DreaMo: Articulated 3D Reconstruction From A Single Casual Video Supplementary Material

1. Advantages of DreaMo's Skeleton Generation

Since predefined skeleton structure in methods like RAC [4] and CASA [2] limits the output space to specific categories such as quadrupeds, we deliberately avoid incorporating such skeleton priors in DreaMo. Specifically, *DreaMo's neural bones are randomly initialized and learned without any predefined skeleton structure*. The human-interpretable skeleton is a valuable byproduct of DreaMo, enabling users to articulate the reconstructed 3D model easily. In contrast, BANMo [3] has neural bones scattering in empty space or converging in less optimal regions. It is hard to find bones correlated to animal parts and unable to perform intuitive articulations. This limits the application of BANMo to only retargeting motions from other videos instead of directly articulating by users.

To assess the practical utility of the generated skeleton, we further provide the rigging results in Figure 2 by plugging DreaMo's generated skeleton and the learned 3D shape into the Blender auto-rigging tool. To build a skeleton with a tree structure required by the rigging tool, we set a neural bone as a parent if it has more connected vertices than nearby bones, with the count determined by the learned skinning weights described in Section 3.4.

2. SDS Loss Optimization

As depicted in Figure 1, we tried several SDS strategies to improve single-video 3D reconstruction: (i) only updating geometry-related parameters, (ii) updating all parameters, (iii) updating all parameters after geometry-related parameters converged, and (iv) updating texture-related parameters after geometry-related parameters converged. We observed that updating texture-related parameters hinders texture development even after the geometry has converged. Therefore, we make the SDS gradients only update the geometry-relevant parameters as discussed in Section 3.2.

3. 3D Metric on PlanetZoo

We evaluate the 3D metric on the CASA PlanetZoo dataset [2], which contains "synthetic" animal videos along with ground-truth meshes for the reconstruction performance evaluation. The Chamfer distance, for BANMo [3] and DreaMo, are 0.40 and 0.35, respectively. The lower Chamfer distance indicates the better reconstruction performance. The results conclude DreaMo still outperforms BANMo. Since PlanetZoo authors did not release the evaluation script, we use BANMo's implementation (including Iterative Closest Point), making the values not directly comparable to CASA.

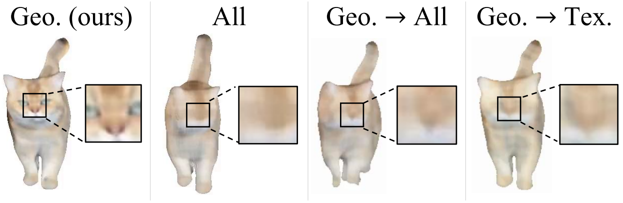

Figure 1. Strategies for SDS loss optimization.

4. Supplementary Video

Please watch the supplementary demo video for a comprehensive comparison between our method DreaMo and the current state-of-the-art method BANMo [3]. The video includes qualitative results for novel view synthesis, 3D shape reconstruction, and novel pose articulation.

5. More Results for 3D Reconstruction

We provide more 3D reconstruction results in Figure 3 to compare between BANMo [3] and our proposed DreaMo. The figure shows that BANMo produces texture artifacts and irregular shapes due to insufficient view coverage. In contrast, benefiting from simultaneously reconstructing training-view frames and hallucinating unobserved regions of the target subjects, DreaMo generates more plausible rendered images and convincing shapes.

6. More Results for Articulating 3D Model

In Figure 4, we provide more examples for controlling DreaMo. Following the 3D model manipulation described in Section 3.1, we manually control the generated skele-

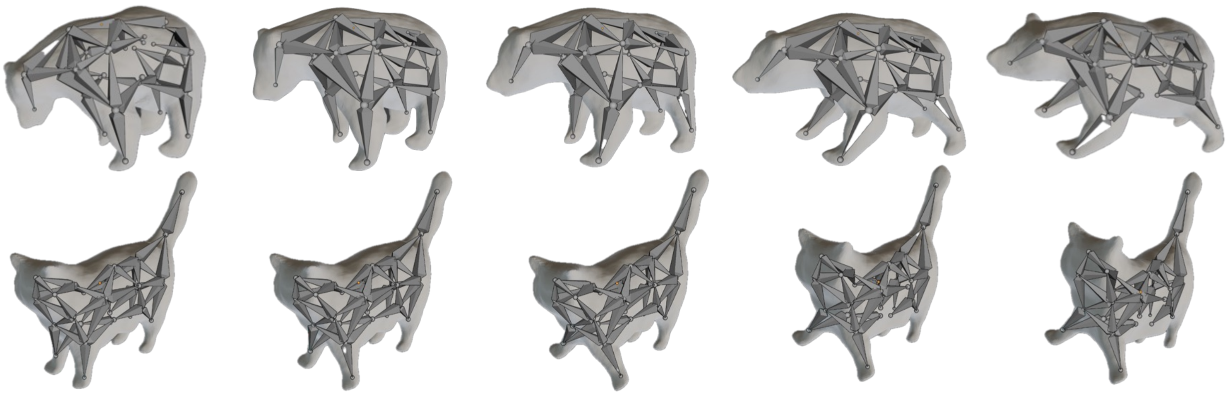

Figure 2. **Utilizing DreaMo's skeleton output for rigging.** DreaMo's skeleton can be used to rig the reconstructed 3D shape obtained from a single Internet video.

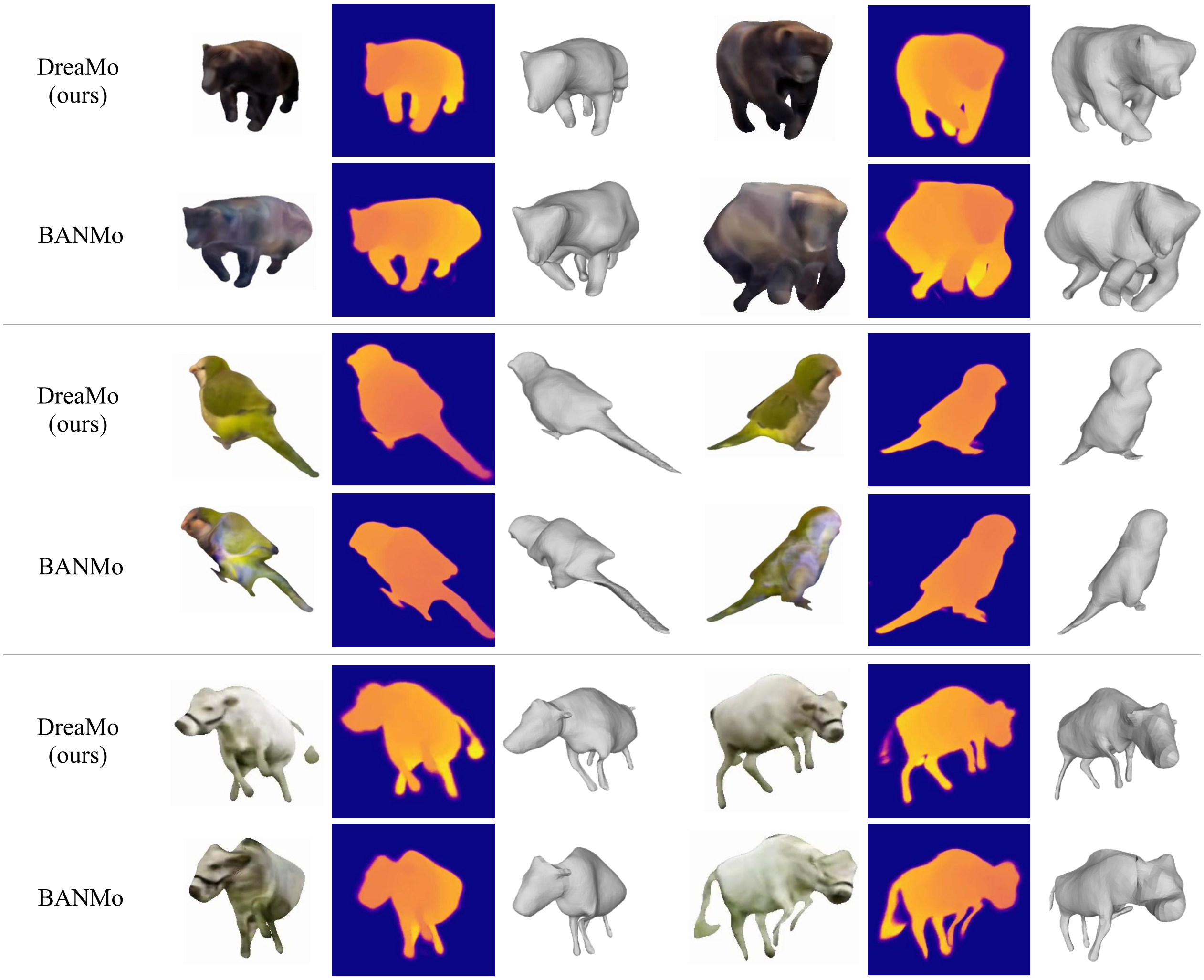

Figure 3. **3D reconstruction comparison among BANMo [3] and our DreaMo.** We show rendered novel-view images, including RGB and depth, and their corresponding reconstructed shapes for each method.

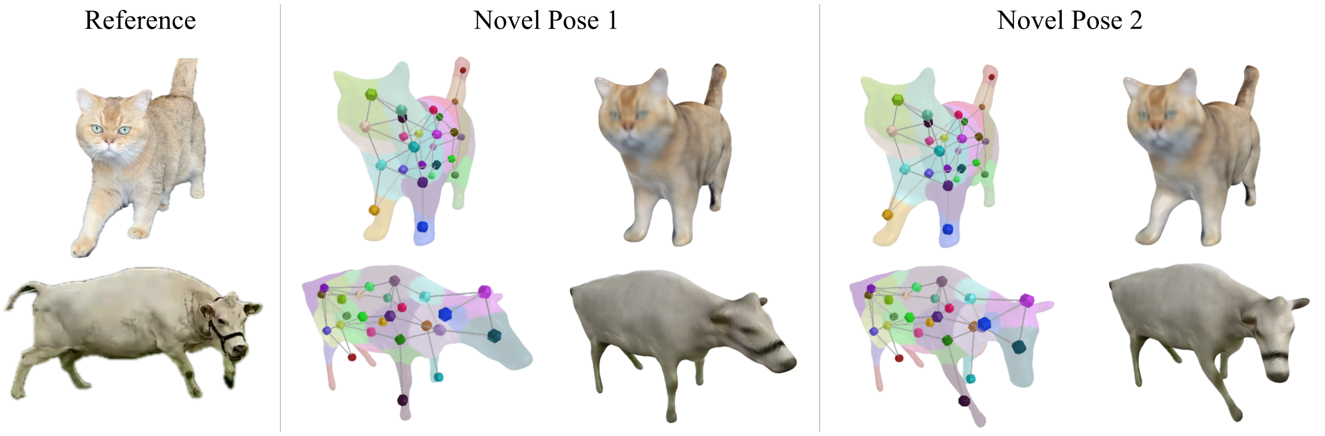

Figure 4. **Manipulating DreaMo by controlling the generated skeletons.** We manually modify the bone positions and warp the associated skin to manipulate the reconstructed model into new poses.

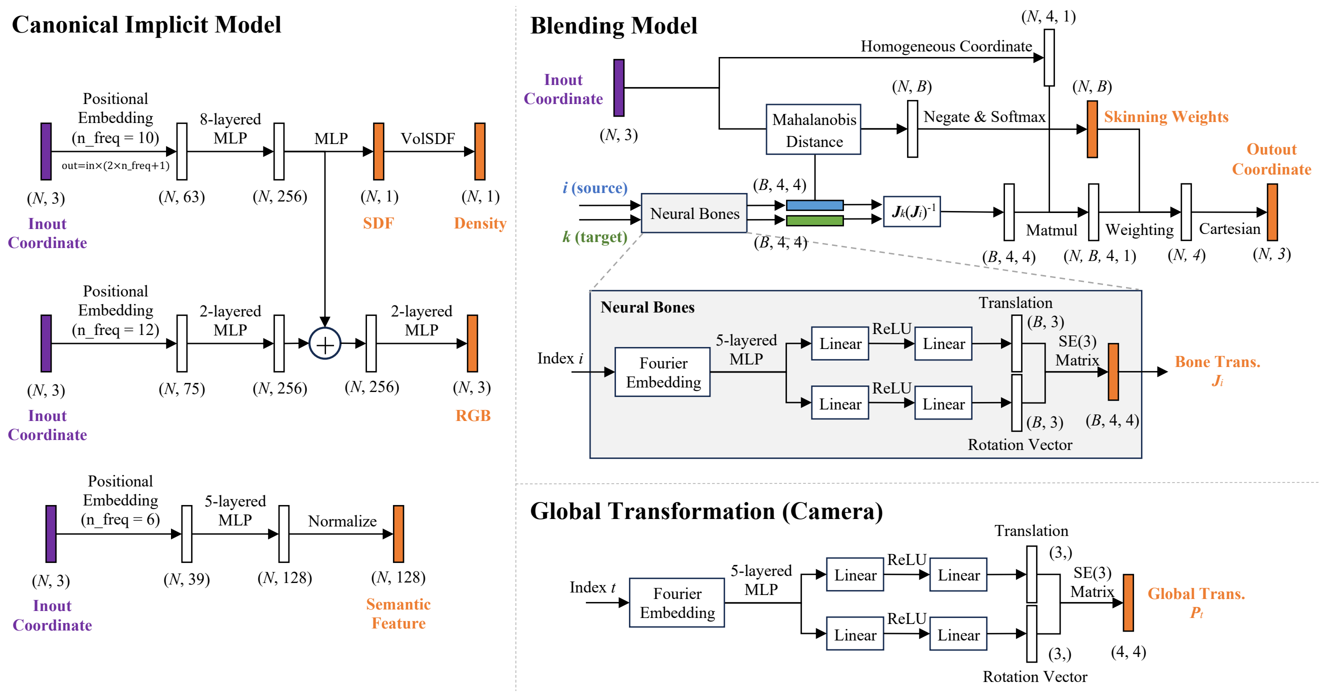

Figure 5. **Network architecture of DreaMo.** We present the network design of the canonical implicit model (left), the blending models (upper right), and the global transformation model (lower right). The warping models, which transform 3D coordinates between canonical and observation spaces, integrate blending and global transformation models, as elaborated in Section 3.1.

ton and transform the skin points (*i.e*. vertices of the mesh) to produce novel poses. With the accurately learned bone placement, skinning weights, and 3D shapes from DreaMo, the skin points can be reasonably transitioned in response to the movement of the neural bones. This results in realistic outcomes for novel poses.

7. Experimental Details

We conduct all the experiments on a single RTX 3090 GPU. Both DreaMo and BANMo are trained for 100 epochs, taking 2.5 and 1.7 hours, respectively. In addition, the training of DreaMo utilizes 22 GB of GPU memory, compared to BANMo which uses 13.5 GB. We represent 3D rotation using quaternions, implemented in PyTorch.

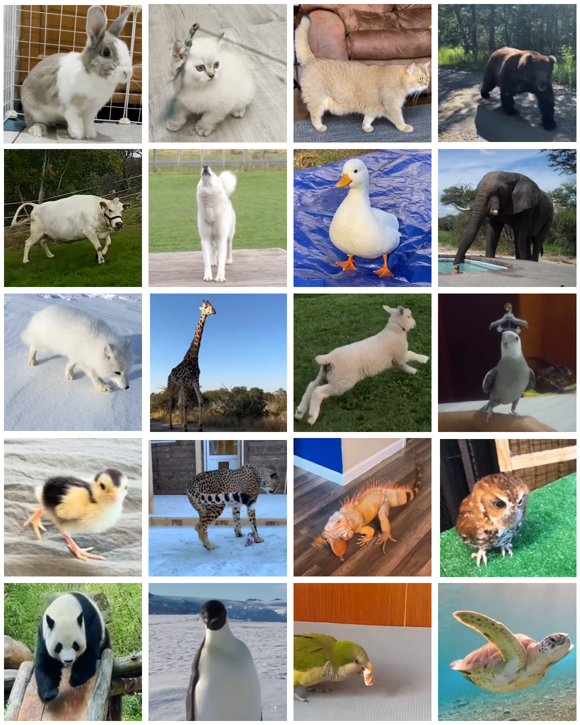

Figure 6. Our self-collected video dataset for diverse species and insufficient view coverage from the Internet. We sample cropped images from the dataset.

8. Implementation Details

8.1. Network Architecture

In Figure 5, we show the network architecture of the canonical implicit model, the blending model, and the global transformation model.

Similar to BANMo [3], we use MLPs for the canonical implicit model, where different frequencies for the positional embedding are employed to model the various degrees of change for density, appearance, and semantic features. Following VolSDF [5], the SDF values can be converted to density by the cumulative distribution function of the Laplacian distribution.

The blending model transforms a given input coordinate into the output coordinate based on the subject's deformation, which the skinning weights and the transformation between the source and target bones can determine. Specifically, the skinning weights are calculated based on the Mahalanobis distance between the input coordinates and the source neural bones. This distance is then negated and normalized by a subsequent softmax layer. On the other hand, the bone positions for each time step can be acquired using frame-wise Fourier embedding [1, 3], followed by an MLP.

Finally, the global transformation, representing the camera transformation, is obtained by feeding the frame-wise Fourier embedding into the subsequent MLP.

8.2. 3D Shape Reconstruction

We perform two steps to reconstruct the subject's shape in a specific video frame. First, using marching cubes, we extract the rest-pose mesh, which represents the 3D shape of the reconstructed subject, from the neural implicit model in canonical space. Afterward, we apply the forward warping model to transform each vertex on the mesh to the observation space at the time step of the given video frame.

9. Dataset

We show the diverse species in the self-collected dataset in Figure 6. It is important to note that the images in Figure 6 have been cropped just for visualization. This cropping does not imply that the target subjects in the training video are large and centered.

10. Limitations and Future Works

Despite DreaMo achieving exciting results, it remains a special case of structure-from-motion methods, which inherently require a certain level of camera baseline and are unable to handle videos with excessively low view coverage. Besides, accurately discovering the correct placement of the neural bones and skinning weights requires a video to demonstrate the movable parts with real-world motions, thus DreaMo cannot hallucinate bones and articulations in the completely invisible regions. We acknowledge these limitations and aim to address them in future work.

References

[1] 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*, 2021. 4

[2] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. *NeurIPS*, 2022. 1

[3] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 1, 2, 4

[4] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 1

[5] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021. 4