ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890874

Generalizable Articulated Object Perception with Superpoints

Qiaojun Yu[1*], Ce Hao[2*], Xibin Yuan[1*], Li Zhang[3], Liu Liu[4], Yukang Huo[5], Rohit Agarwal[6], and Cewu Lu[1+]

***Abstract*— Manipulating articulated objects with robotic arms is challenging due to the complex kinematic structure, which requires precise part segmentation for efficient manipulation. In this work, we introduce a novel superpoint-based perception method designed to improve part segmentation in 3D point clouds of articulated objects. We propose a learnable, part-aware superpoint generation technique that efficiently groups points based on their geometric and semantic similarities, resulting in clearer part boundaries. Furthermore, by leveraging the segmentation capabilities of the 2D foundation model SAM, we identify the centers of pixel regions and select corresponding superpoints as candidate query points. Integrating a query-based transformer decoder further enhances our method's ability to achieve precise part segmentation. Experimental results on the GAPartNet dataset show that our method outperforms existing state-of-the-art approaches in cross-category part segmentation, achieving AP50 scores of 77.9% for seen categories (4.4% improvement) and 39.3% for unseen categories (11.6% improvement), with superior results in 5 out of 9 part categories for seen objects and outperforming all previous methods across all part categories for unseen objects.**

I. Introduction

Articulated objects, such as doors and drawers, are ubiquitous in daily life due to their kinematic connections. As embodied intelligence technology continues to advance, it becomes increasingly important for robots to not only recognize these objects [1]–[4] but also manipulate them effectively by performing tasks like opening doors, closing drawers, or even lifting pot lids. Existing approaches based on reinforcement learning (RL) and imitation learning typically address the manipulation of articulated objects by predicting affordances and generating motion trajectories through learned policies [5]–[9]. However, these methods often face significant challenges in generalizing to unseen objects, particularly when variations in object geometry are introduced, thereby limiting the transferability of the learned skills.

In contrast, by leveraging powerful vision models, part segmentation-based approaches to articulated object modeling offer a more general solution, achieving accurate perception of articulated objects, forming the foundation for successful manipulation [10]–[12]. This precise and efficient perception enables robots to handle complex tasks with greater reliability [13]–[15]. However, while these approaches provide significant advantages, prior part segmentation methods typically segment 3D point clouds into different parts based on individual point clouds [7], [16]–[18]. Although these methods can achieve high modeling accuracy with familiar objects, they struggle to extract transferable information in the face of complex variations in point clouds, significantly reducing segmentation accuracy with unseen objects. Superpoint-based methods [19]–[21] partition point clouds into point sets, known as superpoints, which are groups of neighboring points adapted to local complexity and aggregating geometric information. This superpoint-based approach not only reduces computational overhead but also enhances the model's ability to generalize across diverse object geometries by effectively integrating local geometric features, thereby improving segmentation accuracy with unseen objects.

In this paper, we introduce **G**eneralizable **A**rticulated Object **P**erception with **S**uperpoints (**GAPS**), a novel approach designed to enhance part segmentation in diverse articulated objects within 3D point clouds.GAPS improves superpoint boundary clarity through learnable part-aware superpoint generation techniques, ensuring more distinct superpoints. Building on this, it leverages the 2D foundation model SAM [22] to effectively segment pixel regions, where each region's center uniquely identifies the corresponding 3D superpoints. These superpoints are then used as query points for part segmentation, enabling a more generalizable and adaptable selection of query points. By utilizing a query-based transformer decoder, GAPS achieves precise part segmentation across articulated objects. The main contributions of this paper are as follows:

1) We design the learnable part-aware superpoint generation method that groups point clouds as superpoints based on geometric and semantic similarities. Compared to rule-based superpoint generation, our approach is more effective in handling smaller parts and achieving clearer boundaries.

2) The 2D foundation model SAM segments images into pixel regions, with each center mapping to a unique 3D superpoint. These superpoints act as query points, allowing the transformer decoder to effectively capture local information, enabling GAPS to achieve precise part segmentation across diverse articulated objects.

3) We conduct experiments on the articulated object modeling benchmark GAPartNet [4], where GAPS outperforms existing state-of-the-art part segmentation methods in both

* indicates equal contribution. + indicates corresponding author.

[1]Qiaojun Yu, Xibin Yuan, and Cewu Lu are with Shanghai Jiao Tong University, China, {yqjllxs, 2022yxb, lucewu}@sjtu.edu.cn. [2]Ce Hao is with National University of Singapore, Singapore, cehao@u.nus.edu. [3]Li Zhang is with University of Science and Technology of China, China, zanly20@mail.ustc.edu.cn. [4]Liu Liu is with Hefei University of Technology, China, liuliu@hfut.edu.cn. [5]Yukang Huo is with China Agricultural University, China, huoyukang@cau.edu.cn. [6]Rohit Agarwal is with National Institute of Technology, Durgapur, India, ra.22cs1102@phd.nitdgp.ac.in.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant (62302143), the Anhui Provincial Natural Science Foundation under Grant (2308085QF207), the National Key Research and Development Project of China (2022ZD0160102), the Shanghai Artificial Intelligence Laboratory, and XPLORER PRIZE grants (2021ZD0110704).

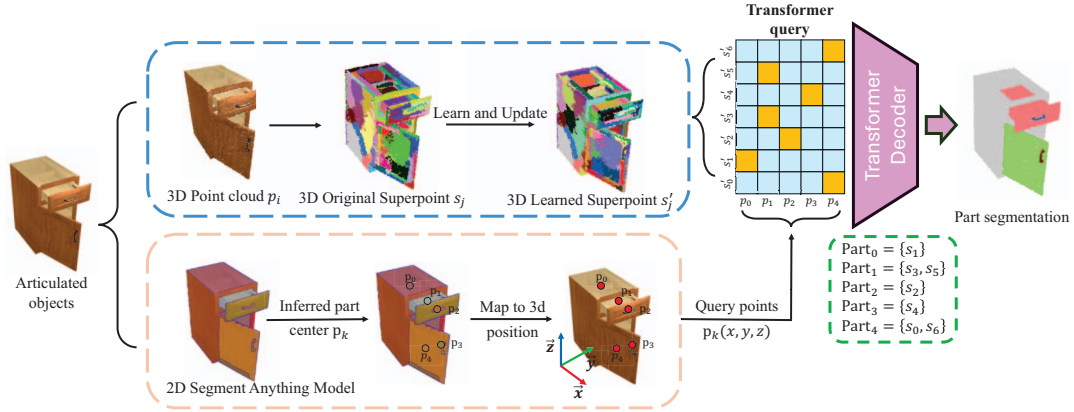

Fig. 1. GAPS segments articulated objects into semantic parts. It leverages both 3D point clouds to cluster superpoints and 2D image segmentation to infer part center, queried by a transformer decoder for part segmentation.

seen objects and unseen cross-category generalization.

II. Preliminary

We formulate the articulated object part segmentation task $\mathcal{T}$ as follows. An articulated object $\mathcal{M}$ consists of K variable movable parts, represented as $\mathcal{M} = \{m_i\}_{i=1}^{K}$. We observe the object $\mathcal{M}$ using RGB-D cameras and project it into a point cloud P with N points, $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$. Superpoints are an over-segmented set of point clouds that adapt to local geometric structures and capture contextual features. Given a point cloud as P with N points as $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$ and its corresponding features $F = \{\mathbf{f}_i \in \mathbb{R}^d\}_{i=1}^{N}$, superpoint generation aims to construct O superpoints as $S = \{\mathbf{s}_i \in \mathbb{R}^3\}_{i=1}^{O}$ and its corresponding features $E = \{\mathbf{e}_i \in \mathbb{R}^d\}_{i=1}^{o}$ from the point cloud P, assigning each point to one of the O superpoint centers with the highest probability. In this way, the superpoints S with corresponding features E can be used to represent the entire point cloud, with each superpoint encoding both local geometric and semantic features.

III. Method

In this section, we present an innovative methodology GAPS, as shown in Fig 1. For a given single-view point cloud of an object, we leverage the Point Transformer-V2 [23] to extract point-wise features, which are then processed through a part-aware superpoint generation module to produce superpoints, resulting in sharper and more well-defined boundaries between superpoints (Section III-A). By utilizing the SAM-guided 2D information to identify corresponding 3D superpoints as candidate 3D query points and combining it with a transformer decoder, we enable accurate part segmentation of articulated objects (Section III-B).

A. Part-aware Superpoint Generation

Superpoints are groups of 3D points semantically clustered based on similar geometric features. In articulated object part segmentation, we leverage these superpoints to capture local geometric information, enhancing the model's generalization ability and improving segmentation accuracy. Unlike previous methods focused on instance segmentation, our task deals with the complexity of articulated part segmentation, where parts vary significantly in size and have intricate connections. To address this, we draw inspiration from SPNet [20] and employ a learnable soft association map to model relationships between points and superpoints. This approach generates part-aware superpoints, effectively addressing the challenges of articulated part segmentation.

Given an articulated object represented by a point cloud $P = \{\mathbf{p}_i \in \mathbb{R}^3\}_{i=1}^{N}$, we initially use hand-crafted features [24] to construct a hard point-superpoint assignment. The superpoint coordinates S and features E are derived as the averages of the coordinates and features of the points assigned to each superpoint. However, this initial over-segmentation, achieved through unsupervised optimization, may not effectively capture fine part instances, leading to issues such as cross-part and nested-part segmentation. To address this, we apply a refinement process that updates both the point-superpoint assignment and the soft association map, improving segmentation accuracy by better handling these complexities. To further enhance computational efficiency, we selectively build the association map using only the nearest 6 superpoints to each point, denoted as $A \in \mathbb{R}^{N\times 6}$. Specifically, a cosine-similarity-like operation, implemented through an MLP, is used to update the point-superpoint association map. The association between the i-th point and j-th superpoint a_{ij}, is updated as follows:

$$a_{ij} = \phi(p_i, s_j)g(p_i) \cdot \psi(f_i, e_j)h(f_i), \tag{1}$$

where $\phi(\cdot,\cdot) : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^c$ and $g(\cdot) : \mathbb{R}^3 \to \mathbb{R}^c$ are two mapping functions in the coordinate space, while $\psi(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^c$ and $h(\cdot) : \mathbb{R}^d \to \mathbb{R}^c$ are two mapping functions in the feature space, all implemented by MLP. Then we normalize the association map of each point:

$$\tilde{a}_{ij} = \frac{\exp(a_{ij})}{\sum_{k=1}^{6} \exp(a_{ik})}. \tag{2}$$

After updating the association map, we use the normalized association scores as weights to update the superpoints' coordinates and features as follows:

$$s_j = \frac{\sum_{i=1}^{N} \tilde{a}_{ij} \cdot p_i}{\sum_{i=1}^{N} \tilde{a}_{ij}}, e_j = \frac{\sum_{i=1}^{N} \tilde{a}_{ij} \cdot f_i}{\sum_{i=1}^{N} \tilde{a}_{ij}}. \tag{3}$$

We assume that points within the same superpoint belong to the same object part, and we use one-hot encoded labels corresponding to the parts. Given the one-hot labels $L = \{l_i \in \mathbb{R}^K\}_{i=1}^N$ for each point, the labels for each superpoint can be computed through a weighted average as follows: $L^s = \{l_j^s = \frac{\sum_{i=1}^N \tilde{a}_{ij} \cdot l_i}{\sum_{i=1}^N \tilde{a}_{ij}}\}_{j=1}^O$. We then reconstruct the point labels $\tilde{L} = \{\tilde{l}_i = \sum_{j=1}^M \tilde{a}_{ij} \cdot l_j^s\}_{i=1}^N$ using normalized association scores. Additionally, pseudo labels $\tilde{L}^s = \{\tilde{l}_j^s = \mathrm{mod}(\sum_{i=1}^N a_{ij} \cdot l_i)\}_{j=1}^M$ can be derived for the superpoints using a voting mechanism based on point-superpoint correlations. The corresponding loss is then defined as follows:

$$\mathcal{L}_{sp} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(l_i, \tilde{l}_i) + \frac{1}{O} \sum_{j=1}^{O} \mathcal{L}(l_j^s, \tilde{l}_j^s), \tag{4}$$

where $\mathcal{L}(\cdot, \cdot)$ is the cross-entropy loss function. This loss encourages part-aware consistency and ensures the superpoint structure aligns with part boundaries.

B. Superpoint-based Part Segmentation

Our approach to part segmentation of articulated objects leverages superpoints, enhanced by integrating SAM-guided 2D information to identify corresponding 3D superpoints as candidate query points. These superpoints, generated from single-view point clouds during the part-aware superpoint generation stage, capture local geometric features, providing rich encoding sensitive to the nuances of articulated parts and robust to scale and size variations. Building on the SPFormer framework [25], we employ a 6-layer query decoder to refine the segmentation process further.

Query Decoder Architecture. Each superpoint, represented by the aggregated coordinates and features of its constituent points, forms the basis of our segmentation approach. The coordinates of these SAM-selected 3D query points are used to generate position embeddings, which serve as queries. The superpoint features, combined with their corresponding position embeddings, are then fed into a query decoder, employing a 6-layer transformer decoder architecture. This structure leverages cross-attention mechanisms, where each superpoint query attends over all points to refine its representation. Let $S = \{s_1, s_2, ..., s_O\}$ be the set of superpoint feature vectors, where $s_j \in \mathbb{R}^d$ is the feature vector for the j-th superpoint, and O is the total number of superpoints. The corss-attention operation for the l-th layer of the decoder can be defined as:

$$\text{CrossAttention}^{(l)}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

where Q, K, and V are the query, key, and value matrices derived from the superpoint features, respectively. The operation computes the attention-weighted sum over the values V, where the attention weights are determined by the compatibility between the queries Q and keys K. The scaling factor $\frac{1}{\sqrt{d_k}}$ ensures stability during training.

The output of the cross-attention layers is then passed through feed-forward networks within each layer of the decoder to progressively refine the superpoint representations. The final superpoint query representation at the l-th layer is given by:

$$s_j^{(l)} = \text{FFN}\left(\text{CrossAttention}^{(l)}(Q^{(l)}, K^{(l)}, V^{(l)})\right) \tag{6}$$

where FFN denotes a feed-forward network that further processes the output of the cross-attention mechanism. Due to the presence of one or more superpoints within a part, each superpoint can query the corresponding part. Unlike bipartite matching, we adopt many-to-one matching [26]. Formally, we use a pairwise matching cost matrix C to evaluate the similarity between the queries and the articulation parts. Using the cost matrix, we assign each query to its corresponding parts.

$$\hat{C}_{im} = \begin{cases} C_{im} & \text{if } i\text{-th query} \in m\text{-th part} \\ +\infty & \text{otherwise} \end{cases} \tag{7}$$

Once the matching is completed, we know the class labels of the queries in advance and compute the cross-entropy loss $\mathcal{L}_{\text{cls}}$ for each query. we compute the segmentation mask loss, which consists of the binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ and the dice loss $\mathcal{L}_{\text{dice}}$ for each matched query and part pair. We compute the BCE score loss $\mathcal{L}_{\text{score}}$ to determine if the IoU of the current query's corresponding part is greater than 50%. Therefore, the overall loss for articulation part segmentation is as follows:

$$\mathcal{L}_{\text{sem}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{score}}\mathcal{L}_{\text{score}} \tag{8}$$

In our paper, the values of the loss function weights are set as follows: $\lambda_{\text{cls}} = 1.5$, $\lambda_{\text{bce}} = 1.25$, $\lambda_{\text{dice}} = 1.0$, and $\lambda_{\text{score}} = 1.0$, respectively.

IV. Experiment

A. Dataset and Evaluation Metrics

We validate the articulated object segmentation using the GaPartNet dataset [4], rendering RGB-D images with annotations in the SAPIEN environment [27]. To evaluate cross-category generalizability, we split the 27 object categories into 17 seen and 10 unseen categories, ensuring all 9 part classes are represented in both. Following the 3D semantic instance segmentation benchmarks in ScanNetV2 [28], we use average precision (AP) as the performance metric for part segmentation, with AP50 (IoU threshold of 50%) assessing both per-part and overall segmentation accuracy.

B. Cross-category Part Segmentation

Table I presents the quantitative comparisons between our method and previous state-of-the-art methods, including PointGroup [29], SoftGroup [30], AutoGPart [31], GAPartNet [4], and SPFormer [25]. Our method surpasses current state-of-the-art approaches in both seen and unseen categories, with superior results in 5 out of 9 categories. While it shows a slight advantage in seen categories, achieving an AP50 of 77.9%, notably, the performance on the slider button category shows an absolute improvement of 11.1% compared to previous

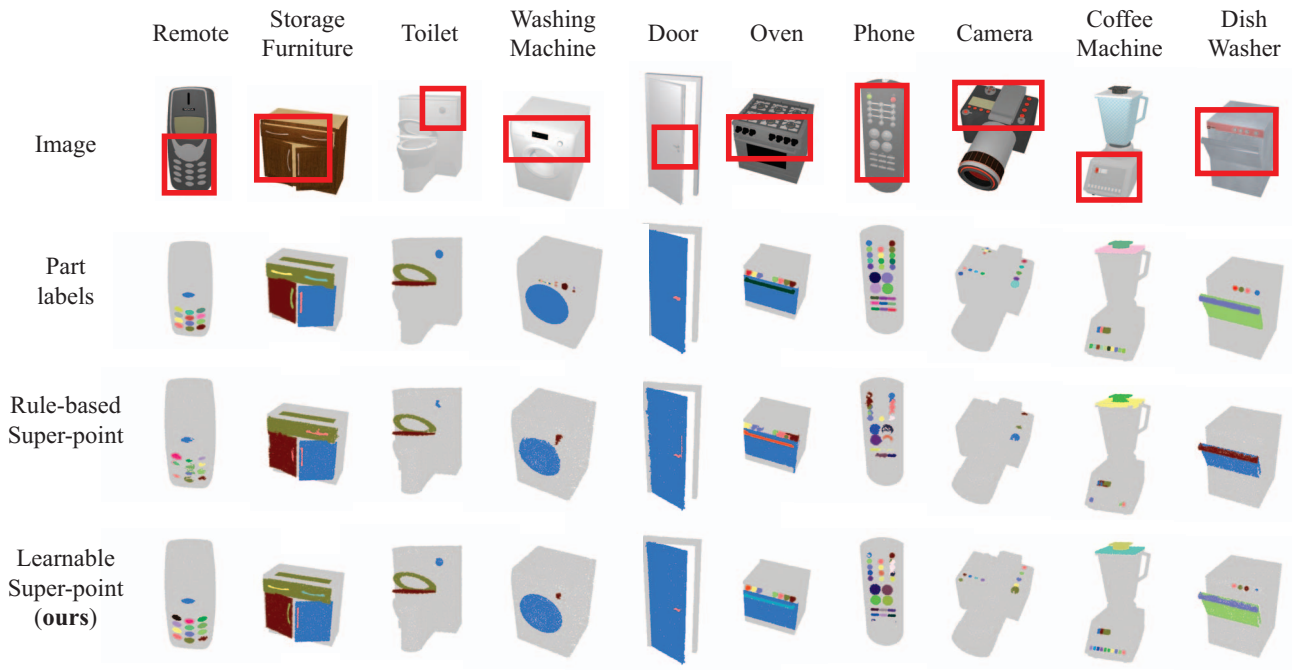

Fig. 2. Experimental results of part segmentation using rule-based and learnable superpoints. The segmented parts are marked in red box.

TABLE I

RESULTS OF PART SEGMENTATION (PER-PART-CLASS AP50 % ↑)

		Ln.F.Hl.	Rd.F.Hl.	Hg.Hl.	Hg.Ld.	Sd.Ld.	Sd.Bn	Sd.Dw.	Hg.Dr.	Hg.Kb.	Avg.AP50
Seen	PG [29]	85.3	23.5	82.4	79.3	86.5	47.3	60.3	90.1	32.5	65.2
	SG [30]	55.6	**92.4**	80.3	75.1	87.6	26.9	50.1	91.5	50.3	67.8
	AGP [31]	84.3	22.5	85.1	75.6	87.5	60.8	59.7	91.4	18.3	65.0
	GAP [4]	89.5	53.9	**90.1**	83.7	**89.5**	56.5	64.1	**93.1**	51.4	74.6
	SPF [25]	53.1	43.5	87.3	76.7	80.6	39.5	52.7	81.2	15.4	58.8
	Ours	**90.4**	62.0	87.5	**91.5**	87.3	**71.9**	**66.4**	89.8	**54.6**	**77.9**
Unseen	PG [29]	30.5	10.9	2.7	27.4	0.0	43.5	58.3	60.5	3.8	26.4
	SG [30]	24.3	7.1	1.8	35.6	0.0	49.2	52.9	67.5	11.4	27.8
	AGP [31]	43.9	6.4	3.6	36.7	0.0	46.2	63.0	60.6	15.7	30.7
	GAP [4]	43.5	37.1	2.8	40.2	3.9	45.4	60.2	63.1	21.2	35.2
	SPF [25]	11.5	8.7	1.5	33.4	0.5	20.2	10.6	45.4	3.0	15.0
	Ours	**46.9**	**41.2**	**5.8**	**42.5**	**4.1**	**51.0**	**65.6**	**71.4**	**25.3**	**39.3**

Ln.=Line. F.=Fixed. Rd.=Round. Hg.=Hinge. Hl.=Handle. Sd.=Slider. Ld.=Lid. Bn.=Button. Dw.=Drawer. Dr.=Door. Kb.=Knob.

PG=PointGroup [29]. SG=SoftGroup [30]. AGP=AutoGPart [31]. GAP=GAPartNet [4]. SPF=SPFormer [25]

TABLE II

RESULTS OF ABLATION STUDIES (PER-PART-CLASS AP50 % ↑)

	Ablation	Ln.F.Hl.	Rd.F.Hl.	Hg.Hl.	Hg.Ld.	Sd.Ld.	Sd.Bn	Sd.Dw.	Hg.Dr.	Hg.Kb.	Avg.AP50
Seen	× SP	56.1	25.6	65.2	63.7	53.5	42.3	46.8	65.1	31.2	49.9
	Para. query	73.6	46.3	86.2	78.1	81.4	51.8	54.6	84.5	24.2	64.5
	Proj. query	85.1	54.5	85.7	70.3	82.5	65.0	59.2	87.6	45.8	70.6
	Ours	**90.4**	**62.0**	**87.5**	**91.5**	**87.3**	**71.9**	**66.4**	**89.8**	**54.6**	**77.9**
Unseen	× SP	32.5	21.5	1.9	13.7	0.0	23.7	37.6	47.4	15.5	21.5
	Para. query	15.3	10.6	1.4	41.3	0.8	35.3	16.9	50.8	9.7	20.2
	Proj. query	35.9	33.5	2.1	36.4	0.3	39.0	55.7	60.3	16.6	31.1
	Ours	**46.9**	**41.2**	**5.8**	**42.5**	**4.1**	**51.0**	**65.6**	**71.4**	**25.3**	**39.3**

× SP= ablate superpoint, Para.query= parameterized query, Proj. query= center-to-point projection query.

methods. Our method performs better in unseen categories, achieving an AP50 of 39.3% and the best results across all part categories, which highlights its enhanced ability to generalize to novel objects.

C. Ablation Study

We conduct ablation studies to validate the effectiveness of the part-aware superpoint generation and SAM-guided 2D information transformer decoder in our method.

In Table II, we sequentially ablate the following: superpoint clustering (replaced with raw point clouds), parameterized queries [25], and point-to-center queries [26]. Results show that using superpoints instead of raw point clouds significantly improves performance for both seen and unseen objects, as superpoint features capture more transferable geometric information. Compared to parameterized queries [25], the 3D position embeddings of queries better integrate local features, resulting in more accurate part segmentation. Additionally, we modified the query point generation method to a point-to-center query [26], which performed well for seen objects but struggled to accurately locate centers in unseen objects. To address this, SAM leverages 2D prior knowledge to precisely locate part centers through back projection.

We visualize the segmentation results in Figure 2, Compared to rule-based superpoints, our learnable superpoint-based queries more effectively integrate local geometric information, enhancing the accuracy of local geometry modeling and improving the overall stability of the model.

V. CONCLUSION

In this paper, we presented Generalizable Articulated Object Perception with Superpoints (GAPS), a novel approach for enhancing part segmentation of articulated objects in 3D point clouds. GAPS employs a learnable, part-aware superpoint generation technique to group points based on geometric and semantic similarities, resulting in clearer boundaries. Furthermore, the method leverages the 2D foundation model SAM to select candidate 3D query points and utilizes a query-based transformer decoder for precise segmentation. GAPS demonstrated state-of-the-art performance on the GAPartNet benchmark, achieving AP50 scores of 77.9% for seen categories and 39.3% for unseen categories, with improvements of 4.4% and 11.6% for seen and unseen categories, highlighting GAPS's generalization capabilities.

REFERENCES

[1] L. Zhang, Z. Han, Y. Zhong, Q. Yu, X. Wu *et al.*, "Vocapter: Voting-based pose tracking for category-level articulated object via inter-frame priors," in *ACM Multimedia 2024*, 2024.

[2] L. Liu, A. Huang, Q. Wu, D. Guo, X. Yang, and M. Wang, "Kpa-tracker: Towards robust and real-time category-level articulated object 6d pose tracking," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3684–3692.

[3] Q. Yu, C. Hao, J. Wang, W. Liu, L. Liu, Y. Mu, Y. You, H. Yan, and C. Lu, "Manipose: A comprehensive benchmark for pose-aware object manipulation in robotics," *arXiv preprint arXiv:2403.13365*, 2024.

[4] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. 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*, 2023, pp. 7081–7091.

[5] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "End-to-end affordance learning for robotic manipulation," *arXiv preprint arXiv:2209.12941*, 2022.

[6] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.

[7] L. Liu, J. Du, H. Wu, X. Yang, Z. Liu, R. Hong, and M. Wang, "Category-level articulated object 9d pose estimation via reinforcement learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 728–736.

[8] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. Guibas, and H. Dong, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," *arXiv preprint arXiv:2106.14440*, 2021.

[9] C. Ning, R. Wu, H. Lu, K. Mo, and H. Dong, "Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] J. Liu, M. Savva, and A. Mahdavi-Amiri, "Survey on modeling of articulated objects," *arXiv preprint arXiv:2403.14937*, 2024.

[11] Y. Weng, H. Wang, Q. Zhou, Y. Qin, Y. Duan, Q. Fan, B. Chen, H. Su, and L. J. Guibas, "Captra: Category-level pose tracking for rigid and articulated objects from point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 209–13 218.

[12] 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.

[13] H. Zhang, B. Eisner, and D. Held, "Flowbot++: Learning generalized articulated objects manipulation via articulation projection," *arXiv preprint arXiv:2306.12893*, 2023.

[14] 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.

[15] Q. Yu, J. Wang, W. Liu, C. Hao, L. Liu, L. Shao, W. Wang, and C. Lu, "Gamma: Generalizable articulation modeling and manipulation for articulated objects," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 670–13 677.

[16] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.

[17] 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.

[18] L. Zhang, Y. Zhong, J. Wang, Z. Min, L. Liu *et al.*, "Rethinking 3d convolution in ℓ_p-norm space," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] D. Robert, H. Raguet, and L. Landrieu, "Efficient 3d semantic segmentation with superpoint transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 195–17 204.

[20] L. Hui, J. Yuan, M. Cheng, J. Xie, X. Zhang, and J. Yang, "Superpoint network for point cloud oversegmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5510–5519.

[21] M. Kolodiazhnyi, A. Vorontsova, A. Konushin, and D. Rukhovich, "Oneformer3d: One transformer for unified point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 943–20 953.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[23] X. Wu, Y. Lao, L. Jiang, X. Liu, and H. Zhao, "Point transformer v2: Grouped vector attention and partition-based pooling," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 330–33 342, 2022.

[24] S. Guinard and L. Landrieu, "Weakly supervised segmentation-aided classification of urban scenes from 3d lidar point clouds," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 151–157, 2017.

[25] J. Sun, C. Qing, J. Tan, and X. Xu, "Superpoint transformer for 3d scene instance segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2393–2401.

[26] J. Lu, J. Deng, C. Wang, J. He, and T. Zhang, "Query refinement transformer for 3d instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 516–18 526.

[27] 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," 2020.

[28] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[29] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, 2020, pp. 4867–4876.

[30] T. Vu, K. Kim, T. M. Luu, T. Nguyen, and C. D. Yoo, "Softgroup for 3d instance segmentation on point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2708–2717.

[31] X. Liu, X. Xu, A. Rao, C. Gan, and L. Yi, "Autogpart: Intermediate supervision search for generalizable 3d part segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 624–11 634.