ArticulatedGS: Self-supervised Digital Twin Modeling of Articulated Objects using 3D Gaussian Splatting

Supplementary Material

1. Experiment Details

Implementation details. Our implementation is primarily based on PyTorch framework[6] and tested in a single RTX3090GPU. We move on to the next optimization stage when the loss value does not decline.

During the reconstruction process, our method requires 5,000 steps for the initialization of Gaussian, 10,000 steps for the isolated-deformation step, 3,000 steps for optimizing motion parameters, and finally, 10,000 steps for the joint optimization step, with the total time ranging from 9 to 12 minutes. In contrast, PARIS requires about 20 minutes to perform 30,000 steps of optimization.

Details and analysis on our real-world results. We acknowledge that our method, similar to the classic 3DGS, requires relatively accurate camera poses. However, this requirement can be mitigated if a more advanced variant of 3DGS, specifically designed to enhance robustness, is employed. For camera pose estimation, we indeed used COLMAP, but it struggled with surfaces that had high glare or low texture. To enhance its performance in these challenging areas, we added QR codes to the surfaces.

1.1. Dataset

Synthetic Datasets. Building on the work of the PARIS algorithm, we employed ten virtual articulated objects from the PartNet-Mobility dataset [1] for our analysis, each state has 100 vantage points that survey the upper hemisphere of the object as the training data and 50 novel views for evaluation. Building upon this foundation, we have expanded our dataset to include an additional seven categories of objects to further validate the efficacy of the algorithm. Following this, we harness the capabilities of Blender[4] to generate RGB visual representations and ascertain the camera parameters and the precise boundaries of the objects within the images, which are essential for compiling our training data.

1.2. Evaluation Metrics

To evaluate the efficacy of our methodology, we have designed a comparative experimental framework that spans three critical aspects: reasoning of motion, precision of geometry, and visual quality of novel views. In the end, we also undertake a comparative assessment of the computational efficiency and memory utilization of our approach against PARIS, which is also a self-supervised technique.

Motion Reasoning. Following the previous works [2, 3, 8], we evaluate the estimated articulation model with the following metrics:

- **Axis Ang Err**(°). The angular error of the predicted joint axis for both revolute and prismatic joints.
- **Axis Pos Err (m)**. The minimum distance between the predicted and ground-truth joint axis for revolute joints.
- **Geo dist (° or m)**. The geodesic distance error of predicted rotations for revolute joints, or Euclidean distance error of translations for prismatic joints.

Object and Part Geometry. We sample 10K points uniformly on the ground truth and predicted meshes of the baselines. In our method, we sample 10K Gaussian kernels as the point cloud randomly. To evaluate the geometry quality, we use bi-directional Chamfer-l_1 distance (CD), and compare **CD-w (mm)** for the whole object, **CD-s (mm)** for the static part and **CD-m (mm)** for movable parts.

Novel View Synthesis. In assessing the fidelity of the appearance model, we quantify the performance using the Peak Signal-to-Noise Ratio (**PSNR**), the Structural Similarity Index (**SSIM**), and Learned Perceptual Image Patch Similarity (**LPIPS(VGG)**) for images rendered from novel perspectives. For each object instance, we generate and render 50 unique viewpoints per state, subsequently computing the mean values to ascertain the overall quality.

1.3. Comparisons details

More visual results. Figure 1 shows the results of various methods in part-segmentation and motion estimation for all objects in the PARIS dataset. We also generated additional data across five categories for comparison, as shown in Figure 2. Multiple examples demonstrate that in the results of the PARIS method, part segmentation, motion estimation, and appearance have all converged to local optima, leading to a significant discrepancy from the ground truth. It can be observed that our method outperforms the PARIS method in terms of segmentation accuracy and reconstruction quality on these new datasets. Figure 3 shows the reconstruction results of our method and PARIS's. From the visualization results, it can be observed that our method achieves higher reconstruction quality and more precise articulation information estimation results.

Figure4 illustrates more predicted articulation sequence based on the reconstructions obtained by different methods.

			Simulation										Real-world	
			Foldchair	Fridge	Laptop	Oven	Scissor	Stapler	USB	Washer	Blade	Storage	Fridge	Storage
Motion	Ang Err	DTA	0.03	0.07	**0.06**	0.22	**0.11**	0.06	**0.11**	0.43	**0.27**	0.06	**2.10**	**18.11**
		Ours	**0.02**	**0.04**	0.07	**0.04**	0.20	**0.03**	0.30	**0.06**	1.99	**0.05**	5.41	41.52
	Pos Err	DTA	0.01	0.01	0.00	0.01	0.02	0.01	0.00	0.01	-	-	0.57	-
		Ours	**0.00**	**0.00**	0.00	0.01	**0.00**	**0.00**	0.00	**0.00**	-	-	**0.06**	-
	Geo Dist	DTA	0.16	0.09	0.08	0.11	0.15	0.05	0.11	0.25	0.00	**0.00**	**1.86**	**0.20**
		Ours	**0.05**	**0.18**	0.55	0.61	**0.15**	**0.06**	**0.33**	**0.34**	**0.09**	0.30	9.38	0.41
Geometry	CD-s	DTA	**0.18**	**0.60**	**0.32**	4.66	0.40	2.65	2.19	**4.80**	**0.55**	4.69	**2.53**	**10.86**
		Ours	0.33	0.90	2.96	**2.06**	**0.34**	**1.68**	**1.02**	6.17	0.69	**2.96**	37.01	50.12
	CD-d	DTA	**0.15**	**0.27**	**0.16**	**0.47**	0.41	2.27	1.34	0.36	**1.50**	**0.37**	**1.14**	**26.46**
		Ours	0.32	0.59	6.23	0.80	0.41	**1.14**	**0.90**	**0.17**	4.75	0.71	43.00	730.45
	CD-w	DTA	**0.27**	**0.70**	**0.35**	4.18	0.43	2.19	1.18	**4.74**	0.36	3.99	**2.19**	**9.33**
		Ours	0.36	0.85	0.87	**1.91**	**0.33**	**1.42**	**0.93**	5.64	**0.34**	**2.09**	13.25	55.57

Table 1. The results on the PARIS dataset, encompassing both synthetic and real data, are presented. We conducted a comparison with the DTA [8] method by examining both the aspects of motion and geometry.

PSNR-b	w/o arap	39.56	PSNR-e	w/o arap	39.54
	w/o geo	37.62		w/o geo	37.64
	w/o bal	**40.23**		w/o bal	38.92
	ours	39.83		ours	**39.82**
SSIM-b	w/o arap	0.966	SSIM-e	w/o arap	0.963
	w/o geo	0.972		w/o geo	0.969
	w/o bal	**0.989**		w/o bal	0.965
	ours	0.989		ours	**0.989**
LPIPS-b	w/o arap	0.035	LPIPS-e	w/o arap	0.037
	w/o geo	0.036		w/o geo	0.035
	w/o bal	**0.031**		w/o bal	0.032
	ours	0.032		ours	**0.032**

Table 2. Visual results of the ablation studies. "w/o bal" represents the results without loss balance in joint optimization step; "w/o geo" is the results without Chamfer distance loss in motion parameter step; ""w/o arap" is the results without ARAP loss.

We also added the DeformGS [9], an excellent and pioneering 4D Gaussian work, for visual comparison. Due to the sparsity of the input temporal states, DeformGS can only reconstruct the appearance of the two input states with relative accuracy. However, it lacks prior knowledge of articulated motion and thus cannot infer the actual transformation process between the two states. PARIS also struggles to accurately estimate motion parameters for certain objects. Even when the segmentation results are close to correct, the motion process still deviates from the actual outcome.

Figure 5 further illustrates the depth of our reconstruction results. The detailed accuracy of the depth highlights the precision of our geometric reconstruction, which is exceptionally beneficial for the task of synthesizing novel views and other downstream applications.

Analysis on visual quality. Table **??** presents the comparisons on visualization quality. Our approach outperforms PARIS across all evaluated metrics. This is partly due to the fact that the appearance model of PARIS is constructed based on the Instant-NGP [5] method, which, even with accurate motion estimation, has a visual quality ceiling that is lower than that of 3D-GS.

Analysis on PARIS real-world results. The performance drop on the two real-world objects provided by PARIS is mainly due to the inaccurate annotations on the camera pose and object masks. Learning-based methods like Ditto are less sensitive to such noisy input. However, our method consistently obtains better results compared to PARIS, including the extra five real-world objects we collected.

1.4. Comparisons to DTA

Table 1 shows the comparisons between our method and DTA [8]. We directly utilized the data shown in the DTA paper. From the results, it is evident that even without the input of depth values, our method outperforms DTA in estimating object motion across multiple metrics. However, due to DTA's incorporation of depth data and image feature extraction as priors, it achieves commendable results when dealing with real-world object data that contain significant noise. Additionally, in terms of algorithmic efficiency, our approach achieves similar reconstruction results in approximately one-third of the time required by DTA.

1.5. More ablation studies

Initial mask threshold. When optimizing motion parameters, a lower threshold as 0.1, is effective for most objects. However, when the moving parts of an object possess high symmetry, and half of the object is obscured from view in the initial state, optimization of motion may converge to a local optimum, as shown in Fig. 6. At this point, the higher threshold we set will come into effect, increasing the discrepancy between the local optimum and the true state, thereby achieving better optimization results.

Figure 1. Qualitative results of part segmentation and motion estimation for some of the synthetic objects derived from the PARIS dataset.

Design of objective function. To further justify the design of our objective functions, we conducted several ablation studies, including: "w/o arap" for the results without ARAP loss during deformation prediction, "w/o geo" for results without Chamfer distance loss in motion parameter step, and "w/o bal" for the results without loss balance in joint optimization step. Experiments show that by adding those designs, the results are consistently better. The quantitative results are reported in the supplementary material.

Tab. 2 shows the results of our ablation study. It should be noted that for many objects with few changes between two states, the original loss is already sufficient, and these losses will not provide much assistance

Table 3. Average performance on the PARIS dataset.

	Simulation			Real-world		
	default	w/ depth	switch	default	w/ depth	switch
Ang Err	0.28	0.25	0.29	23.47	23.95	23.51
Pos Err	0.00	0.00	0.00	0.05	0.08	0.06
Geo Dist	0.27	0.20	0.27	4.90	3.29	4.95
CD-s	1.91	1.55	1.91	43.57	42.55	43.59
CD-d	1.60	1.41	1.60	386.73	394.62	386.85
CD-w	1.47	1.62	1.48	34.41	33.35	34.56

The switch of *Start* & *End* States. Switching the state order will not affect the results of our method, as shown in Table 3 with the setting denoted as *swith*. In fact, the results provided in the paper are obtained with a random order.

Injest depth as input. We tried injesting depth in a straightforward manner, where we directly use the depth maps for additional supervision to 3DGS's depth rendering output, and the results are shown in Table3 with the setting denoted as *w/ depth*. We can see that there is only slight improvement in this naive implementation, and we find that many existing RGB-D 3DGS methods could be employed to further enhance the performance, such as combining RGB-D information with 2D-GS to obtain better geometry accuracy, which we will leave for future work.

2. Limitation analysis & future works

Our method has certain limitations, which also point the way for future work. Our methodology may not consistently deliver satisfactory outcomes for every pair of object states. This is particularly true in scenarios such as: when the movable part is disproportionately small relative to the entire object; or when the color of the movable part is closely similar to the object's other parts. When reconstructing real-world objects, our method is also susceptible to the interference of coordinate alignment and lighting conditions.

Two failure cases are shown in Figure 7. Besides, when there is a deviation in the world coordinates of the two states, the deviated part can be directly estimated as the object's movement, which poses a substantial challenge to our method, especially when dealing with real-world objects with inaccurate camera pose estimation. In the reconstruction of real objects, significant differences in lighting between corresponding parts of the object in two states can also have a substantial negative impact on the outcome. Our method currently can only handle objects with a single movable part at one time, and the type of joint is limited to prismatic and revolute.

A promising direction for future research is the development of self-supervised techniques to accurately estimate the unknown quantity of movable components within an object. Additionally, there is a significant opportunity to advance our reconstruction capabilities to handle

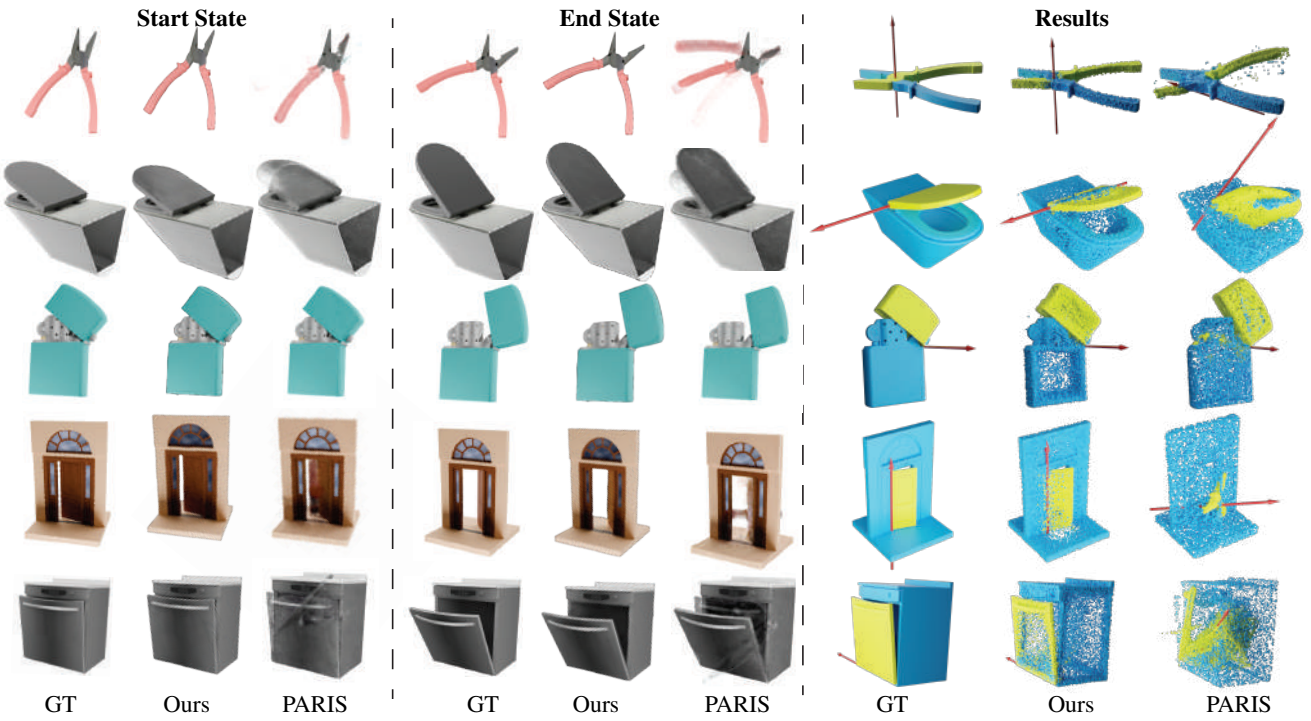

Figure 2. Illustration of the effects of applying our method and the PARIS algorithm to our dataset. It includes the appearance at the beginning and end states, along with part segmentation and motion estimation for both methods.

more intricate mechanical configurations of articulated objects.

References

[1] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 1

[2] Zhenyu Jiang, Cheng-Chun Hsu, and Yuke Zhu. Ditto: Building digital twins of articulated objects from interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5616–5626, 2022. 1

[3] Jiayi Liu, Ali Mahdavi-Amiri, and Manolis Savva. Paris: Part-level reconstruction and motion analysis for articulated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 352–363, 2023. 1

[4] Tony Mullen. *Mastering blender*. John Wiley & Sons, 2011. 1

[5] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 1

[7] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 7

[8] Yijia Weng, Bowen Wen, Jonathan Tremblay, Valts Blukis, Dieter Fox, Leonidas Guibas, and Stan Birchfield. Neural implicit representation for building digital twins of unknown articulated objects. *arXiv preprint arXiv:2404.01440*, 2024. 1, 2

[9] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2

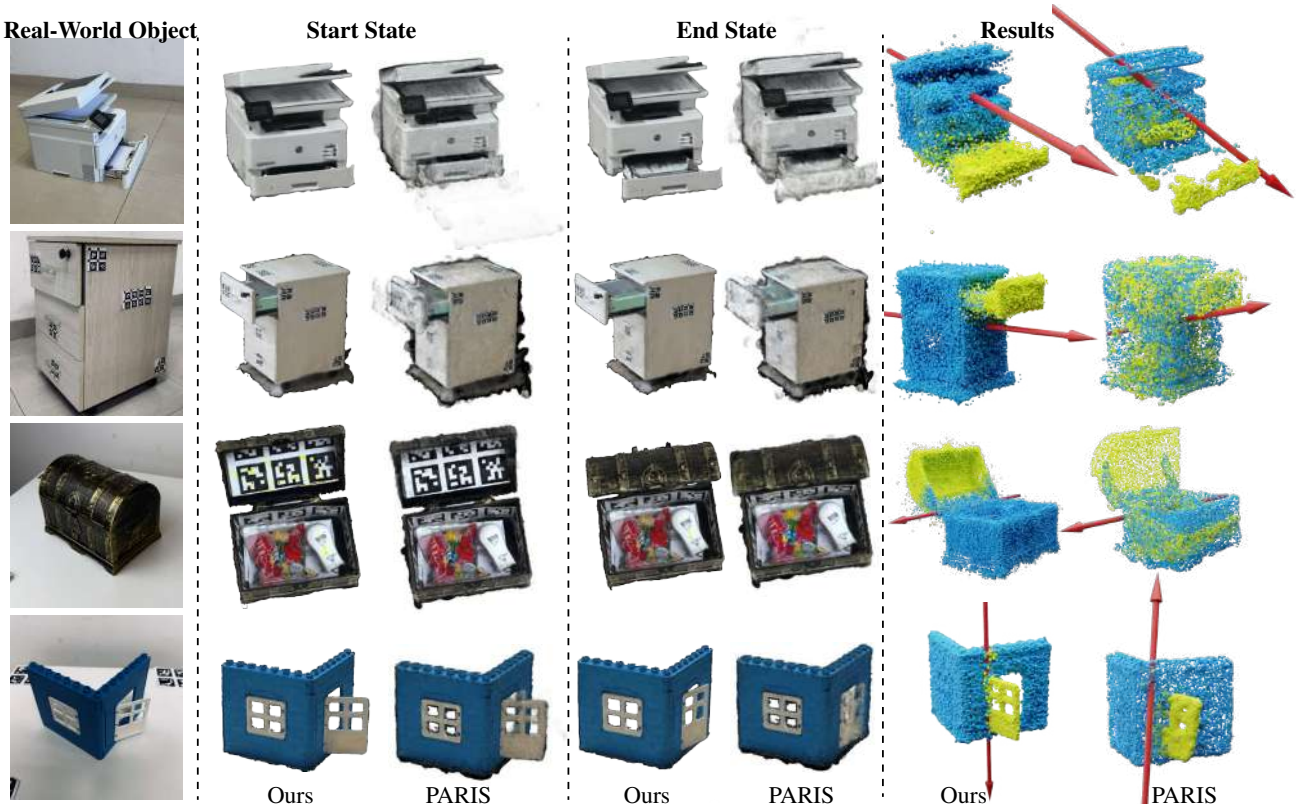

Figure 3. Illustration of applying our method and the PARIS algorithm to real-world objects.

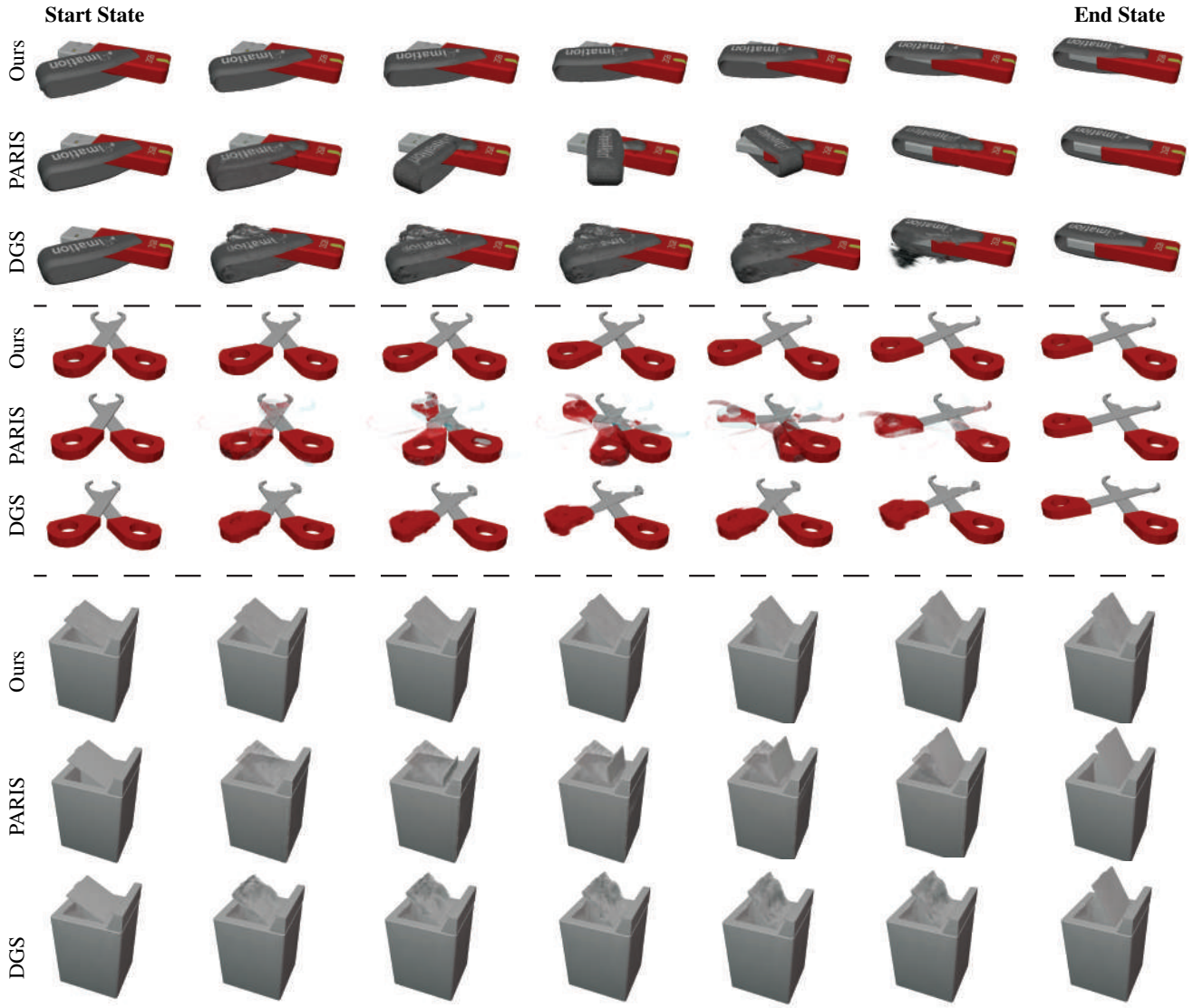

Figure 4. Illustration of unseen states inference. It reveals transformation between two known states while ours yields consistent visual outcomes.

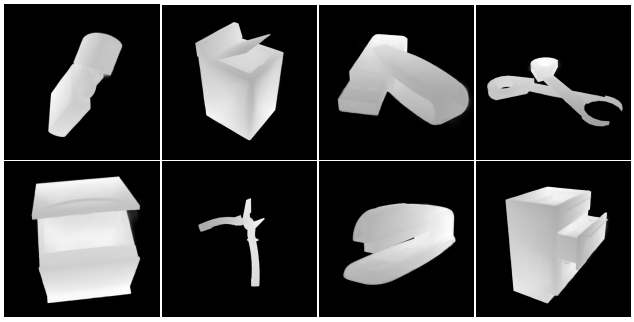

Figure 5. Depth visualization. We rendered the depth maps of the synthetic objects from Shape2Motion [7] dataset.

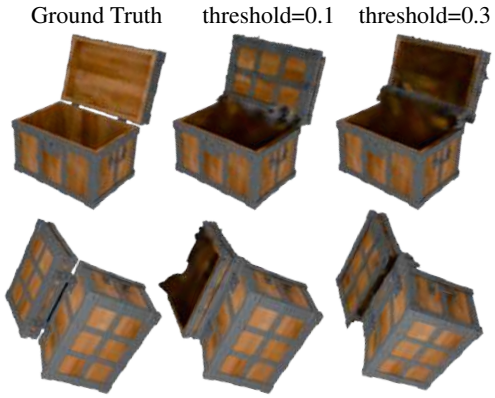

Figure 6. Illustration of our method using different thresholds when optimizing the motion parameters.

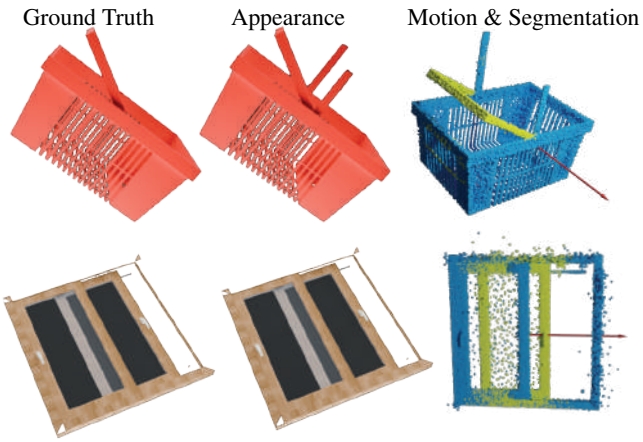

Figure 7. On the left, the basket, due to the small volume of movement, under the influence of DeformNet, its Gaussian is directly obscured into the static part, and the overall loss did not decrease significantly. The subsequent optimization process continued with this erroneous result, leading to an incorrect outcome. On the right, the window, due to the overlapping window frame positions with similar colors, resulted in low distinctiveness, thus learning an erroneous segmentation result.