Distilling Neural Fields for Real-Time Articulated Shape Reconstruction

Jeff Tan, Gengshan Yang, Deva Ramanan
Robotics Institute, Carnegie Mellon University
Webpage: https://jefftan969.github.io/dasr | Email: jefftan@andrew.cmu.edu | Github: https://github.com/jefftan969/dasr

1. Introduction

How to reconstruct **articulated** objects from imagery in **real-time**?

Input: RGB stream Output: Articulated 3D model

Challenge: Without multiple views or depth sensors, it is hard to infer accurate shape/motion from casual videos.

2. Related Work

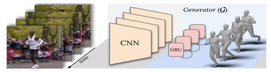
Feed-forward **parametric body models** are fast, but it's hard to get 3D templates for arbitrary objects.

(*requires expensive 3D registration and scanning!*)

Body Templates

SMPL (M. Loper, CVPR 2015)

Motion Predictors

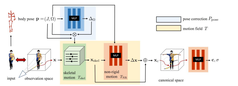


VIBE (M.Kocabas, CVPR 2019)

Using **differentiable rendering optimization** to jointly learn shape and motion is accurate, but far too slow.

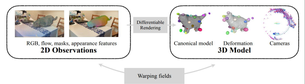
(*takes hours per video!*)

Humans



HumanNeRF (C. Weng, CVPR 2022)

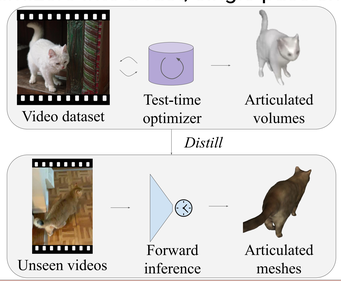
Quadrupeds



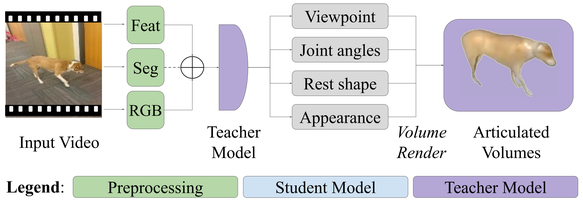
BANMo (G. Yang, CVPR 2022)

3. Method

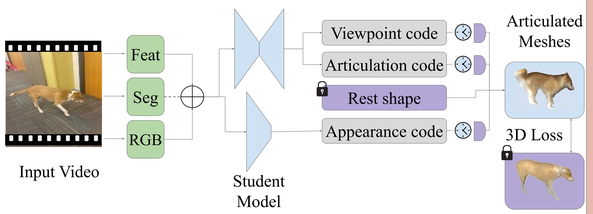
Our method **distills** slow-but-accurate shape and motion optimizers into a fast, single-pass shape predictor.



Step 1: Train a **teacher model** to optimize for shape and motion *offline*, given input videos from a category (50+).



Step 2: Train a **student model** to output forward shape predictions, supervised *in 3D* by the teacher's output



4. Results

Conclusion: Category-specific dynamic NeRFs can be distilled into feed-forward shape predictors, enabling real time shape prediction from casual videos at scale (600x speedup). **For video results, please see the project page.**

Figure 1:

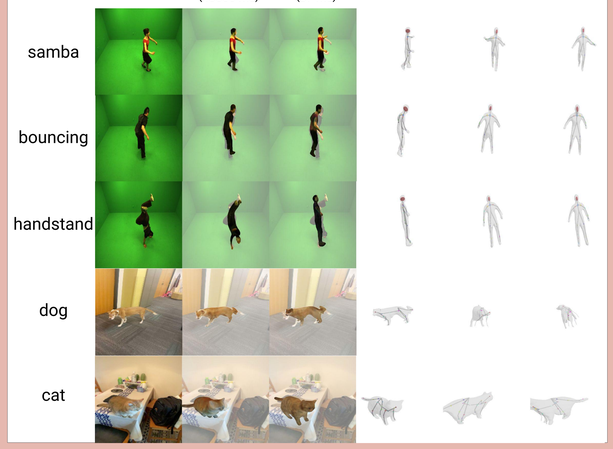


Figure 2: Comparison vs. BARC, a real-time forward dog model.

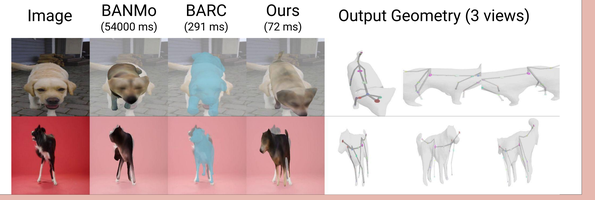


Figure 3: Failure cases on challenging inputs and poses.

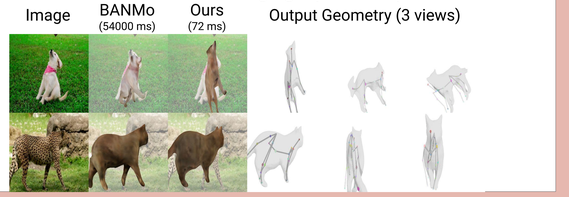
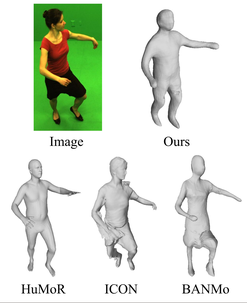


Table 1: 3D chamfer distance (cm, ↓) and F-score (%, ↑) for humans, averaged over all frames. We also report average inference time per frame (ms) on an RTX-3090 GPU.

Method	Time	T_samba				D_bouncing				D_handstand			
		CD	F@1%	F@2%	F@5%	CD	F@1%	F@2%	F@5%	CD	F@1%	F@2%	F@5%
Ours	**67**	11.54	83.0	59.8	28.0	16.26	71.3	44.6	18.1	26.46	54.6	32.2	13.1
HuMoR	42000	**10.32**	**88.3**	60.8	26.0	11.75	85.1	56.6	23.4	30.24	46.4	25.1	9.7
ICON	63000	10.43	85.9	**62.3**	**29.7**	**9.77**	**88.3**	**65.6**	**31.0**	16.02	72.5	48.2	20.4
BANMo	43000	11.56	82.7	57.0	25.3	10.90	86.2	64.9	29.8	**15.22**	**75.5**	**50.7**	**21.8**

Our method approaches the quality of BANMo and baselines, and is **600x faster**.

Figure 4: Qualitative comparison on humans.



5. Future Work

How to build general articulated body models from thousands of Internet videos?