LASR: Learning Articulated Shape Reconstruction from a Monocular Video

Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, William T. Freeman, Deva Ramanan, Ce Liu

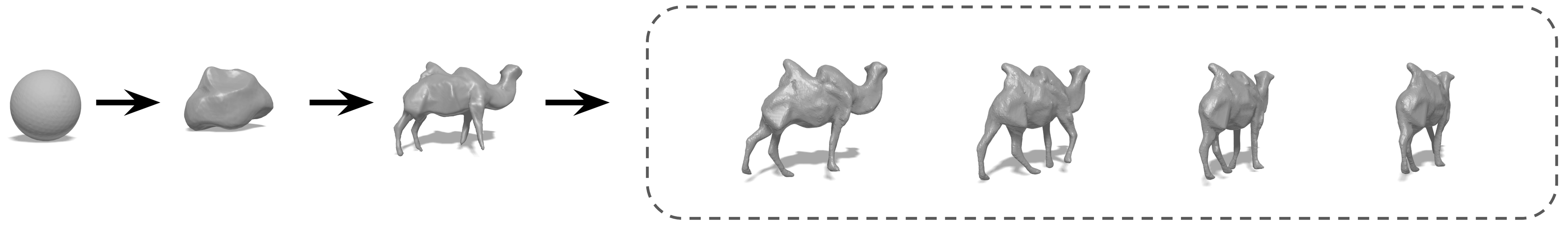
Research at Google

LASR reconstructs articulated 3D shapes from a single video by differentiable rendering with data-driven motion correspondence priors.

Code at http://lasr-google.github.io/.

Problem setup

Video of an object from arbitrary category



Articulated 3D shape reconstruction

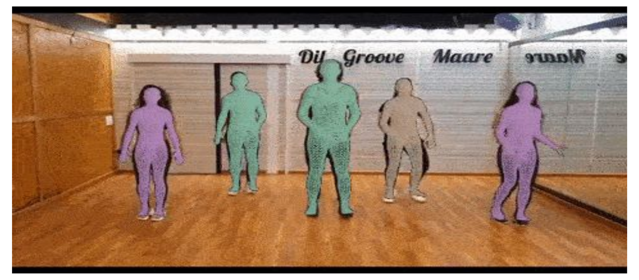
Challenges

"3 x N + camera pose + intrinsics" for SfM

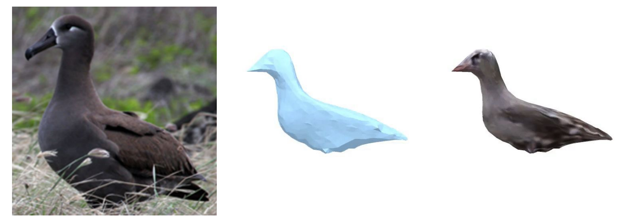
- "2 x N points x T" vs "3 x N x T + camera pose + intrinsics"
  - The problem is under-constrained due to the time-varying 3D shape.
  - Difficult to obtain accurate 2D point trajectories.

Related work

Template-based



Parametric shape-pose models (SMPL, SMAL) e.g., SMPLify, VIBE [1]



Weakly-supervised category shape reconstruction e.g., CMR, A-CSM, UMR, UCMR [2]

NRSfM, e.g., ND-NRSfM, C3DPO [3]

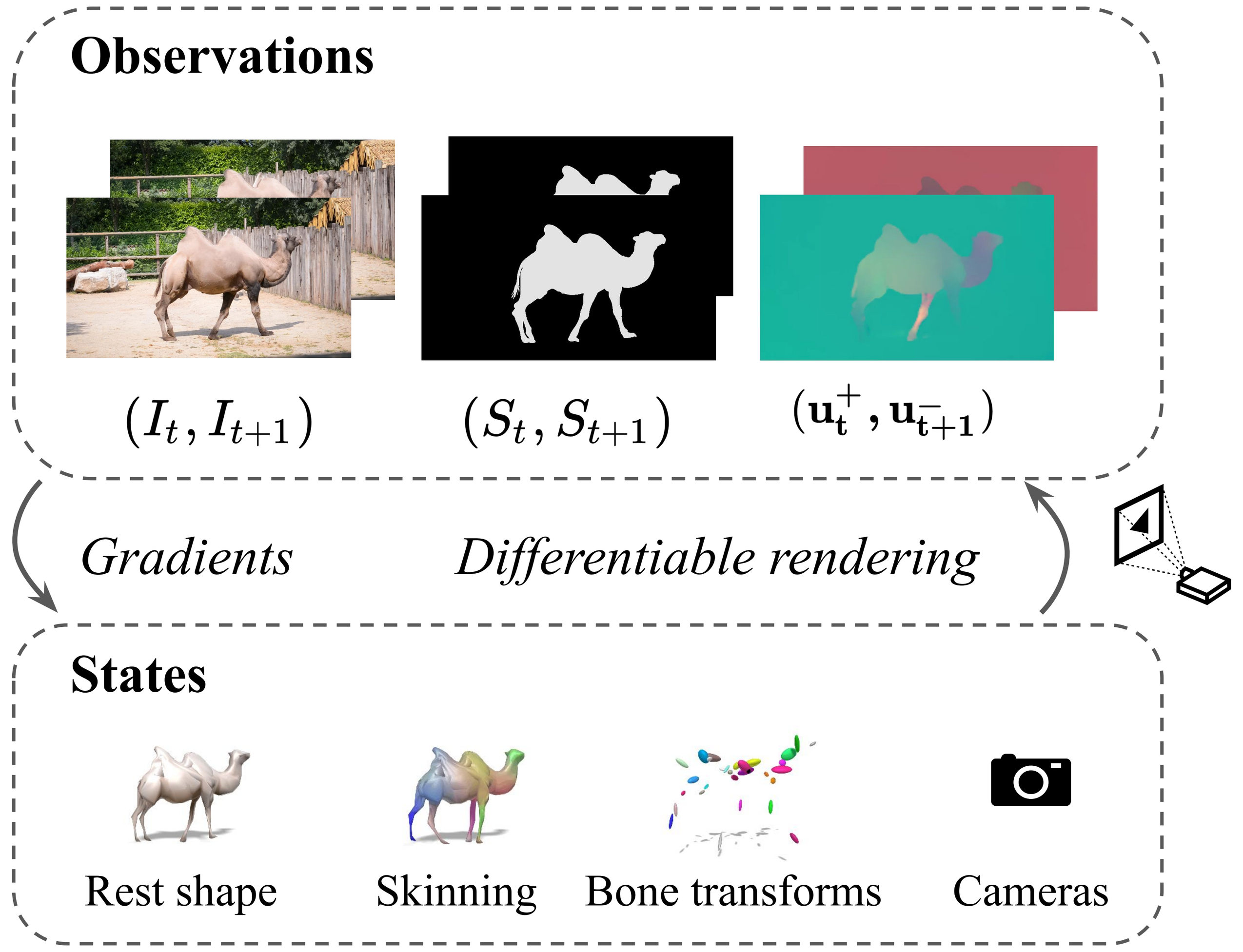
LASR (Ours)

Template-free

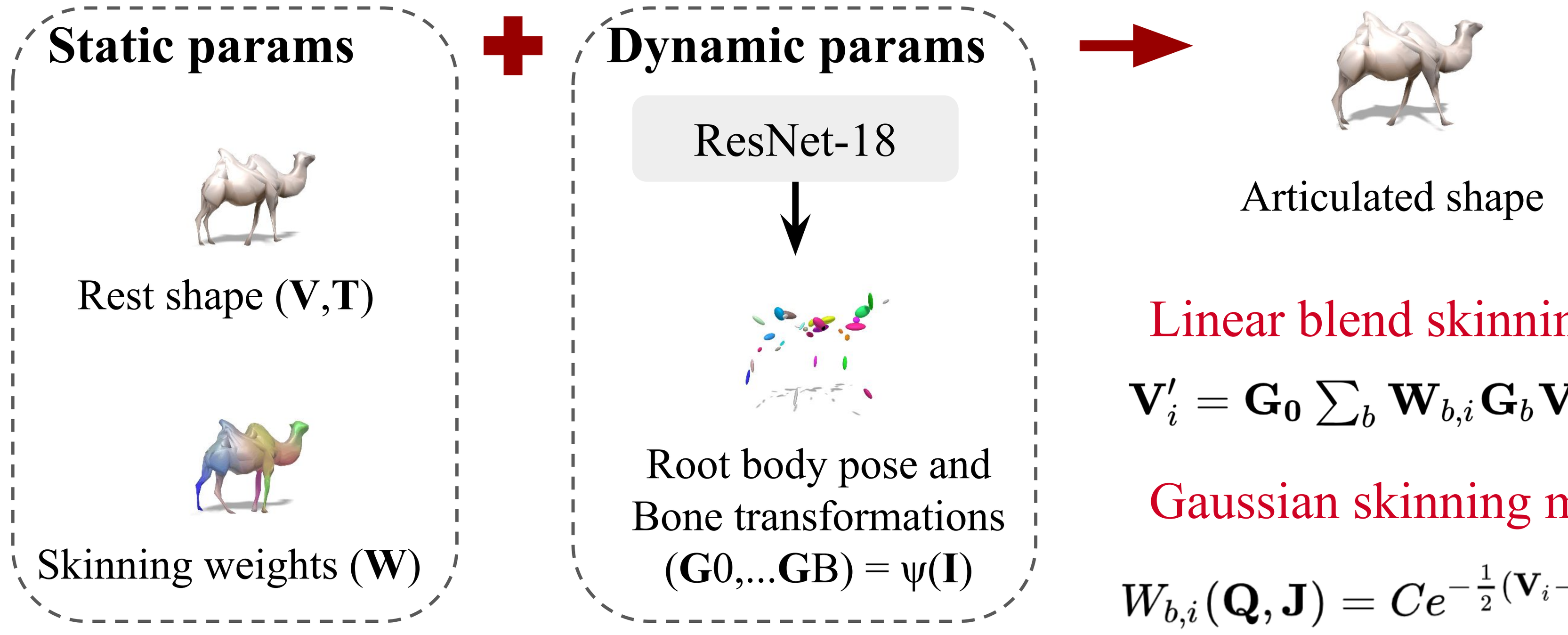
[1] Muhammed, Athanasiou, and Black. "VIBE: Video inference for human body pose and shape estimation." CVPR 2020.
[2] Goel, Kanazawa, and Malik. "Shape and viewpoint without keypoints." ECCV 2020.
[3] David, et al. "C3DPO: Canonical 3d pose networks for non-rigid structure from motion." ICCV 2019.

Approach

Key idea: diff. rendering to fit pixels, masks, and *flow*



Articulation modeling



Linear blend skinning:

$$V'_i = G_0 \sum_b W_{b,i} G_b V_i$$

Gaussian skinning model:

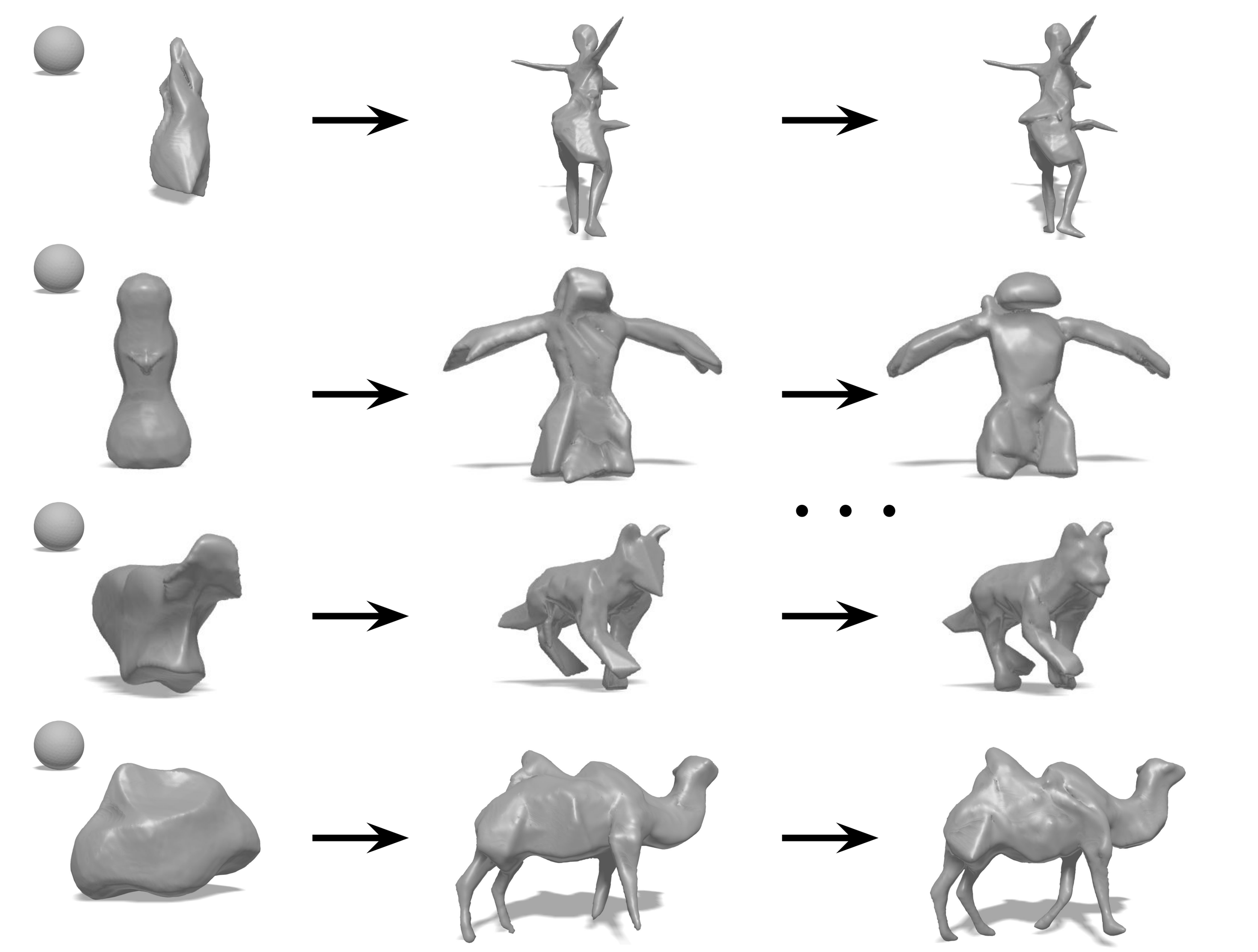
$$W_{b,i}(Q, J) = Ce^{-\frac{1}{2}(V_i - J_b)^T Q_b (V_i - J_b)}$$

Objective

$$L(V, T, W, K, \psi) = L_{recon} + L_{reg} \begin{cases} L_{recon} = L_{mask} + L_{flow} + L_{color} \\ L_{reg} = L_{shape} + L_{ARAP} + L_{least\text{-}motion} + L_{symmetry} \end{cases}$$

intrinsics

Coarse-to-fine optimization



Results