PPR: Physically Plausible Reconstruction from Monocular Videos:
SUPPLEMENTARY MATERIALS

Gengshan Yang   Shuo Yang   John Z. Zhang   Zachary Manchester   Deva Ramanan
Carnegie Mellon University

A. Dynamic Scene Reconstruction

As mentioned in the submission, we collected a RGBD-pet dataset containing videos of a cat and a dog, captured by an iPad with RGBD sensor. We use the RGB stream for reconstruction. To evaluate the dynamic scene reconstruction accuracy, although one would want to use the complete scene geometry as ground-truth, it is difficult to obtain for in-the-wild dynamic scenes. Instead, we render the depth and evaluate against the depth from LiDAR sensors as a proxy.

Depth Metrics. Following Eigen *et al.* [2], we compute the root mean squared error (RMSE) for both rendered depth and disparity (inverse depth) maps. To find the unknown global scale factor, we align the median value of the rendered depth with the ground truth similar to Luo *et al.* [6]:

$$s_i = \operatorname*{median}_x \left\{ D_i^{pred}(x) / D_i^{ground-truth}(x) \right\}. \quad (1)$$

Table 1: **Comparison of scene reconstruction on `RGBD-pet`.** We report root-mean-square-error (RMSE, ↓) on rendered depth and disparity (inverse depth) maps, averaged over all frames. DPT-omnidata [1, 8] trains transformer-based depth predictors on a mix of multiple depth datasets. BANMo* [11] applies differentiable rendering to reconstruct deformable objects, and we follow NeuMan [4] to fit the object scale to a ground plane. PPR out-performs DPT-omnidata on the cat sequence, and out-performs BANMo* on both sequences.

Method	cat		dog	
	depth	disparity	depth	disparity
DPT-omnidata	0.620	0.201	**0.165**	**0.027**
BANMo*	0.181	0.149	0.232	0.061
PPR	**0.179**	**0.139**	0.216	0.041

Results. The results are shown in Tab. 1. We first interpret the results of DPT-omnidata. Leveraging depth priors learned from large-scale training data, DPT-omnidata performs very well for the dog sequence. However, it fails to produce accurate depth estimates for the cat sequence, possibly due to the uncommon top-down view angle of the video. PPR produces much better results on the cat sequence because it relies on *multiview* constraints that is more robust than depth priors. BANMo with ground fitting computes a *rough* relative scale between the object and the scene. As a result, the object still appears floating in many frames, producing less accurate depth estimations. In contrast, PPR couples differentiable physics optimization with differentiable rendering to jointly solve for the object scale and its global movements, which successfully reduces errors on the dynamic scene reconstruction task.

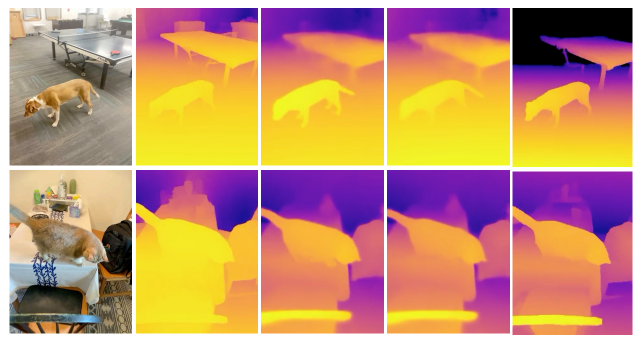

Figure 1: **Comparison of scene reconstruction on `RGBD-pet`.** Pixels with ground-truth depth greater than 10 meters are not captured by the depth sensor, and therefore removed from evaluation (marked as black). BANMo*: BANMo with ground plane fitting.

B. Additional Implementation Details

Regularization Terms. During differentiable rendering optimization, we apply shape and motion regularization terms as follows. We use 3D cycle loss to encourage the forward and backward warping fields $\mathcal{W}$ to be consistent with each other [5, 11]. We additionally apply an eikonal loss [3, 12] to both scene and object fields, which enforces the reconstructed signed distances to represent a surface:

$$\mathcal{L}_{\text{eikonal}} = (\|\nabla_{\mathbf{X}} \mathbf{MLP}_{\text{SDF}}(\mathbf{X})\| - 1)^2, \quad (2)$$

where we force the first order gradient of predicted SDF to have unit norm. Eikonal regularization helps produce

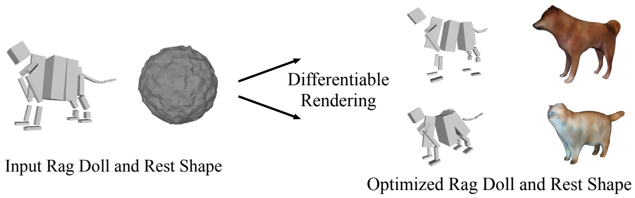

Figure 2: **Optimization of Rag Doll Model.** We start with a general rest shape (a unit sphere), and a known skeleton topology of the rag doll model. During optimization, both the shape and the rag doll model (joint locations and generalized mass of each link) are specialized to fit the input videos.

well-defined mesh when running marching cubes on the implicitly-defined surface.

Rag Doll Optimization. To optimize the object fields, we start with a general rest shape (a unit sphere) and a known skeleton topology of the rag doll model. During optimization, both the shape and the rag doll model (joint locations and generalized mass of each link) are specialized to fit the input videos. Please see Fig. 2 for the visualization of rest shapes and rag doll models.

Contact Plane Fitting. We assume the potential contact bones of a skeleton (the "feet") are known, and the contact locations are visible. The algorithm is as follows:

Input: Scene points $\mathbf{P} \in \mathbb{R}^{N\times 3}$, scene-to-camera transforms $\mathbf{G}_{s\rightarrow c} \in \mathbb{R}^{T\times 4\times 4}$ over T frames, camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$, and object "feet" trajectories in the camera space $\mathbf{J} \in \mathbb{R}^{T\times B\times 3}$.
'**Output**: Contact plane parameters $\mathbf{A} = (\mathbf{n}, d)$.
Parameters: Number of plane hypotheses $K = 5$, threshold $T_1 = 0.01$.

Step 1: Fit Multiple Planes
For k in 1:K
Fit a plane $\mathbf{A}^k$ to $\mathbf{P}$ using RANSAC with threshold T_1.
Set inlier points of $\mathbf{A}^k$ as $\mathbf{P}^k$, and remove those from $\mathbf{P}$.
Step 2: Find the Plane in Contact
Project scene points to images: $\mathbf{p} = \mathbf{K}\mathbf{G}_{s\rightarrow c}\mathbf{P} \in \mathbb{R}^{T\times N\times 2}$.
Project "feet" points to images: $\mathbf{q} = \mathbf{K}\mathbf{J}_c \in \mathbb{R}^{T\times B\times 2}$.
For k in 1:K
Score $\mathbf{A}^k$ by "feet"-to-$\mathbf{P}^k$ distance over frames and "feet":
$$\text{d}=\sum_{t=1}^{T}\sum_{j=1}^{B}\min(||\mathbf{p}_t^k - \mathbf{q}_t^j||).$$
Return $\mathbf{A}^k$ with the lowest total "feet"-to-$\mathbf{P}^k$ distance.

Under those assumptions, the contact plane does not have to occupy the majority of the background, and cameras do not have to point forward. Our algorithm works for the videos we tested on (included in the supplementary page), but breaks: (1) when the contact points are hard to define (e.g., cat lying sideways), or (2) when the object makes contact with multiple planes in a video.

Gradient Clipping. We find that differentiable physics introduces unstable gradients to the optimization, causing a high final reconstruction loss. Therefore, we clip outlier gradients to an empirical value $c = 0.1$:

$$\nabla_\phi L_{\text{DP}} = \begin{cases} \nabla_\phi L_{\text{DP}} & \text{if } \|\nabla_\phi L_{\text{DP}}\| \leq c \\ \frac{c}{\|\nabla_\phi L_{\text{DP}}\|}\nabla_\phi L_{\text{DP}} & \text{if } \|\nabla_\phi L_{\text{DP}}\| > c \end{cases} \tag{3}$$

where L_{DP} is the differentiable physics loss in Eq. (11) and ϕ is the physics parameters.

C. Additional Results

Comparison with animal body models. Creating accurate body models for animals is difficult due to lack of 3D data containing diverse animal shape, appearance, and pose. In the following, we show a visual comparison with BARC [10], a state-of-the-art dog body model in Fig. 3. The video comparison can be found on the supplement website.

Figure 3: **Comparison with BARC.** BARC fails to reconstruct the sharp ears of the dog, and puts the legs into the wrong positions, while PPR faithfully reconstructs them.

Roll-out Performance. In Fig. 4, we show qualitative results of simulating the physical system (rag doll model) for various time windows. Within the time window T in training, the simulation is almost always stable. When simulating a time window greater than T, the controller might fail to track the motion.

We posit that it is because the error in the states of the rag doll model accumulates over time [9]. The PD controller is not able to generalize to never-before scenarios. One potential direction to improve this is to ask the controller to reason about future time horizons (instead of the direct next step) [7].

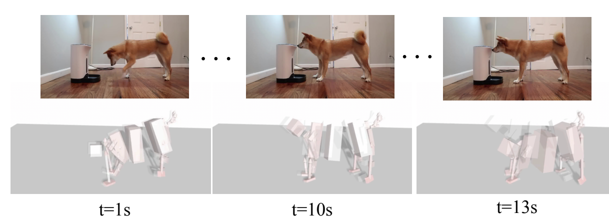

Figure 4: **Simulation over long time window.** We perform physics optimization with a window size of 2.4s. The controller keeps track of the target for 10s, and diverged at around 13s. Red: simulated character. Gray: reference character.

References

[1] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 1

[2] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1

[3] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 1

[4] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *ECCV*, pages 402–418. Springer, 2022. 1

[5] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 1

[6] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. 39(4), 2020. 1

[7] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), July 2021. 2

[8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021. 1

[9] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011. 2

[10] Nadine Rueegg, Silvia Zuffi, Konrad Schindler, and Michael J. Black. BARC: Learning to regress 3D dog shape from images by exploiting breed information. In *CVPR*, pages 3876–3884, 2022. 2

[11] Gengshan Yang, Minh Vo, Neverova Natalia, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 1

[12] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021. 1