SegmentAnything-Based Approach to Scene Understanding and Grasp Generation

Songting Liu[1], Zhezhi Lei[2⋆], Haiyue Zhu[3⋆⋆], Jun Ma[4], and Zhiping Lin[1]

[1] School of Electrical & Electronic Engineering, Nanyang Technological University (NTU), Singapore 639798
{lius0081,ezplin}@ntu.edu.sg
[2] School of Electrical & Computer Engineering, National University of Singapore (NUS), Singapore 117583
lei_zhezhi@u.nus.edu
[3] Singapore Institute of Manufacturing Technology (SIMTech), Agency for Science,Technology and Research (A*STAR), 2 Fusionopolis Way, Singapore 138634
zhu_haiyue@simtech.a-star.edu.sg
[4] Robotics and Autonomous Systems Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China
jun.ma@ust.hk

Abstract. Autonomous robot grasping in multi-object scenarios poses significant challenges, requiring precise grasp candidate detection, determination of object-grasp affiliations. To solve these challenges, this research presents a novel approach to address these challenges by developing a dedicated grasp detection model called GraspAnything, which is extended from SegmentAnything model. The GraspAnything model, based on SegmentAnything (SAM) model, receives bounding boxes as prompts and simultaneously outputs the mask of objects and all possible grasp poses for parallel jaw gripper. A grasp decoder module is added to the SAM model to enable grasp detection functionality. Experiment results have demonstrate the effectiveness of our model in grasp detection tasks. The implications of this research extend to various industrial applications, such as object picking and sorting, where intelligent robot grasping can significantly enhance efficiency and automation. The developed models and approaches contribute to the advancement of autonomous robot grasping in complex, multi-object environments.

Keywords: Autonomous Robot Grasping · Grasp Detection · Multi-Object Environments.

1 INTRODUCTION

Robot grasping has been a fundamental challenge in the field of robotics and automation. The ability to accurately detect and grasp objects is crucial for

⋆ The first two authors contributed equally to this work
⋆⋆ Corresponding Author

various applications, such as manufacturing, warehousing, and home assistance. Traditional approaches to robot grasping often rely on predefined object models and carefully engineered features, limiting their flexibility and adaptability to new objects and environments.

Recent advancements in deep learning and computer vision have paved the way for more intelligent and versatile robot grasping systems. Models like GG-CNN [2], Dex-Net [3] and GR-ConvNet [4] have demonstrated remarkable capabilities in grasp pose detections based on RGB and depth map, enabling operations on unseen objects with high success rate. However, the task of autonomous robot grasping in multi-object scenarios remains a significant challenge, as it requires not only precise grasp candidate detection but also an understanding of object-grasp affiliations.

The motivation behind this research stems from the need to develop intelligent robot grasping systems that can operate effectively in complex, multi-object environments. Existing robot grasping methods are only trained to predict all possible grasp poses but ignoring objects, hence requiring a separate object detection model to detect the location of target object. This limitation hinders their practical applicability in dynamic environments where novel objects may be encountered. Therefore, there is a strong motivation to develop an open-vocabulary grasping system that can adapt to new objects without the need for retraining. [5–7].

In this research, we focus on grasp detection task and propose a grasp detection model called GraspAnything which can recognize the pose of objects, especially in complex environments

2 RELATED WORK

Robot grasping detection is a crucial component in the development of autonomous robotic systems capable of manipulating objects in various environments. The goal of grasp detection is to identify suitable grasping points or regions on an object that allow a robot to securely and stably hold the object for manipulation tasks.

Traditionally, grasp detection methods relied on analytical approaches based on geometric and physical models of objects and robot grippers. These methods often required precise knowledge of object shapes, sizes, and material properties, limiting their applicability in dynamic and unstructured environments. With the advent of deep learning, data-driven approaches have emerged as a prominent solution for grasp detection. These methods leverage large datasets of grasping examples to learn generalizable models that can predict suitable grasping points or regions on novel objects [8, 9].

Both the 2D rectangle grasping and 6-DoF pose grasping have been extensively explored. To address the multi-object scenarios, both 2D heatmaps and 3D equivalents are proposed to localize the graspness as the first stage, and then for corresponding pose regression in the second stage. The grasp detection is also treated as a special formation of object detection problem, and most works

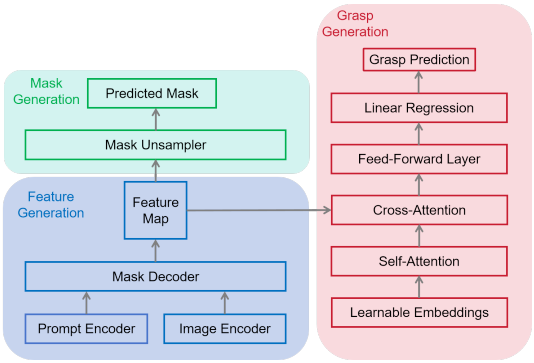

Fig. 1. The architecture of the GraspAnything model

build the grasp detectors based on two-stages object detection models such as Faster-RCNN [10].

Various sensor formats are explored as the input for the prediction network, such as RGB image, depth map, RGB-D image and point cloud. Many methods take depth image as input for grasp inference, but they rely on additional sensors and their accuracy is highly dependent on the quality of the input depth image [11, 12]. In this work, we only focus on 2D rectangle grasping detection based on RGB image input.

3 METHOD

3.1 Feature Generation

The GraspAnything model's segmentation function is built upon Meta's SegmentAnything (SAM) model [1], which has demonstrated remarkable performance in instance segmentation tasks. The SAM model is designed to generate high-quality instance masks from simple prompts, such as bounding boxes or points, making it a powerful tool for object segmentation.

Figure 1 shows the architecture of our GraspAnything model. The segmentation branch of the model consists of four main components: an image encoder, a prompt encoder, a mask decoder, and a mask upsampler. The image encoder is a vision transformer (ViT) that extracts features from the input image. The prompt encoder processes the input prompts, such as bounding boxes or points, and generates corresponding embeddings. The mask decoder takes the image features and prompt embeddings as input and produces mask latent in high dimension but low resolution. The mask upsampler is to upsample the mask latent to the final mask output which is of the same size of input image.

During training, we set all model parameters to be trainable. To prevent model from forgetting mask prediction ability, we also apply mask prediction loss during training:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\widehat{y_i}\right) + \left(1-y_i\right)\log\left(1-\widehat{y_i}\right)\right]$$
$$L_{Dice} = 1 - \frac{2\times\sum_{i=1}^{N} y_i\widehat{y_i}}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N}\widehat{y_i}^2}$$
$$L_{mask} = \lambda_{BCE}L_{BCE} + \lambda_{Dice}L_{Dice}$$

where y_i is the ground truth mask value for each pixel, $\widehat{y_i}$ is the predicted mask value for each pixel, N is the total number of pixels (in our case is always 1024×1024), λ_{BCE} and λ_{Dice} are loss coefficients for binary cross-entropy loss and dice loss, respectively.

3.2 Grasp Generation

We extend the functionality of SAM to grasp detection by adding a grasp detection branch. A detailed breakdown of the grasp detection branch is in Figure 1. Given the mask decoder output F_m, we use it as the context for cross attention blocks in the grasp decoder. To be specific, given the learnable embeddings $E_0 \in R^{N_q\times 256}$:

$$E_i = FF_i\left(CA_i\left(SA_i\left(E_{i-1}\right),\ \ F_m,\ F_m\right)\right),\ i\in[1,\ N_{layers}]$$

where N_q is a hyperparameter denoting the maximum number of grasp boxes to detect, N_{layers} is the number of transformer blocks in the grasp decoder, SA stands for self-attention module, CA stands for cross-attention module, and FF stands for the feed-forward layer. With the final output $E_{N_{layers}}$, grasp predictions are calculated as:

$$\left[\widehat{\mathbf{x}},\ \widehat{\mathbf{y}},\ \widehat{\mathbf{o}},\ \widehat{\mathbf{w}},\ \widehat{\theta},\widehat{\mathbf{c}}\right] = \sigma(Linear\left(E_{N_{layers}}\right)),\ [\widehat{\mathbf{x}},\ \widehat{\mathbf{y}},\ \widehat{\mathbf{o}},\ \widehat{\mathbf{w}},\ \widehat{\theta},\widehat{\mathbf{c}}]\ \in\mathbf{R}^{6\times N_q}$$

where $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$ are the center coordinates of the grasp boxes, $\widehat{\mathbf{o}}$ is the predicted gripper opening distance, $\widehat{\mathbf{w}}$ is the allowed range for the gripper to shift, $\widehat{\theta}$ is the predicted rotation angle of the gripper regarding to horizontal axis, $\widehat{\mathbf{c}}$ is the predicted confidence scores, σ stands for the sigmoid activation function, *Linear* is the final regression layer.

To calculate loss, we create a 1-to-1 matching between ground truth grasp boxes and the predicted ones. Following DETR, we use Hungarian Match for this matching problem. First, we construct a cost matrix $C,\ C\in R^{N_q\times n}$, where n is the number of ground truth grasp annotations. Each element C_{ij} of the matrix represents the cost of assigning the i-th prediction to the j-th ground truth object. We form the cost matrix as follows:

$$\begin{aligned} C_{ij} = {} & \lambda_c \times \text{CrossEntropyLoss}\left(\widehat{c_i}, c_j\right) \\ & + \lambda_{bbox} \times \text{L1Loss}\left(\left[\widehat{x_i},\widehat{y_i},\widehat{o_i},\widehat{w_i},\widehat{\theta_i}\right],[x_j,y_j,o_j,w_j,\theta_j]\right) \\ & + \lambda_{giou} \times \text{GIOULoss}\left(\left[\widehat{x_i},\widehat{y_i},\widehat{o_i},\widehat{w_i}\right],[x_j,y_j,o_j,w_j]\right) \end{aligned}$$

After the cost matrix is obtained, we use the Hungarian algorithm to find the assignment of predictions to ground truth objects that minimizes the total cost. Formally, we seek to find a bijection $f{:}P{\to}G$ that minimizes the total cost: $\min_f \sum_{i=1}^{N} C_{i,\ f(i)}$, where P is the set of predictions, G is the set of ground truth objects, and $f(i)$ gives the ground truth object assigned to the i-th prediction. Once the optimal assignment is found, compute the loss for the model based on this assignment:

$$\begin{aligned} L_{grasp} = \lambda_c \times \text{CrossEntropyLoss}\left(\widehat{\mathbf{c}}, \mathbf{c}\right) \\ + \lambda_{bbox} \times \text{L1Loss}\left(\left[\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \widehat{\mathbf{o}}, \widehat{\mathbf{w}}, \widehat{\theta}\right], [\mathbf{x}, \mathbf{y}, \mathbf{o}, \mathbf{w}, \theta]\right) \\ + \lambda_{bbox} \times \text{GIOULoss}\left([\widehat{\mathbf{x}}, \widehat{\mathbf{y}}, \widehat{\mathbf{o}}, \widehat{\mathbf{w}}], [\mathbf{x}, \mathbf{y}, \mathbf{o}, \mathbf{w}]\right) \end{aligned}$$

4 RESULTS

We built GraspAnything upon SegmentAnything ViT-B with 12-layered, 768-dimensional image encoder. The additional grasp decoder is a 6-layered, 256-dimensional transformer decoder.

The GraspAnything model was trained on combined dataset of OCID-grasp [14], VMRD [13], Jacquard [13] and MetaGraspNet [16] for 500k steps, with batch size of 16. Learning rate of the grasp decoder extension is set to 1×10^{-4} and the pretrained SegmentAnything modules to 1×10^{-5}. For evaluations, we further finetune the model on each single dataset for 100k steps to improve performance on the dataset domain and fit to the dataset distribution.

We evaluate the performance of our system on OCID-grasp and Jacquard using grasp success rate which measures the ability of the model to predict accurate grasp boxes for a given object.

The results are presented in Table 1. Figure 2 shows various detection results on images from these two datasets. With completely different structure from previous models which mostly based on heatmap prediction or anchor-based prediction, our model achieves grasp success rates that are nearly equal to the state-of-the-art (SOTA) models on both datasets, while keeping the flexibility in object selection and minimizing the misallocation from predicted grasps to the desired object. This demonstrates the effectiveness of our approach in accurately predicting grasp boxes for objects in cluttered environments.

5 CONCLUSION

This research successfully tackles the complexities inherent in autonomous robot grasping within multi-object environments. By extending the SegmentAnything model to create the novel GraspAnything model, we have developed a dedicated solution that excels in identifying precise grasp candidates. Experimental results have demonstrated effectiveness of GraspAnything in grasp detection tasks, thereby affirming its potential for broad application in robotic grasping systems. This advancement represents a step forward in enhancing the efficiency and reliability of autonomous robotic interactions in complex scenarios.

Fig. 2. Example predictions on the OCID-grasp and Jacquard dataset. Left of each pair: original images and box prompts for desired object to grasp. Right of each pair: Top-3 grasps and instance masks predicted by the model.

Table 1. Evaluation of grasping performance

Dataset	Methods	Grasp Success Rate
OCID-grasp	GG-CNN	63.4
	GR-ConvNet	74.1
	EfficientGrasp RGB-D [17]	76.4
	Det-Seg-Refine [14]	89.02
	GSMR-CNN [18]	**91.9**
	GraspAnything (Ours)	87.3
Jacquard	Zhou [19]	81.95
	Depierre [15]	85.74
	Song [20]	91.5
	Det-Seg-Refine	**92.65**
	GraspAnything (Ours)	89.4

References

1. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
2. Morrison, D., Corke, P., Leitner, J.: Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. arXiv preprint arXiv:1804.05172 (2018)
3. Mahler, J., Matl, M., Liu, X., Li, A., Gealy, D., Goldberg, K.: Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In: 2018 IEEE International Conference on robotics and automation (ICRA), pp. 5620–5627. IEEE (2018)
4. Kumra, S., Joshi, S., Sahin, F.: Gr-convnet v2: A real-time multi-grasp detection network for robotic grasping. Sensors **22**(16), 6208 (2022)
5. Morgan, A.S., Wen, B., Liang, J., et al.: Vision-driven compliant manipulation for reliable, high-precision assembly tasks. arXiv preprint **arXiv:2106.14070** (2021)
6. Di Lillo, P., Arrichiello, F., Di Vito, D., Antonelli, G.: BCI-controlled assistive manipulator: developed architecture and experimental results. IEEE Transactions on Cognitive and Developmental Systems **13**(1), 91-104 (2020)
7. Zhang, H., Lan, X., Bai, S., et al.: A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. In: Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 6435-6442 (2019)

8. Dong, M., Zhang, J.: A review of robotic grasp detection technology. Robotica **2023**, 1-40 (2023)
9. Caldera, S., Rassau, A., Chai, D.: Review of deep learning methods in robotic grasp detection. Multimodal Technologies and Interaction **2**(3), 57 (2018)
10. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)
11. Ainetter, S., Böhm, C., Dhakate, R., et al.: Depth-aware object segmentation and grasp detection for robotic picking tasks. arXiv preprint **arXiv:2111.11114** (2021)
12. Redmon, J., Angelova, A.: Real-time grasp detection using convolutional neural networks. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp. 1316-1322 (2015)
13. Zhang, H., Lan, X., Zhou, X., Tian, Z., Zhang, Y., Zheng, N.: Visual manipulation relationship network for autonomous robotics. In: 2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids), pp. 118–125. IEEE (2018)
14. Ainetter, S., Fraundorfer, F.: End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb. In: 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 13452–13458. IEEE (2021)
15. Depierre, A., Dellandréa, E., Chen, L.: Jacquard: A large scale dataset for robotic grasp detection. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3511–3516. IEEE (2018)
16. Gilles, M., Chen, Y., Zeng, E.Z., Wu, Y., Furmans, K., Wong, A., Rayyes, R.: MetaGraspNetV2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping. IEEE Transactions on Automation Science and Engineering (2023)
17. Cao, H., Chen, G., Li, Z., Feng, Q., Lin, J., Knoll, A.: Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation. IEEE/ASME Transactions on Mechatronics (2022)
18. Holomjova, V., Starkey, A.J., Meißner, P.: GSMR-CNN: An End-to-End Trainable Architecture for Grasping Target Objects from Multi-Object Scenes. In: 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 3808–3814. IEEE (2023)
19. Zhou, X., Lan, X., Zhang, H., Tian, Z., Zhang, Y., Zheng, N.: Fully convolutional grasp detection network with oriented anchor box. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 7223–7230. IEEE (2018)
20. Song, Y., Gao, L., Li, X., Shen, W.: A novel robotic grasp detection method based on region proposal networks. Robotics and Computer-Integrated Manufacturing **65**, 101963 (2020)