2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

You Only Scan Once: A Dynamic Scene Reconstruction Pipeline for 6-DoF Robotic Grasping of Novel Objects

Lei Zhou[1,*], Haozhe Wang[1,2], Zhengshen Zhang[1], Zhiyang Liu[1], Francis EH Tay[1], and Marcelo H. Ang Jr[1]

***Abstract*— In the realm of robotic grasping, achieving accurate and reliable interactions with the environment is a pivotal challenge. Traditional methods of grasp planning methods utilizing partial point clouds derived from depth image often suffer from reduced scene understanding due to occlusion, ultimately impeding their grasping accuracy. Furthermore, scene reconstruction methods have primarily relied upon static techniques, which are susceptible to environment change during manipulation process limits their efficacy in real-time grasping tasks. To address these limitations, this paper introduces a novel two-stage pipeline for dynamic scene reconstruction. In the first stage, our approach takes scene scanning as input to register each target object with mesh reconstruction and novel object pose tracking. In the second stage, pose tracking is still performed to provide object poses in real-time, enabling our approach to transform the reconstructed object point clouds back into the scene. Unlike conventional methodologies, which rely on static scene snapshots, our method continuously captures the evolving scene geometry, resulting in a comprehensive and up-to-date point cloud representation. By circumventing the constraints posed by occlusion, our method enhances the overall grasp planning process and empowers state-of-the-art 6-DoF robotic grasping algorithms to exhibit markedly improved accuracy.**

I. INTRODUCTION

In the realm of robotic manipulation, the ability to grasp and interact with objects in dynamic and complex environments remains a cornerstone challenge. Achieving effective grasping hinges on the fusion of accurate scene understanding and real-time adaptability, which has traditionally posed significant hurdles for existing methodologies [1], [2]. Previous efforts in grasp planning and scene reconstruction have primarily gravitated towards either partial point cloud utilization or static scene representations, each marked by inherent limitations in capturing the dynamic nature of real-world scenarios. Robotic tasks involving some form of 3D visual perception greatly benefit from a complete knowledge of the working environment.

Grasp generation methods [3]–[7], reliant on partial point clouds back-projected from depth image, introduce their own set of constraints. The quality of grasp generation directly hinges upon the accuracy and completeness of the partial views captured. Owing to occlusions and partial observability [8], the objects (or parts) suitable for grasping are invisible from a single viewpoint. The resultant grasp plans lack a comprehensive understanding of the scene, subsequently compromising both accuracy and diversity of robotic grasping.

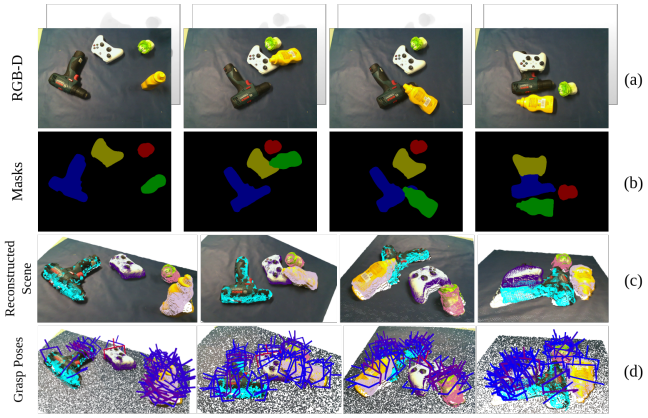


Fig. 1: Dynamic scene reconstruction and grasp generation. (a) RGB-D images are captured by an RGB-D camera as it scans the grasping workspace. (b) A *Video-segmentation Module* segments the graspable objects in the scene. (c) Using the RGB-D images and masks from (a) and (b), we reconstruct the meshes of the graspable objects and merge them with the original partial point cloud to create a full point cloud of the workspace. (d) Finally, a *Grasp Pose Predictor* is used to generate the valid grasps based on the reconstructed full point cloud.

By predicting the missing part of the object point cloud, the full shape of an object can be recovered to generate more diverse grasps [9]–[11]. However, these methods introduce uncertainty to the generated grasps at the same time as the generated points are unreliable compared to the initially observed point cloud.

By leveraging multi-view input, static scene reconstruction methods, exemplified by NeRF-based [12]–[14] and TSDF-based [15]–[17] approaches, have demonstrated commendable efficacy in reconstructing environments to achieve more accurate and diverse grasp generation. However, these methods rely on static snapshots, inherently incapable of adapting to changes after scanning. Consequently, though the marriage of such methods with robotic grasping tasks has demonstrated the ability to generate more diverse grasps compared to taking partial point cloud as input, its applicability is impeded in robotic manipulation environments marked by constant change.

To bridge these disparate domains and unlock a new paradigm in robotic grasping, we present *You Only Scan Once* (YOSO), a novel approach that harmonizes the benefits of both static scene reconstruction and partial point cloud-

*Corresponding Author.
[1]Authors are with the Advanced Robotics Centre, National University of Singapore, 117608, Singapore. {leizhou, wang_haozhe, zhengshen_zhang, zhiyang}@u.nus.edu, {mpetayeh, mpeangh}@nus.edu.sg
[2]Haozhe Wang is with the Integrative Sciences and Engineering Programme, National University of Singapore Graduate School, 119077, Singapore.

based grasp planning. As shown in Fig. 1, our proposed pipeline introduces a dynamic scene reconstruction methodology that operates in real-time to complete the object point cloud in the scene for subsequent robotic grasping tasks. Unlike conventional static methods that need to repetitively scan the scene when it is changed, our approach only scans the scene once to generate mesh for each novel object in the scene. After that, it dynamically tracks the object pose and transforms the generated object mesh back into the scene to encompass the evolving environment.

Through comprehensive testing, we have assessed the effectiveness of our approach. Our results indicate the substantial improvement in grasping accuracy achieved by providing more complete scene understanding for the grasp planning process while operating at near real-time. Our contributions can be summarized as follows:

1) We propose a novel and modularized pipeline, YOSO, for dynamic scene reconstruction tailored to the context of robotic grasping tasks.
2) We evaluate a pre-trained state-of-the-art (SOTA) grasp generation model on our reconstructed scene and demonstrate that replacing the input partial point cloud with a more informative reconstructed scene from YOSO pipeline enables it to surpass its current SOTA evaluation results on the GraspNet-1Billion benchmark.
3) We also extend the GraspNet-1Billion dataset to include the completed point cloud of each scene. This addition aims to establish a theoretical upper limit of performance for models when provided with a fully visible point cloud of a scene within the dataset.

II. RELATED WORKS

A. Grasping Methods Utilizing Partial Point Clouds

In recent years, the field of robotic grasping has witnessed significant advancements in leveraging partial point clouds as a critical sensory input for grasp planning. Several notable works have contributed to the exploration of this area, each with its unique approach and methodologies. Fang *et al.* proposed both a popular benchmark dataset for general object grasping consisting of partial point cloud scenes with more than 1 billion annotated grasps, as well as a baseline model to generate 6-Degree-of-Freedom (6-DoF) grasps from partial point clouds [3]. Wang *et al.* improved upon the baseline by adding a graspness model which utilizes geometry cues to distinguish graspable areas in cluttered scenes [18]. Sundermeyer *et al.* proposed Contact-GraspNet [8], which treats the 3D points of a partial point cloud as potential grasp contacts. By rooting the full 6-DoF grasp pose and width in the observed point cloud, the dimensionality of the grasp representation can be reduced to 4-DoF, which greatly facilitates the learning process. Ma and Huang [4] proposed a Scale Balanced Learning loss and an Object Balanced Sampling strategy to address the challenge of generating accurate grasp poses for small-scale samples. However, the reliance on partial point clouds in robotic grasping still faces significant challenges. The primary issue is the incomplete data these point clouds provide, capturing only visible object surfaces. This limitation can obscure vital details about an object's shape and orientation and potentially lead to suboptimal grasps.

B. Grasping Methods Utilizing Single-view Shape Completion

In order to obtain a more complete and informative geometric understanding of the scene, some previous works took partial point cloud as input to predict the missing parts of objects [19], [20]. Lundell *et al.* represented objects as voxels and trained a deep learning network for object completion, along with Carlo (MC) sampling enhanced by dropout techniques [21]. Subsequently, grasps are predicted on the completed object shapes. However, this voxel-based approach significantly increases memory usage and computational time. Recent efforts [10], [22] have shifted towards using point cloud representations to streamline shape completion and grasp generation processes, aiming for more time-efficient testing phases. Although employing implicit object representations has been shown to accelerate shape completion to 0.7 seconds per object, as demonstrated by [23], such speed remains unsatisfactory for real-time applications, particularly in environments with multiple objects.

Additionally, the efficacy of those shape completion methods is hindered by their limited ability to generalize to novel object categories not covered in the training data, often leading to less precise reconstructions. In contrast, our proposed pipeline focuses on the reconstruction of novel object shapes, advocating for the use of multi-view inputs over single-view input. This approach not only enriches data reliability but also ensures a more accurate and comprehensive understanding of the scene.

C. Static Scene Reconstruction Methods

1) TSDF-based Methods: The field of robotic grasping has also witnessed significant advancements through the utilization of Truncated Signed Distance Field (TSDF) representations in scene reconstruction. TSDF is a volumetric approach that represents the scene as a voxel grid. Each voxel stores the distance to the nearest surface. Several works have leveraged this simple and efficient technique to enhance 6-DoF grasp generation and overcome various challenges encountered in robotic manipulation tasks [14]–[16].

While TSDF-based methods have undeniably advanced the field of robotic grasping, it is imperative to recognize their limitations. Notably, TSDF-based techniques can be computationally expensive, particularly when dealing with large-scale scenes, due to memory and computation requirements for updating the TSDF representation.

2) NeRF-based Methods: Neural Radiance Field (NeRF)-based methods for static scene reconstruction [24], [25] has shown potential in capturing and rendering intricate 3D scenes from different perspectives. These approaches are valued for their ability to produce high-quality reconstructions, proving beneficial in areas such as robotic grasping

[26]. NeRFs are particularly adept at handling complex scenarios, such as those involving non-Lambertian materials and challenging lighting [12]–[14], offering impressive visual quality in the representations they create.

However, it's important to recognize the limitations associated with NeRF-based reconstruction methods, especially when considering their application in dynamic tasks like 6-DoF robotic grasping. Similar to TSDF-based methods, these NeRF-based methods primarily accommodate static scenes, which limits their direct applicability in environments subject to change, such as those involving moving objects. In an effort to bridge this gap, continual NeRF training, as implemented in Evo-NeRF [13], has been developed to facilitate rapid updates to the NeRF model shortly after each grasp. Nevertheless, this process necessitates the re-scanning of the workspace and the acquisition of a small set of images to support continual NeRF training, still assuming that the environment remains unaltered after scanning. Such a requirement makes it less practical in dynamic scenes.

III. METHOD

The proposed pipeline is structured into two distinct stages, denoted as **Stage I** and **Stage II** in Fig. 2. Unlike conventional static scene reconstruction methods that necessitate repetitive re-scanning when changes occur, our approach only perform a single scene scan to register for novel objects in **Stage I**, subsequently enabling dynamic scene reconstruction in **Stage II**. Hence, we have coined the term *"You Only Scan Once"* to encapsulate the essence of our methodology.

In **Stage I**, a camera mounted to the robotic manipulator's wrist scans the scene along a predefined motion trajectory, encompassing the hemispherical region above the workspace. This scanning procedure yields a monocular RGB-D video, referred to as the reference video. Each object within this reference video is registered within a memory pool, wherein their respective poses and images are stored. Concurrently, a separate parallel thread generates the object meshes.

Transitioning to **Stage II**, we strategically pause the computationally intensive mesh reconstruction process while maintaining real-time tracking of object pose changes relative to the initial frame of the reference video. This adjustment enables the remainder of our pipeline to operate nearly in real-time. By transforming the generated object meshes into camera coordinates, accounting for their 6D poses, we effectively address occluded regions and complete missing object parts. Consequently, we reconstruct a point cloud representation of the scene, subsequently employed as input for the grasp pose prediction network to estimate the 6-DoF pose of the gripper.

Furthermore, our pipeline has been meticulously designed in a modular fashion, comprising three primary components: *Video-segmentation Module* (III-A), *Object Pose Tracker and Mesh Generator* (III-B), and *Grasp Pose Predictor* (III-C). This modular design approach empowers us to enhance overall performance through the integration of advanced algorithms within each module.

It's noteworthy that this pipeline relies on only two key assumptions. The first assumption is that objects should be visible in the initial few frames, acknowledging the constraints of the algorithms incorporated in Section III-A and Section III-B. The second assumption pertains to having access to the segmentation mask of target objects in the initial frame. Beyond these considerations, no additional information is necessitated for the pipeline's operation.

A. Video-segmentation Module

To prepare masks denoted as $\boldsymbol{M}^t$ corresponding to objects in each frame of the video $\boldsymbol{I}^t$ for the two subsequent modules, we employ the unified long-term video object segmentation framework, XMem [27]. In **Stage I**, we utilize the camera to scan the entire scene and capture a reference video. This reference video is then input into the XMem model, which subsequently generates masks corresponding to the target objects for each frame. In **Stage II**, we process the image of each frame captured by the camera through the XMem model for real-time segmentation, enabling us to track the masks of the target objects. It's important to note that in **Stage II**, we also provide the XMem model with the segmented masks obtained in **Stage I**. This additional input enables the model to establish long-term memory, enhancing its segmentation performance.

B. 6D Object Pose Tracker and Mesh Generator

In pursuit of our objective, which involves generating meshes for novel objects and seamlessly integrating them into the scene to complete missing parts, we incorporate the approach outlined in BundleSDF [28]. This approach allows for the simultaneous tracking of object poses and the generation of object meshes aligned with the initial camera pose. Consequently, the object pose ξ_i^t in frame t can be considered as the object's pose within the camera coordinates.

During **Stage I**, the RGB-D images $\boldsymbol{I}^t$ and their corresponding segmentation masks $\boldsymbol{M}^t$, obtained from the *Video-segmentation Module*, are combined to form an input frame $\boldsymbol{F}^t = [\boldsymbol{I}^t, \boldsymbol{M}^t]$. The primary objective in this stage is to reconstruct the mesh O_i for each object while ensuring alignment with the camera pose from the initial frame.

In **Stage II**, a pose tracker is employed to estimate the relative pose change $\Delta\xi_i^t$ of each object with respect to the pose ξ_i^0 in the initial frame. Notably, this initial frame's pose also serves as the object's pose in the camera coordinates.

1) 6D Pose Tracker for Novel Object: Building on the work of Wen *et al.* in [28], our approach includes feature matching in RGB between the current frame $\boldsymbol{F}^t$ and the previous frame $\boldsymbol{F}^{t-1}$. To facilitate long-term pose tracking, $\boldsymbol{F}^t$ is designated as a keyframe if the relative feature difference between the two frames surpasses a predefined threshold. These keyframes are subsequently stored in a *Keyframe Memory Pool*. Then, for each new frame, a comparison is made with the nearest K keyframes in the memory pool. This process facilitates online pose graph optimization, which further refines the pose estimation results and mitigates pose drift.

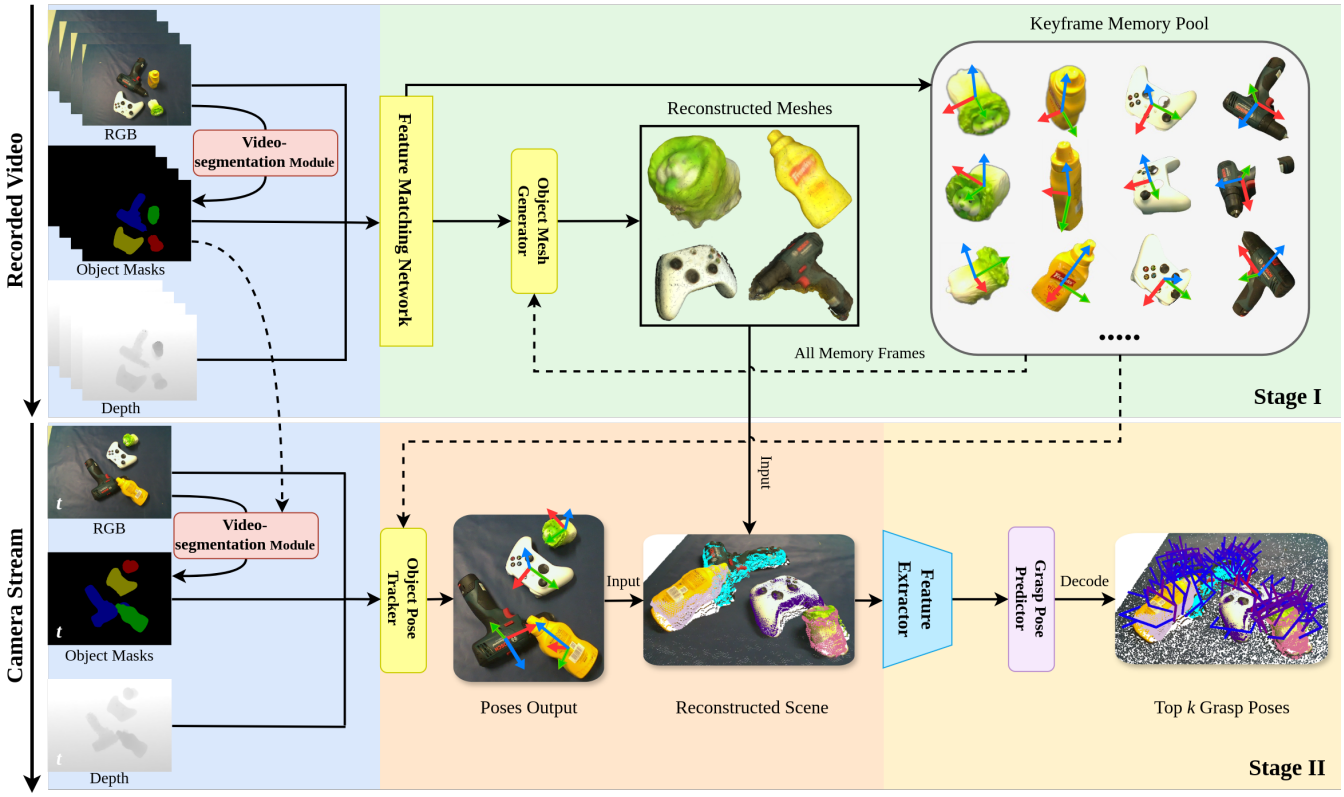


Fig. 2: Overview of the proposed pipeline. **Stage I**: Given a monocular RGB-D video, object masks are segmented using a *Video-segmentation Module*. Subsequently, feature matching is performed in the *Object Pose Tracker and Mesh Generator* module to simultaneously track object pose and reconstruct object mesh. Keyframes with informative historical observations are stored in the memory pool to facilitate pose tracking in both stages. **Stage II**: In testing, given an RGB-D image, the masks of the objects in the workspace are segmented out and the object pose is estimated by taking the *Keyframe Memory Pool* as a reference. Subsequently, the reconstructed meshes are transformed into camera coordinates with the estimated object pose. Taking this reconstructed scene point cloud, grasp generation is performed to generate the top k grasp poses for real-world experiments. The dotted lines represent the supplementation of historical information.

2) NeRF-based Mesh Generator: Following [28], a *NeRF-based Mesh Generator* is employed to train an object-centric neural signed distance field (SDF). This SDF learns both the multi-view consistent 3D shape and appearance of the object. Given that the mesh generation process is relatively computationally intensive, it is strategically frozen during testing to ensure the real-time performance of our pipeline.

C. 6-DoF Grasp Pose Predictor

In **Stage II**, given a partial point cloud $\boldsymbol{P}^t$ back-projected from depth image , the reconstructed object point cloud is first transformed into the camera coordinates based on the tracked poses:

$$O_i^t = \Delta\xi_i^t O_i. \tag{1}$$

Subsequently, the scene is reconstructed by merging the observed scene point cloud and the transformed object point cloud:

$$\boldsymbol{P}^t_{merge} = [\boldsymbol{P}^t, \boldsymbol{O}^t], \tag{2}$$

where $\boldsymbol{O}^t = [O_1^t, O_1^t ..., O_M^t]$ is the reconstructed point clouds of all target objects and M represents number of objects.

In our pipeline, we incorporate Scale-balanced GraspNet [4] as our baseline, which is one of the SOTA 6-DoF grasp pose prediction networks. Taking point cloud and segmentation mask from the previous module as input, it predicts a set of grasp poses $\mathbf{G}$, which can be further interpreted as the 6-DoF pose of a gripper.

Grasp Pose Predictor aims to predict the orientation and translation of the gripper under the camera coordinates, as well as the width of the gripper. We represent the grasp pose $\mathbf{G}$ following [3] as:

$$\mathbf{G} = [\mathbf{R}, \quad \mathbf{t}, \quad w], \tag{3}$$

where $\mathbf{R} \in \mathbb{R}^{3\times3}$ denotes the gripper orientation, $\mathbf{t} \in \mathbb{R}^{3\times1}$ denotes the center of grasp and $w \in \mathbb{R}$ denotes the gripper width that is suitable for grasping the target object.

IV. EXPERIMENTS

A. Benchmark and Metric

GraspNet-1Billion [3] is a well-acknowledged dataset in 6-DoF robotic grasping, which includes 190 cluttered scenes captured in the real world by Kinect/Realsense camera in 256 views. Images are obtained by moving a robotic arm along predefined paths, covering 256 unique viewpoints on a quarter sphere. To assess the quality of grasps generated

TABLE I: A comparison with SOTA methods on GraspNet-1Billion dataset.

Model	Seen			Similar			Novel		
	AP	**AP**$_{0.8}$	**AP**$_{0.4}$	**AP**	**AP**$_{0.8}$	**AP**$_{0.4}$	**AP**	**AP**$_{0.8}$	**AP**$_{0.4}$
PointNet GPD [5]	25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74
GraspNet Baseline [3]	27.56	33.43	16.95	26.11	34.18	14.23	10.55	11.25	3.98
Li *et al.* [29]	36.55	47.22	19.24	28.36	36.11	10.85	14.01	16.56	4.82
RGB Matters [30]	27.98	33.47	17.75	27.23	36.34	15.60	12.25	12.45	5.62
SB Baseline [4]	58.95	68.18	54.88	52.97	63.24	46.99	22.63	28.53	12.00
YOSO (Ours)	**61.22**	**71.40**	**55.79**	**59.21**	**70.94**	**52.52**	**25.60**	**32.43**	**13.43**

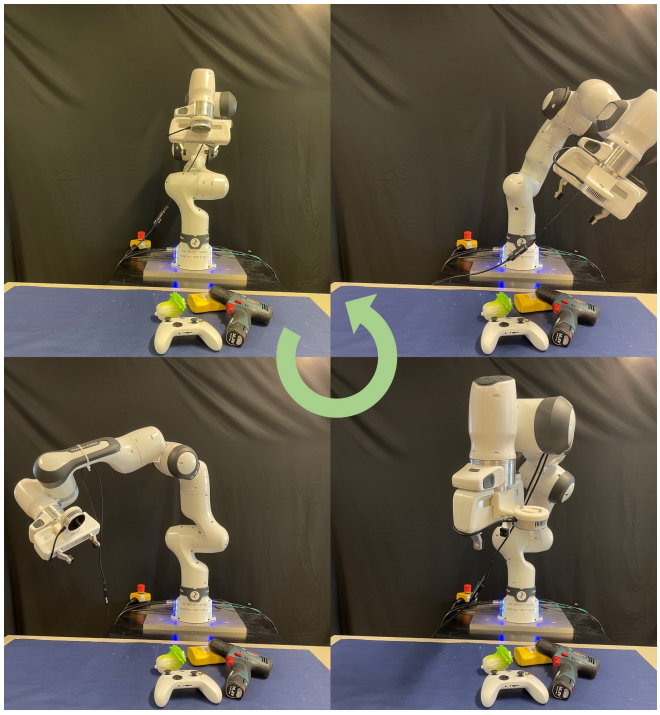


Fig. 3: Configuration of real-world experiment.

in complex, cluttered environments, we employ a precision-based evaluation metric. Following the methodology outlined in [3], $\mathbf{AP}_\mu$ is computed to signify the average *Precision@k* across a range of k values spanning from 1 to 50 with friction μ, and **AP** is obtained by the average of $\mathbf{AP}_\mu$, where μ varies from 0.2 to 1.2.

B. Implementation Details

To demonstrate that the reconstructed scene point cloud generated by our pipeline has a more informative understanding of objects in the scene to overcome occlusion from a single view and further facilitates the grasp generation process, we conduct experiments on the GraspNet-1Billion dataset by reconstructing meshes of all objects in a scene throughout all 256 views and estimate the object pose of each object in each view. Then grasp poses are generated on the reconstructed scene with the pre-trained model from Scale-balanced GraspNet [4]. However, some objects are severely occluded or even fully invisible in the first few or even half of the frames in a scene, which violates one of the assumptions of our pipeline that the object should be visible in the first few frames, it is inevitable that few of objects can not be tracked or reconstructed with BundleSDF [28]. In experiments, we manually remove objects with visually unacceptable mesh reconstruction of pose tracking results.

In our real-world experiment, an Intel RealSense L515 RGB LiDAR depth camera is mounted on a Franka Emika Panda robot arm as shown in Fig. 3. The robotic arm scans the workspace by moving the end effector along a motion trajectory that covers the hemispherical area above the workspace. All three modules run on an NVIDIA RTX3090 GPU.

V. RESULTS AND ANALYSIS

A. Comparison with State-of-the-art Methods

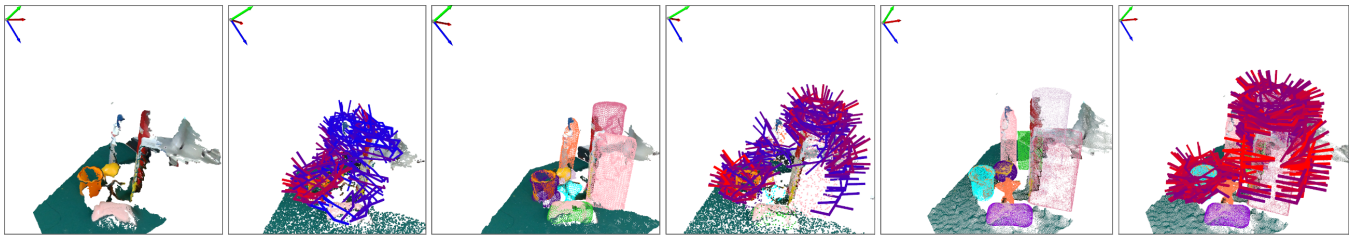
We present a comprehensive analysis of the experimental outcomes on the GraspNet-1Billion benchmark, spanning all three distinct test sets (seen, similar, and novel), as shown in Table I. Notably, evaluation result reveals that our method yields substantial improvements in grasp generation accuracy when compared to current SOTA models that take partial point clouds as input. Instead, missing parts of objects are recovered through our YOSO pipeline, which provides reconstructed point clouds to the grasp pose prediction network and significantly boosts the accuracy of grasp generation.

B. Effect of Scene Reconstruction on Grasp Generation Accuracy

For robotic grasping and manipulation tasks, having a complete and detailed representation of the environment and objects within it is vital. This entails the reconstruction of occluded parts of objects that are absent in a single-view partial point cloud. The YOSO pipeline comprises two pivotal stages: **Stage I**, which focuses on generating detailed object meshes, and **Stage II**, which is dedicated to tracking object poses to seamlessly reintegrate the generated meshes into the overall scene. The quality of the perfect scene reconstruction hinges on the flawless execution of both pose tracking and mesh generation processes.

To assess the potential of our proposed pipeline, we enhance the GraspNet-1Billion dataset by creating a detailed visible point cloud for each scene. We compile these visible points from different perspectives to form a complete scene-level point cloud, using ground truth segmentation masks from 256 unique views and corresponding object CAD models.

We then use the pre-trained Scale-Balanced model [4] to systematically assess grasp generation accuracy across three different input conditions: partial point clouds, scene point clouds reconstructed via YOSO, and our complete scene



(a) (b) (c) (d) (e) (f)

Fig. 4: Qualitative comparison of grasp prediction with partial point cloud and reconstructed scene on GraspNet-1Billion dataset. Color varies from red to blue to represent the grasp quality from high to low. (a). Partial point cloud back-projected from depth image. (b). Grasps that are generated on a partial point cloud. (c). Reconstructed scene from YOSO pipeline. (d). Grasps that are generated on the reconstructed scene. (e). Complete scene-level point cloud. (f). Grasps generated on the complete scene-level point cloud.

TABLE II: Comparison of grasp generation accuracy between different qualities of input point cloud. Partial represents partial point cloud back-projected from a depth image. YOSO (Ours) represents scene point cloud reconstructed by our YOSO pipeline. Fully Visible represents scene-level fully visible point cloud, which is regarded as a perfect reconstructed scene.

AP / Input PC	Seen	Similar	Novel
Partial	58.95	52.97	22.63
YOSO (Ours)	**61.22**	**59.21**	**25.60**
Fully Visible	69.76	65.58	29.73

TABLE III: A breakdown of the time required to execute each module in the pipeline.

	XMem	BundleSDF	Scale-balanced GraspNet	Total
Time	33.3ms	76ms	250ms	359.3ms

reconstruction from ground truth data, noted as Fully Visible. The results, presented in Table II, suggest that providing more detailed input point clouds can potentially improve the performance of grasp prediction networks. However, while YOSO provides substantial improvements, as shown in Table I, there is still a notable gap when compared to the outcomes achieved with complete scene reconstructions.

Our pipeline's modular design allows for future improvements, which may include enhancing the *Video-segmentation Module* and improving the *Object Pose Tracker and Mesh Generator* components with more sophisticated algorithms, all while maintaining real-time processing capabilities.

In our qualitative evaluation using the GraspNet-1Billion dataset, we examine the impact of different levels of scene reconstruction detail on grasp generation, as illustrated in Fig. 4. The analysis indicates that as the quality of scene reconstruction increases, the diversity and accuracy of the generated grasps improve correspondingly.

C. Inference Time Analysis

In our proposed pipeline, a scene is only scanned once in **Stage I** to record RGB-D video for target object registration, which can be regarded as preparation phase. Therefore, we focus more on analyzing the inference time of each module during testing phase (**Stage II**), including Sec. III-A *Video-segmentation Module* (XMem), Sec. III-B *Object Pose Tracker* (BundleSDF) and Sec. III-C *Grasp Pose Predictor* (Scale-Balanced GraspNet), as shown in Table III. For *Video-segmentation Module* and and *Grasp Pose Predictor*, they takes input as a whole for prediction. For *Object Pose Tracker*, the incorporated BundleSDF can only process one instance each time. When handling a single object, the overall inference of our pipeline amounts to 359.3 ms, in which scene reconstruction (combining XMem and BundleSDF) occupying merely 109.3 ms. Therefore the scene reconstruction runs at 9.15 frames per second (FPS) and the whole YOSO pipeline runs at 2.8 FPS. For scenarios with M target objects, overall inference becomes $76 * M + 283.3$ ms if object tracking is performed sequentially, which is still faster than 1 FPS if $M < 10$. Alternatively, object pose tracking can be performed in multiple threads, leading to similar overall inference as single-object situation.

VI. CONCLUSIONS

In this research work, we introduce our innovative dynamic scene reconstruction pipeline, YOSO, tailored for 6-DoF robotic grasping of novel objects. This pipeline streamlines the process by capturing the workspace in a single scan, generating object meshes, and storing essential key features for each novel object. During the testing phase, it continues to track the object's pose, seamlessly integrating the generated mesh back into the scene. By taking the reconstructed scene point cloud as input, accuracy and diversity of grasping strategies predicted by our incorporated grasp generation algorithm are significantly enhanced. Our pipeline reconstruct the scene at near real-time while providing more informative scene point cloud to the grasp generation algorithms, which is more applicable in real-world robotic grasping. Our comprehensive evaluation, conducted against the perfect scene visible point cloud data, demonstrates the promising potential for enhancing scene reconstruction quality, thereby achieving higher accuracy in grasp generation.

ACKNOWLEDGMENT

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project #A18A2b0046).

REFERENCES

[1] H. Duan, P. Wang, Y. Huang, G. Xu, W. Wei, and X. Shen, "Robotics dexterous grasping: The methods based on point cloud and deep learning," *Frontiers in Neurorobotics*, vol. 15, 2021. 1

[2] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, pp. 1–22, 2023. 1

[3] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 444–11 453. 1, 2, 4, 5

[4] M. Haoxiang and D. Huang, "Towards scale balanced 6-dof grasp detection in cluttered scenes," in *Conference on Robot Learning (CoRL)*, 2022. 1, 2, 4, 5

[5] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "PointNetGPD: Detecting grasp configurations from point sets," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 1, 5

[6] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds," *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3619–3625, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214612215 1

[7] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, "Gpr: Grasp pose refinement network for cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4295–4302. 1

[8] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 438–13 444. 1, 2

[9] J. Lundell, F. Verdoja, and V. Kyrki, "Beyond top-grasps through scene completion," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 545–551. 1

[10] S. S. Mohammadi, N. F. Duarte, D. Dimou, Y. Wang, M. Taiana, P. Morerio, A. Dehban, P. Moreno, A. Bernardino, A. Del Bue *et al.*, "3dsgrasp: 3d shape-completion for robotic grasp," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3815–3822. 1, 2

[11] D. Hidalgo-Carvajal, H. Chen, G. C. Bettelani, J. Jung, M. Zavaglia, L. Busse, A. Naceri, S. Leutenegger, and S. Haddadin, "Anthropomorphic grasping with neural object shape completion," *IEEE Robotics and Automation Letters*, 2023. 1

[12] J. Ichnowski*, Y. Avigal*, J. Kerr, and K. Goldberg, "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Conference on Robot Learning (CoRL)*, 2020. 1, 3

[13] J. Kerr, L. Fu, H. Huang, Y. Avigal, M. Tancik, J. Ichnowski, A. Kanazawa, and K. Goldberg, "Evo-nerf: Evolving nerf for sequential robot grasping of transparent objects," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 353–367. [Online]. Available: https://proceedings.mlr.press/v205/kerr23a.html 1, 3

[14] Q. Dai, Y. Zhu, Y. Geng, C. Ruan, J. Zhang, and H. Wang, "Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 2, 3

[15] M. Breyer, J. J. Chung, L. Ott, S. Roland, and N. Juan, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*, 2020. 1, 2

[16] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," 2021. 1, 2

[17] K. Liu, D. Kulic, and J. Ichnowski, Eds., *Volumetric-based Contact Point Detection for 7-DoF Grasping*, ser. Proceedings of Machine Learning Research, vol. 205. PMLR, 14–18 Dec 2023. 1

[18] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 944–15 953. 2

[19] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 498–12 507. 2

[20] X. Wang, M. H. Ang, and G. H. Lee, "Voxel-based network for shape completion by leveraging edge generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 13 189–13 198. 2

[21] J. Lundell, F. Verdoja, and V. Kyrki, "Robust grasp planning over uncertain shape completions," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1526–1532. 2

[22] A. Rosasco, S. Berti, F. Bottarel, M. Colledanchise, and L. Natale, "Towards confidence-guided shape completion for robotic applications," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 580–586. 2

[23] M. Humt, D. Winkelbauer, U. Hillenbrand, and B. Bäuml, "Combining shape completion and grasp prediction for fast and versatile grasping with a multi-fingered hand," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*. IEEE, 2023, pp. 1–8. 2

[24] 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," in *ECCV*, 2020. 2

[25] 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. 2

[26] L. Wang, R. Guo, Q. Vuong, Y. Qin, H. Su, and H. Christensen, "A real2sim2real method for robust object grasping with neural surface reconstruction," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2023. 3

[27] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *ECCV*, 2022. 3

[28] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Muller, A. Evans, D. Fox, J. Kautz, and S. Birchfield, "Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2023, pp. 606–617. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00066 3, 4, 5

[29] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 3571–3578. 5

[30] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images," in *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2021. 5