fig. 1(b), we follow NeRF [9] to use an 8-layer MLP as our state-conditional canonical space model. At object state i, we concatenate the positionally encoded canonical points $\gamma\left(\mathbf{x}_{\mathrm{c}}\right)$ with the human-object state embedding $\mathbf{O}_{\mathrm{c}}^{i}$ and pass them to the canonical space MLP. In this canonical MLP, we adopt a skip connection that concatenates the input to the fifth layer. We use the ReLU activation after each fully connected layer, with the exception of the prediction of color, for which we employ the Sigmoid activation function.

3. Additional Results

Additional Ablation on State-Level Embedding. In a dynamic scene with N object states, we define N learnable state embeddings to represent state changes such as human pick up or put down objects. Therefore, N ($N \leqslant 7$ in main paper's Tab. 1) is much smaller than the number of frames ($[300, 400]$). Our state-level embeddings are designed to integrate state information through each state period, and thus enable more consistent modelling of scene changes for each period than frame-level embeddings [13, 14]. As validated in Fig. 2, when human picks up the backpack at frame 70, our state-conditional scene model can immediately remove the backpack starting from frame 71, while the frame-conditional model fails to model such a sudden change and renders wrong and incomplete backpacks on multiple frames.

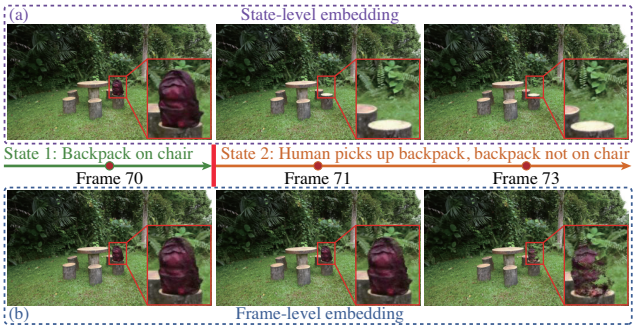

Figure 2: Ablation on the state-level embedding.

Method	Ours			NeuMan [4]		HyperNeRF [11]	Nerfies [10]	D^2NeRF [18]	K-Planes [2]
	1st stage	2nd stage	3rd stage	1st stage	2nd stage				
No. of GPUs	4	4	4	3	1	2	4	1	1
Training time (hours)	32	34	52	80	95	39	35	5.7	5.3

Table 1: Training time comparison on the HOSNeRF dataset against baselines.

Training Time Comparison on the HOSNeRF Dataset. Tab. 1 presents the training time of all methods on our HOSNeRF dataset. To ensure a fair comparison with the state-of-the-art (SOTA) approaches, we employ their highest configurations. Our three-stage training of HOSNeRF requires a total of five days, whereas NeuMan's [4] two-stage training demands over seven days. Due to the absence of distributed training support and the need for CPU computing, NeuMan's [4] second stage training takes 95 hours. In contrast, although the training time for D^2NeRF [18] and K-Planes [2] is less than 6 hours, their performances are significantly inferior on our challenging dataset, as evidenced by Tab. 2 and Fig. 4 of the main paper.

Optimized State-Conditional Canonical Spaces from Our HOSNeRF. Fig. 3 illustrates the state-conditional canonical spaces learned by our HOSNeRF on the HOSNeRF dataset. As shown in the figure, our proposed state-conditional dynamic human-object model can effectively represent different human-object states, and can reconstruct both the human bodies and objects with photorealistic details, enabling both 360° dynamic novel view synthesis and novel object / human pose manipulations. In addition, our complete HOSNeRF is able to render clean human-object canonical spaces based on coarse human-object masks.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[2] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. *arXiv preprint arXiv:2301.10241*, 2023.

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[4] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 402–418. Springer, 2022.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[6] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166, 2009.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[8] 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.

[9] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[10] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[11] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[12] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

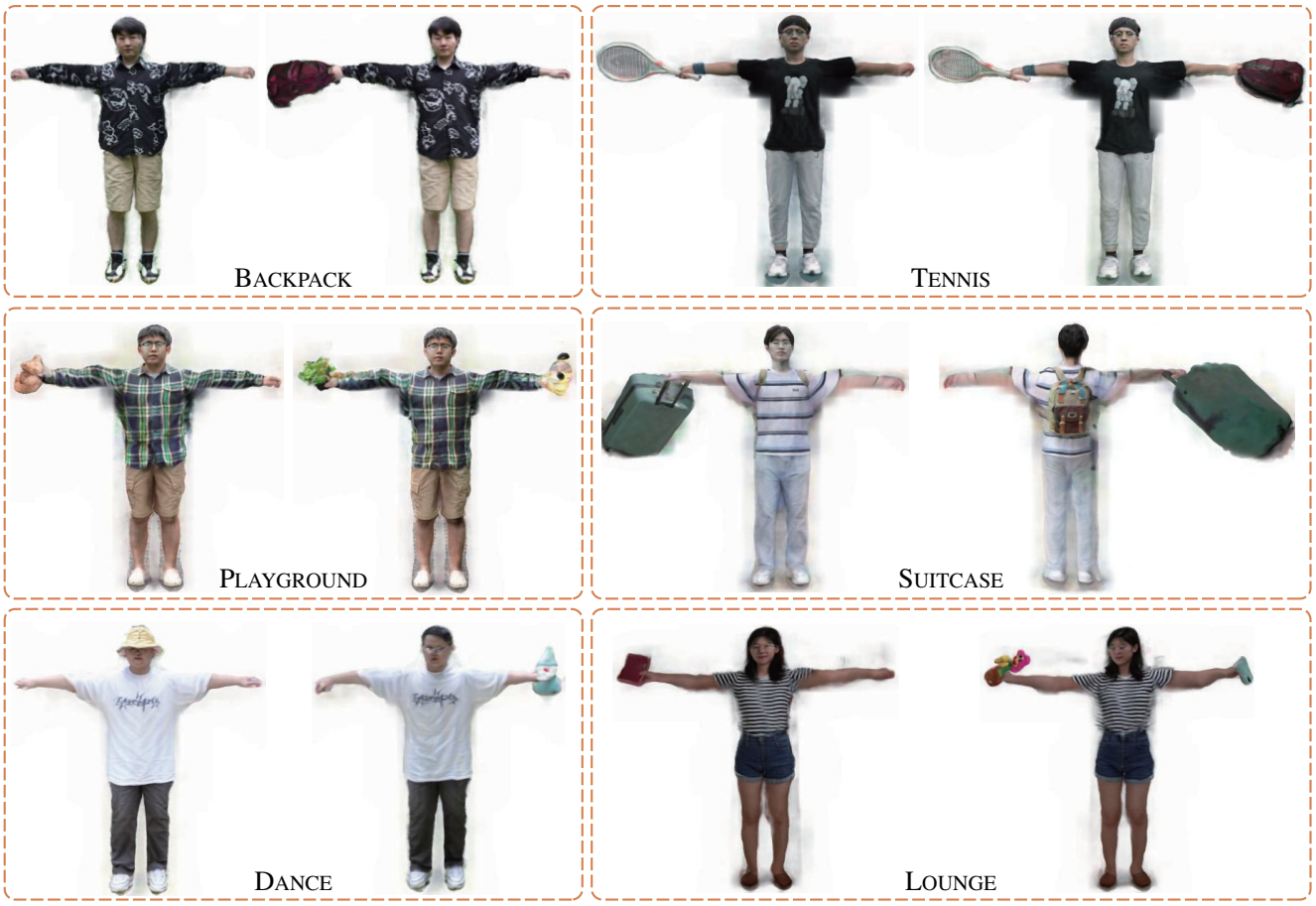

Figure 3: Optimized state-conditional canonical spaces of HOSNeRF on our HOSNeRF dataset.

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[13] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021.

[14] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021.

[15] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11179–11188, 2021.

[16] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022.

[18] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. Dˆ 2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in Neural Information Processing Systems*, 35:32653–32666, 2022.