MARS: Multimodal Active Robotic Sensing for Articulated Characterization

Hongliang Zeng, **Ping Zhang***, **Chengjiong Wu**, **Jiahua Wang**, **Tingyu Ye** and **Fang Li**
South China University of Technology, Guangzhou, China
scutzenghongl@gmail.com, pzhang@scut.edu.cn,

Abstract

Precise perception of articulated objects is vital for empowering service robots. Recent studies mainly focus on point cloud, a single-modal approach, often neglecting vital texture and lighting details and assuming ideal conditions like optimal viewpoints, unrepresentative of real-world scenarios. To address these limitations, we introduce MARS, a novel framework for articulated object characterization. It features a multi-modal fusion module utilizing multi-scale RGB features to enhance point cloud features, coupled with reinforcement learning-based active sensing for autonomous optimization of observation viewpoints. In experiments conducted with various articulated object instances from the PartNet-Mobility dataset, our method outperformed current state-of-the-art methods in joint parameter estimation accuracy. Additionally, through active sensing, MARS further reduces errors, demonstrating enhanced efficiency in handling suboptimal viewpoints. Furthermore, our method effectively generalizes to real-world articulated objects, enhancing robot interactions. Code is available at https://github.com/robhlzeng/MARS.

1 Introduction

In an era increasingly marked by the integration of robotic assistance into everyday scenarios, conducting research on the precise perception of articulated objects, such as kitchen utensils and personal devices, is of paramount importance. These objects often have intricate joints and multiple moving parts, presenting unique and complex challenges in robotic perception and manipulation.

By precisely perceiving the joint parameters of articulated parts, robots enhance their effectiveness in manipulation planning for these objects, considering parameters such as joint position, orientation, and part state. We reviewed previous research [Yi *et al.*, 2018; Yan *et al.*, 2020; Jain *et al.*, 2021; Jiang *et al.*, 2022; Chu *et al.*, 2023] and identified limitations in existing methods. Firstly, Many studies perceive joint characteristics using a single point cloud modality, neglecting valuable information of color and texture data. Secondly, different joints, such as revolute and prismatic, necessitate separate perception networks, which can limit practical applications. Lastly, current work often assumes the availability of an ideal observation viewpoint, disregarding scenarios where the target part is obstructed or invisible. However, robots frequently face suboptimal viewing angles, hindering complete object observation.

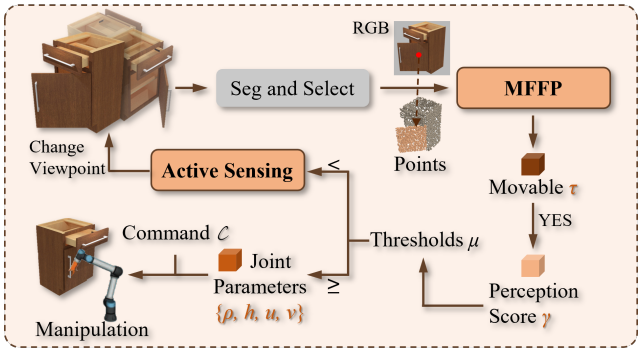

Figure 1: MARS uses active sensing to find optimal viewpoints for observing articulated objects, predicting precise joint parameters from RGB and point cloud for command-based robot planning.

To address these challenges, MARS implements a multimodal fusion strategy with a novel multi-layer dueling module, efficiently extracting and combining feature maps from diverse convolutional layers into a more effective, information-dense image feature representation. Subsequently, MARS integrates multi-scale RGB features and point cloud features via a transformer encoder-based fusion module. Moreover, MARS standardizes the description of joint parameters, allowing for the perception of various joint types through a single network. Addressing the issues posed by suboptimal viewing angles, MARS features a Reinforcement Learning (RL) driven active sensing strategy. This approach authorizes dynamic camera position adjustment, ensuring acquisition of the most informative viewpoint.

Ultimately, we introduce a comprehensive task flow for robotic perception and manipulation of articulated objects, as shown in Fig. 1. The process starts with the robot selecting an object part, then inputting corresponding RGB images and point cloud data into the perception network to determine the

*Corresponding author

part's movability and identify joint parameters. If deemed movable and the perception score exceeds a threshold, the robot plans a manipulation sequence based on the identified joint parameters as well as the operation command. Conversely, when suboptimal viewing angles are detected, the robot adjusts its viewpoint for further assessment, ensuring reliable, precise interactions with the articulated objects.

To rigorously test our model, experiments were conducted on Sapien simulation platform [Xiang *et al.*, 2020] using PartNet-Mobility [Mo *et al.*, 2019]. Our method showed significant advancements, evident in notable performance improvements over existing benchmarks. Key contributions are summarized as follows.

- We developed a multimodal feature fusion technique that significantly enhances point feature representation by integrating multi-scale image details, thus substantially improving representational capability.
- We developed an active sensing technique grounded in reinforcement learning, empowering the robot to autonomously optimize its camera position to capture the most informative viewpoint.
- MARS achieved state-of-the-art in joint parameter estimation for diverse articulated objects. It effectively perceives both revolute and prismatic joints using a single network, validated in real-world applications.

2 Related Works

Articulated Object Characterization. To enhance robotic capabilities in perceiving and manipulating articulated objects, a wealth of simulators [Todorov *et al.*, 2012; Xiang *et al.*, 2020; Szot *et al.*, 2021] and open datasets [Chang *et al.*, 2015; Mo *et al.*, 2019; Wang *et al.*, 2019; Geng *et al.*, 2023] have emerged as essential resources. These tools have facilitated advancements in 3D reconstruction [Li *et al.*, 2020; Bozic *et al.*, 2021; Yang *et al.*, 2021], joint parameter estimation [Jain *et al.*, 2021; Shi *et al.*, 2021; Zeng *et al.*, 2021; Jiang *et al.*, 2022; Chu *et al.*, 2023], and the prediction of interactive positions and trajectories [Mo *et al.*, 2021; Wu *et al.*, 2021; Wang *et al.*, 2022] for articulated objects.

In joint parameter estimation, prior work [Yi *et al.*, 2018; Abbatematteo *et al.*, 2019; Shi *et al.*, 2021; Jain *et al.*, 2021; Siarohin *et al.*, 2021; Jiang *et al.*, 2022] has leveraged multi-view observations from ongoing monitoring as visual inputs, capitalizing on the rich visual cues provided by the changing joint states. However, this method complicates data collection, limiting its practicality in real-world robotics. Furthermore, many studies [Yan *et al.*, 2020; Jiang *et al.*, 2022; Chu *et al.*, 2023] rely exclusively on point cloud data, neglecting the valuable features available in RGB imagery that can enhance joint parameter estimation. Our methodology focuses on employing a single RGB image along with point cloud data to achieve accurate perception of joint parameters.

Multimodal Feature Fusion. In the realm of multimodal feature fusion, especially in combining RGB and point cloud data, current research predominantly focuses on 3D object detection [Ku *et al.*, 2018; Sindagi *et al.*, 2019; Zhu *et al.*, 2021; Piergiovanni *et al.*, 2021; Wu *et al.*, 2022; Zhang *et al.*, 2022]. However, due to differences in input data, these fusion methods often struggle to be directly applicable and effective in joint analysis. A notable example similar to our approach is EPNet [Huang *et al.*, 2020], which enhances point cloud features by integrating image features into the point cloud domain, utilizing both local and global contexts of images. In contrast, our method employs a competitive mechanism to dynamically prioritize the significance of image features across various scales.

Active Sensing. In practical applications, the challenging task of obtaining optimal imaging angles for robots [Ammirato *et al.*, 2017; Zhao *et al.*, 2022] underscores the need for active sensing. This technique entails dynamically adjusting sensor positions or the manipulation environment to enhance data acquisition, proving essential for certain tasks [Han *et al.*, 2019; Mattamala *et al.*, 2021; Safronov *et al.*, 2021], especially in scenarios where initial views are insufficient due to partial occlusions or limited visibility of articulated objects. In current studies, researchers commonly assume that high-quality images and point clouds can be obtained from an ideal viewpoint [Jain *et al.*, 2021; Yang *et al.*, 2021; Jiang *et al.*, 2022; Chu *et al.*, 2023], but such assumptions frequently fail to align with the reality. Our approach successfully adapts to complex real-world application scenarios by incorporating active sensing techniques.

3 Method

We introduce MARS, a framework designed for estimating joint parameters in articulated objects. As depicted in Fig. 2, MARS consists of two primary components: Multimodal Feature Fusion Perception (MFFP) and Active Sensing (AS).

The MFFP component of MARS utilizes ResNet18 [He *et al.*, 2016] and PointNet++ [Qi *et al.*, 2017] as backbone networks to efficiently extract features from the input RGB image and point cloud data. The Multi-Layer Dueling Module (MLDM) is designed to strategically extract and weigh image features across different scales. Following this, image and point cloud features are merged at the feature level via a specialized fusion module. A decoding module then processes these integrated features to produce outputs for joint parameters and a perception score, reflecting the efficacy of the current viewpoint.

The AS module operates based on the perception score and a set threshold. If the score is below this threshold, it triggers a viewpoint change. A new observation position is determined from the action space, followed by a repeated perception process. This approach enables real-time adjustments for optimal data acquisition and improved joint parameter estimation in complex or obstructed scenarios.

3.1 MLDM

Employing ResNet, we capturing a range of articulated object RGB features from local to global, as depicted in Fig. 3a. For each feature map $f_r^i \in \mathbb{R}^{C^i \times H^i \times W^i}$ with C^i channels and spatial dimensions $H^i \times W^i$, and for point cloud features $f_p^j \in \mathbb{R}^{N^j \times K}$ comprising N^j points each of K dimensions, we perform point-wise convolution and subsequent pooling operations to aggregate the features (as shown in Fig. 3b).

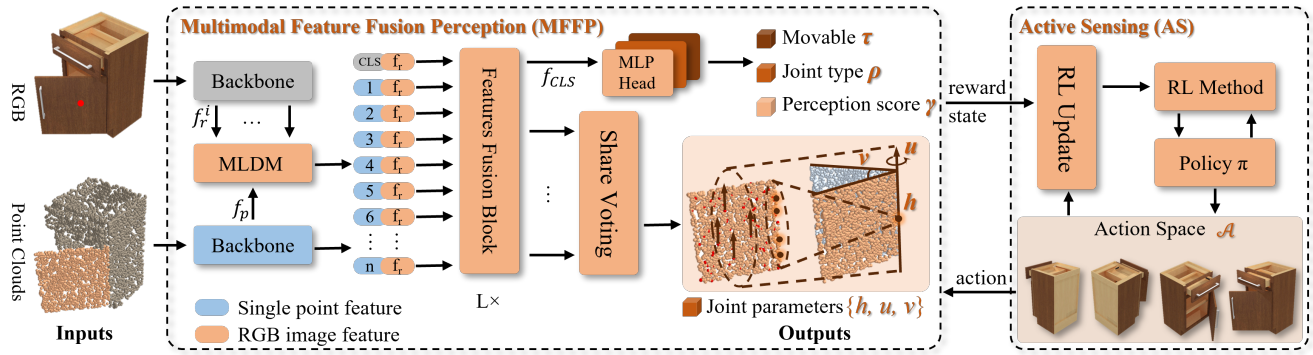

Figure 2: MARS Framework with MFFP and AS components. MFFP integrates RGB and point cloud data, utilizing MLDM for adaptive RGB feature scaling (see Fig. 3) and a Fusion Block for combining features, aiding in joint parameters prediction. AS adjusts viewpoints under suboptimal conditions, enhancing perception accuracy in real-world scenarios.

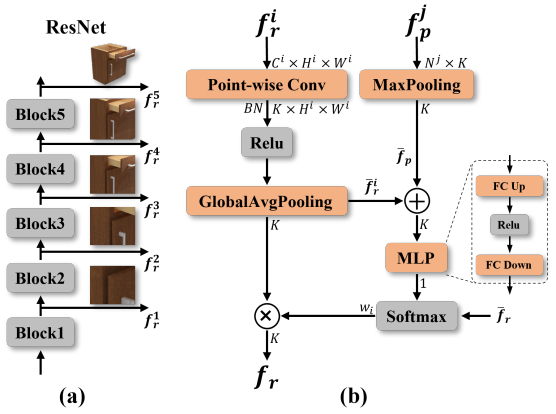

Figure 3: MLDM Architecture for RGB Image Feature Aggregation. (a) Image feature maps f_r^i are extracted at multiple scales from ResNet blocks. (b) These feature maps are first combined with point cloud feature f_p^j, upon which adaptive weights w_i are computed to form the final weighted image feature representation f_r.

The aggregated image feature $\bar{f}_r^i \in \mathbb{R}^K$ is computed as follows:

$$\bar{f}_r^i = \mathbf{g}\left(\delta\left(\mathcal{B}\left(\mathrm{PWConv}(f_r^i)\right)\right)\right), \tag{1}$$

where $\mathbf{g}$ is the global average pooling, δ is the ReLU function [Nair and Hinton, 2010], and $\mathcal{B}$ represents BN [Ioffe and Szegedy, 2015], with the kernel size for the PWConv being $K \times 1 \times 1$. For point cloud feature aggregation, the max pooling operation $\mathbf{m}$ is utilized:

$$\bar{f}_p = \mathbf{m}\left(f_p^j\right). \tag{2}$$

We combine the aggregated features $\bar{f}_r^i$ and $\bar{f}_p$, and use a Multi-Layer Perceptron (MLP) to determine the weights for the final RGB image feature $f_r \in \mathbb{R}^K$. The computation is given by:

$$f_r = \sum_{i=1}^{n}\left(\bar{f}_r^i \otimes \sigma\left(\theta_w\left(\bar{f}_r^i + \bar{f}_p\right)\right)\right), \tag{3}$$

where σ represents the softmax function, and θ_w denotes the learnable parameters within the MLP. Comprising two linear layers with a ReLU activation between them, the MLP scales up and then reduces the dimensions of the fused feature to compute the weight $w_i \in \mathbb{R}^1$. The operation $\otimes$ indicates element-wise multiplication. The weights from all blocks are then combined to form the final weight distribution w.

3.2 Feature Fusion Block

This component leverages transformer encoders [Vaswani *et al.*, 2017], intentionally excluding positional embedding due to the intrinsic spatial information of the RGB and point cloud features. The input tokens $t^j = f_p^j \oplus f_r \in \mathbb{R}^{N^j \times 2K}$ for the fusion module are formed by concatenating image feature to each point feature. A CLS token t^{CLS} is also incorporated to encapsulate global features, which is pivotal for the model's generalization. The Feature Fusion Block outputs a global feature $f^{\mathrm{CLS}} \in \mathbb{R}^{2K}$ and local features $\widetilde{f}_p^j \in \mathbb{R}^{N^j \times 2K}$ for each point, expressed as:

$$\{f^{\mathrm{CLS}}, \widetilde{f}_p^j\} = \mathbf{F}_L(\mathbf{F}_{L-1}(\ldots \mathbf{F}_1(t^{\mathrm{CLS}}, t^j)\ldots)), \tag{4}$$

where $\mathbf{F}_l$ is the l-th layer of the fusion block.

3.3 Articulation Decoders

We utilize the global feature f_{CLS} to determine the movability of the chosen rigid part, represented as a binary variable $\tau \in 0, 1$, where $\tau = 0$ signifies an immovable part and $\tau = 1$ a movable part. An MLP head dedicated to this task decodes this parameter:

$$\tau = \theta_{mov}\left(f_{\mathrm{CLS}}\right) \tag{5}$$

Joint Parameters. Upon determining the movability of the target part, we assume a fully closed position as the initial state to estimate the joint parameters. These include the joint type $\rho \in 0, 1$, where $\rho = 0$ indicates a revolute joint and $\rho = 1$ a prismatic joint, the joint position $h \in \mathbb{R}^3$, orientation $u \in \mathbb{R}^3$, and the current state $v \in \mathbb{R}$. Unlike previous models that omit position predictions for prismatic joints [Jiang *et al.*, 2022; Chu *et al.*, 2023], our framework accommodates both revolute and prismatic joints, identifying the centroid of the movable part as the position for the prismatic joint. The joint type ρ is deduced from f_{CLS} using an MLP head. For the joint parameters, a shared voting module leverages each point

$p^j \in \mathbb{R}^{N^j \times 3}$ and its corresponding feature $\widetilde{f}_p^j \in \mathbb{R}^{N^j \times 2K}$ for point-wise voting to infer h^j, u^j, and v^j. The final joint parameters are then the mean of these votes:

$$\begin{aligned} \rho &= \theta_{type}\left(f_{\text{CLS}}\right), \\ \{h^j, u^j, v^j\} &= \theta_{para}\big(p^j, \widetilde{f}_p^j\big), \\ \{h, u, v\} &= \frac{1}{N}\{\sum_{j=1}^{N} h^j, \sum_{j=1}^{N} u^j, \sum_{j=1}^{N} v^j\}. \end{aligned} \tag{6}$$

Perception Score. To assess the quality of perception results, we use an additional MLP head with f_{CLS} as the input. This module predicts the likelihood of successful perception $\gamma \in (0, 1)$ in the following manner:

$$\gamma = \zeta(\theta_{score}(f_{\text{CLS}})), \tag{7}$$

where ζ is the sigmoid activation function, ensuring γ falls between 0 and 1. A success threshold of 0.5 is set for binary decision-making within the module.

Loss Functions. To supervise the predictions of movability and joint type, binary cross-entropy loss functions $\mathcal{L}_{mov}$ and $\mathcal{L}_{type}$ are utilized. For joint state prediction, an L1 norm loss function $\mathcal{L}_{state}$ is employed. As the perception score prediction has been transformed into a binary classification task for assessing viewpoint optimality, it is supervised using binary cross-entropy loss $\mathcal{L}_{score}$. To penalize the discrepancy in orientation between the estimated joint and the ground truth $\hat{u}$, which is a unit vector, the loss $\mathcal{L}_{ori}$ is defined as:

$$\mathcal{L}_{ori} = \frac{1}{N}\sum_{j=1}^{N} \arccos\left(\hat{u} \cdot u^j\right), \tag{8}$$

The loss $\mathcal{L}_{pos}$ penalizes the distance between the estimated projection point h^j and the actual joint axis $\hat{u}$:

$$\mathcal{L}_{pos} = \frac{1}{N}\sum_{j=1}^{N} ||(h^j - \hat{h}) \times \hat{u}||. \tag{9}$$

Training Steps. In the initial training phase, we pre-train the model parameters using the movability loss $\mathcal{L}_{mov}$ along with the corresponding decoding head. This step is crucial for ensuring efficient learning of movability prediction. For joint parameter prediction, the model is trained to minimize the differences between the predicted and the actual ground truth values. The cumulative loss for this training phase is given by:

$$\mathcal{L}_{para} = \mathcal{L}_{type} + \mathcal{L}_{ori} + \mathcal{L}_{pos} + \mathcal{L}_{state} \tag{10}$$

In the final training stage, parameters except for the perception scoring decoder are fixed, with the training focused on optimizing the scoring decoder using the loss $\mathcal{L}_{score}$.

3.4 The RL Policy For Active Sensing

Training of a conditional RL policy, using a DQN approach [Mnih *et al.*, 2015], follows the completion of perception module training. This policy, aimed at active viewpoint optimization, is trained within a simulated environment. Success in a training iteration occurs when the perception score exceeds 0.5; over 5 action steps indicate failure. Details of the RL policy are provided in the following section.

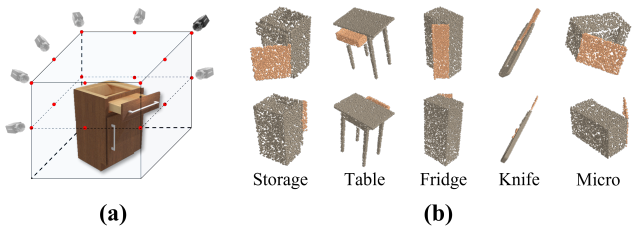

Figure 4: (a) Sixteen discrete positions constitute the discrete action space for viewpoint selection. (b) Comparative analysis of point cloud quality obtained from different viewpoints.

State Space. In the state space, a pre-trained perception network processes an RGB image and point cloud from the current viewpoint, yielding the global feature f_{CLS}. Additionally, the position of camera is represented as a one-hot encoded vector $x \in \mathbb{R}^{16}$, indicating one of 16 distinct positions. The state input for the RL algorithm is formed by the combination of f_{CLS} and this one-hot encoded position x.

Action Space. We established a discrete action space $\mathcal{A} \in \{0, 1\}^{16}$ with 16 viewpoints around the object. This action space aligns with the coordinate framework of the object, as depicted in Fig. 4a. To better simulate a real environment, some locations will be randomly set as unreachable.

Reward Design. The per-step rewards r_{step} are determined by two primary criteria: perception score and point cloud quantity variation. To balance these aspects, the step reward is defined as:

$$r_{step} = \lambda_s r_{score} + \lambda_n r_{num}, \tag{11}$$

where λ_s and λ_n are weight coefficients that adjust the relative importance of r_{score} and r_{num} in the reward. Here, $r_{score} = s - s'$ represents the change in perception scores before and after an action, and the point cloud variation reward is given by:

$$r_{num} = \frac{n - n'}{n'} - 1, \tag{12}$$

where n denotes the number of point clouds. Additionally, at the end of a round, a positive reward of $+10$ is assigned for successful task completion. Conversely, a negative reward of -10 is incurred if the task fails due to exceeding the action step limit.

3.5 Command-Based Point Cloud Manipulation

We use perceived joint parameters for point cloud manipulations based on operational commands $\mathcal{C}$. The system modifies the joint state in response to $\mathcal{C}$, involving adjustments such as angle and position changes. Specifically, $\mathcal{C}$ specifies the target state for the joint, and we calculate the difference between this target and the current perceived joint state as the operational compensation $\Delta v = \mathcal{C} - v$. Based on Δv, joint orientation u, and joint position h, we compute a rotation-translation matrix $\mathcal{M}$ to represent the current operation. The matrix is calculated as follows:

$$\mathcal{M} = \begin{bmatrix} R(\Delta v, u) & T(\Delta v, h) \\ 0 & 1 \end{bmatrix}, \tag{13}$$

Methods	Revolute										Prismatic				Avg
	Laptop	Chair	Pliers	Safe	Eyeglass	Fridge	Scissor	Door	Micro	Oven	Table	Storage	Window	Knife	
Errors of Joint Position															
RPM-Net	0.12	0.10	0.13	0.16	0.18	0.14	0.17	0.25	0.16	0.21	-	-	-	-	0.16/-
ANCSH	0.08	-	-	-	0.06	-	-	-	0.06	0.12	-	-	-	-	0.08/-
Ditto	0.03	**0.02**	**0.03**	0.06	0.04	0.06	0.04	0.06	**0.03**	0.03	-	-	-	-	0.04/-
Cart	0.03	0.03	**0.03**	0.06	**0.03**	**0.04**	0.03	0.06	0.05	0.03	-	-	-	-	0.04/-
EPNet*	0.03	0.03	0.04	**0.05**	0.05	0.05	0.03	0.07	**0.03**	0.04	0.15	0.25	0.04	0.03	0.04/0.12
MFFP(Ours)	**0.02**	**0.02**	**0.03**	**0.05**	**0.03**	0.05	**0.02**	**0.05**	**0.03**	**0.02**	**0.13**	**0.21**	**0.02**	**0.01**	**0.03/0.09**
w/o MLDM	0.03	0.03	**0.03**	0.06	0.04	0.05	**0.02**	0.06	**0.03**	0.04	0.15	0.22	0.04	0.03	0.04/0.11
Errors of Joint Orientation															
RPM-Net	10.73°	11.69°	7.83°	12.64°	20.52°	6.27°	4.85°	16.38°	7.42°	5.28°	17.37°	9.63°	7.84°	5.92°	10.36°/10.19°
ANCSH	3.92°	-	-	-	5.87°	-	-	-	3.64°	3.15°	-	7.74°	-	-	4.15°/7.74°
Ditto	1.52°	2.54°	1.42°	2.98°	1.87°	2.21°	1.33°	1.96°	1.30°	1.70°	1.83°	4.74°	2.73°	1.76°	1.88°/2.77°
Cart	1.45°	2.78°	1.23°	2.04°	1.82°	**2.03°**	1.26°	2.48°	1.26°	1.78°	1.78°	4.03°	2.56°	1.69°	1.81°/2.52°
EPNet*	0.94°	1.69°	1.15°	1.53°	1.27°	4.38°	1.16°	1.79°	0.83°	1.02°	2.37°	4.14°	0.18°	0.16°	1.58°/1.71°
MFFP(Ours)	**0.08°**	**0.44°**	**0.06°**	**0.24°**	**0.39°**	3.36°	**0.05°**	**1.51°**	**0.28°**	**0.10°**	**0.42°**	**2.96°**	**0.04°**	**0.04°**	**0.65°/0.87°**
w/o MLDM	0.56°	1.78°	0.95°	1.24°	1.72°	4.16°	0.68°	3.23°	2.24°	0.87°	1.08°	4.27°	0.25°	0.56°	1.74°/1.54°
Errors of Joint State															
RPM-Net	15.38°	17.99°	14.66°	9.96°	18.85°	16.45°	8.24°	25.78°	19.91°	14.24°	0.27	0.31	0.38	0.11	16.15°/0.27
ANCSH	10.15°	-	-	-	15.74°	-	-	-	14.72°	9.03°	-	0.24	-	-	12.41°/0.24
Ditto	3.98°	4.98°	3.89°	4.18°	7.73°	6.14°	3.05°	13.53°	6.80°	3.86°	0.05	**0.07**	0.16	0.02	5.81°/0.08
Cart	3.64°	5.53°	3.14°	4.20°	6.43°	5.83°	2.79°	**11.07°**	**3.74°**	3.18°	**0.04**	0.08	0.17	0.02	4.96°/0.08
EPNet*	1.96°	3.79°	2.37°	3.53°	3.79°	**5.25°**	2.53°	13.49°	5.53°	2.64°	0.09	0.08	0.18	0.02	4.49°/0.09
MFFP(Ours)	**1.21°**	**3.20°**	**1.98°**	**2.38°**	**3.13°**	5.30°	**1.94°**	12.03°	4.88°	**1.34°**	0.05	**0.07**	**0.13**	**0.01**	**3.74°/0.07**
w/o MLDM	1.79°	3.78°	2.05°	2.94°	4.24°	5.80°	2.39°	15.48°	5.51°	2.28°	0.06	0.08	0.20	0.02	4.63°/0.09

Table 1: Quantitative evaluation of joint parameters estimation across 14 categories. The best results are in bold, with our method demonstrating superior performance in most categories.

where $R(\Delta v, u)$, the rotation matrix, is derived from Δv and joint orientation u. $T(\Delta v, h)$, the translation vector, is calculated from Δv and joint position h.

4 Experimental Evaluation

We evaluated perception capabilities of MARS for articulated objects. Quantitative assessment across various object categories confirmed accuracy of the MFFP module in estimating joint parameters. Integration of active sensing for viewpoint optimization significantly enhanced algorithm performance. Visual demonstrations of command-based point cloud manipulation and qualitative showcasing of method effectiveness on real-world objects were also conducted.

4.1 Experimental Setup

Datasets. For evaluation, we utilized the SAPIEN simulator [Xiang *et al.*, 2020] and PartNet-Mobility dataset [Mo *et al.*, 2019], selecting 14 common articulated objects (10 with revolute and 4 with prismatic joints). In the simulator, these objects, with randomized joint states and camera positions, generated various viewpoint samples (see Fig. 4b). Post movability prediction module training, immovable parts data was removed, resulting in 10K training, 1K testing, and 1K validation samples for each category for perception network training. Image data was captured at 600 $\times$ 600 resolution using an RGB-D camera.

Baselines. We benchmarked our method against six approaches. RPM-Net [Yan *et al.*, 2020] employs recurrent neural networks for predicting object motion from point clouds. ANCSH [Li *et al.*, 2020] focuses on joint parameter estimation in canonical object space. Ditto [Jiang *et al.*, 2022] utilizes multi-view for articulated object understanding. Cart [Chu *et al.*, 2023], the current state-of-the-art (SOTA), specializes in joint parameter estimation. To facilitate prediction of the current joint state by Cart, a 'Closed' command was issued in each evaluation. EPNet [Huang *et al.*, 2020], a multimodal fusion approach for 3D object detection, was adapted by replacing our MLDM module with its LiDAR-guided Image Fusion (LI-Fusion) module, enabling EPNet* to estimate joint parameters. Additionally, we included an ablated version of MARS lacking MLDM for comparison. Lastly, we proportionally increased poor viewpoints data for all methods to ensure a fair comparison.

Evaluation Metrics. Our investigation primarily focuses on the estimation errors of articulated object joint parameters. Specifically, we measure errors in estimating joint orientation, joint position, and the current joint state of the selected part relative to its fully closed initial state.

RL Environment. To train and validate active sensing strategies, a reinforcement learning environment was constructed. In each training round, an object is randomly imported into the simulation environment, and a camera posi-

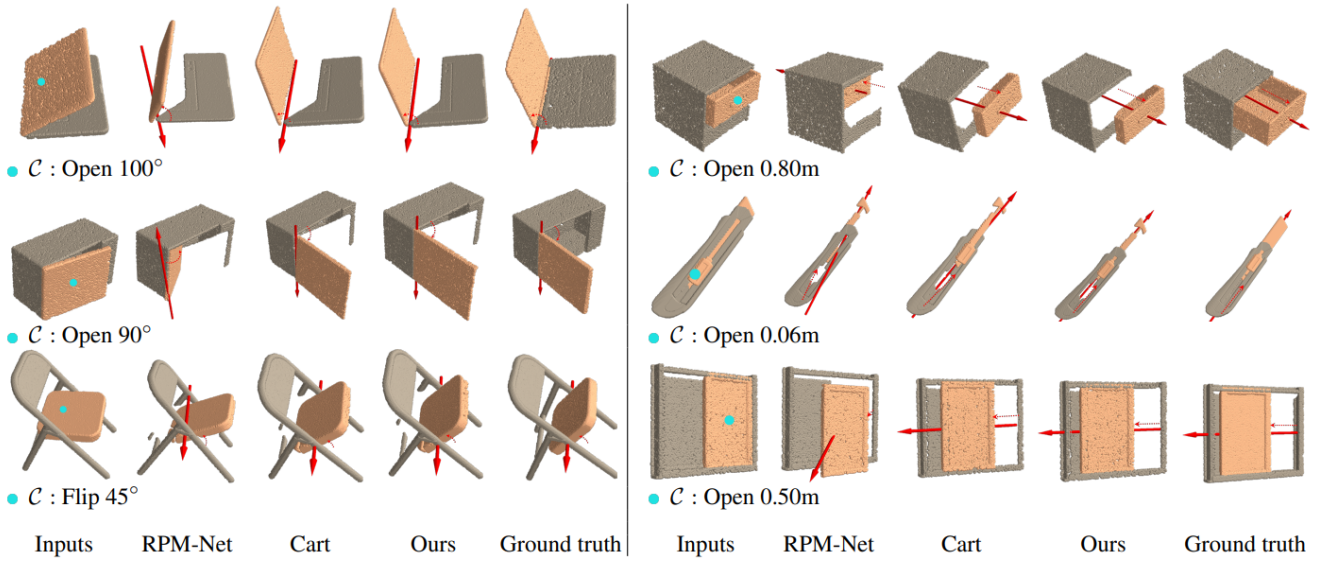

Figure 5: Comparison of point cloud-level manipulation visualizations, where blue dots represent selected parts and $\mathcal{C}$ denotes the current manipulation command.

tion is initialized within the action space with a step size limit of 5. For testing purposes, the step size is set to 1. If the initial sensing score falls below a predefined threshold, the robot selects a new observation position based on the action strategy.

4.2 Main Results.

Comparison of joint parameters estimation. Table 1 shows the joint parameter estimation results from our quantitative evaluation, where our method surpasses the SOTA in most categories and closely matches it in the rest. This superior performance can be attributed to our robust feature representation, a result of the multimodal fusion approach that effectively utilizes the rich information in RGB images to enhance point cloud features. Additionally, in comparison to EPNet*, our MLDM exhibits enhanced multi-scale feature extraction capabilities, significantly improving articulated object perception. It's important to note that joint state estimation is the most challenging, with the largest errors. Our method leads in performance, yet encounters a 3.74° error in revolute joints and 0.07m in prismatic joints. The large error mainly arises from poor observation perspectives in the acquired data, as depicted in Figure 4b. Practically, robots often encounter such angles, highlighting the importance of active sensing to adjust the viewpoint.

Ablation Studies. Table 1 demonstrates evaluation of the MLDM module impact on performance. Across all categories, the complete version outperformed the ablated version in joint parameter estimation. Notably, even the ablated version marginally surpassed other approaches, highlighting the effectiveness of multimodal fusion with RGB data combined with limited point cloud input.

Harmonized Training for Mixed Joint Types. To address the need for separate models for revolute and prismatic joints in current approaches, joint parameter representation was standardized by defining prismatic joint positions at centroids of movable parts. As demonstrated in Table 2, with balanced training samples for both joint types, 100% accuracy was achieved in joint type prediction. Performance slightly declined compared to models trained on each joint type separately, yet notable results were still achieved in quantitative joint parameter evaluation.

Class		Type acc	Position	Orientation	State
Revolute	Micro	100%	0.03	0.35°	4.05°
	Oven	100%	0.03	0.27°	1.37°
Prismatic	Table	100%	0.14	2.63°	0.06
	Storage	100%	0.25	4.57°	0.07

Table 2: Results from harmonized training on mixed joint types

Visualization results. We visualized and evaluated point cloud level manipulations by first specifying the desired joint state as a command, such as 'open 100°'. Then, we executed the corresponding manipulation on the point cloud of the selected part using the rotation-translation matrix computed as per Equation (13). Fig. 5 compares the performance of RPM-Net [Yan *et al.*, 2020], Cart [Chu *et al.*, 2023], and our method against the ground truth, with blue dots indicating the selected parts. Overall, our method more accurately reflects the real situation post-manipulation. In comparison, Cart tends to have larger state estimation errors, resulting in poorer maneuvering performance. RPM-Net often incorrectly predicts the joint direction, leading to significant discrepancies with actual ground conditions.

4.3 Enhancing Results through Active Sensing

We compared the performance of our method with and without the active sensing module. To expedite testing, we pre-collected samples from each articulated object instance at 16

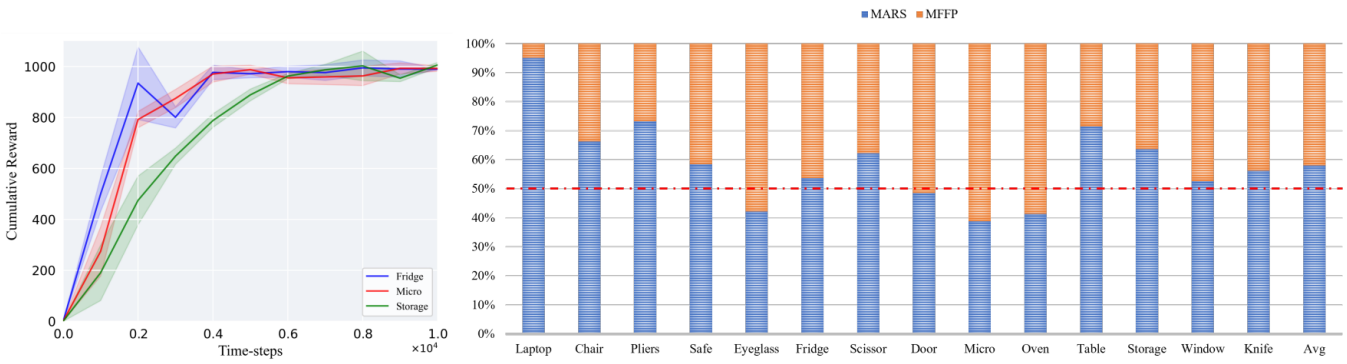

Figure 6: (a)Reward trends across there categories;(b)Bar chart showing joint state error reduction using active sensing in MARS versus MFFP, with MFFP normalized to 100% and the red line at 50%.

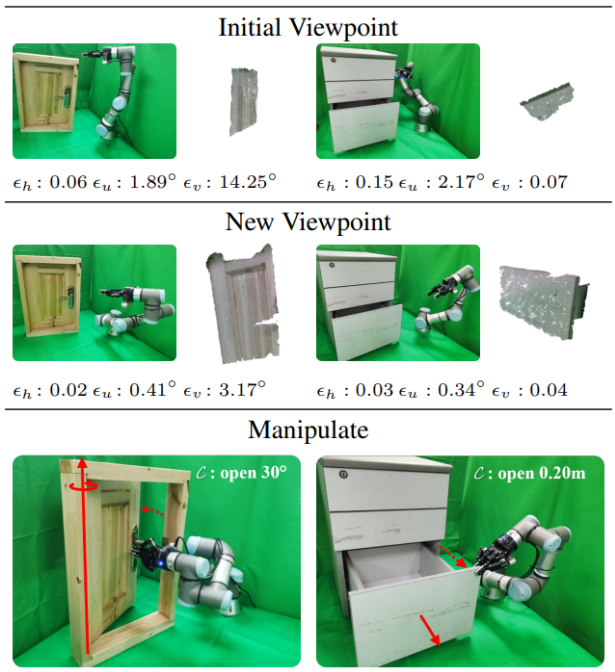

Figure 7: Real world Experiments.

camera positions after randomizing joint states. For each instance, the initial camera viewpoint provided a perception result to establish MFFP performance. A single active sensing adjustment informed by the perception score followed, marking the outcome of this camera position change as MARS performance. Fig. 6a charts the learning curves for active sensing in three categories, with cumulative rewards increasing and eventually converging over time, a pattern echoed across other categories not depicted. Fig. 6b presents a performance comparison of the two models, where the estimation error for the most challenging joint state in each articulated object class is normalized to 100% for MFFP. Bar graphs subsequently illustrate the proportional error reduction achieved by MARS. Overall, MARS achieved an average error reduction of approximately 50% across all categories, particularly notable in categories of articulated objects with larger sizes and inherent self-occlusion issues. This analysis underscores the efficacy of active sensing in enhancing the accuracy of our method, demonstrating its vital role in complex joint parameter estimation.

4.4 Real-world Experiments

To validate the generalizability of our method in real-world settings, we selected two articulated objects: a door and a table with drawers. As depicted in Fig. 7, we used an Intel RealSense RGB-D camera on a mobile robot to capture RGB and point cloud data of these objects. Segmentation of the rigid parts was achieved using 3D U-Net [Choy *et al.*, 2019]. The perception scoring module assessed the input viewpoint quality. If the score fell below the threshold, active sensing guided the robot to a new position ID for additional data acquisition. Once the perception score exceeded the threshold, the robot formulated its action plan based on the perception results and specified commands. Interaction positions between the robot and the object were manually assigned, leading to the successful manipulation of the target object.

5 Conclusion

In this paper, we introduced MARS, a multimodal framework specifically designed for accurately sensing joint parameters of articulated objects. Central to this framework is the MLDM, an innovative approach for adaptive multiscale feature fusion that significantly enhances image feature representation. MARS utilizes a transformer encoder, devoid of positional embedding, to effectively integrate RGB features with point cloud data.The significant advancement in this work is the reinforcement learning-based active perception strategy, empowering robots to autonomously seek new perspectives and substantially improve practical applicability in response to inadequate perception. Future research aims to enhance MARS by seeking more powerful point cloud representation capabilities and improving algorithmic generalization to cover a broader range of articulated objects.

Acknowledgments

This work was supported by the Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016).

References

[Abbatematteo *et al.*, 2019] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *Proceedings of the 3rd Conference on Robot Learning*, 2019.

[Ammirato *et al.*, 2017] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Košecká, and Alexander C. Berg. A dataset for developing and benchmarking active vision. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1378–1385, 2017.

[Bozic *et al.*, 2021] Aljaz Bozic, Pablo Palafox, Michael Zollhofer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1450–1459, 2021.

[Chang *et al.*, 2015] 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.

[Choy *et al.*, 2019] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019.

[Chu *et al.*, 2023] Ruihang Chu, Zhengzhe Liu, Xiaoqing Ye, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, and Jiaya Jia. Command-driven articulated object understanding and manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8813–8823, 2023.

[Geng *et al.*, 2023] 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.

[Han *et al.*, 2019] Xiaoning Han, Huaping Liu, Fuchun Sun, and Xinyu Zhang. Active object detection with multistep action prediction using deep q-network. *IEEE Transactions on Industrial Informatics*, 15(6):3723–3731, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2020] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[Jain *et al.*, 2021] 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)*, pages 13670–13677. IEEE, 2021.

[Jiang *et al.*, 2022] 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.

[Ku *et al.*, 2018] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[Li *et al.*, 2020] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3706–3715, 2020.

[Mattamala *et al.*, 2021] Matias Mattamala, Milad Ramezani, Marco Camurri, and Maurice Fallon. Learning camera performance models for active multi-camera visual teach and repeat. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14346–14352, 2021.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Mo *et al.*, 2019] 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)*, June 2019.

[Mo *et al.*, 2021] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[Piergiovanni *et al.*, 2021] AJ Piergiovanni, Vincent Casser, Michael S Ryoo, and Anelia Angelova. 4d-net for learned multi-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15435–15445, 2021.

[Qi *et al.*, 2017] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30:5099–5108, 2017.

[Safronov *et al.*, 2021] Evgenii Safronov, Nicola Piga, Michele Colledanchise, and Lorenzo Natale. Active perception for ambiguous objects classification. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4437–4444, 2021.

[Shi *et al.*, 2021] Yahao Shi, Xinyu Cao, and Bin Zhou. Self-supervised learning of part mobility from point cloud sequence. In *Computer Graphics Forum*, volume 40, pages 104–116. Wiley Online Library, 2021.

[Siarohin *et al.*, 2021] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021.

[Sindagi *et al.*, 2019] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019.

[Szot *et al.*, 2021] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021.

[Todorov *et al.*, 2012] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033. IEEE, 2012.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2019] 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.

[Wang *et al.*, 2022] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *European Conference on Computer Vision*, pages 90–107. Springer, 2022.

[Wu *et al.*, 2021] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *arXiv preprint arXiv:2106.14440*, 2021.

[Wu *et al.*, 2022] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427, 2022.

[Xiang *et al.*, 2020] 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 Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Yan *et al.*, 2020] 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 preprint arXiv:2006.14865*, 2020.

[Yang *et al.*, 2021] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15980–15989, 2021.

[Yi *et al.*, 2018] Li Yi, Haibin Huang, Difan Liu, Evangelos Kalogerakis, Hao Su, and Leonidas Guibas. Deep part induction from articulated object pairs. *arXiv preprint arXiv:1809.07417*, 2018.

[Zeng *et al.*, 2021] Vicky Zeng, Tabitha Edith Lee, Jacky Liang, and Oliver Kroemer. Visual identification of articulated object parts. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2443–2450. IEEE, 2021.

[Zhang *et al.*, 2022] Zehan Zhang, Yuxi Shen, Hao Li, Xian Zhao, Ming Yang, Wenming Tan, ShiLiang Pu, and Hui Mao. Maff-net: Filter false positive for 3d vehicle detection with multi-modal adaptive feature fusion. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 369–376. IEEE, 2022.

[Zhao *et al.*, 2022] Qianfan Zhao, Lu Zhang, Lingxi Wu, Hong Qiao, and Zhiyong Liu. A real 3d embodied dataset for robotic active visual learning. *IEEE Robotics and Automation Letters*, 7(3):6646–6652, 2022.

[Zhu *et al.*, 2021] Ming Zhu, Chao Ma, Pan Ji, and Xiaokang Yang. Cross-modality 3d object detection. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 3772–3781, 2021.