LASR: Learning Articulated Shape Reconstruction from a Monocular Video

Gengshan Yang[1][*], Deqing Sun[2], Varun Jampani[2], Daniel Vlasic[2], Forrester Cole[2], Huiwen Chang[2],
Deva Ramanan[1], William T. Freeman[2], and Ce Liu[2]

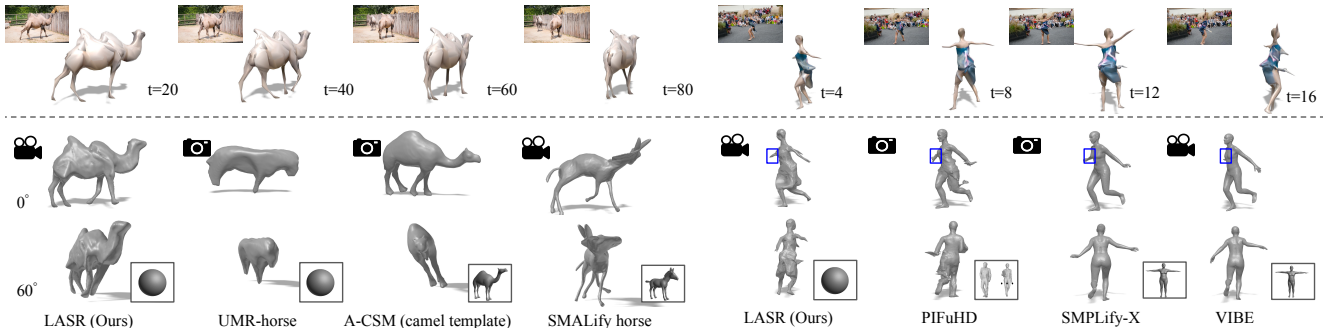[1]Carnegie Mellon University,   [2]Google Research

Figure 1. **Top**: Sample input video frames and articulated shapes recovered by our method (LASR). **Bottom**: Comparison with existing methods, where the input to each method (either video or image) is denoted at the top left, and the shape template being used is denoted at the bottom right of each result. Many existing approaches on nonrigid shape reconstruction heavily rely on category-specific 3D shape templates, such as SMPL for human [33, 35] and SMAL for quadrupeds [6, 58]. In contrast, LASR jointly recovers the object shape, articulation, and camera parameters from a monocular video without using category-specific shape templates. By relying on generic shape and motion priors, LASR applies to a wider range of nonrigid shapes and yields high-fidelity 3D reconstructions: It recovers both humps of the camel, which are missing from other methods. It also recovers the silk ribbon of the dancer (as denoted by the blue box), which confuses SMPLify-X and VIBE as the right arm. Please refer to video results on our project page.

Abstract

Remarkable progress has been made in 3D reconstruction of rigid structures from a video or a collection of images. However, it is still challenging to reconstruct nonrigid structures from RGB inputs, due to its under-constrained nature. While template-based approaches, such as parametric shape models, have achieved great success in modeling the "closed world" of known object categories, they cannot well handle the "open-world" of novel object categories or outlier shapes. In this work, we introduce a template-free approach to learn 3D shapes from a single video. It adopts an analysis-by-synthesis strategy that forward-renders object silhouette, optical flow, and pixel values to compare with video observations, which generates gradients to adjust the camera, shape and motion parameters. Without using a category-specific shape template, our method faithfully reconstructs nonrigid 3D structures from videos of human, animals, and objects of unknown classes. Our code is available at `lasr-google.github.io`.

1. Introduction

Perceiving and modeling the geometry and dynamics of 3D entities is an open research problem in computer vision and has numerous applications. One fundamental challenge is the under-constrained nature of the problem: from limited 2D image measurements, there exist multiple interpretations of the geometry and motion of the 3D world.

A recent and promising trend for addressing this challenge is exploiting data priors, which have proven quite successful for high-level vision tasks, such as image classification and object detection [13, 31]. However, in contrast to high-level vision tasks, it is often costly to obtain 3D annotations for real-world entities. For example, SMPL [33] is learned from thousands of registered 3D scans of human. SMAL [58] is learned from scans of animal toys and a manually rigged mesh model. It involves nontrivial efforts to collect such data for an arbitrary object category. Therefore, existing methods often fail to capture objects of novel or unknown classes, and hallucinate an average 3D structure based on the category shape prior, as shown in Fig. 1.

Interestingly, remarkable progress has been made in the field of SLAM and structure-from-motion without relying

---

on strong shape priors by taking advantage of *multiview* data recordings. However, such results are limited to static scenes. We explore an intermediate regime between these two extremes: *Can one reconstruct an articulated shape from video data without relying on template priors?*

Why videos? To reconstruct 3D object shape from images, prior work learns category-specific shape models either from 3D data [17, 41] or from 2D supervision, such as object silhouette and keypoints in a large image collection [10, 18, 23, 30]. However, 3D data are generally difficult to acquire at a large scale due to sensor design. Although it is easier to collect images of the same category, enforcing multiview constraints is often challenging, due to ambiguities of associating 2D observations across instances and under different viewpoints [11, 39]. Video serves as an alternative to depth scans and image collections – videos are easier to acquire, and provide well-defined multiview constraints on the 3D shape of the same instance.

Why optical flow? To solve the inverse problem, prior work discussed various forms of 2D constraints or supervision, such as object silhouette, texture, 2D keypoints, and semantic parts [4, 18, 23, 30]. Why should motion be treated as a first-class citizen? Besides that optical flow naturally encodes correspondences, it provides more fine-grained information than keypoints as well as semantic parts. Different from long-range point tracks, which is the classic input for NRSfM [42], optical flow can be obtained more reliably [47, 52] over two consecutive frames.

Why not nonrigid SfM? NRSfM deals with a problem similar to ours: given a set of 2D point trajectories depicting a deformable object in a collection of images, the goal is to recover the 3D object shape and pose (i.e., relative camera position) in each view. Usually, trajectories of 2D points are factorized into low-rank shape and motion matrices [8, 19, 26] without using 3D shape templates. Although NRSfM is able to deal with generic shapes, it requires reliable long-term point tracks or keypoint annotations, which are challenging to acquire densely in practice [42, 44, 46].

Proposed approach: Instead of inferring 3D shapes from category-specific image collections or point trajectories, we build an articulated shape model from a monocular video of an object. Recent progress in differentiable rendering allows one to recast the problem as an analysis-by-synthesis task: *we solve the inverse graphics problem of recovering the 3D object shape (including spacetime deformations) and camera trajectories (including intrinsics) to fit video observations, such as object silhouette, raw pixels, and optical flow*. An overview of the pipeline is shown in Fig. 2.

Contributions: We propose a method for articulated shape reconstruction from a monocular video that does not require a prior template or category information. It takes advantage of dense two-frame optical flow to overcome the inherent ambiguity in the nonrigid structure and motion estimation

Table 1. Related work in nonrigid shape reconstruction. [(1)]Model-based optimization and regression methods. [(2)]Category-specific mesh reconstruction. [(3)]Template-free approaches. S: single view. V: video or multi-view data. I: images. J2: 2D joints. J3: 3D joints. M: 2D masks. V3: 3D scans. C: camera matrices. F: optical flow. MF: multi-frame optical flow. quad: quadruped animals. [†]:Only representative categories are listed.

	Method	category	template	test-input	train
(1)	SMPLify [7]	human	SMPL	S:J2,M	None
	VIBE [25]	human	SMPL	V:I	J2,J3
	SMALify [6]	quadx5	SMAL	V:J2,M	None
	SMALR [57]	quadx12	SMAL	S:J2,M	None
	SMALST [56]	zebra	SMAL	S:I	J2,V3
	WLDO [5]	dog	SMAL	S:I	J2,M
(2)	CMR [23]	bird†	SfM-hull	S:I	J2,M,C
	UCMR [18]	bird†	cate-mesh	S:I	M
	UMR [30]	bird†	None	S:I	M
	IMR [48]	animals	cate-mesh	S:I	M
	A-CSM [27]	animals	cate-mesh	S:I	M
	VMR [29]	animals	cate-mesh	V:M	None
(3)	PIFuHD [41]	human	None	S:I	V3
	NRSfM [2, 12]	any	None	V:J2	None
	N-NRSfM [44]	any	None	V:MF,M	None
	WSD [10]	dolphin†	cate-mesh	V:J2,M	None
	A3DC [38]	any	None	V:stroke	None
	LASR (Ours)	any	None	V:F,M	None

problem. It automatically recovers a nonrigid shape under the constraints of rigid bones under linear-blend skinning. It combines coarse-to-fine re-meshing with soft-symmetric constraints to recover high-quality meshes. Our method demonstrates state-of-the-art reconstruction performance in the BADJA animal video dataset [6], strong performance against model-based methods on humans, and higher accuracy on two animated animals than A-CSM [27] and SMALify [6] that use shape templates.

2. Related Work

Below and in Tab. 1, we discuss related work of nonrigid shape recovery according to priors being used (shape template, category prior, or generic shape and motion priors).

Model-based reconstruction: Model-based reconstruction leverages a parametric shape model to solve the under-constrained 3D shape and pose estimation problem. A large body of work in 3D human and animal reconstruction uses such parametric shape models [33, 35, 51, 57, 58], which are learned from 3D scans of human or animal toys [33, 58], and allow one to recover the 3D shape given very few annotations at test time (2D keypoints and silhouettes). Recently, model-based regression methods are developed to predict model-specific shape and pose parameters from a single image or video [3, 5, 25, 56], usually trained with ground-truth 3D data generated by such parametric shape models. Although model-based methods achieve great success in reconstructing "closed-world" objects of known category and
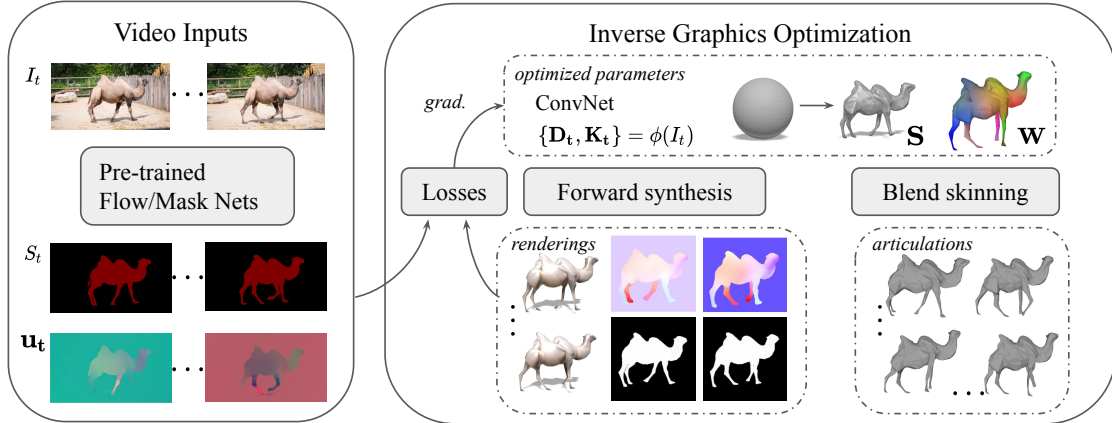
Figure 2. Method overview. Given a monocular video of an object, we jointly recover the object's rest shape $\mathbf{S}$, skinning weights $\mathbf{W}$, articulation $\mathbf{D_t}$, and camera parameters $\mathbf{K_t}$ by solving an inverse graphics problem through gradient-based optimization. The object rest shape is represented by a mesh (initialized from a sphere) and articulated at each frame under linear blend skinning (Sec. 3.2). The time-varying parameters, including focal length, object root transformation and articulation, are parameterized by a neural basis, i.e., convolutional network given image inputs (Sec. 3.4). At each iteration, we randomly sample pairs of consecutive frames and forward-render object silhouette, texture, and two-frame optical flow using a differentiable renderer (Sec. 3.1). The renderings are compared against raw pixels, segmentation and optical flow estimated by off-the-shelf networks to generate gradients signals and update the model (Sec. 3.3).

rich 3D data, it is challenging to apply to unknown object categories, or categories with limited 3D data.

Category mesh reconstruction: A recent trend is to reduce supervision from 3D or multi-view capturing to 2D annotations, such as keypoints and object silhouettes [18, 23, 30]. Such methods often take advantage of category priors, including a collection of images from the same category, and category-specific shape templates [27, 48]. Recent progress makes single-view reconstruction of birds and other common categories possible without 3D annotation. However, the single view reconstruction is usually coarse and lacks instance-specific details. Recent work adapts category-specific models to a test-time video [29], but still does not handle objects of unknown classes.

Template-free reconstruction: Among the template-free methods, PIFu [40, 41] learns to predict an implicit shape representation for clothed human reconstruction, but requires ground-truth 3D shapes to train. A3DC [38] reconstructs articulated animals from videos, but requires involved user stroke interactions. Without requiring 3D data or user interactions, NRSfM factorizes a set of 2D keypoints or point trajectories into the 3D object shape and pose in each view assuming "low-rank" shape or deformation [2, 12, 19]. Recently, deep networks have been applied to learn such factorization of specific categories from 2D annotations [26, 34, 50]. Close to our approach, Neural Dense NRSfM (N-NRSfM) [44] learns a video-specific shape and deformation model from dense 2D point tracks. However, such methods are limited by the accuracy of 2D trajectory inputs, which is challenging to estimate in real-world sequences when large motion occurs [42, 44, 46].

3. Approach

Problem: Given a monocular video $\{I_t\}$ with an object of interest (indicated by a segmentation mask $\{S_t\}$), we tackle the nonrigid 3D shape and motion estimation problem, which includes estimating (1) $\mathbf{S}$: the rest shape of the object, (2) $\mathbf{D_t}$: the time-varying articulations as well as the object root body transformations, and (3) $\mathbf{K_t}$: the camera intrinsics.

Overview: Figure 2 illustrates the overview of our method. Motivated by recent progress in differentiable rendering and self-supervised shape learning [23, 32], we cast the nonrigid 3D shape and motion estimation problem as an analysis-by-synthesis task. Despite the under-constrained nature of this problem, we hypothesize that, by giving appropriate video measurements, a "low-rank" shape and motion can be solved up to an unknown scale. Model parameters $\mathbf{X} = \{\mathbf{S}, \mathbf{D_t}, \mathbf{K_t}\}$ are updated (via gradient descent) to minimize the difference between the rendered output $\mathbf{Y} = \mathbf{f}(\mathbf{X})$ and ground-truth video measurements $\mathbf{Y}^*$ at *test* time (Sec. 3.1). To deal with the fundamental ambiguities in object shape, deformation and camera motion, we seek (1) a "low-rank" but expressive parameterization of deformation (Sec. 3.2), (2) rich constraints provided by optical flow and raw pixels, and (3) appropriate regularization of object shape deformation and camera motion (Sec. 3.3).

3.1. Forward-synthesis model

We first introduce the forward synthesis model. Given a frame index t and model parameters $\mathbf{X}$, we synthesize the measurements of the corresponding frame pair $\{t, t+1\}$, including color images renderings $\{\hat{I}_t, \hat{I}_{t+1}\}$, object silhou-

ettes renderings $\{\hat{S}_t, \hat{S}_{t+1}\}$ and forward-backward optical flow renderings $\{\hat{\mathbf{u}}_t^+, \hat{\mathbf{u}}_{t+1}^-\}$.

Rendering pipeline: We represent object shape as a triangular mesh $\mathbf{S} = \{\bar{\mathbf{V}}, \bar{\mathbf{C}}, \mathbf{F}\}$ with vertices $\bar{\mathbf{V}} \in \mathbb{R}^{N \times 3}$, vertex colors $\bar{\mathbf{C}} \in \mathbb{R}^{N \times 3}$ and a fixed topology $\mathbf{F} \in \mathbb{R}^{M \times 3}$. To model time-varying articulations $\mathbf{D_t}$, we have

$$\mathbf{V}_t = \mathbf{G_{0,t}}(\bar{\mathbf{V}} + \Delta\mathbf{V_t}) \tag{1}$$

where $\Delta\mathbf{V_t}$ is a per-vertex motion field applied to the rest vertices $\bar{\mathbf{V}}$, and $\mathbf{G_{0,t}} = (\mathbf{R_0} \mid \mathbf{T_0})_\mathbf{t}$ is an object root body transformation matrix (index 0 is used to differentiate from bone transformations indexed from 1 in Sec. 3.2). Finally, we apply a perspective projection $\mathbf{K_t}$ before rasterization, where principal point (p_x, p_y) is assumed to be constant and focal length f_t varies over time to deal with zoom-in/out.

Shaders: We render object silhouette and color images with a differentiable renderer [32]. Color images are rendered given per-vertex appearance $\bar{\mathbf{C}}$ and constant ambient light. To synthesize the forward flow $\mathbf{u}_t^+$, we take surface positions $\mathbf{V}_t$ corresponding to each pixel in frame t, compute their locations $\mathbf{V}_{t+1}$ in the next frame, then take the difference of their projections:

$$\begin{pmatrix} u_{x,t}^+ \\ u_{y,t}^+ \end{pmatrix} = \begin{pmatrix} \mathbf{P}_t^{(1)}\mathbf{V}_t/\mathbf{P}_t^{(3)}\mathbf{V}_t \\ \mathbf{P}_t^{(2)}\mathbf{V}_t/\mathbf{P}_t^{(3)}\mathbf{V}_t \end{pmatrix} - \begin{pmatrix} \mathbf{P}_{t+1}^{(1)}\mathbf{V}_{t+1}/\mathbf{P}_{t+1}^{(3)}\mathbf{V}_{t+1} \\ \mathbf{P}_{t+1}^{(2)}\mathbf{V}_{t+1}/\mathbf{P}_{t+1}^{(3)}\mathbf{V}_{t+1} \end{pmatrix}, \tag{2}$$

where $\mathbf{P}^{(i)}$ denotes the ith row of the projection matrix $\mathbf{P}$.

3.2. Articulation Modeling

Unknowns vs constraints: Similar to NRSfM, we analyse the number of unknowns and constraints to solve the inverse problem. Given T frames of a video, we have

$$\#\,\text{Unknowns} = \underset{(\bar{\mathbf{V}})}{3N} + \underset{(\Delta\mathbf{V})}{3NT} + \underset{(\mathbf{R_0},\mathbf{T_0})}{6T} + \underset{(\mathbf{K})}{(T+2)}, \tag{3}$$

which grows linearly with the number of vertices. Motivated by NRSfM [12] that uses low-rank shape and motion basis to deal with the exploding solution space, we seek an expressive but low-rank representation of shape and motion.

Linear-blend skinning: Instead of modeling deformation as per-vertex motion $\Delta\mathbf{V_t}$ [18, 23, 30], we adopt a linear-blend skinning model (LBS) [27, 28] to constrain vertex motion by blending B rigid "bone" transformations $\{\mathbf{G_1}, \cdots, \mathbf{G_B}\}$, which reduces the number of parameters and makes optimization easier. Besides bone transformations, the LBS model defines a skinning weight matrix $\mathbf{W} \in \mathbb{R}^{B \times N}$ that attaches the vertices of rest shape vertices $\bar{\mathbf{V}}$ to the set of bones. Each vertex is transformed by linearly combining the weighted bone transformations in the object coordinate frame and then transformed to the camera coordinate frame,

$$\mathbf{V}_{i,t} = \mathbf{G_{0,t}}(\textstyle\sum_b \mathbf{W}_{b,i}\mathbf{G}_{b,t})\bar{\mathbf{V}}_i \tag{4}$$

where i is the vertex index, b is the bone index. Unlike A-CSM [27] that only learns articulation, we learn skinning weights and time-varying bone transformations jointly.
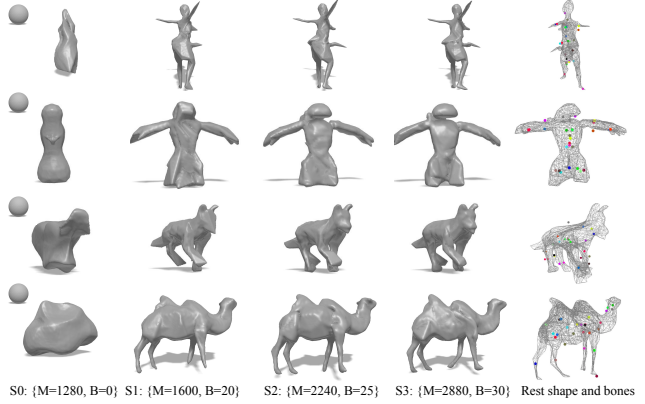

S0: {M=1280, B=0}   S1: {M=1600, B=20}   S2: {M=2240, B=25}   S3: {M=2880, B=30}   Rest shape and bones

Figure 3. Coarse-to-fine reconstruction from S0 to S3. The learned centers of 3D Gaussians are shown as colored dots.

Parametric skinning model: Inspired by the work on surface editing and local 3D shape learning [15, 45], we model the skinning weights as a mixture of Gaussians with B components. The probability of assigning vertex i to component b is given as

$$W_{b,i} = C e^{-\frac{1}{2}(\mathbf{v}_i - \mathbf{J}_b)^T \mathbf{Q}_b (\mathbf{v}_i - \mathbf{J}_b)}, \tag{5}$$

where $\mathbf{J}_b \in \mathbb{R}^3$ is the center of b-th Gaussian, $\mathbf{Q}_b$ is the corresponding precision matrix that determines the orientation and radius of a Gaussian, and C is a normalization factor that ensures the probabilities of assigning a vertex to different Gaussians sum up to one. $\mathbf{W} \rightarrow \{\mathbf{Q}, \mathbf{J}\}$ is optimized. Note that the mixture of Gaussian models not only reduces the number of parameters for skinning weights from NB to $9B$, but also guarantees smoothness, the benefits of which are empirically validated in our experiments (Tab. 4). The number of shape and motion parameters now becomes

$$\#\,\text{Unknowns} = \underset{(\bar{\mathbf{V}})}{3N} + \underset{(\mathbf{G_{1\ldots B}})}{6BT} + \underset{(\mathbf{J},\mathbf{Q})}{9B} + \underset{(\mathbf{R_0},\mathbf{T_0})}{6T} + \underset{(\mathbf{K})}{(T+2)}, \tag{6}$$

which grows linearly w.r.t. the number of frames and bones.

3.3. Self-supervised Learning from a Video

We exploit rich supervision signals from dense optical flow and raw pixels, as well as shape and motion regularizers to further constrain the problem.

Reconstruction Losses: The supervision for our analysis-by-synthesis pipeline includes silhouette loss, optical flow loss, texture loss, and perceptual loss. Given a pair of rendered outputs $(\hat{S}_t, \hat{I}_t, \hat{\mathbf{u}}_t)$ and measurements $(S_t, I_t, \mathbf{u}_t)$, the inverse graphics loss is computed as,

$$L_{\text{IG}} = \beta_1 ||\hat{S}_t^i - S_t||_2^2 + \beta_2 \sigma_t ||\hat{\mathbf{u}}_t^i - \mathbf{u}_t||_2 + \beta_3 ||\hat{I}_t^i - I_t||_1 \\ + \beta_4 pdist(\hat{I}_t, I_t) \tag{7}$$

where $\{\beta_1, \cdots, \beta_4\}$ are weights empirically chosen, σ_t is the normalized confidence map for flow measurement, and $pdist(\cdot, \cdot)$ is the perceptual distance [54] measured by an

AlexNet pretrained on ImageNet. Applying L2 norm loss to optical flow is empirically better than squared L2 loss, and we hypothesize the reason being that the former is more tolerant to outliers in the observed flow fields.

Shape and motion regularization: We exploit generic shape and temporal regularizers to constrain the problem. A Laplacian operator is applied to the rest mesh to enforce smooth surfaces,

$$L_{\text{shape}} = ||\bar{\mathbf{V}}_i - \frac{1}{|N_i|} \sum_{j \in N_i} \bar{\mathbf{V}}_j||^2. \qquad (8)$$

Motion regularization includes an ARAP (as-rigid-as-possible) deformation term and a least deformation term. The ARAP term encourages natural deformation [45, 48],

$$L_{\text{ARAP}} = \sum_{i=1}^{V} \sum_{j \in N_i} |\ ||\mathbf{V}_{i,t} - \mathbf{V}_{j,t}||_2 - ||\mathbf{V}_{i,t+1} - \mathbf{V}_{j,t+1}||_2\ |. \qquad (9)$$

The least deformation term encourages the deformation from the rest shape to be small [23],

$$L_{\text{least-motion}} = \sum_{i=1}^{V} ||\mathbf{V}_{i,t} - \bar{\mathbf{V}}_i||_2, \qquad (10)$$

which discourages arbitrarily large deformations and reduces ambiguities in joint object root body pose and articulation recovery.

Soft-symmetry constraints: To exploit the reflectional symmetry structure exhibited in common objects, we pose a soft-symmetry constraint along the symmetry plane $(\mathbf{n}^*, 0)$ at an arbitrary frame t^*. The symmetry plane is initialized from visual inspection and jointly optimized. We encourage the rest shape to be similar to its reflection,

$$L_{\text{symm-shape}} = L_{\text{cham}}(\{\bar{\mathbf{V}}, \mathbf{F}\}, \{\mathbf{H}\bar{\mathbf{V}}, \mathbf{F}\}) \qquad (11)$$

where $\mathbf{H} = \mathbf{I} - 2\mathbf{n}_*\mathbf{n}_*^T$ is the Householder reflection matrix, and the Chamfer distance (L_{cham}) is computed as bidirectional pixel-to-face distances. For the centers of Gaussian control points $\mathbf{J}$, we also have

$$L_{\text{symm-bone}} = L_{\text{cham}}(\bar{\mathbf{J}}, \mathbf{H}\bar{\mathbf{J}}). \qquad (12)$$

The total loss is a weighted sum of all losses with the weights empirically chosen and fixed for all experiments.

3.4. Implementation Details

Neural basis for time-varying parameters: Instead of optimizing explicit time-varying parameters $\{\mathbf{D_t}, \mathbf{K_t}\}$, we parameterize those as predictions from a convolutional network (ResNet-18 [21]) given an input image $\mathbf{I}_t$,

$$\psi_w(\mathbf{I}_t) = (\mathbf{K}, \mathbf{G}_0, \mathbf{G}_1, \mathbf{G}_2, \cdots, \mathbf{G}_B)_t, \qquad (13)$$

where one parameter is predicted for focal length, four parameters are predicted for each bone rotation parameterized by quaternion, and three numbers are predicted for each translation, adding to $1 + 7(B + 1)$ numbers in total at each frame. The weights are initialized with ImageNet [13] pretraining and then optimized by LASR for each test video.
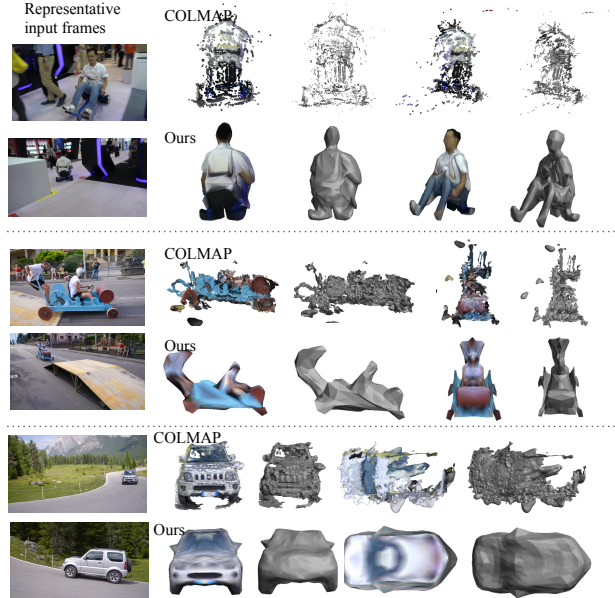

Figure 4. Visual comparison on near-rigid DAVIS sequences: scooter, soapbox, and car. COLMAP reconstructs only the visible rigid part, while our method faithfully reconstruct both the rigid object and near-rigid person.

Intuitively, the network learns a joint basis for cameras and poses that is empirically much easier to optimize than the raw parameters (Tab. 4).

Silhouette and flow measurements Our approach assumes that a reliable segmentation of the foreground object is given, which can be manually annotated [36], or estimated using instance segmentation and tracking methods [24, 55]. Our method requires reasonable optical flow estimation, which can be provided by state-of-the-art flow estimators [47, 52] trained on a mixture of datasets [1]. Notably, LASR recovers from some bad flow initialization and obtains better long-term correspondences (Tab. 2).

Coarse-to-fine reconstruction We adopt a coarse-to-fine strategy to reconstruct high-quality meshes inspired by Point2Mesh [20]. **S0**: We first assume a rigid object and optimize the rest shape and cameras $\{\mathbf{S}, \mathbf{G_{0,t}}, \mathbf{K_t}\}$ for 20 epochs. The rest shape is initialized from a subdivided icosahedron projected onto a sphere. **S1-S3:** We perform iso-surface extraction and re-meshing [22] to fix mesh self-intersections and long edges. After remeshing, the number of vertices and the number of bones increase, as shown in Fig. 3. The centers of Gaussian control points are initialized by running K-means on the vertices coordinates. We then jointly optimize all parameters $\{\mathbf{S}, \mathbf{D_t}, \mathbf{K_t}\}$ for 10 epochs. The above procedure is repeated three times (S1-S3).

4. Experiments

Setup: Due to the difficulty of obtaining 3D ground truth for nonrigid objects in the real world, we evaluate 2D key-
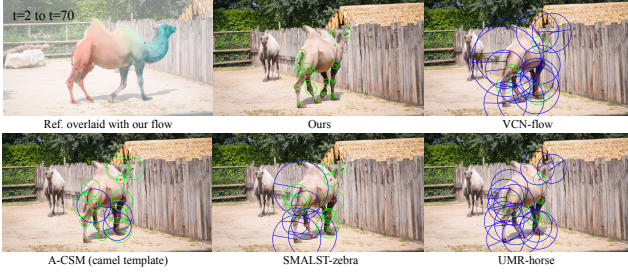
Figure 5. Keypoint transfer between frame 2 and frame 70 of the camel video. The distance between tranferred keypoint and target annotation is represented by the radius of circles. A correct transfer is marked with green and a wrong transfer is marked with blue. Our method transfers keypoint more accurately than baselines.

point transfer accuracy as a proxy of 3D reconstruction quality on real videos. We additionally evaluate 3D reconstruction accuracy on objects with ground-truth meshes.

4.1. 2D Keypoint Transfer on Animal Videos

Dataset: We test our method on an animal video dataset, **BADJA** [6], which provides nine real animal videos with 2D keypoint and mask annotations, derived from the DAVIS video segmentation dataset [36] and online stock footage. It includes three videos of dogs, two videos of horsejump, and one video of camel, cow, bear as well as impala. We report quantitative results on one video per-category and show the reconstruction of the rest in the sup. mat.

Metric: To approximate the accuracy of 3D shape and articulation recovery, we adopt **percentage of correct keypoint transfer** (PCK-T) [23, 27, 53] metric. Given a reference and target image pair with 2D keypoint annotations, the reference keypoint is transferred to the target image, and labeled as "correct" if the transferred keypoint is within some threshold distance $d_{th} = 0.2\sqrt{|S|}$ from the target keypoint, where $|S|$ is the area of the ground-truth silhouette [6]. In practice, we transfer points by re-projection from the reference frame to the target frame given the articulated shape and camera pose estimations. If the back-projected keypoint lies outside the reconstructed mesh, we re-project its nearest neighbor that intersects the mesh. The accuracy is averaged over all T(T-1) pairs of frames.

Baselines: We compare with state-of-the-art methods for animal reconstruction and refer to Tab. 1 for a taxonomy. **SMALST** [56] is a model-based regressor trained for zebras. It takes an image as input and predicts shape, pose and texture for the SMAL [58] model. **UMR** [30] is a category-specific shape estimator trained for several categories, including birds, horses and other categories that have a large collection of annotated images. We report the performance of the horse model since the models of other animal categories are not available. **A-CSM** [27] learns a category-specific canonical surface mapping and articulations from an image collection. At test time, it takes an image as input

Table 2. 2D Keypoint transfer accuracy on BADJA. [1]Model-based regression. [2]Category-specific reconstruction. [3]Free-form reconstruction. Methods with† do not reconstruct 3D shape. Results with* indicates the method is not designed for such category. Best results are underlined, and bolded if reconstruct a 3D shape.

Method	camel	dog	cows	horse	bear
[1]SMALST [56]	49.7*	12.8*	59.7*	10.4*	67.2*
[2]A-CSM [27]	60.2*	24.5*	65.7*	21.5	39.7*
[2]UMR [30]	35.1*	38.5*	68.1*	32.4	56.9*
[3]N-NRSfM [44]	67.8	17.9	70.0	8.7	60.2
[3]LASR (Ours)	**81.9**	**65.8**	**83.7**	**49.3**	**85.1**
[3] +Auto-mask	78.9	59.5	82.7	42.2	82.6
†Static	51.9	13.0	55.5	8.8	58.6
†Detector [6]	73.3	66.9	89.2	26.5	83.1
†OJA [6]	87.1	66.9	94.7	24.4	88.9
†Flow-VCN [52]	47.9	25.7	60.7	14.4	63.8

and predicts the articulation parameters of a rigged template mesh. It provides 3D templates for 27 animal categories and an articulation model for horses, which is used throughout the experiments. **SMALify** [6] is a model-based optimization approach that fits one of five categories (including cat, dog, horse, cow and hippo) of SMAL models to a video or a single image. We provide all the video frames with ground-truth keypoint and mask annotations. Close to our setup, **N-NRSfM** [44] trains a video-specific model for object shape, deformation and camera parameters from multi-frame optical flow estimations [14]. Finally, we include a detection-based method, **OJA** [6], which trains an hourglass network to detect animal keypoints (indicated by Detector), and post-process the joint cost maps with a proposed optimal assignment algorithm. The results of PCK are taken from the paper [6] without recomputing PCK-T.

Results: Qualitative results of 3D shape reconstruction are shown in Fig. 1 and Fig. 6, where we compare with UMR, A-CSM and SMALify on the camel, bear and dog video. Quantitative results of keypoint transfer are shown in Tab. 2. Given that all 3D reconstruction baselines are category-specific and might not provide the exact model for some categories (such as camel), we pick up the best model or template for each animal video. Compared with 3D reconstruction baselines, LASR is better for all categories, even on the categories the baselines are trained for (e.g., LASR: 49.3 vs UMR: 32.4 on horsejump-high). Replacing the GT masks with an object segmentor, PointRend [24], the performance of LASR ('+Auto-mask' in Tab. 2) drops, but is still better than all the reconstruction baselines. Compared to detection-based methods, our accuracy is higher on the horsejump video, and close to the baseline on other videos. LASR also shows a large improvement compared to the initial optical flow (81.9% vs 47.9% for camel), especially between long-range frames as shown in Fig. 5.
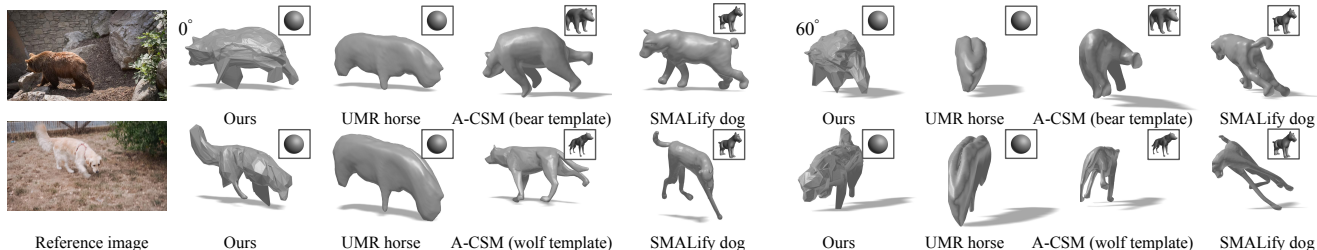
Figure 6. Comparison on BADJA bear and dog videos. The reconstruction of the first frame of the video is shown from two viewpoints. Compared to UMR that also does not use a shape template, LASR reconstructs more fine-grained geometry. Compared to A-CSM and SMALify that use a shape template, LASR recovers instance-specific details, such as the fluffy tail of the dog, and a more natural pose.

Table 3. Mesh reconstruction error in terms of Chamfer distance on our animated object dataset. To ensure comparable results over objects, ground-truth shapes are rescaled such that the maximum distance between vertices is 10. Best results are bolded. "-" means a method does not apply to a particular sequence.

Method	dancer ↓	dog ↓	horse ↓	golem ↓
SMPLify-X [35]	0.26	-	-	-
VIBE [25]	**0.22**	-	-	-
A-CSM [27]	-	0.38	0.26	-
SMALify [6]	-	0.51	0.41	-
PIFuHD [41]	0.28	-	-	-
UMR [30]	-	0.44	0.42	-
LASR (Ours)	0.35	**0.28**	**0.23**	0.16

4.2. Mesh Reconstruction on Articulated Objects

Dataset: To evaluate mesh reconstruction accuracy, we collect a video dataset of five articulated objects with ground-truth mesh and articulation, including one dancer video from AMA (Articulated Mesh Animation dataset) [49], one German shepherd video, one horse video, one eagle video and one stone golem video from TurboSquid. We also include a rigid object, Keenan's spot to evaluate performance on rigid object reconstruction and ablation for the S0 stage.
Metric: Most prior work on mesh reconstruction assumes given camera parameters. However, both the camera and the geometry are unknown in our case, which leads to ambiguities in evaluation, including scale ambiguity (exists for all monocular reconstruction) as well as the depth ambiguity (exists for weak perspective cameras as used in UMR, A-CSM, VIBE, etc.). To factorize out the unknown camera matrices, we align two meshes with a 3D similarity transformation solved by ICP. Then, the bidirectional Chamfer distance is adopted as the evaluation metric. We follow prior work [16, 37] to randomly sample 10k points uniformly from the surface of predicted and ground-truth meshes, and compute the average distance between the nearest neighbor for each point in the corresponding point cloud.
Baselines: Besides A-CSM, SMALify, and UMR for animal reconstruction, we compare with SMPLify-X, VIBE, and PiFUHD for human reconstruction. **SMPLify-X** [35]

Table 4. Ablation study with mesh reconstruction error.

S0	ref.	[(1)]w/o flow	[(2)]w/o L_{can}	[(3)]w/o CNN
spot	0.05	0.55	0.61	0.63
S0-S3	ref.	[(4)]w/o LBS	[(5)]w/o C2F	[(6)]w/o GMM
dog	0.28	0.68	0.59	0.34

is a model-based optimization method for expressive human body capture. We use the female SMPL model for the dancer sequence, and provide the keypoint inputs estimated from OpenPose [9]. **VIBE** [25] is a state-of-the-art model-based video regressor for human pose and shape inference. **PIFuHD** is a state-of-the-art free-form 3D shape estimator for clothed humans. It takes a single image as input and predicts an implicit shape representation, which is converted to a mesh by the marching cube algorithm. To compare with SMALify on dog and horse, we manually annotate 18 keypoints per-frame, and initialize with the corresponding shape template.

Results The visual comparison on human and animals are shown in Fig. 1 and Fig. 7 respectively. We report the quantitative results in Tab. 3. On the dog video, our method is better than all the baselines (0.28 vs A-CSM: 0.38), possibly because A-CSM and UMR are not trained specifically for dogs (although A-CSM uses a wolf template), and SMALify cannot reconstruct a natural 3D shape from limited keypoint and silhouette annotations. For the horse video, our method is slightly better than A-CSM, which uses a horse shape template, and outperforms other baselines. For the dancer sequence, our method is not as accurate as baseline methods (0.35 vs VIBE: 0.22), which is expected given that all baselines either use a well-designed human model, or have been trained with 3D human mesh data, while LASR does not have access to 3D human data. For the stone golem video, our method is the only one that reconstructs a meaningful shape. Although the stone golem has a similar shape to a human's, OpenPose does not detect joints correctly, leading to the failure of SMALify-X, VIBE and PiFUHD.

Qualitative results on DAVIS videos: To examine the performance on arbitrary real-world objects, we use five
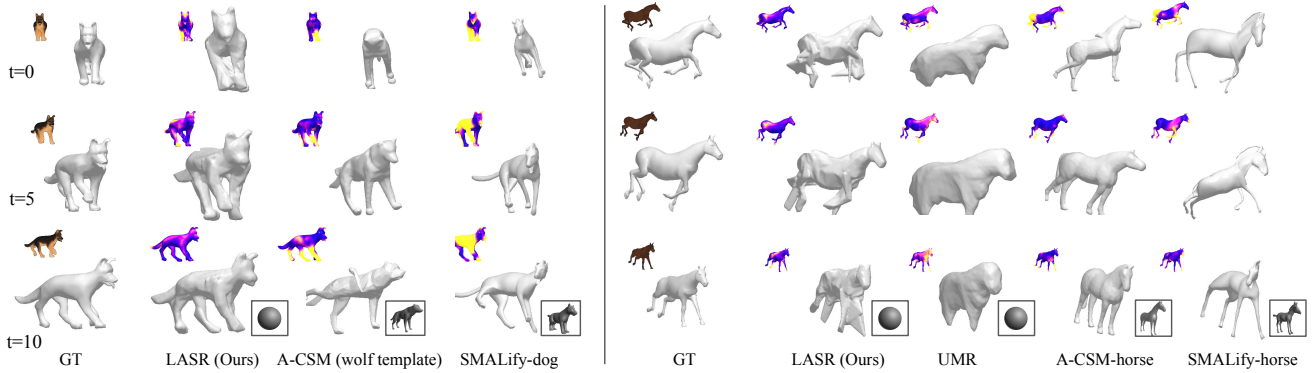
Figure 7. Shape and articulation reconstruction results on synthetic dog and horse sequences. We also visualize Chamfer distances measured from the ground truth to the reconstructed mesh on top of each result, and yellow indicates high error. Compared to UMR, LASR successfully reconstructs the four legs of the horse. Compared to template-based methods (A-CSM and SMALify), LASR successfully reconstructs the instance-specific details (ears and tails of the dog) and recovers a more natural articulation. The reference is shown at the left corner and the template mesh used is shown in the bottom right boxes.
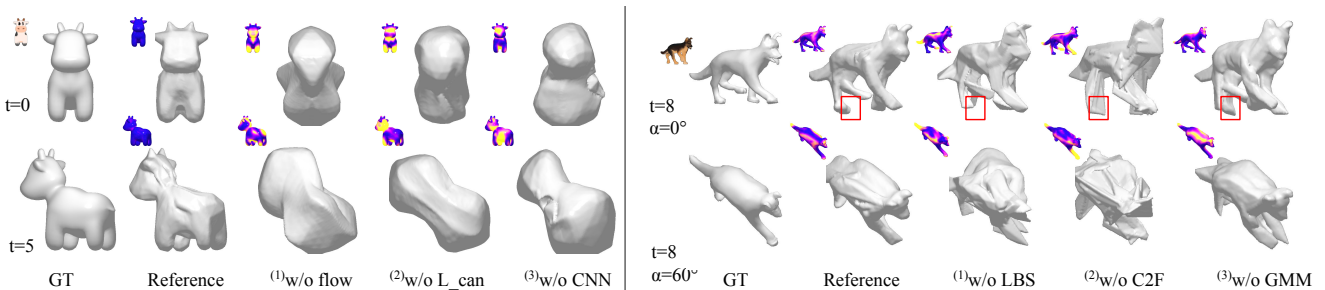

Figure 8. Left: Ablation study on camera and rigid shape optimization (**S0**). Removing the optical flow loss introduces large errors in camera pose estimation and therefore the overall geometry is not recovered. Removing the canonicalization loss leads to worse camera pose estimation, and therefore the symmetric shape constraint is not correctly enforced. Finally, if we directly optimize the camera poses without using a convolutional network, it converges much slower and does not yield an ideal shape within the same iterations. Right: Ablation study on articulated shape optimization (**S1-S3**). We show the reconstructed articulated shape at the middle frame (t=8) from two viewpoints. Without LBS model, although the reconstruction looks plausible from the visible view, it does not recover the full geometry due to the redundant deformation parameters and lack of constraints. Without coarse-to-fine re-meshing, fine-grained details are not recovered. Replacing GMM skinning weights (9xB parameters) with an NxB matrix leads to extra limbs and tails on the reconstruction.

DAVIS videos, including dance-twirl, scooter-board, soap-box, car-turn, mallard-fly, and a cat video captured by us and segmented by PointRend. The comparison with COLMAP [43], a template-free SfM-MVS pipeline, is shown in Fig. 4. More results are available in the sup. mat.

Ablation study: We investigate the effect of different design choices on the rigid "spot" and animated dog sequences. The videos are rendered into T=15 frames given ambient light and a camera rotating around the object by 90 degrees at zero elevation. Besides color images, we render silhouette and optical flow as the supervision. Results are shown in Fig. 8 and quantitative results are reported in Tab. 4. In terms of camera parameter optimization and rigid shape reconstruction (**S0**), we find it beneficial to use [1] optical flow as supervision signals, [2] canonicalization of symmetry plane, and [3] CNN as an implicit representation for camera parameters. For articulated shape reconstruction (**S1-S3**), it is critical to use [4] linear blend skin-ning, [5] coarse-to-fine re-meshing, and [6] parametric skinning model.

Limitations: Empirically, LASR struggles to estimate surfaces that are not visible in any input view and fails at heavy occlusions that are missed by mask annotations. Its efficiency also needs improvement, as it takes less than one hour for rigid objects and a few hours for nonrigid shapes on a single GPU.

5. Conclusion

We present LASR, a template-free approach for articulated shape reconstruction from a monocular video. LASR faithfully reconstructs individual objects from diverse categories (such as human, camel, dog, bear, etc.) without relying on category-specific shape templates, making it applicable to a wide range of scenarios. We hope that LASR will enable more progress in articulated shape reconstruction.

References

[1] Robust vision challenge. www.robustvision.net, 2020. 5

[2] Antonio Agudo, Melcior Pijoan, and Francesc Moreno-Noguer. Image collection pop-up: 3d reconstruction and clustering of rigid and non-rigid categories. In *CVPR*, pages 2607–2615, 2018. 2, 3

[3] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd Pfrommer, Marc Schmidt, and Kostas Daniilidis. 3D bird reconstruction: a dataset, model, and shape recovery from a single view. In *ECCV*, 2020. 2

[4] Simon Baker, Takeo Kanade, et al. Shape-from-silhouette across time part i: Theory and algorithms. *IJCV*, 62(3):221–247, 2005. 2

[5] Benjamin Biggs, Ollie Boyne, James Charles, Andrew Fitzgibbon, and Roberto Cipolla. Who left the dogs out: 3D animal reconstruction with expectation maximization in the loop. In *ECCV*, 2020. 2

[6] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures great and smal: Recovering the shape and motion of animals from video. In *ACCV*, pages 3–19. Springer, 2018. 1, 2, 6, 7

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 2

[8] Christoph Bregler, Aaron Hertzmann, and Henning Biermann. Recovering non-rigid 3d shape from image streams. In *CVPR*, volume 2, pages 690–696. IEEE, 2000. 2

[9] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 7

[10] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *PAMI*, 35(1):232–244, 2012. 2

[11] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. In *NeurIPS*. 2016. 2

[12] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *IJCV*, 107(2):101–122, 2014. 2, 3, 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 5

[14] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. A variational approach to video registration with subspace constraints. *IJCV*, 2013. 6

[15] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, pages 4857–4866, 2020. 4

[16] Justin Johnson Georgia Gkioxari, Jitendra Malik. Mesh r-cnn. *ICCV*, 2019. 7

[17] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, pages 9785–9795, 2019. 2

[18] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoints without keypoints. In *ECCV*, 2020. 2, 3, 4

[19] Paulo FU Gotardo and Aleix M Martinez. Non-rigid structure from motion with complementary rank-3 spaces. In *CVPR*, pages 3065–3072. IEEE, 2011. 2, 3

[20] Rana Hanocka, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *arXiv preprint arXiv:2005.11084*, 2020. 5

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[22] Jingwei Huang, Hao Su, and Leonidas Guibas. Robust watertight manifold surface generation method for shapenet models. *arXiv preprint arXiv:1802.01698*, 2018. 5

[23] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2, 3, 4, 5, 6

[24] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, pages 9799–9808, 2020. 5, 6

[25] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, June 2020. 2, 7

[26] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *ICCV*, pages 1558–1567, 2019. 2, 3

[27] Nilesh Kulkarni, Abhinav Gupta, David F Fouhey, and Shubham Tulsiani. Articulation-aware canonical surface mapping. In *CVPR*, pages 452–461, 2020. 2, 3, 4, 6, 7

[28] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000. 4

[29] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *NeurIPS*, 2020. 2, 3

[30] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. *ECCV*, 2020. 2, 3, 4, 6, 7

[31] 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 *ECCV*, pages 740–755. Springer, 2014. 1

[32] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *ICCV*, Oct 2019. 3, 4

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 1, 2

[34] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *ICCV*, pages 7688–7697, 2019. 3

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1, 2, 7

[36] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 5, 6

[37] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 7

[38] Bernhard Reinert, Tobias Ritschel, and Hans-Peter Seidel. Animated 3d creatures from single-view video by skeletal sketching. In *Proceedings of Graphics Interface 2016*, GI 2016, pages 133–141. Canadian Human-Computer Communications Society / Société canadienne du dialogue humain-machine, 2016. 2, 3

[39] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. In *NeurIPS*, 2018. 2

[40] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, pages 2304–2314, 2019. 3

[41] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2, 3, 7

[42] Peter Sand and Seth Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72, 2008. 2, 3

[43] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *ACCV*, 2016. 8

[44] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural dense non-rigid structure from motion with latent space constraints. In *ECCV*, pages 204–222. Springer, 2020. 2, 3, 6

[45] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM SIGGRAPH 2007 papers*, pages 80–es. 2007. 4, 5

[46] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, pages 438–451. Springer, 2010. 2, 3

[47] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. 2020. 2, 5

[48] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections, 2020. 2, 3, 5

[49] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *SIGGRAPH 2008*, pages 1–9. 2008. 7

[50] Chaoyang Wang, Chen-Hsuan Lin, and Simon Lucey. Deep nrsfm++: Towards 3d reconstruction in the wild. *arXiv preprint arXiv:2001.10090*, 2020. 3

[51] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019. 2

[52] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. In *NeurIPS*, pages 794–805, 2019. 2, 5, 6

[53] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392. IEEE, 2011. 6

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 4

[55] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, pages 13066–13073, 2020. 5

[56] Silvia Zuffi, Angjoo Kanazawa, Tanya Berger-Wolf, and Michael Black. Three-d safari: Learning to estimate zebra pose, shape, and texture from images "in the wild". In *ICCV*, pages 5358–5367. IEEE. 2, 6

[57] Silvia Zuffi, Angjoo Kanazawa, and Michael J. Black. Lions and tigers and bears: Capturing non-rigid, 3D, articulated shape from images. In *CVPR*, 2018. 2

[58] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, July 2017. 1, 2, 6