Active Perception for Autonomous Multi-View Geometrically Consistent Data Collection in Local Environments

Khiem Phi[1*], Anh Tung Ho[2*], Aditya Patankar[2], Nilanjan Chakraborty[2], and IV Ramakrishnan[1]

***Abstract*— Robust deployment of robots in specific environments such as homes or nursing facilities requires reliable annotated data for objects in those environments. Despite the impressive progress in computer vision algorithms with the advent of deep learning and foundation models, there remains a bottleneck in generating high-quality, environment-specific data. In this paper, we formulate data collection as an active perception problem, where the robot purposefully moves to acquire informative observations. We present a system in which a robot follows a hemispherical trajectory to capture multi-view images of a scene. From a single seed annotation, provided via vision-language models using point or language prompts, our method leverages robot kinematics, camera intrinsics, and depth sensing to propagate annotations across views, producing a dense, 3D-consistent multi-view dataset. This dataset is then used to train a lightweight, deployable perception model tailored to the local environment. Across 5.3k images spanning 32 cluttered tabletop scenes and 30 object categories, models trained with our method outperform zero-shot Grounding-DINO + SAM by up to 33.5 mAP@50–95 while running at 58 FPS, demonstrating the effectiveness of purposeful robot motion for collecting reliable perception data.**

I. Introduction

Robots are increasingly deployed in cluttered and partially known environments such as homes, hospitals, warehouses, and laboratories [1]–[4]. Reliable manipulation in these settings requires perception systems that can robustly recognize and segment objects under varying viewpoints, occlusions, and sensor noise. Although recent foundation models, including Grounding-DINO [5] and the Segment Anything (SAM) family [6]–[8] demonstrate strong zero-shot generalization, they typically depend on repeated user prompting and substantial computational resources, limiting their practicality for on-board active robotic vision systems. Moreover, their performance often degrades in environment-specific scenarios due to domain shifts not captured in their training data. Lightweight supervised models, such as the single-stage YOLO [9] and the two-stage R-CNN [10], [11] families, are more suitable for active robot vision systems but require environment-specific annotated data to achieve reliable performance. Collecting and annotating such data remains time-consuming and labor-intensive, posing a significant barrier to rapid deployment [12].

Existing approaches to autonomous data collection often focus on task-specific applications, such as marker detection [13], [14] or 3D reconstruction [15]–[17], and do not address the problem of semantic labeling. Methods that do incorporate semantic labeling typically rely on passive video-based propagation, where labels are transferred across frames without exploiting robot motion or geometry. For example, foundation models such as SAM-2 and SAM-3 [7], [8] perform temporal mask propagation through appearance-based tracking, which are computationally expensive and remain susceptible to feature drift, occlusion failures, and inconsistent labeling across viewpoints. In our work, instead of treating the robot as a passive user of previously trained perception models, we seek to leverage the robot's embodiment to curate its own data, enabling it to effectively learn perception models tailored to its local environment.

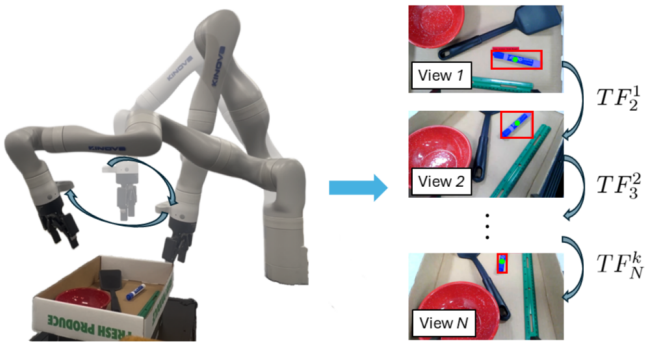

Fig. 1: Overview of the proposed autonomous data collection pipeline. A robot arm sweeps a hemispherical viewpoint trajectory over a tabletop workspace, capturing N views. Segmentation masks initialized from a single user prompt are propagated across frames via geometric transformations TF_{k+1}^k, producing a densely annotated multi-view dataset without further human involvement.

Therefore, we introduce an autonomous robot system that can actively collect data and generate reliable object detection and segmentation annotations with minimal human effort by propagating a single human annotation across multiple viewpoints. The annotated dataset can then be used for fine-tuning or retraining of perception models for the specific environments where the robots would be deployed.

II. Methodology

To address this challenge, we proposed a two-stage robot framework consisting of active data collection and label propagation between viewpoints which exploits the robot's embodiment and geometry, as shown in Fig. 1.

In Stage I, the manipulator equipped with an RGB-D sensor mounted to the end-effector follows a hemispherical viewpoint trajectory Π and captures a synchronized RGB-D observation $\mathcal{O}_t$ and end-effector pose TF_t at each point $\pi_t \in \Pi$. The trajectory is locally optimized via 2-opt Trav-

[1]The authors are with the Department of Computer Science, Stony Brook University, USA.
[2]The authors are with the Department of Mechanical Engineering, Stony Brook University, USA.
*These authors contributed equally to this work.

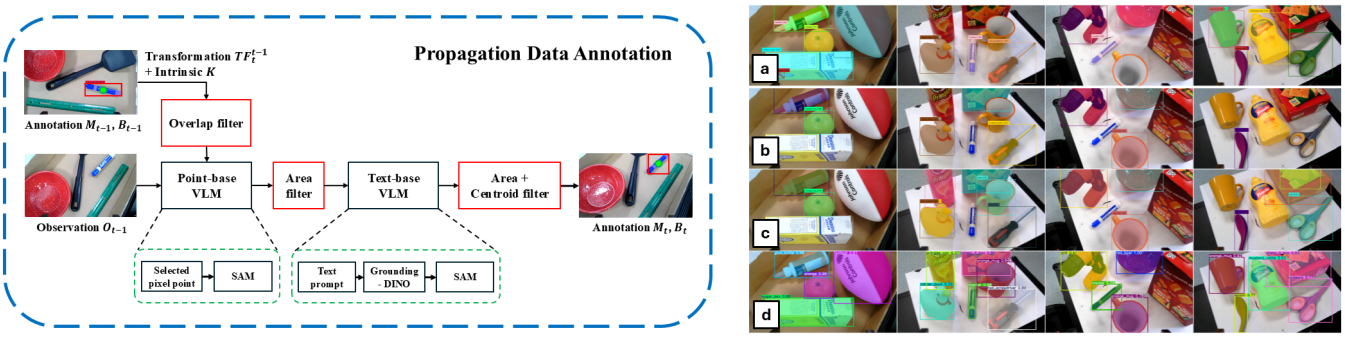

Fig. 2: Left: Annotation propagation across views using foundation models with geometric consistency filtering. Right: Qualitative comparison of perception pipelines: (a) VLM zero-shot, (b) VLM one-shot, (c) VLM multi-shot, and (d) YOLO26-Nano trained on our autonomously generated dataset.

eling Salesman Problem (TSP) solver such that consecutive poses are spatially close enough to enable reliable annotation transfer across frames.

In Stage II, seed labels from a single user prompt u on image I_0 are propagated sequentially along the trajectory Π. Using known camera geometry including transformations TF_t, intrinsics K, and depth D_t, 2D masks M_t^k from frame t are reprojected into 3D space and then projected into frame $t+1$, which then serves as the new reference to propagate forward to the next frame $t+2$. This process iteratively propagates annotations across views, constructing a dataset $\mathcal{D} = (I_t, B_t^k, M_t^k)$ consisting image, bounding boxes, and segmentation masks, for training the perception model f_θ without further human intervention. To ensure geometric consistency, each label transfer is validated using three filtering criteria for both text-based and point-based prompts, as illustrated in left image of Fig. 2.

Firstly, an overlap filter ensures sufficient geometric overlap by requiring that reprojected points contain valid pixels located within image bounds. The median of these points is then used as prompts to generate a candidate mask, which is validated against the most reliable previous mask using an overlap ratio constraint. If the candidate fails this check, a text-based prompt is used to recover a valid segmentation without additional user input. Finally, a 3D centroid distance filter rejects geometrically inconsistent masks, and only candidates passing all criteria are accepted for propagation.

III. Experimental Evaluation

We evaluate our framework on a Kinova Gen3 7-DOF manipulator equipped with an eye-in-hand RGB-D camera across 32 cluttered tabletop scenes containing 30 object classes. For each scene, the robot executes a hemispherical trajectory and propagates annotations over up to 300 viewpoints, resulting in a dataset of 5.3k RGB-D images with corresponding camera poses.

To assess label quality, we train a lightweight YOLO26-seg Nano model for joint object detection and segmentation and compare its performance against three zero-shot VLM baselines (Grounding-DINO-Large [5] + SAM-V1 [6]). These baselines include zero-shot (full category set), one-shot (scene-specific categories), and multi-shot (per-category queries) settings to progressively reduce ambiguity.

Method	Box@50	Box@50-95	Mask@50	Mask@50-95	FPS
VLM Zero-Shot	11.5	11.3	11.5	11.4	3
VLM One-Shot	27.0	26.5	26.9	26.4	3
VLM Multi-Shot	40.8	39.9	40.7	39.9	0.5
Ours (Nano)	**87.8**	**84.6**	**78.1**	**73.4**	**58**

TABLE I: Comparison of detection and segmentation performance (mAP in %) between Grounding-DINO + SAM VLM baselines and YOLO26-Nano trained on autonomously labeled data. Experiments are conducted on a single RTX-4090 GPU.

IV. Results and Discussion

Table I summarizes perception performance of YOLO26-Nano trained on our dataset and VLM baselines. Zero-shot, one-shot, and multi-shot VLMs achieve below 12, around 27, and 40 mAP@50–95, respectively, while YOLO26-Nano outperforms the best baseline by nearly $2\times$ at runs at 58 FPS, which is approximately $20\times$ faster.

Qualitative results in Fig. 2 further highlight these differences. YOLO26-Nano trained on our dataset produces accurate detections and precise masks across cluttered, occluded scenes, whereas VLM-based methods suffer from inconsistent boundaries, missed detections, and reduced confidence.

We evaluate the importance of TSP-based viewpoint ordering Π against a random ordering baseline. Random ordering leads to large viewpoint jumps, resulting in only 25% successful label propagation due to projection failures. In contrast, TSP ordering ensures smooth transitions, successfully executing 210/300 viewpoints and achieving an average 85% label transfer rate across scenes.

V. Conclusion

We present an active perception framework where a robot executes a hemispherical trajectory to collect multi-view observations and propagates a single seed annotation into a dense, geometrically consistent dataset using robot kinematics, camera intrinsics, and depth sensing. This purposeful data collection enables lightweight models to far exceed zero-shot baselines. YOLO11-Nano trained on our generated data outperforms Grounding-DINO + SAM by up to 33.5 mAP@50–95 while running at 58 FPS. Future work will integrate motion planning and manipulation to enable active scene rearrangement, further improving data diversity toward fully autonomous dataset generation.

REFERENCES

[1] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[2] Z. He, X. Zhang, S. Jones, S. Hauert, D. Zhang, and N. F. Lepora, "Tacmms: Tactile mobile manipulators for warehouse automation," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4729–4736, 2023.

[3] T. Dai, S. Vijayakrishnan, F. T. Szczypiński, J.-F. Ayme, E. Simaei, T. Fellowes, R. Clowes, L. Kotopanov, C. E. Shields, Z. Zhou, *et al.*, "Autonomous mobile robots for exploratory synthetic chemistry," *Nature*, vol. 635, no. 8040, pp. 890–897, 2024.

[4] G. Ren, T. Wu, T. Lin, L. Yang, G. Chowdhary, K. Ting, and Y. Ying, "Mobile robotics platform for strawberry sensing and harvesting within precision indoor farming systems," *Journal of Field Robotics*, vol. 41, no. 7, pp. 2047–2065, 2024.

[5] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *European conference on computer vision*. Springer, 2024, pp. 38–55.

[6] 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.

[7] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[8] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.

[9] R. Sapkota, R. H. Cheppally, A. Sharda, and M. Karkee, "Yolo26: key architectural enhancements and performance benchmarking for real-time object detection," *arXiv preprint arXiv:2509.25164*, 2025.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2013. [Online]. Available: https://arxiv.org/abs/1311.2524

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017. [Online]. Available: https://arxiv.org/abs/1703.06870

[12] H. W. Liao, C. Klugmann, D. Kondermann, and R. Mahmood, "Minority reports: Balancing cost and quality in ground truth data annotation," *arXiv preprint arXiv:2504.09341*, 2025.

[13] V. Ilin, I. Kalinov, P. Karpyshev, and D. Tsetserukou, "Deepscanner: a robotic system for automated 2d object dataset collection with annotations," in *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2021, pp. 01–08.

[14] D. De Gregorio, A. Tonioni, G. Palli, and L. Di Stefano, "Semiautomatic labeling for deep learning in robotics," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 611–620, 2019.

[15] A. T. Ho, P. K. Lee, and H. Yun, "Utilizing multiple point cloud scenes for precise robotic bin-picking tasks," in *Proceedings of the Spring/Fall Conference of the Korean Society of Mechanical Engineers (KSME)*. The Korean Society of Mechanical Engineers, 2024, pp. 301–302.

[16] C. Mineo, D. Cerniglia, V. Ricotta, and B. Reitinger, "Autonomous 3D geometry reconstruction through robot-manipulated optical sensors," *The International Journal of Advanced Manufacturing Technology*, vol. 116, pp. 1895–1911, Sept. 2021. [Online]. Available: https://doi.org/10.1007/s00170-021-07432-5

[17] S. Isler, R. Sabzevari, J. Delmerico, and D. Scaramuzza, "An information gain formulation for active volumetric 3d reconstruction," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 3477–3484.