CS598 Research Survey - Articulated Shape

Shaowei Liu
shaowei3@illinois.edu

Hao Zhang
haoz19@illinois.edu

Abstract

Articulated shapes, characterized by their ability to undergo complex deformations and motions through joint rotations, play a pivotal role in various domains of computer vision and graphics. This survey provides a comprehensive overview of modeling, techniques and applications for articulated shapes, focusing on four key categories: articulated objects, hands, humans and animals.

1. Articulated Object

1.1. Introduction

Articulated objects consist of rigid bodies, connected through joints that allow motion between them. An articulated object can be described as a kinematic chain, where each link is connected to another via specific types of joints. Two primary joint types used: the *revolute joint* and the *prismatic joint*. Daily articulated objects are shown in 1.

A *revolute joint*, also known as a hinge joint, allows rotation between two connected links around a single fixed axis. The motion of a revolute joint can be described by an angular parameter, denoted as θ, which represents the angle of rotation about the axis. Given a rotation matrix $R(\theta)$ and a translation vector $\mathbf{t}$, the transformation of the link is:

$$T(\theta) = \begin{bmatrix} R(\theta) & \mathbf{t} \\ 0 & 1 \end{bmatrix}$$

where $R(\theta)$ is a 3×3 rotation matrix describing the orientation of the link, and $\mathbf{t}$ is a 3×1 vector representing the position of the link's origin.

A *prismatic joint* allows linear motion between two connected links along a specified axis. The position of the moving link is determined by a displacement parameter, denoted as d, which represents the distance traveled along the axis. If the translation along the joint's axis is $\mathbf{d}$, the transformation matrix is:

$$T(d) = \begin{bmatrix} I_3 & \mathbf{d}(d) \\ 0 & 1 \end{bmatrix}$$

Figure 1. Articulated objects in daily life.

where I_3 is the 3×3 identity matrix, and $\mathbf{d}(d)$ is the 3×1 displacement vector as a function of the linear displacement parameter d.

An articulated object can be modeled as a kinematic chain composed of multiple links connected by a combination of revolute and prismatic joints. The number of independent parameters required to specify the configuration of the object is referred to as the *degrees of freedom* (DOF). For example, a robotic arm with n revolute joints will have n degrees of freedom, corresponding to the n angles of rotation.

The position and orientation of the end-effector (the final link in the chain) can be computed by chaining the transformations of each link along the kinematic chain:

$$T_{\text{end-effector}} = T_1(\theta_1)T_2(\theta_2)\dots T_n(\theta_n)$$

where each $T_i(\theta_i)$ or $T_i(d_i)$ represents the transformation for the i-th joint.

1.2. Modeling

Most work tackle the problem of inferring articulable models in a category-agnostic manner by specifying number of parts or given kinematics [1, 2, 10, 11, 16, 18, 24, 29, 33,

41, 42, 44, 48, 49, 53]. Here we introduce 3 works support category-agnostic articualted object modeling.

MultiBodySync MultiBodySync [14] paper presents a framework for for jointly solving multi-body segmentation and motion estimation from 3D scans using a synchronization framework. The core of the method is the alignment and synchronization of multiple 3D scans using iterative optimization. However, the work do not infer the kinematic tree connecting these parts together and parts can undergo 6DOF transformation relative to one another.

Watch It Move Watch-It-Move [34] propose an unsupervised method to discover 3D joints in articulated objects. They explicitly represent each object's part as an ellipsoid, and use MLP to output color and a residual SDF to perform volume rendering as supervision from multi-view video of an articulated object. However, their method doesn't incorporate the articulated object joint constraint (revolute or prismatic) in modeling, thus the final motions could lead to unrealistic motion and wrong kinematic tree.

Reart Reart [27] formulate the problem of articulable object modeling from 4D point cloud as an energy minimization problem. The optimization is divided into a relaxation stage that reasons about a 6-DOF piece-wise rigid model without kinematic constraints and a projection stage that casts the solution to a valid kinematic tree (all joints satisfy 1-DoF constraints). The method outputs an animatable 3D model which can be retargeted to novel poses.

1.3. Evaluation

The evaluation usually includes the part segmentation evalution (Iou and Random Index [5]), motion reconstruction error, and retarget error of novel poses. For kinematic tree evaluation, [27] uses Tree Edit Distance [55] that measures the similarity between predicted kinematic tree and the ground truth kinematic tree.

1.4. Future work

In [27], they uses 4D point cloud sequences as input. However, in real-world RGBD images are much easier to capture. It will be interesting to replace the input with mutliview RGB images as input to build the animatable articulated model from a single or few photos recording different states of an object.

2. Animal Shape

2.1. Introduction

Animal shape modeling involves capturing the complex articulated structures and deformations of various species, which are represented as kinematic chains of segments and joints. This approach has applications in animation, virtual and augmented reality, and other fields requiring realistic animal representations. In this context, articulated models must represent flexible joints, body parts, and textures accurately. Figure 2 illustrates examples of animal shape representations used in practical applications.

2.2. Introduction

Numerous studies take a category-agnostic approach to modeling animal shapes, enabling versatility across diverse species by specifying the structure and kinematic constraints. Here, we discuss several recent advancements in category-agnostic animal shape modeling.

CASA [47] is a category-agnostic method that reconstructs skeletal shapes from monocular videos without prior templates. It retrieves a rough skeletal template for animals using pretrained language-vision models and refines joint angles and skinning weights through optimization. CASA demonstrates robust skeletal animation capabilities across various species, handling occlusions and diverse poses effectively.

BANMo [50] constructs high-fidelity animatable 3D models from multiple casual videos of deformable objects. Integrating canonical embeddings, neural radiance fields, and neural blend skinning, BANMo avoids pre-built templates, enabling pose-based articulations from real-world data. This approach supports a wide range of animal and human models and can render realistic images from novel viewpoints.

MagicPony [46] predicts 3D articulated shapes from single images using an implicit-explicit hybrid model. By leveraging a neural signed distance function and DINO-ViT features for viewpoint disambiguation, this method can produce articulated 3D shapes with minimal supervision, even capturing complex details in artistic representations. MagicPony is effective for animals with challenging pose variations and generalizes well across different visual styles.

3D-Fauna [25] develops a comprehensive deformable model for over 100 animal species using only 2D Internet images, without the need for predefined templates. Introducing the Semantic Bank of Skinned Models (SBSM), it generalizes across species, capturing intricate details even in rare animals with limited data. The model supports single-image 3D mesh reconstruction, making it suitable for animation and rendering.

Hi-LASSIE [52] extends 3D part discovery to animals using sparse in-the-wild images, supporting high-quality skeletal animation from minimal data. It achieves this through 3D skeleton estimation and optimization without predefined templates, making it an effective choice for diverse animal shapes and motions.

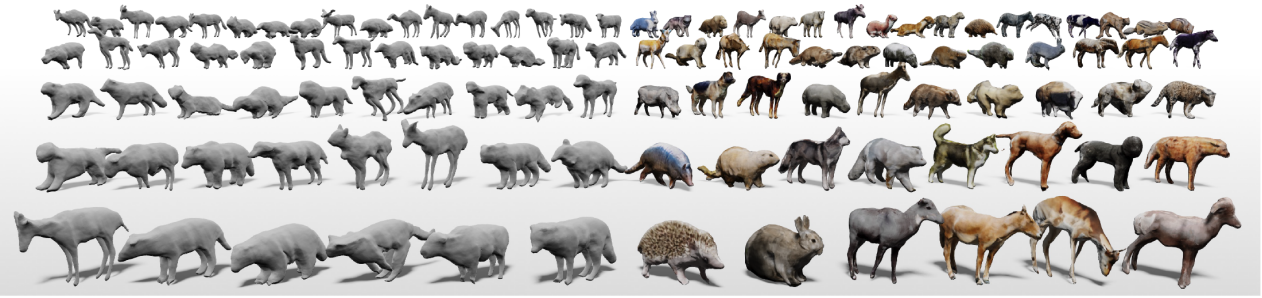

Figure 2. Examples of animal shape representations.

2.3. Evaluation

Evaluation of animal shape models primarily includes the following metrics:

- **Keypoint Transfer Accuracy**: Measures the accuracy of transferring keypoints from one instance to another within the model, evaluating the consistency of predicted joint and key positions across different poses.
- **Mean Intersection over Union (mIoU)**: Assesses part segmentation accuracy by comparing predicted and ground truth part masks, indicating the degree of overlap.
- **Chamfer Distance**: Used to evaluate shape fidelity by measuring the average distance between predicted and ground truth 3D points, reflecting the precision of shape reconstruction.

These metrics collectively offer a comprehensive assessment of the model's ability to accurately capture the shape, structure, and articulation of animal forms, supporting quantitative evaluation across various species and poses.

3. Human Shape

3.1. Introduction

Human shape modeling aims to accurately represent human body structure, pose, and surface details, which are essential for applications in animation, virtual and augmented reality, and medical simulations. These models often incorporate complex articulation and surface deformation, relying on advanced representations that capture body shape across various poses and viewpoints.

3.2. Introduction

Recent work on human shape modeling has focused on different approaches to balance detail, efficiency, and flexibility. Below, we discuss several significant advancements.

SMPL SMPL (Skinned Multi-Person Linear Model) [31] is a widely-used parametric model that represents human shape as a mesh deformed by pose and identity blend shapes. SMPL achieves compatibility with industry-standard rendering engines, making it ideal for applications requiring realistic animations and low memory usage. It offers efficient linear blend skinning, allowing realistic body deformation while preserving compatibility with graphics pipelines.

VIBE VIBE (Video Inference for Body Pose and Shape Estimation) [22] integrates motion capture data for training and leverages temporal networks for producing kinematically plausible human motions from monocular videos. Using adversarial learning, VIBE enhances the realism of motion sequences in various settings, making it suitable for dynamic video-based applications. It achieves state-of-the-art results on several 3D pose estimation benchmarks.

GART GART (Gaussian Articulated Template Model) [23] introduces an efficient model using a mixture of 3D Gaussians to represent articulated human shapes from monocular video. It combines the flexibility of Gaussian splatting with a skeleton-based template for real-time rendering, making it useful for applications requiring both accuracy and speed. The model further introduces latent bones to capture additional details, such as clothing dynamics.

GoMAvatar GoMAvatar [43] employs a Gaussians-on-Mesh approach to create high-fidelity human avatars. It uses Gaussian splatting for real-time rendering and supports complex articulations by integrating deformable meshes. The method is memory-efficient and achieves high frame rates, making it suitable for real-time applications in VR/AR and simulation.

Humans in 4D Humans in 4D [8] leverages a fully transformer-based network for tracking and reconstructing human meshes from video. By employing a transformer backbone, it improves upon previous methods in handling challenging poses and occlusions, achieving robust tracking across temporal sequences. This approach enhances performance in video-based applications, including action recognition.

GaussianAvatar GaussianAvatar [13] models human shape using dynamic 3D Gaussians for animatable avatars. It combines explicit representations with a dynamic appearance network, supporting high-fidelity avatar modeling from monocular videos. The joint optimization of motion and appearance significantly improves the accuracy of dynamic features, such as clothing wrinkles, enhancing visual realism in the final output.

3.3. Evaluation

Evaluation of human shape models includes several metrics to assess different aspects of model performance. The primary metrics are as follows:

- **Keypoint Transfer Accuracy**: Evaluates the accuracy of transferring joint positions across frames, providing insight into motion consistency.
- **Mean Intersection over Union (mIoU)**: Measures part segmentation accuracy by comparing predicted and ground truth segmentation masks, reflecting how well the model captures part boundaries.
- **Chamfer Distance**: Quantifies shape fidelity by calculating the average distance between reconstructed and ground truth 3D points, assessing the precision of shape reconstruction.
- **Peak Signal-to-Noise Ratio (PSNR)**: Used to assess reconstruction quality by measuring the similarity between the reconstructed and ground truth images, with higher values indicating better quality.
- **Structural Similarity Index Measure (SSIM)**: Measures perceived structural similarity between the reconstructed and reference images, focusing on luminance, contrast, and structural information.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Assesses perceptual similarity by comparing feature embeddings of image patches, with lower values indicating higher perceptual similarity.

These metrics are applied on several benchmark datasets, including *ZJU-MoCap*, *PeopleSnapshot*, and *3DPW*, to provide a comprehensive evaluation of the model's ability to capture shape, articulation, and texture consistency across various poses and lighting conditions.

3.4. Future Work

Future directions in human shape modeling may involve integrating RGBD sequences or expanding to more generalizable models that work robustly with fewer input sources. This shift could improve model reliability for in-the-wild applications, including real-time applications, while maintaining high fidelity in both shape and motion details.

4. Articulated hand

Similar to SMPL [30], the parametric hand model named MANO [37] built upon the SMPL body model. The MANO hand model represents the 3D hand as a mesh with a fixed topology, where the vertices of the mesh, $\mathbf{V} \in \mathbb{R}^{N\times 3}$, are deformed according to both pose and shape parameters. The shape of the hand is controlled by a low-dimensional shape vector $\boldsymbol{\beta} \in \mathbb{R}^{|\beta|}$. This vector describes the variations in hand shape across individuals. The articulation of the hand is controlled by a pose vector $\boldsymbol{\theta} \in \mathbb{R}^{|\theta|}$, where $\boldsymbol{\theta}$ represents the joint angles for the 16 joints in the hand. The deformation of the hand mesh is expressed as a learned pose blend shape $\mathbf{B}_P$: $\mathbf{V}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\beta}) + \mathbf{B}_P(\boldsymbol{\theta})$. The final position of each vertex is determined using linear blend skinning (LBS). each vertex is linearly blended as:

$$\mathbf{V}' = \sum_{j=1}^{K} w_j(\mathbf{J}_j(\boldsymbol{\theta})\mathbf{V})$$

where w_j are the skinning weights for each joint, and $\mathbf{J}_j(\boldsymbol{\theta})$ is the transformation matrix for joint j based on the pose parameters.

4.1. Modeling

Parametric hand model couldn't capture fine-grained hand details especially when hand is interacting with objects. Recent works addresses the issue by introducing implicit functions based on the explicit mesh representation.

Grasping Field Grasping Field [19] represents the hand through an implicit function learned via neural networks, where each point in space can be queried to determine whether it belongs to the hand or not. This field is learned using a neural network, which takes as input the position of a point in 3D space, $\mathbf{x} \in \mathbb{R}^3$, and outputs the likelihood of that point being on the surface of the hand or inside the hand volume.

HALO HALO [20] introduces a skeleton-driven neural occupancy representation for modeling articulated hands, leveraging the hand's kinematic skeleton to condition the neural network, allowing it to deform the implicit field based on the hand's articulation and shape. The joint angles $\boldsymbol{\theta}$ of skeleton control the deformation of the occupancy field, ensuring that the implicit shape of the hand is articulated according to the skeleton's structure.

4.2. Applications

Hand modeling has lots of application in animation, games, and augmented and virtual reality [12, 15, 40, 45]. We show two most important application in hand-object interaction and grasp generation.

Hand-object interaction Hand-object interaction refers to jointly model 3D hand and object [6, 9, 26, 39, 51]. [9, 26] introduces 3D hand reconstruction and 6D object estimation from monocular image input using the MANO

model, [5, 39, 51] further model their contact relations given the 3D hand and object. By considering the contact information, we could get a better reconstruction of hand in occlusion. [35] is the latest work generalize MANO hand pose estimation to in-the-wild settings.

4.3. Grasp generation

Another interesting application of hand is grasp generation [4, 17, 28, 38]. Given an object, we need to generate a proper hand grasping the object to achieve goals. This has gain extensive attention across both robotic hand manipulation [4, 32], animation [3, 7], digital human synthesis [54], and physical motion control [21, 36].

4.4. Future work

[28, 38] have proposed to use C-VAE to model the grasp generation, it will be interesting to generalize the model to be in-the-wild and applicable to arbitrary objects. Adopt diffusion for grasp generation and modeling hand-hand interaction could also be interesting future directions.

References

[1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *CoRL*, 2019. 1

[2] Hameed Abdul-Rashid, Miles Freeman, Ben Abbatematteo, George Konidaris, and Daniel Ritchie. Learning to infer kinematic hierarchies for novel object instances. In *ICRA*. IEEE, 2022. 1

[3] Christoph W Borst and Arun P Indugula. Realistic virtual grasping. In *IEEE Virtual Reality 2005*, pages 91–98. IEEE, 2005. 5

[4] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, pages 2386–2393. IEEE, 2019. 5

[5] Xiaobai Chen, Aleksey Golovinskiy, and Thomas Funkhouser. A benchmark for 3d mesh segmentation. *TOG*, 28(3), 2009. 2, 5

[6] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, pages 231–248. Springer, 2022. 4

[7] George ElKoura and Karan Singh. Handrix: animating the human hand. In *ACM SIGGRAPH/Eurographics*, pages 110–119, 2003. 5

[8] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 3

[9] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 4

[10] David S Hayden, Jason Pacheco, and John W Fisher. Nonparametric object and parts modeling with lie group dynamics. In *CVPR*, 2020. 1

[11] Eric Heiden, Ziang Liu, Vibhav Vineet, Erwin Coumans, and Gaurav S Sukhatme. Inferring articulated rigid body dynamics from rgbd video. *arXiv*, 2022. 1

[12] Markus Höll, Markus Oberweger, C. Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. *VR*, pages 175–182, 2018. 4

[13] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 634–644, 2024. 4

[14] Jiahui Huang, He Wang, Tolga Birdal, Minhyuk Sung, Federica Arrigoni, Shi-Min Hu, and Leonidas J Guibas. Multibodysync: Multi-body segmentation and motion estimation via 3d scan synchronization. In *CVPR*, 2021. 2

[15] Wolfgang Hürst and Casper Van Wezel. Gesture-based interaction via finger tracking for mobile augmented reality. *Multimedia Tools and Applications*, 62(1):233–258, 2013. 4

[16] Ajinkya Jain, Rudolf Lioutikov, Caleb Chuck, and Scott Niekum. Screwnet: Category-independent articulation model estimation from depth images using screw theory. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021. 1

[17] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11107–11116, 2021. 5

[18] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *CVPR*, 2022. 1

[19] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 4

[20] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *3DV*, pages 11–21. IEEE, 2021. 4

[21] Jun-Sik Kim and Jung-Min Park. Physics-based hand interaction with virtual objects. In *ICRA*, pages 3814–3819. IEEE, 2015. 5

[22] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3

[23] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19876–19887, 2024. 3

[24] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *CVPR*, 2020. 1

[25] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9752–9762, 2024. 2

[26] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 4

[27] Shaowei Liu, Saurabh Gupta, and Shenlong Wang. Building rearticulable models for arbitrary 3d objects from 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21138–21147, 2023. 2

[28] Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, and Shenlong Wang. Contactgen: Generative contact modeling for grasp generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20609–20620, 2023. 5

[29] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *ICCV*, 2019. 1

[30] 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, 2015. 4

[31] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[32] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 5

[33] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *ICCV*, 2021. 1

[34] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3d joints for re-posing of articulated objects. In *CVPR*, 2022. 2

[35] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 5

[36] Nancy S Pollard and Victor Brian Zordan. Physically based grasping control from example. In *ACM SIGGRAPH/Eurographics*, pages 311–318, 2005. 5

[37] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 4

[38] Omid Taheri, Vasileios Choutas, Michael J Black, and Dimitrios Tzionas. Goal: Generating 4d whole-body motion for hand-object grasping. In *CVPR*, pages 13263–13273, 2022. 5

[39] Tze Ho Elden Tse, Zhongqun Zhang, Kwang In Kim, Ales Leonardis, Feng Zheng, and Hyung Jin Chang. S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *ECCV*, pages 568–584. Springer, 2022. 4, 5

[40] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *IEEE Trans. Ind. Electron.*, 50:676–684, 2003. 4

[41] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *CVPR*, 2019. 2

[42] Fangyin Wei, Rohan Chabra, Lingni Ma, Christoph Lassner, Michael Zollhöfer, Szymon Rusinkiewicz, Chris Sweeney, Richard Newcombe, and Mira Slavcheva. Self-supervised neural articulated shape and appearance models. In *CVPR*, 2022. 2

[43] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2069, 2024. 3

[44] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *ICCV*, 2021. 2

[45] Min-Yu Wu, Pai-Wen Ting, Yahui Tang, En-Te Chou, and L. Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *J. Vis. Commun. Image Represent.*, 70:102802, 2020. 4

[46] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 2

[47] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. *Advances in Neural Information Processing Systems*, 35:28559–28574, 2022. 2

[48] Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval. *arXiv*, 2021. 2

[49] Zihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver Van Kaick, Hao Zhang, and Hui Huang. Rpm-net: recurrent prediction of motion and parts from point cloud. *arXiv*, 2020. 2

[50] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2

[51] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, pages 11097–11106, 2021. 4, 5

[52] Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, and Varun Jampani. Hi-lassie:

High-fidelity articulated shape and skeleton discovery from sparse image ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4853–4862, 2023. 2

[53] Li Yi, Haibin Huang, Difan Liu, Evangelos Kalogerakis, Hao Su, and Leonidas Guibas. Deep part induction from articulated object pairs. *arXiv*, 2018. 2

[54] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: neural manipulation synthesis with a hand-object spatial representation. *TOG*, 40(4):1–14, 2021. 5

[55] Kaizhong Zhang and Dennis Shasha. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6), 1989. 2