REACTO: Reconstructing Articulated Objects from a Single Video

Supplementary Material

Appendix A provides more details of our work, such as implementation details, loss functions, and details of the dataset. In Appendix B and Appendix C, we provide more experimental results, including the comparison between different methods and additional ablation studies. Finally, we describe limitations of our method in Appendix D.

A. More details of REACTO

A.1. Loss functions

In addition to the reconstruction losses and sparse skinning loss presented in the main paper, we incorporate a 3D cycle loss [4, 6, 9] to ensure consistency between observation-to-canonical and canonical-to-observation deformations. We also leverage the consistency losses as implemented in PPR [10] to ensure coherence between object and bones. Additionally, we apply an implicit geometric regularization term [3] as follows:

$$\mathcal{L}_{eikonal} = \sum_{\mathbf{X}} (\|\nabla \mathbf{MLP}_{\mathrm{SDF}}(\mathbf{X})\|_2 - 1)^2. \quad (1)$$

A.2. Implementation details

The sparse skinning loss is incorporated after 2,000 iterations when we can extract a coarse 3D mesh, which is essential for locating the nearest vertices using KNN. The hyperparameter γ, a temperature factor for softmax, decreases from 1 to 0.1 in the first 2,000 iterations and remains constant in the subsequent iterations. η and ζ in Geodesic point assignment are set to 0.3 and 0.2 respectively. The optimization takes approximately 30 minutes on a single NVIDIA RTX 6000 Ada GPU (requires only 10G). The weights of loss terms are tuned to have similar initial magnitudes, ensuring a balanced optimization process. We conduct all the experiments following PPR's setting, all the comparing methods are optimized for 4,000 iterations.

A.3. Dataset details

For the real-world videos employed in this work, the frame counts are indicated in Table 1. These videos were casually captured using a phone camera.

The synthetic videos (*laptop*, *faucet*, and *box*) used in the appendix undergo the same processing steps as discussed in the main paper, originating from the PartNet-Mobility dataset [1, 5, 8]. Each video is rendered with 100 frames.

B. Additional comparison results

In this section, we show the qualitative comparison results of our method with BANMo [9], MoDA [6], and PPR [10] on *USB* and *stapler* in Figure 1. BANMo and MoDA struggle with complete shape reconstruction and always produce non-smooth surfaces. The results of PPR are smoother but with inaccuracies in motion modeling. In contrast, our REACTO outperforms these methods, excelling in the shape and motion reconstruction of articulated objects.

Additionally, we also compare our method (rig on bones) with BANMo, MoDA, and PPR using rigging on joints (rigs positioned at the ends of each rigid part and the hinge joints) in Table 2 and Figure 3. Our method consistently outperforms them both qualitatively and quantitatively. The rigid parts often appear influenced by more than one joint, resulting in seams (BANMo on *real-scissors*, PPR on *real-scissors* and *real-stapler*), and distortion (PPR on *real-faucet* and *real-laptop*). Some examples are highlighted in red.

C. Additional ablation studies

In this section, we extend our ablation studies to include additional categories from the PartNet-Mobility dataset [1, 5, 8], such as *laptop*, *faucet*, and *box*, as presented in Table 3.

When comparing our method with other deformation models, including displacement field [7], Real-NVP [2], and rigid skinning, our approach consistently outperforms them quantitatively. For the qualitative comparison on *real-handle* and *faucet* as shown in Figure 3, our method produces high-fidelity results, while other deformation models always exhibit artifacts during motion.

Furthermore, we explore different rigging strategies, comparing rig on bones with rig on joints, as shown in Table 3 and Figure 4. Our bone-based rigging proves more suitable for modeling the motion of general articulated objects.

D. Limitations

One limitation of this work is the reconstruction quality on the unseen side. Our method may struggle to accurately reconstruct parts that are not visible. For example, when considering a pair of glasses, if only one side of the lens is visible, the 3D reconstructed results for the lens may appear recessed. The reconstruction results improve when both sides are visible, as demonstrated in Figure 5. This factor may also lead to relatively low performance in quantitative evaluations when compared to the ground truth.

Another limitation is illustrated in Figure 6 in the main paper and Figure 1, where the gray part of the *USB* should not be connected altogether. However, our method fails to learn this hollow component.

Table 1. **The number of frame of all real-world videos in this work.**

Data	*real-stapler*	*real-scissors*	*real-faucet*	*real-laptop*	*real-nail clipper*	*real-glasses*	*real-handle*	*real-clamp*
Frame	64	41	132	240	49	203	51	64

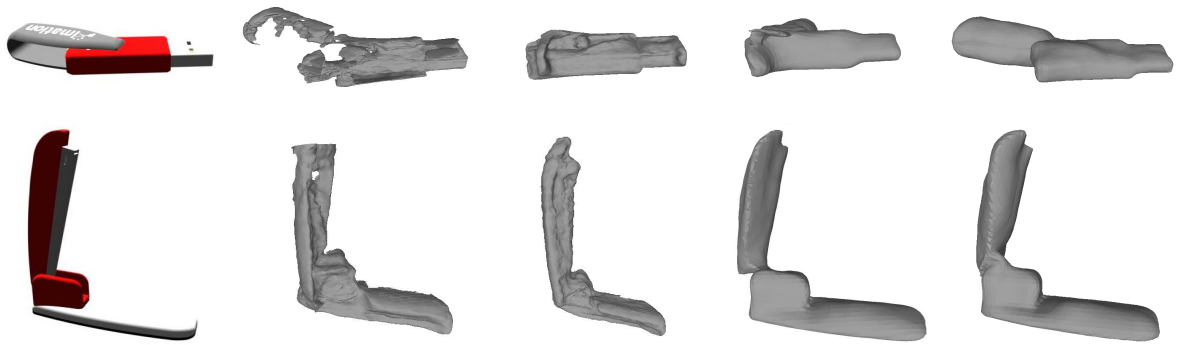

Figure 1. **Qualitative comparison of our method with BANMo [9], MoDA [6] and PPR [10] on *USB* and *stapler*.** BANMo and MoDA struggle with complete shape reconstruction and always produce non-smooth surfaces. The results of PPR are smoother but with inaccuracies in motion modeling. In contrast, REACTO outperforms these methods, excelling in the shape and motion reconstruction of articulated objects.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[2] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Ben-

Table 2. **Quantitative comparison between different methods.** Our method defines rig on bones, BANMo, MoDA, and PPR define rig on joints.

Method	USB			stapler			scissors		
	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)
BANMo	16.8	73.0	48.6	20.9	59.8	39.8	14.9	72.8	41.8
MoDA	17.3	71.1	43.7	18.1	67.6	40.9	15.5	71.4	40.1
PPR	16.7	70.8	44.6	18.4	57.6	24.0	16.2	69.2	37.2
Ours	**15.3**	**78.6**	**51.5**	**14.3**	**75.5**	**42.7**	**14.0**	**78.2**	**43.9**

Table 3. **Quantitative ablation studies.** We evaluate different settings on synthetic data and measure their performance using Chamfer distance (cm, ↓) and F-score(%, ↑) as the metrics. Our method outperforms the displacement field, Real-NVP, rigid skinning and rig on joints across various data.

Method	laptop			faucet			box		
	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)	CD(↓)	F(10%, ↑)	F(5%, ↑)
Displacement	30.1	34.6	16.2	29.8	26.8	8.6	22.9	49.0	21.2
Real-NVP	23.4	57.3	32.1	18.7	59.1	30.5	21.7	54.7	25.6
Rigid	24.9	59.8	38.7	14.7	73.7	38.1	25.8	54.3	33.5
Rig on joints	27.4	54.9	34.6	18.8	60.6	32.4	25.6	55.4	30.8
Ours	**13.8**	**80.0**	**53.1**	**14.0**	**76.5**	**41.5**	**18.0**	**66.2**	**39.8**

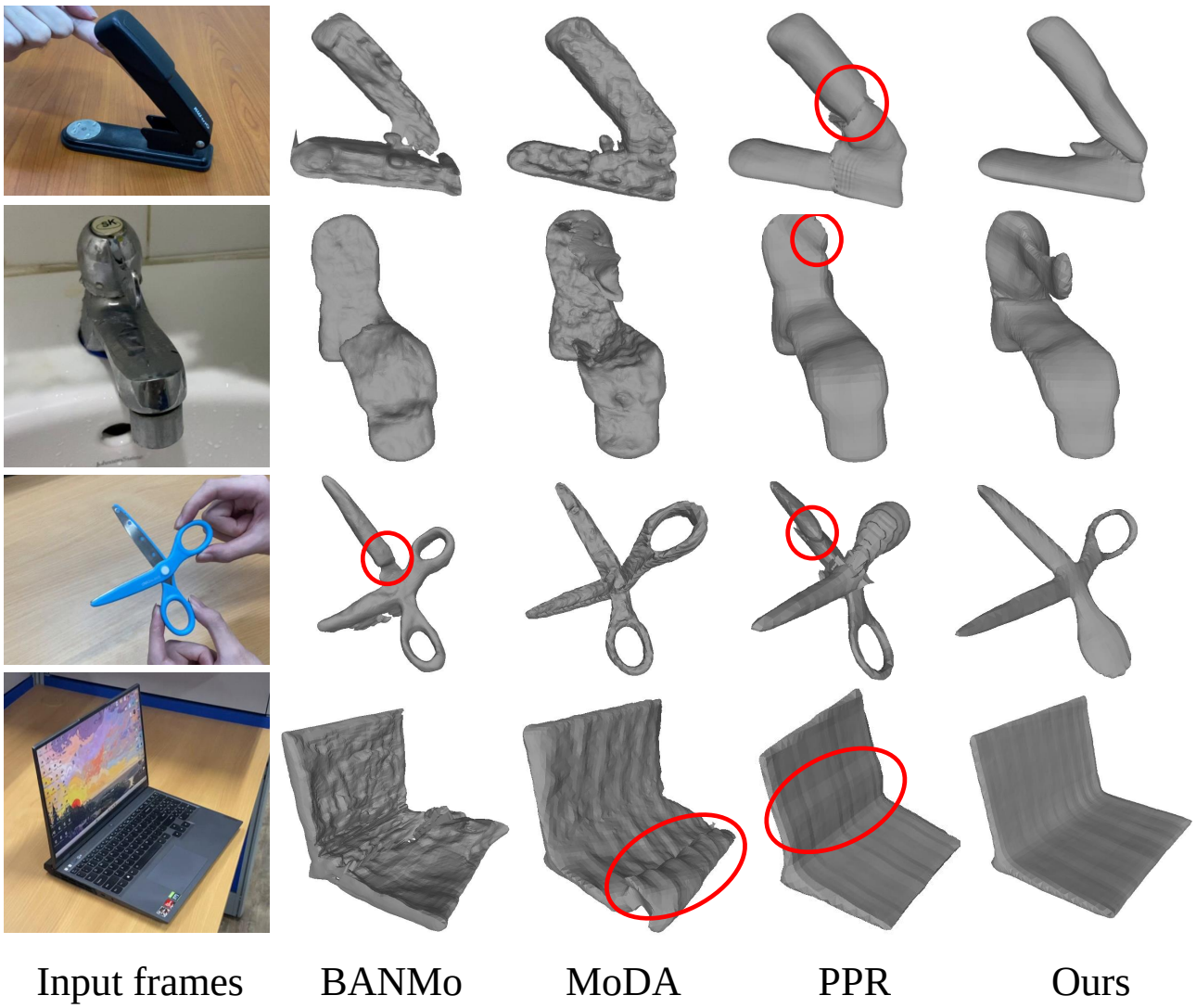

Figure 2. **Qualitative comparison of our method (rig on bones) with BANMo [9], MoDA [6] and PPR [10] defining rig on joints.** When defining rig on joints for these methods (3 joints for *real-stapler*, *real-faucet*, and *real-laptop*, 5 for *real-scissors*), the rigid parts often appear influenced by more than one joint, resulting in seams (BANMo on *real-scissors*, PPR on *real-scissors* and *real-stapler*), and distortion (PPR on *real-faucet* and *real-laptop*). Some examples are highlighted in red.

gio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 1, 4

[3] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 1

[4] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1

[5] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] Chaoyue Song, Tianyi Chen, Yiwen Chen, Jiacheng Wei, Chuan Sheng Foo, Fayao Liu, and Guosheng Lin. Moda: Modeling deformable 3d objects from casual videos. *arXiv preprint arXiv:2304.08279*, 2023. 1, 2, 3

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15816–15826, 2022. 1, 4

[8] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao

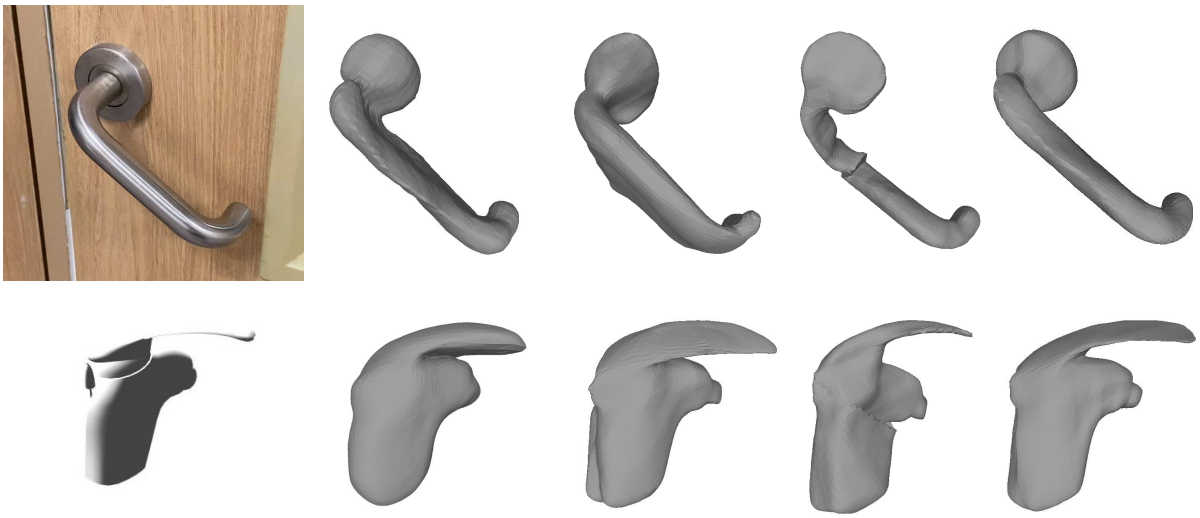

Figure 3. **Ablation studies on deformation models.** We compare our method with other deformation models, including displacement field [7], Real-NVP [2], and rigid skinning on *real-handle* and *faucet*. Displacement field and Real-NVP struggle to maintain the object shape during motion, and rigid skinning often introduces unwanted discontinuities. In contrast, our method consistently reconstructs high-fidelity results.

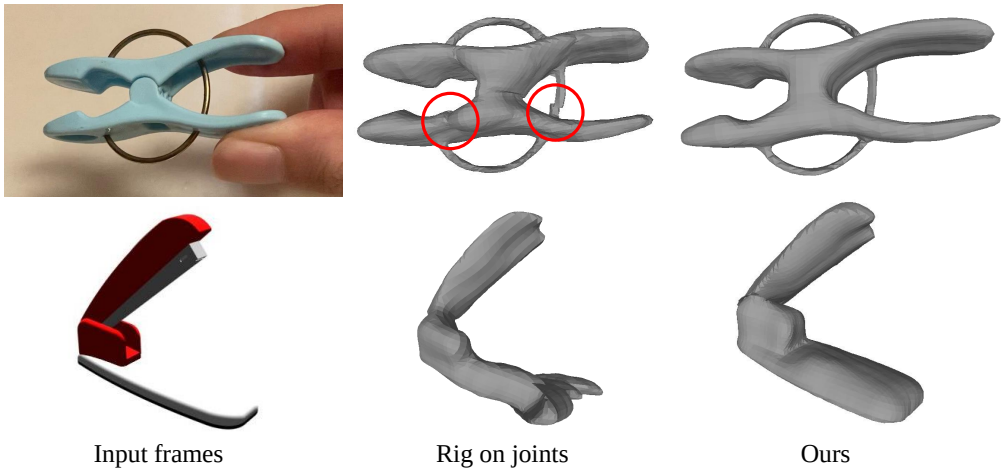

Figure 4. **Rig on joints vs Rig on bones (ours).** Defining the rig on joints (5 joints for *real-clamp* and 3 joints for *stapler*) may result in bending shapes and discontinuities (in the red circle). Our rig on bones design enhances the rigidity and motion integrity of each component.

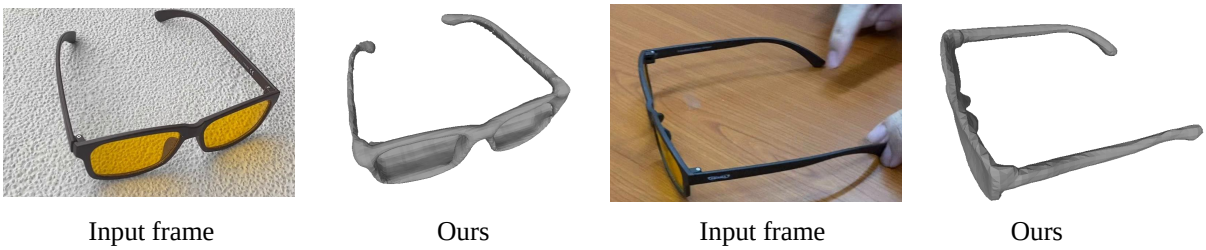

Figure 5. If only one side of the lens is visible, the 3D reconstructed results for the lens may appear recessed. The reconstruction results improve when both sides are visible

Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[9] 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. 1, 2, 3

[10] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3914–3924, 2023. 1, 2, 3