

SM³: Self-supervised Multi-task Modeling with Multi-view 2D Images for Articulated Objects

Haowen Wang¹, Zhen Zhao², Zhao Jin², Zhengping Che², Liang Qiao¹,
Yakun Huang¹, Zhipeng Fan¹, Xiuquan Qiao^{1,†}, and Jian Tang^{2,†}

Abstract—Reconstructing real-world objects and estimating their movable joint structures are pivotal technologies within the field of robotics. Previous research has predominantly focused on supervised approaches, relying on annotated datasets to model articulated objects within limited categories. However, these approaches fall short of effectively addressing the diversity present in the real world. To tackle this issue, we propose a self-supervised interaction perception method, referred to as SM³, which leverages multi-view RGB images captured before and after interaction to model articulated objects, identify the movable parts, and infer the parameters of their rotating joints. By constructing 3D geometries and textures from the captured 2D images, SM³ achieves integrated optimization of movable part and joint parameters during the reconstruction process, obviating the need for annotations. Furthermore, we introduce the MMart dataset, an extension of PartNet-Mobility, encompassing multi-view and multi-modal data of articulated objects spanning diverse categories. Evaluations demonstrate that SM³ surpasses existing benchmarks across various categories and objects, and its adaptability in real-world scenarios has been thoroughly validated.

I. INTRODUCTION

Recreating real-world objects in a virtual environment and predicting how the open parts of an item will move is a critical step in helping robots complete daily household tasks [1], [2], [3], [4]. The design process of an object often requires designers to create its geometric shape and motion structure of movable parts from scratch. The same is true for robot learning systems. By using methods such as implicit neural representations (INR) [5], [6], [7] and neural radiance fields (NeRF) [8], [9], [10], it has been possible to achieve a relatively complete reconstruction of static objects in the environment. However, accurately modeling articulated objects with movable components remains a challenging problem in this field. The endeavor calls for advancements in part-level geometric modeling and analysis of the motion of articulated joints.

Recent studies [11], [12] adopted PointNet-like structures [13] to predict movable parts and their associated motion parameters directly from the point cloud of articulated objects. However, these methods usually highly rely on the complete point clouds of objects, resulting in strict

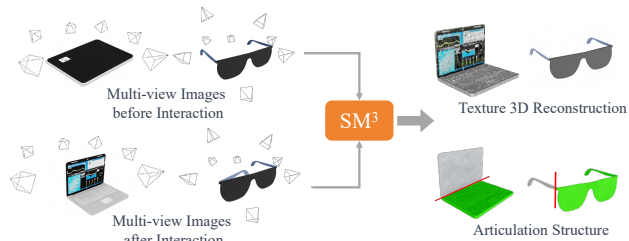


Fig. 1: Our proposed SM³ enables textured 3D reconstruction and articulation structure estimation solely from multi-view images captured before and after object interaction.

requirements on equipment and difficult implementation in the real world. To address this issue, some techniques [14], [15] have embraced the Normalized Object Coordinate Space (NOCS) approaches [16], [17] to canonicalize object scales and joints, enabling the reconstruction of articulated objects or joint parameter estimation from a single-view image. These methods typically underperform with novel shapes and exhibit certain limitations.

The recently introduced Ditto [18] offers a fresh perspective, demanding point clouds before and after object interaction to decode its motion and geometric characteristics. It employs implicit neural representations for 3D object reconstruction and integrates part segmentation and joint parameters through joint optimization. However, it still necessitates relatively complete point cloud inputs and is nearly incapable of extending to previously unseen object categories. Moreover, these methods require supervised training, relying on expensive datasets that include articulated objects with precise 3D part models and comprehensive articulation annotations.

In light of the aforementioned challenges, we propose a novel approach called Self-supervised Multi-task Modeling with Multi-view 2D Images for Articulated Objects (SM³), uniquely characterized by its independence from any annotated data and 3D models. Utilizing multi-view RGB images of articulated objects taken pre- and post-interaction, SM³ achieves textured 3D reconstruction, segmentation of the movable part, and the estimation of corresponding rotational joint parameters, as shown in Fig. 1. Our 3D reconstruction methodology builds upon the Nvdiffrac [19] framework, employing a deformable tetrahedral grid and computing post-rendering image losses for objects before and after the interaction. This tetrahedral structure serves as the foundation for subsequent movable part segmentation and joint parameter

¹State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China {hw.wang, LiangQ, ykhuang, fzp, qiaoxq}@bupt.edu.cn

²Midea Group, China {zhaozhen8, jinzhao1, chezp, tangjian22}@midea.com, chezhengping@gmail.com
Work done during Haowen Wang's internship at Midea Group.

[†]Corresponding authors: Xiuquan Qiao and Jian Tang.

optimization. To generate reliable constraints for integrated optimization and mitigate convergence issues, we design two algorithmic workflows that analyze geometric structure differences between objects in their pre- and post-interaction states, respectively. These workflows facilitate the generation of movable part segmentation priors and rotational joint position and direction candidates. Furthermore, we introduce a patch-split method to refine the original image loss, thereby enhancing segmentation accuracy. After several steps in the process, we render the pre-interaction tetrahedral grid of the movable parts as they rotate through the joints, converting them to RGB images. To finalize the optimization of the accuracy of movable part segmentation and rotational joint parameters, we calculate the loss against post-interaction object-sampled RGB images. Given the scarcity of multi-view RGB image datasets specifically designed for articulated objects, we have meticulously curated a pioneering dataset named MMArt, which is an in-depth redevelopment of the PartNet Mobility dataset [20] through simulation-based enhancements and modifications. This multi-view, multi-modal, and multi-state dataset encompasses diverse articulated objects spanning multiple categories. It caters to our evaluation requirements and readily adapts to facilitate the training and testing of other methods within this domain.

Our contributions can be summarized as follows:

- We introduce SM³, a pioneering self-supervised multi-task method that simultaneously learns textured 3D reconstruction, movable part segmentation, and rotating joint parameters for articulated objects.
- We devise two algorithmic workflows to effectively generate active part segmentation priors and joint candidates to further refine accuracy through optimization strategies.
- We present the MMArt dataset that supports comprehensive evaluations for articulated object modeling.
- Our experimental results demonstrate the effectiveness of SM³ in accurately modeling articulated objects, surpassing recent SOTA methods by a significant margin.

II. RELATED WORK

1) *3D Reconstruction for Articulated Objects*: The 3D reconstruction of articulated objects demands the overall geometric reconstruction and necessitates precise modeling of movable parts. While some methods [14], [11] utilize PointNet-like architectures [13] to recognize movable parts from input point clouds, they fall short of performing reconstruction. With the evolution of implicit neural networks [5], [21], A-SDF [22] leverages them alongside joint angle parameters to achieve geometric reconstruction while segmenting parts. Ditto [18] employs them in combination with point clouds in two states to construct 3D meshes for each part. However, both approaches require complete 3D models for supervised training and struggle to extend to unseen categories. Wei et al. [23] adopted NeRF’s concept to perform object reconstruction through multi-view image self-supervision and articulated joint motion by modifying learned joint encodings. Yet, its implicit representation hin-

ders part segmentation and effective joint parameter identification. Our method, building on Nvdiffrac’s [19] explicit reconstruction framework, enables unsupervised reconstruction while accurately identifying the movable parts of the reconstructed object.

2) *Motion Structure Estimation*: Analyzing the inter-movement relationships of parts based on their hinges is one of the key challenges in modeling articulated objects. Recently, some methods such as [24], [25], [26] have focused on joint estimation. ANCSH [14] and OMAD [27] predict segmentation and joint parameters from single-view point cloud but struggle with untrained objects and neglect object reconstruction. Ditto [18] leverages implicit decoders with point cloud inputs from pre- and post-interaction states to achieve geometric reconstruction of objects while also performing part segmentation and hinge parameter estimation. However, it still struggles to handle objects of unseen categories effectively. Meanwhile, above methods require extensive annotated data for articulation, which are challenging to obtain in real-world scenarios. In contrast, our self-supervised approach, utilizing multi-view RGB images, seamlessly integrates textured object reconstruction, component segmentation, and joint prediction.

3) *Datasets for Articulated Objects*: Currently, several datasets offer 3D models of articulated objects. RPM-Net [28] offers 949 joint objects across 43 categories. RBO [29] provides RGB-D videos of 14 articulated objects, while Shape2Motion [11] boasts 2,440 joint objects in 45 categories. PartNet [30] emphasizes semantic segmentation, whereas PartNet-Mobility [20] enriches PartNet and ShapeNet [31] with 2,346 joints across 46 categories. Nonetheless, these datasets lack multi-view and multi-modal data. To address this gap, we introduce a new dataset to enrich training and testing tasks in this domain.

III. METHODOLOGY

A. Overview

The goal of our work is to actively construct textured virtual geometric models of real-world articulated objects and estimate the articulation structure. An overview of our framework is shown in Fig. 2. For an articulated object in the scene, we initially capture n RGB images from varied viewpoints, denoted as $\mathcal{I}_{\text{pre}} = \{I_i\}_n$. After actively manipulating the object with rotation of its movable part, we obtain another set of m images, $\mathcal{I}_{\text{post}} = \{I_j\}_m$. Our method processes these two sets of images to construct a 3D texture model of the articulated object, as well as estimating the rotational joints and segmenting the movable part.

To build the geometry, we use a method based on the deformable tetrahedral grid [32], [33], [19] to render the model $\mathcal{M} = (\mathcal{V}, \mathcal{F})$. Here, $\mathcal{V} \in \mathbb{R}^{n_v \times 3}$ and $\mathcal{F} \in \mathbb{N}^{n_f \times 4}$ represent the vertex and the tetrahedral face of the model, respectively, where n_v and n_f are the number of vertices and faces. In addition, the movable part is represented using a vertex subset $\mathcal{V}^{\text{mov}} \subseteq \mathcal{V}$.

For joint construction, we follow the definition method in line with previous works [14], [18], where kinematic

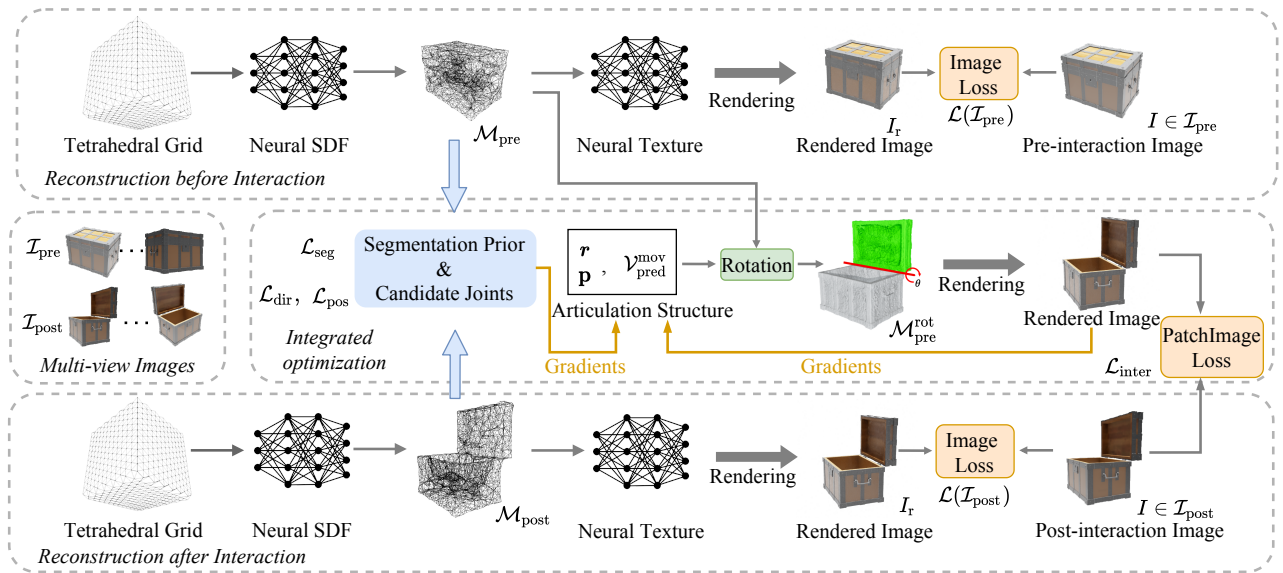


Fig. 2: Architecture Overview of proposed SM³. Leveraging multi-view RGB images \mathcal{I}_{pre} and $\mathcal{I}_{\text{post}}$, the method deforms tetrahedral grids into \mathcal{M}_{pre} and $\mathcal{M}_{\text{post}}$ as detailed in Section III-B. It then extracts segmentation priors $\mathcal{V}_{\text{pred}}^{\text{mov}}$ and candidate joints \mathbf{r}, \mathbf{p} , following the processes in Sections III-C and III-D. Using \mathbf{r} and \mathbf{p} , it rotates \mathcal{M}_{pre} 's active components to generate $\mathcal{M}_{\text{pre}}^{\text{rot}}$, optimizing parameters end-to-end against $\mathcal{I}_{\text{post}}$ as described in Section III-E.

constraints between object components are represented by parameterizing the joints. A revolute joint is characterized by its rotation axis direction $\mathbf{r} \in \mathbb{R}^3$ and a pivot point $\mathbf{p} \in \mathbb{R}^3$ on this axis. Notably, both the rotation axis and pivot point remain invariant to interactions.

B. Reconstruction Based on Tetrahedron

To enhance the prediction of object geometric reconstruction, aiding subsequent joint and part segmentation, we employ Nvdiffrac [19] and construct an approach leveraging differentiable tetrahedrons to optimize the topology and texture of the initial tetrahedra directly, as shown in Fig. 2.

Given a single image I from a set of multi-view sampled images, an implicit neural network determines the deformation directions v and the Signed Distance Function (SDF) values s for each vertex of the initial tetrahedral grid:

$$(v, s) = f_{\phi}(I), \quad (1)$$

where f_{θ} is the multi-layer perception with parameters ϕ .

Utilizing v and s , the tetrahedral grid \mathcal{M} refines to capture the object's geometry better. Concurrently, the texture image T and lighting information L are inferred from I . With these, a differentiable rasterizer [19] renders an RGB image:

$$I_r^c = R(\mathcal{M}, T, L; \Phi), \quad (2)$$

where I_r^c is the rendered RGB image for a given camera pose c and Φ are the parameters of the rasterizer.

The optimization aims to minimize the image loss between the rendered images I_r and the input images I :

$$\mathcal{L}_{\text{render}} = \frac{1}{nhw} \sum_{c=1}^n \sum_{i=1}^h \sum_{j=1}^w (I_{r,ij}^c - I_{ij}^c), \quad (3)$$

where h and w denote the dimensions of each input image, $I_{r,ij}^c$ and I_{ij}^c are pixel values at position (i, j) in the rendered and input images, respectively, for camera pose c . Through iterative propagation across all perspective images $\mathcal{I} = \{I_i\}_n$, we obtain the tetrahedral grid representing the object's geometry and its associated texture.

Leveraging this reconstruction approach, we derive two 3D geometric models: \mathcal{M}_{pre} from the pre-interaction images \mathcal{I}_{pre} and $\mathcal{M}_{\text{post}}$ from the post-interaction images $\mathcal{I}_{\text{post}}$.

C. Movable Part Segmentation Prior

For articulated object modeling, it is pivotal to derive segmentation priors for the movable part. We ascertain the overlap between the 3D models \mathcal{M}_{pre} (pre-interaction) and $\mathcal{M}_{\text{post}}$ (post-interaction) to determine segmentation prior, as shown in Fig. 3.

Specifically, we analyze vertices from the tetrahedral models of both states, \mathcal{V}_{pre} and $\mathcal{V}_{\text{post}}$. For a vertex v_i in \mathcal{V}_{pre} , it is labeled as static (0) if its distance to the nearest vertex in $\mathcal{V}_{\text{post}}$ is below a threshold τ , and movable (1) otherwise. This is expressed as:

$$\text{label}(v_i) = \begin{cases} 0 & \text{if } \min_{v_j \in \mathcal{V}_{\text{post}}} \text{dist}(v_i, v_j) < \tau, \\ 1 & \text{otherwise,} \end{cases} \quad (4)$$

where we set $\tau = 0.04$ in our study.

During segmentation refinement, we rectify potential errors from overlaps using neighborhood analysis. For each vertex, its label is determined by the majority label of its neighboring vertices. If the majority of a vertex's neighbors are labeled as static, then the vertex itself is labeled as static. Conversely, if most neighbors are movable, the vertex is labeled as movable.

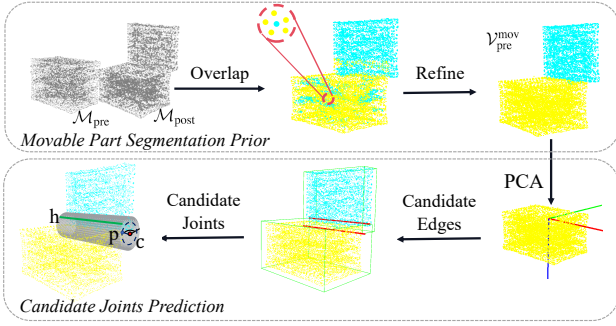


Fig. 3: Algorithmic workflow for Movable Part Segmentation Prior and Candidate Joints Prediction.

From the described workflow, we derive the segmentation prior for the movable part of the pre-interaction object’s tetrahedral grid vertices, denoted as $\mathcal{V}_{\text{pre}}^{\text{mov}}$. In the optimization phase (Section III-E), we use a cross-entropy loss $L_{\text{seg}}(\mathcal{V}_{\text{pred}}^{\text{mov}}, \mathcal{V}_{\text{pre}}^{\text{mov}})$ to guide the prediction of movable part vertices, where $\mathcal{V}_{\text{pred}}^{\text{mov}}$ represents the movable part vertices predicted during the integrated optimization stage.

D. Candidate Joints Prediction

Identifying the revolute joints in real-world articulated objects is challenging. Our method, shown in Fig. 3, assumes joints align with the object’s primary planes [34].

By applying Principal Component Analysis (PCA) to the static part’s point cloud from Section III-C, we establish a local coordinate system. Within this, we compute Axis-Aligned Bounding Boxes (AABBs) for both movable and static parts. Candidate joint positions are identified by matching parallel edges from both parts’ bounding boxes, ranked using:

$$\text{rank} = \text{rank}_{\text{dist}} + w \times (\text{rank}_{e_1} + \text{rank}_{e_2}). \quad (5)$$

where $\text{rank}_{\text{dist}}$ is based on the midpoint distance between parallel edges, thus ensuring proximity. rank_{e_1} and rank_{e_2} quantify points along edges surrounded by both static and movable parts, ensuring edge proximity to the parts’ interface. The weight w balances these contributions.

The top-ranked edge pair suggests a guiding cylinder for the joint’s position and direction. Utilizing the cylinder’s normal vector \mathbf{h} and center \mathbf{c} , we construct a Cartesian coordinate system to constrain the predicted joint’s pivotal point \mathbf{p} and direction \mathbf{r} through designated loss functions:

$$\mathcal{L}_{\text{pos}} = \max(0, \|\mathbf{p} - \mathbf{c}\|_2 - R), \quad (6)$$

and

$$\mathcal{L}_{\text{dir}} = 1 - \frac{\mathbf{r} \cdot \mathbf{h}}{\|\mathbf{r}\| \|\mathbf{h}\|}, \quad (7)$$

ensuring alignment with our assumptions. where R is the radius of the cylinder. These losses ensure the predicted joint’s position and orientation align with our assumptions.

E. Integrated Optimization

To integrate optimization of the predicted joint parameters, specifically the direction vector \mathbf{r} and the pivotal point

\mathbf{p} , alongside movable part segmentation, we employ rigid rotation transformations on the movable components.

For each vertex in the movable component of the pre-interaction tetrahedral grid, denoted as $\mathbf{v} \in \mathcal{V}_{\text{pre}}^{\text{mov}}$, we compute its post-rotation position \mathbf{v}' using the formula:

$$\mathbf{v}' = \mathbf{R}(\mathbf{r}, \theta)(\mathbf{v} - \mathbf{p}) + \mathbf{p}. \quad (8)$$

Here, $\mathbf{R}(\mathbf{r}, \theta)$ is the rotation matrix derived from the Rodrigues formula [35], θ signifies the rotation angle. Through the above operation, we transform the tetrahedra from their pre-interaction state \mathcal{M}_{pre} to their post-interaction state $\mathcal{M}_{\text{pre}}^{\text{rot}}$.

Subsequently, analogous to Section III-B, we render the deformed tetrahedral structure into an RGB image and compute the loss concerning the captured RGB image post-interaction. Importantly, during the optimization process, we refrain from propagating gradients to the SDF values and offset vector predictions of the tetrahedral vertices, preserving the geometric topology of the reconstructed object.

In tandem with our approach, we segment the image into patches and introduce the PatchImage Loss. By computing the loss for each patch and subsequently aggregating them, we derive the comprehensive image loss. This strategy adeptly mitigates the influence of minor noise or misalignments in the rotation axis, facilitating a granular examination of distinct image regions. The integrated loss function is articulated as:

$$\begin{aligned} \mathcal{L}_{\text{inter}} &= \frac{1}{n} \sum_{c=1}^n \sum_{\text{patches}} \frac{1}{hw} \sum_{i=1}^h \sum_{j=1}^w \Delta R_{ij}^c \\ \Delta R_{ij}^c &= R(\mathcal{M}_{\text{pre}}^{\text{rot}}, T, L; \Theta)_{ij}^c - I_{ij_{\text{post}}}^c \end{aligned} \quad (9)$$

The overall loss function is formulated as follows:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{pos}} + \lambda_3 \mathcal{L}_{\text{dir}} + \lambda_4 \mathcal{L}_{\text{inter}} \quad (10)$$

where λ_1 , λ_2 , λ_3 , and λ_4 equal 1, 2, 2, and 10, respectively, in this study. Utilizing this loss computation, we achieve integrated optimization of movable part and joint parameters, guided by geometric priors and sampled image supervision.

IV. THE MMART DATASET

Several existing datasets [28], [30], [20] provide component models and joint parameters for articulated objects. However, a clear void exists for datasets dedicated to multi-view 2D image-driven 3D reconstruction of these entities. Bridging this void, we create the Multi-Modal Articulated Objects Dataset (MMArt), an extension of the PartNet-Mobility [20] dataset, enriched using the Isaac Sim simulation environment [36].

Specially, we meticulously selected thirteen categories: box, dishwasher, door, eyeglasses, laptop, lighter, microwave, safe, stapler, storage furniture, toilet, oven, and washing machine from the PartNet-Mobility [20] dataset. Within each of these categories, we carefully chose eight unique objects for inclusion. To ensure high data fidelity, we utilized Nvidia Isaac Sim [36], a robust robotics simulation platform and

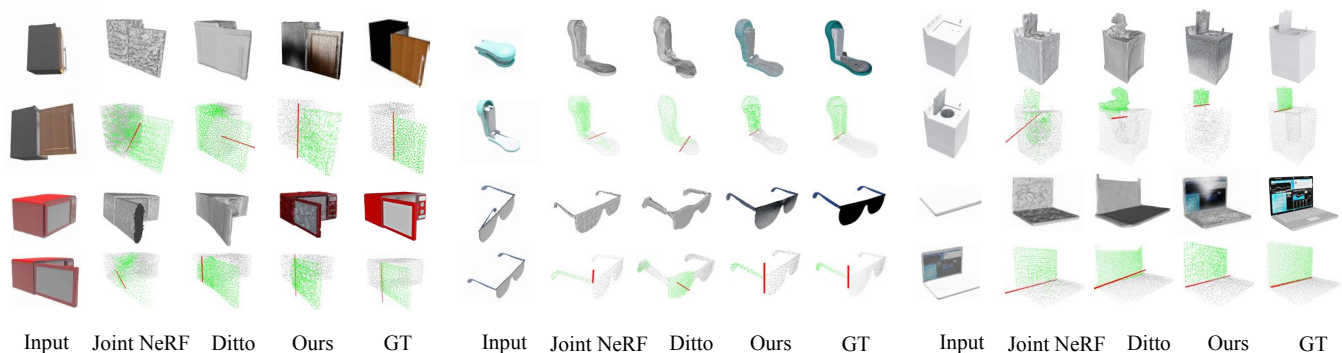


Fig. 4: Visualization of 3D reconstruction and articulation. Movable parts are shown as green point clouds for clarity.

synthetic data generation tool, to sample from the PartNet-Mobility models. We incorporated natural lighting using Nvidia’s ray tracing capabilities and simulated the Intel RealSense D435 as the RGB-D sensor within the environment, aiming to mimic real-world conditions. Each object was imported into our Isaac Sim environment, with one rotational joint axis, typically the primary axis, and its associated movable parts selected.

To achieve comprehensive sensor coverage, we implemented a systematic placement strategy. Specifically, we uniformly distributed 128 RGB-D sensors over a sphere, capturing a wide range of viewing angles, with the object positioned at the sphere’s center. The camera poses were oriented to align with the central orthogonal coordinate system of the object. Each camera generated an RGB image, a depth map, and a corresponding object mask. Subsequently, the joint axis underwent a rotation of either 90 degrees or to its maximum joint limit, transitioning the object to its post-interaction state. The sensor data capture process was then reiterated as described above.

The MMArt dataset emerges as a comprehensive multi-view, multi-modal, and multi-state resource for articulated objects, laying a robust groundwork for training and evaluation in multi-modal interactive object reconstruction tasks. Additionally, its design allows for easy conversion into specialized single-object or single-modal datasets, thereby expanding its utility and reinforcing its relevance for various tasks within this domain.

V. EXPERIMENTS

A. Evaluation Metrics

1) *Geometric Reconstruction*: To assess the quality of the reconstructed mesh for articulated objects, we employ the Chamfer Distance (CD) metric. For evaluation purposes, all baseline methods are uniformly sampled to 4096 points. We also compute the CD metric exclusively for the point cloud associated with the movable component, providing insight into the method’s precision in delineating movable parts.

2) *Joint Parameter*: Consistent with the evaluation metrics used in Ditto [18], we gauge the precision of the predicted rotational joint through a dual-metric approach. Initially, we determine the Angular Deviation (Angle Error), which captures the variance in direction between the

forecasted axis direction and the actual ground truth. Subsequently, we quantify the Axis Displacement (Position Error), a metric that pinpoints the minimal spatial separation between the anticipated rotation axis and the ground truth, while also accounting for the pivot point’s specific position.

B. Baselines

We introduce Ditto [18], a method with tasks similar to ours, and designed two multi-view image-based methods for a comprehensive evaluation of our approach.

1) *Ditto*: Ditto [18] models articulated objects based on single-view point clouds before and after interaction, primarily using implicit decoders to predict occupancy, segmentation, and joint parameters with respect to the query points. The model undergoes training with ground truth 3D structures and joint parameters. For our evaluation, we utilized a pre-trained Ditto model, further enhancing it by training on additional categories within an MMArt-transformed dataset. Subsequently, we tested it across 13 categories using point clouds derived from depth maps, aiming to rigorously assess its capabilities in a diverse range of scenarios.

2) *Arti Nvdiff*: Arti Nvdiff, developed to enrich comparisons in end-to-end method, similarly harnesses Nvdiffrec [19] to facilitate object reconstruction from multi-view RGB images via grid deformation. Subsequent to this initial reconstruction, it employs joint parameters to execute a rigid rotation of the movable part. The refinement of joint parameters and the movable part is accomplished by comparing the grid’s rendering with post-interaction multi-view images.

3) *Joint NeRF*: Joint NeRF is rooted in NeRF-based reconstruction [8] of objects both pre- and post-interaction. We employ the Marching Cube to derive meshes from radiance fields. Movable components are segmented as delineated in Section III-C. Leveraging the keypoint algorithm [37], we identify candidate joints. The most fitting candidate, ascertained by the least RGB rendering discrepancy against the ground truth, dictates the ultimate joint parameter.

C. Main Results

1) *Joint Parameter*: From the quantitative results in Table I, our method outperforms other approaches in both numerical accuracy and stability. Arti Nvdiff is at a clear

	Method	Box	Laptop	Door	Safe	Microwave	Dishwasher	Storage	Eyeglasses	Staples	Washer	Oven	Fridge	Toilet	Mean
Angle Error	Ditto [18]	3.425	3.463	1.975	2.551	2.517	4.011	4.031	3.072	3.115	1.412	4.514	5.342	2.776	2.947
	Arti Nvdiff	27.084	19.984	43.692	24.754	32.572	21.120	16.203	40.912	35.192	64.348	27.890	32.309	53.427	32.253
	Joint NeRF	7.245	0.801	0.795	13.580	11.984	9.039	19.253	82.298	49.945	59.434	6.023	40.561	62.124	27.929
	Ours	0.922	0.784	1.001	1.419	1.218	1.272	1.357	1.905	1.033	1.025	0.923	0.816	1.192	1.059
Position Error	Ditto [18]	0.203	0.050	0.177	0.201	0.179	0.276	0.186	0.073	0.161	0.224	0.173	0.291	0.230	0.158
	Arti Nvdiff	3.328	2.547	2.082	3.478	7.952	5.323	5.391	6.457	2.310	3.418	4.599	2.974	4.723	4.314
	Joint NeRF	0.337	0.084	0.010	0.493	0.512	0.329	0.957	3.429	2.605	4.253	3.507	3.185	6.632	2.025
	Ours	0.021	0.061	0.069	0.149	0.134	0.202	0.116	0.124	0.125	0.187	0.109	0.152	0.048	0.101

TABLE I: Evaluations on the absolute errors of the articulation parameter estimation.

	Method	Box	Laptop	Door	Safe	Microwave	Dishwasher	Storage	Eyeglasses	Staples	Washer	Oven	Fridge	Toilet	Mean
Whole Models	Ditto [18]	0.230	0.206	0.194	0.181	0.158	0.512	0.273	0.164	0.141	0.747	0.221	0.207	0.316	0.304
	Joint NeRF	0.142	0.085	0.088	0.109	0.105	0.113	0.112	0.107	0.097	0.100	0.095	0.106	0.120	0.106
	Ours	0.091	0.110	0.089	0.087	0.120	0.081	0.099	0.103	0.088	0.079	0.090	0.074	0.097	0.086
	Ditto [18]	0.159	0.121	0.075	0.324	0.057	1.547	0.355	0.242	0.093	6.633	0.489	0.166	0.603	0.755
Movable Parts	Arti Nvdiff	1.502	0.993	1.663	0.938	1.179	0.984	0.846	1.114	1.657	0.858	1.340	1.226	1.478	1.056
	Joint NeRF	0.865	0.522	0.508	0.743	0.512	0.962	1.136	0.968	1.044	0.624	0.793	1.114	0.988	0.787
	Ours	0.035	0.038	0.034	0.031	0.022	0.029	0.034	0.026	0.027	0.029	0.030	0.043	0.046	0.033

TABLE II: Evaluations on the Chamfer Distance of the 3D reconstruction of articulated objects.

Case #	Component	Ang. Err.	Pos. Err.	CD-Mov
1	Baseline	32.253	4.314	1.056
2	#1 + Priori-Segmentation	13.157	1.505	0.912
3	#1 + Candidate Joints	2.147	0.172	0.219
4	#2 + Candidate Joints	1.302	0.163	0.116
5	#4 + PatchImage	1.059	0.101	0.033

TABLE III: Ablation studies results of different modules.

disadvantage compared to other methods, indicating that end-to-end implicit articulated object modeling still faces convergence challenges during training. Joint NeRF is unable to achieve joint optimization makes it overly reliant on the accuracy of keypoint detection, leading to failures on objects with complex geometric structures. The results in Fig. 4 reveal that, due to intra-class morphological variations, Ditto [18] misplaces the joint direction for some objects.

2) *Geometric Reconstruction*: Table II highlights our method’s superior geometric reconstruction, surpassing Ditto and closely rivaling Joint NeRF, which leverages NeRF [8]. For movable part segmentation, the end-to-end Arti Nvdiff still struggles, while the segmentation results of Joint NeRF entirely depend on its joint estimation accuracy. Our approach excels in movable part segmentation, achieving up to 96% better chamfer distance than Ditto [18], thanks to our joint optimization strategy and image loss patch optimization, significantly boosting segmentation precision. Fig. 4 demonstrates our method’s clean segmentation with minimal noise.

D. Ablation Study

Table III demonstrates the effectiveness of each component in the proposed SM³ framework. Case #1 corresponds to the structure of Arti Nvdiff. Cases #2 and #3 highlight the significant role of Priori-Segmentation and Candidate Joints, derived from the geometric differences between pre- and post-interaction objects, in estimating the object’s motion structure. Compared to the former, Candidate Joints play a more pronounced role due to their effective constraint on the 6-DoF joint position and direction in Case #4. Case #5 shows

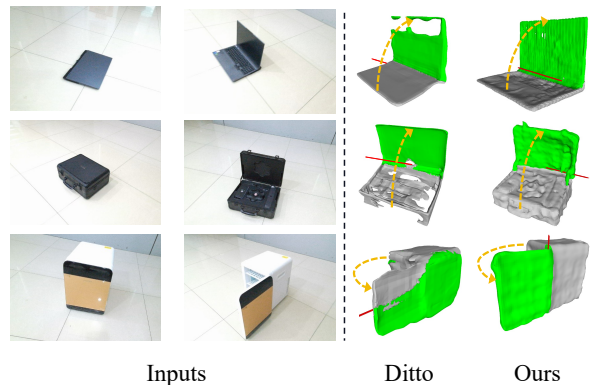


Fig. 5: Visualization results for real-world articulated objects.

that the PatchImage Loss further enhances the method’s performance, especially for movable part segmentation, by computing a more detailed image loss.

E. Real-world Results

To validate the universality of our method, we tested it in real-world scenarios. Avoiding any potential for additional information, we solely used an iPhone 14 to capture videos centered on the object and extracted 64 images from different viewpoints. Subsequently, using these images and the corresponding poses from Colmap, we applied our proposed SM³ for modeling. As shown in Fig. 5, our method consistently delivers outstanding performance across all objects.

VI. CONCLUSION

We introduce the SM³ framework, a pioneering solution that, given only multi-view images captured before and after object interaction, achieves detailed textured 3D reconstruction and motion structure analysis of articulated objects without any labels or 3D models. This encompasses the segmentation of movable components and the estimation of joint parameters. Our approach outperforms supervised methods across all categories and objects. Furthermore, we present the MMArt dataset, tailored for training and testing tasks related to articulated objects.

REFERENCES

- [1] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. Lu, "Akb-48: A real-world articulated object knowledge base," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 809–14 818.
- [2] M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, "Articulated object interaction in unknown scenes with whole-body mobile manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1647–1654.
- [3] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 190–21 200.
- [4] G. Schiavi, P. Wulkop, G. Rizzi, L. Ott, R. Siegart, and J. J. Chung, "Learning agent-aware affordances for closed-loop interaction with articulated objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5916–5922.
- [5] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.
- [6] Z. Chen, A. Tagliasacchi, and H. Zhang, "Learning mesh representations via binary space partitioning tree networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- [7] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 123–16 133.
- [8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [9] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.
- [10] Q. Xu, Z. Xu, J. Philip, S. Bi, Z. Shu, K. Sunkavalli, and U. Neumann, "Point-nerf: Point-based neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5438–5448.
- [11] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. 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*, 2019, pp. 8876–8884.
- [12] Z. Yan, R. Hu, X. Yan, L. Chen, O. Van Kaick, H. Zhang, and H. Huang, "Rpm-net: recurrent prediction of motion and parts from point cloud," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–15, 2019.
- [13] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
- [14] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3706–3715.
- [15] N. Heppert, M. Z. Irshad, S. Zakharov, K. Liu, R. A. Ambrus, J. Bohg, A. Valada, and T. Kollar, "Carto: Category and joint agnostic reconstruction of articulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 201–21 210.
- [16] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2642–2651.
- [17] H. Wang, Z. Fan, Z. Zhao, Z. Che, Z. Xu, D. Liu, F. Feng, Y. Huang, X. Qiao, and J. Tang, "Dtf-net: Category-level pose estimation and shape reconstruction via deformable template field," *arXiv preprint arXiv:2308.02239*, 2023.
- [18] Z. Jiang, C.-C. Hsu, and Y. Zhu, "Ditto: Building digital twins of articulated objects from interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5616–5626.
- [19] J. Munkberg, J. Hasselgren, T. Shen, J. Gao, W. Chen, A. Evans, T. Müller, and S. Fidler, "Extracting triangular 3d models, materials, and lighting from images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8280–8290.
- [20] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 097–11 107.
- [21] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.
- [22] J. Mu, W. Qiu, A. Kortylewski, A. Yuille, N. Vasconcelos, and X. Wang, "A-sdf: Learning disentangled signed distance functions for articulated shape representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 001–13 011.
- [23] F. Wei, R. Chabra, L. Ma, C. Lassner, M. Zollhöfer, S. Rusinkiewicz, C. Sweeney, R. Newcombe, and M. Slavcheva, "Self-supervised neural articulated shape and appearance models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 816–15 826.
- [24] B. Abbatematteo, S. Tellex, and G. Konidaris, "Learning to generalize kinematic models to novel objects," in *Proceedings of the 3rd Conference on Robot Learning*, 2019.
- [25] A. Jain, R. Lioutikov, C. Chuck, and S. 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, pp. 13 670–13 677.
- [26] A. Jain, S. Giguere, R. Lioutikov, and S. Niekum, "Distributional depth-based estimation of object articulation models," in *Conference on Robot Learning*. PMLR, 2022, pp. 1611–1621.
- [27] H. Xue, L. Liu, W. Xu, H. Fu, and C. Lu, "Omad: Object model with articulated deformations for pose estimation and retrieval," *arXiv preprint arXiv:2112.07334*, 2021.
- [28] Z. Yan, R. Hu, X. Yan, L. Chen, O. Van Kaick, H. Zhang, and H. Huang, "Rpm-net: recurrent prediction of motion and parts from point cloud," *arXiv preprint arXiv:2006.14865*, 2020.
- [29] R. Martín-Martín, C. Eppner, and O. Brock, "The rbo dataset of articulated objects and interactions," *The International Journal of Robotics Research*, vol. 38, no. 9, pp. 1013–1019, 2019.
- [30] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 909–918.
- [31] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.
- [32] J. Gao, W. Chen, T. Xiang, A. Jacobson, M. McGuire, and S. Fidler, "Learning deformable tetrahedral meshes for 3d reconstruction," *Advances In Neural Information Processing Systems*, vol. 33, pp. 9936–9947, 2020.
- [33] T. Shen, J. Gao, K. Yin, M.-Y. Liu, and S. Fidler, "Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6087–6101, 2021.
- [34] D. Lee, H. Seo, D. Kim, and H. J. Kim, "Aerial manipulation using model predictive control for opening a hinged door," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1237–1242.
- [35] B. Wandt, M. Rudolph, P. Zell, H. Rhodin, and B. Rosenhahn, "Canonpose: Self-supervised monocular 3d human pose estimation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 294–13 304.
- [36] NVIDIA, "Isaac sim." [Online]. Available: www.developer.nvidia.com/isaac-sim
- [37] J. Li and G. H. Lee, "Usip: Unsupervised stable interest point detection from 3d point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 361–370.