2024 33rd IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 26-30, 2024. Pasadena, California, USA.

2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) | 979-8-3503-7502-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/RO-MAN60168.2024.10731387

A Deep Learning-based Grasp Pose Estimation Approach for Large-Size Deformable Objects in Clutter

Minghao Yu[1], Zhuo Li[1], Zhihao Li[1], Junjia Liu[1], Tao Teng[1] and Fei Chen[†1], *Senior Member, IEEE*

***Abstract*— Deformable objects especially large-size deformable objects grasping is unappreciated but widespread in industrial applications (e.g., clothes recycling). While it encounters several challenges, for example, the existing methods didn't take large-size deformable objects into account, no typical boundary of deformable objects. To solve the challenges, we proposed a grasp detection framework consisting of a self-trained object detection network, an instance segmentation module, and a grasp pose generation pipeline. The experiments were successfully conducted on the industrial laundry mock-up with an 88.9% success ratio. The experiments result indicates the effectiveness of the proposed framework on spatial-constrained large-size deformable objects grasping in clutter.**

I. Introduction

Robotic gripping serves as a cornerstone in industrial automation, significantly advancing the efficiency of factories by taking over repetitive and hazardous tasks from human workers. With the integration of deep learning advancements in image processing and object detection, robots have showcased remarkable capabilities in handling a variety of tasks [1]–[4]. However, while the technology for grasping normal-sized, rigid objects is well-established, the challenge of handling deformable objects remains an area of active research and development.

There are several methods considered the deformable objects grasping. Liu et al [5], [6] conducted stir-fry semi-deformable objects on a bimanual robot system and further manipulated deformable object: dough rolling. Hang Yin et al [7] surveyed more than 100 relevant studies and synthesized insights from analytical and data-driven methodologies. Isabella Huang et al [8] researched the interaction between deformable objects by physical simulation and created a data set containing 34 objects, 6800 grasp evaluations, and 1.1M grasp measurement. Han et al [9] proposed a grasping architecture for rigid grippers based on the transformer, created a fruit grasping data set, and conducted online experiments. However, these methods have two limitations: 1) only considered normal-size objects; 2) only considered table-top setup; 3) didn't consider the cluttered situation. For industrial applications like laundry (as shown in Fig. 1.), whose grasping targets are large-size objects with spatial-constrained working environments, also be appreciated because the human labor and working environments are even worse than the adopted applications like production flow lines and logistics.

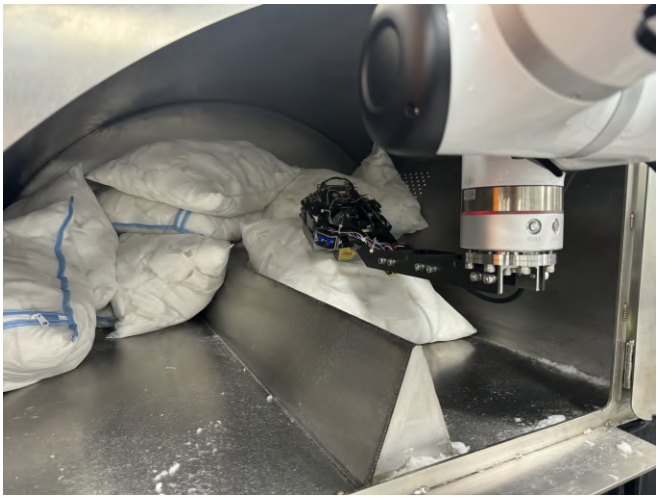

Fig. 1. **Robot arm approaching grasp target in the washing machine.** By inputting an RGB-D image, the proposed framework can detect, and segment the objects and give a reasoning and feasible 6-DoF grasp pose.

Integrating deep learning-driven perception with robotic grasping necessitates precise object detection, solid image processing [10]–[16], instance segmentation [17], and the generation of suitable 6-DoF (Degrees of Freedom) grasp poses. Recent studies have incorporated offline object detection modules and force-analysis-based methods for grasping objects [18], [19]. These approaches leverage calculations and force analysis to develop models for grasping target objects, such as predicting and calculating the stability of a grasp based on the object's appearance and geometry.

As a result, the grasping of large-size deformable objects in industrial applications, which are conducted in spatial-constrained clutter environments, needs to be calculated according to the 3D model of target objects and use the collision checking algorithm to select feasible grasp poses from the grasp poses data set created from the force-analysis repetitively. On the other hand, these approaches require a complete and accurate 3D model of the objects that appear in the environment, which in practice, is hard to acquire. Thus, these approaches suffer excessive failure when encountering novel and deformable objects. Though simulation can accelerate the training and generalization, these kinds of object

*This work was supported in part by the Research Grants Council of the Hong Kong SAR under Grant 14211723, 14222722 and C7100-22GF and in part by InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics. *(Corresponding author: Fei Chen.)*

[1]Minghao Yu, Zhuo Li, Zhihao Li, Junjia Liu, Tao Teng and Fei Chen are with the Department of Mechanical and Automation Engineering, T-Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong (e-mail: mhyu@mae.cuhk.edu.hk; zli@mae.cuhk.edu.hk; zhihaoli@mae.cuhk.edu.hk; jjliu@mae.cuhk.edu.hk; tao.teng@cuhk.edu.hk; f.chen@ieee.org).

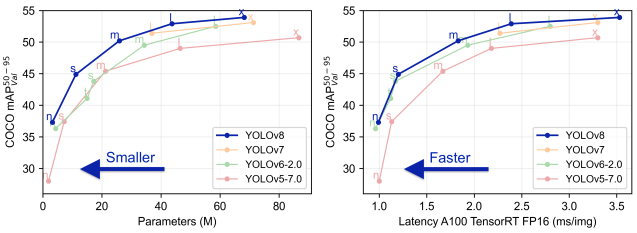

Fig. 2. **YOLO Family Comparison.** Diagram based in [20]

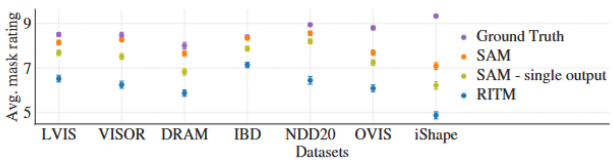

Fig. 3. Mask quality ratings by human annotators.

detection systems, which take in the RGB image, require more effort to narrow the gap between sim and real.

These limitations motivate us to propose a deep learning-based large-size deformable objects grasp pose estimation framework which can be used in cluttered situations by single-shot.

The principal contributions of this paper are outlined below:

- This paper proposes a grasp pose estimation framework based on deep learning for large-size deformable objects in clutter.
- This paper trains a neural network based on YOLOv8 for large-size deformable object detection.
- This paper proposes a grasp pose estimation module that can provide reasoning and feasible 6-DoF grasp pose for large-size deformable object grasping in a spatial-constrained environment.
- This paper verifies the proposed framework in a real industrial environment and the object detection accuracy reaches 88.9%.

II. Related Works

A. YOLOv8 Network

YOLOv8 is the newest neural network in the YOLO (You Only Look Once) family used for object detection tasks and YOLO processes images in a single pass through a convolutional neural network (CNN). Unlike traditional object detection algorithms that give region suggestions and then classify regions individually, the YOLO network first, separates the input image into a grid and outputs the predictions of bounding boxes and objects' classification from the grid cells. Compared to other YOLO families, their performance improved not only solely on the accuracy but also the balance of accuracy and speed. As illustrated in Fig. 2, this is the results of mAP, several parameters and FLOPs tested on the COCO Val 2017 data set. The accuracy and perception speed of YOLOv8 are higher than the other YOLO family, which is suitable for deployment on industrial applications.

B. Segment Anything (SAM)

Segment Anything [21] is a deep learning model trained by Meta researchers. It's a zero-shot pre-trained instance segmentation model trained on the SA-1B data set, which consists of more than 11M diverse images and more than 1.1B segmentation masks. As illustrated in Fig. 3, the Segment Anything is compared with other mask tools, and the confidence intervals for mean mask ratings are 95%. The human annotators evaluate the quality of SAM's output masks are significantly better than the currently most robust method, RITM (Reviving Iterative Training with Mask Guidance for Interactive Segmentation). Abortively, SAM's single mask output module has relatively lower ratings, though still higher than RITM. SAM's mean ratings are between 7 and 9, which corresponds to the qualitative rating standard. These results suggest that SAM's pre-trained model has the ability to segment valid masks from a single point.

C. 6-DoF Grasping Pose Generation of Deformable Objects

CNNs (Convolutional Neural Networks) have achieved outstanding detection results across various tasks [22]–[25]. Several works have been done on 6-Dof grasp pose generation via deep learning [26]–[30]. The deformable objects grasping problem requires an object detection module, like template matching based on a 2D image and an instance segmentation module, that perceives the target deformable object before grasping. Traditional Siamese Networks based on RGB images are also well-explored in single-shot detection for robotic grasping tasks. We combined the depth data and RGB images for object perception.

III. Problem Statement

A. Assumptions

The large-size deformable objects follows the assumptions: 1) objects larger than the generalized object datasets such as: YCB [31], and 2) objects smaller than the industrial laundry bags.

Our approach following the assumptions: 1) self-designed needle gripper suction grasping with known geometry parameters and 2) one RGB-D camera with known intrinsics.

For suction grasping, the sucker gripper (self-designed needle gripper in our case) approaches along the normal vector of the surface, which also matches human preference and ergonomics. Thus we add one more hypothesis: 3) The normal vector of the grasping surface served as the approach vector of the suction grasping.

B. Problem Definitions

Camera Status. The camera status can be represented by $\boldsymbol{c}$=($\boldsymbol{T}^{\boldsymbol{c}}$,$\boldsymbol{i}$), in which $\boldsymbol{T}^{\boldsymbol{c}}$ represents the extrinsic and $\boldsymbol{i}$ represents the intrinsic we know.

Point Clouds. $\boldsymbol{P}$ represents the point clouds we reconstructed from the depth image.

Suction Grasps. g=$(g_1, g_2 ... g_n)$ represents suction grasp points in 3D space. Each suction grasp can be illustrate as follows: g_i=(s, Z, r), i=1,2...n, s represents the grasp point s=(s_x, s_y, s_z), which provides the location point of suction grasping. Normally grasp point s is on the surface of the point clouds P. Z represents the unit normal vector of grasp point grasp point s. r represents the approaching vector of the suction needle gripper r=(r_x, r_y, r_z).

IV. Methodology

We proposed a 6-DoF target-driven grasp detection framework for spatial-constrained large-size deformable objects grasping in clutter. Our framework is to 1) Accurately detect the target object (previously seen or unseen) based on incomplete visual data. 2) generate the feasible 6-DoF grasp pose based on the combination of RGB-D data and point cloud.

A. Framework Overview

As shown in Fig. 4, our grasping detection framework consists of three modules: a large-size object detection module, a cluttered deformable objects instance segmentation module, and a spatial-constrained grasp pose generation module. By giving an RGB-D image input, the RGB image is forwarded to the large-size object detection module for the classification and output the bounding box for each object in the image. At the same time, it will forward to the cluttered deformable object instance segmentation module for image encoding. After the bounding boxes are given, the cluttered deformable object instance segmentation module will segment each object's mask according to the bounding box. For grasp stability estimation, we combine the object classification confidence thresholds above 0.8 and the maximum mask area to select the grasping order. Then according to the target mask, the point cloud P reconstructed from the depth image is separated into the masked target point cloud P_s and the surrounding obstacles P', then we find the center of P_s and search the nearest point from it on the mesh which outputs the grasp point s. By using the point and the surrounding 20 points of it to formulate a plane, we can use the normal vector of the plane as the Z axis of the grasp pose which matches the Z axis of camera frame. and the X, Y axes are decided by transforming the camera frame to the target frame. After the transformation matrix T is got, it can be forwarded to do further motion. Our framework can be seen as the combination of peripheral visual and a foveated visual of the object.

B. Large-size Object Detection

We adopted the YOLOv8 neural network in our large-size object detection module and trained it with the data set created by us, which contained the sole cluttered large-size deformable objects, wash bags in this case. The module extracts the features of the input RGB image and classifies the objects as wash bags. The masked RGB image is output with bounding boxes' confidence above 0.8.

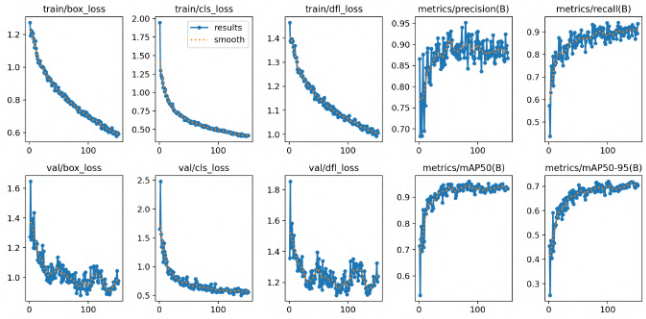

Fig. 4. Object detection neural network training result.

C. Cluttered Deformable Object Instance Segmentation

The pre-trained instance segmentation model Segment Anything (SAM) [9], is adopted in our cluttered deformable object instance segmentation module, which is trained on a hybrid of cluttered rigid and deformable objects data set. The input RGB image is encoded and forwarded to the image embedding. combining the bounding boxes output from the large-size object detection module with the text prompt, we can get the zero-shot accurate instance segmentation masks and forward to the next module.

D. Spatial-constrained Grasp Pose Generation

Our grasping module first generates the grasp pose, which takes in the reconstructed point cloud P and the masked RGB image, the output mask is decided by the object classification confidence and grasping area, and the threshold is confidence above 0.8 with maximum grasping area. The masked RGB image is then combined with the point cloud P to get the grasp target point cloud P_s and forward to do the KD-Tree nearest neighbor search from the P_s center to find the point on surface as the grasp point $s = (s_x, s_y, s_z)$ and get translation matrix T_s as follows:

$$T_s = \begin{bmatrix} 1 & 0 & 0 & s_x \\ 0 & 1 & 0 & s_y \\ 0 & 0 & 1 & s_z \end{bmatrix} \tag{1}$$

Considering the grasp orientation, we can still use the KD-Tree search to find the surround points to formulate the grasping area and use its unit normal vector $Z = (Z_x, Z_y, Z_z)$ as the grasp orientation's Z axis, the rotation angle θ can be decided by the following equation:

$$\theta = \arcsin \frac{Z \times Z'}{Z \cdot Z'} \tag{2}$$

Unit vector Z' represents the camera frame's Z axis, by finding the rotation angle, we can get the rotation matrix R by using the Rodrigues' equation:

$$R = \cos\theta \cdot I + \sin\theta \cdot \begin{bmatrix} 0 & -Z_z & Z_y \\ Z_z & 0 & -Z_x \\ -Z_y & Z_x & 0 \end{bmatrix} + (1 - \cos\theta) \cdot Z \cdot Z^t \tag{3}$$

And the overall transformation matrix T can be decided by:

$$T = T_s \times R \tag{4}$$

Fig. 5. Deformable object instance segmentation results.

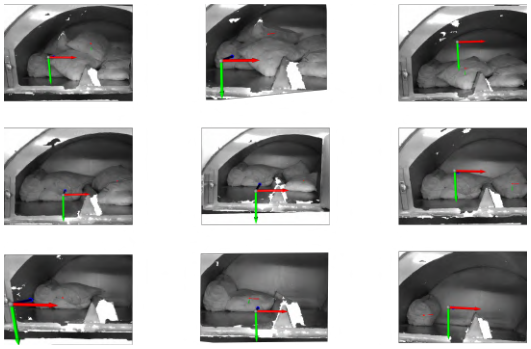

Fig. 6. Spatial-constained grasp pose generation results.

E. Data Acquisition and Model Training

The entire framework operates under a self-supervised training paradigm. As we discussed above, the instance segmentation module is adopted from Segment Anything (SAM) which is robust and validated on different data sets. So, the main influence factor of the system is the bounding box output. We trained the object detection module with 1K RGB images and 1 classification of different shapes and light of object with ground truth labels. During the label of the data set, the markup of each image is performed and the boundary of each bag is highlighted. The data set was divided into the train, valid and test samples in the ratio of 7:2:1. Since the data set is small, to avoid over-fitting, we tuned the parameters with a weight decay of 0.0007 and learning rate of 0.000688. The network training has been conducted for 200 epochs and 8 batches on the Nvidia RTX 3070 laptop GPU and the result is shown in Fig. 5.

V. Experiments

A. Implementation Details

The upper computer system of the experiments adopts a laptop with an Intel i7-11700H CPU, 3070 laptop graphic card, and 32GB RAM with Ubuntu 20.04 OS. The camera we used Intel Realsense D435i RGB-D camera. The mock-up consists of 9 bags in total, each time we will fetch one out with a self-designed needle gripper if the object is detected and segmented appropriately with a feasible grasp pose. The experiments are designed to answer three questions: 1) Is the perception module capable of consistently identifying the target object in various situations like corner stacking, and weak illumination by one shot, 2) Can the perception module output the reasoning and feasible target object mask for generate grasp pose, 3) Can the grasp pose generation module generates the reasoning and feasible grasp pose.

B. Large-size Object Detection

Firstly, we used a zero-shot object detection model Owl-Vit to test on the mock-up with 9 wash bags inside. The results are shown in Fig. 8 a). Despite its' impressive accuracy in normal-size object detection, its success rate is literally low at 44.4% in large-size object detection. The object detection results by our trained YOLOv8 neural network are shown in Fig. 8 b). Each bounding box represents a wash bag classified by the trained network. Except for a multiple detection, which takes three wash bags as one. Other rounds all consist of single bounding boxes which can be used for instance segmentation and further grasp. The results comparison between Owl-Vit and our trained model is shown in Table 1.

C. Cluttered Deformable Object Instance Segmentation

The instance segmentation results are shown in Fig. 6. Though the boundary of deformable objects varies, our instance segmentation module performed well and its' accuracy and robustness can be validated in this experiment. By collecting the data on cluttered objects in narrow spaces like corners, the spatial-constrained situations are also been tackled well according to the result. The only wrong segmentation result is due to the wrong bounding box output from the large-size object detection module.

D. Spatial-constrained Grasp Pose Generation

The grasp pose generation results are shown in Fig. 7. We used the cluttered deformable object instance segmentation results to get the accordingly target point cloud, which we can use to find the grasp pose. As we can see in the results, the grasp pose generation module first outputs a reasonable and feasible grasping candidate which is the suitable grasp target in clutter. Secondly, The grasp point of each grasp candidate appears on the outer surface with more visual area than other surfaces which means its grasp stability is higher, compared to other surfaces. Finally, the approaching pose is along the normal vector of the grasp point and is transformed from the camera pose, so it's easy to calculate the transformation matrix for the robot to do motion planning.

E. Industrial Laundry Bag Grasping

The result of grasping the industrial laundry bag is shown in Fig. 8. We mounted an Intel realsense D405 camera on the needle gripper for the grasp pose generation. Once the grasp pose is validated, the collision-free motion planning module will conduct the grasping as the Fig. 8 shown.

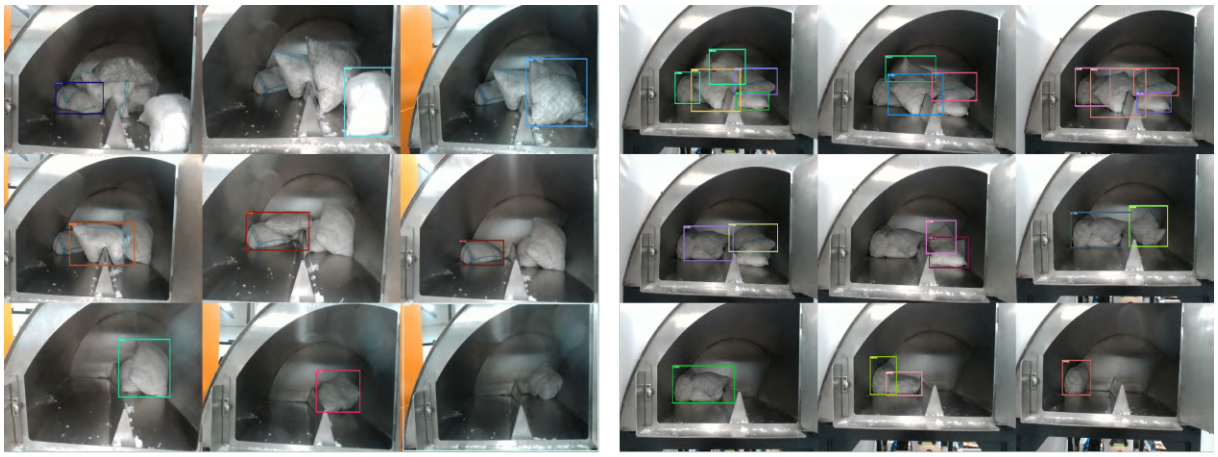

Fig. 7. Owl-Vit object detection result a) and trained YOLOv8 object detection result b).

TABLE I

OBJECT DETECTION RESULTS COMPARISON OF OWL-VIT AND OUR TRAINED YOLOV8 NETWORK ON CUSTOM LARGE-SIZE DEFORMABLE OBJECTS DATA SET.

Model	Wash Bags Amount	Successful Detection	Failed Detection	Failed Reason	Success Rate
Owl-Vit	9	4	5	Multiple detection, wrong object detecion	44.4%
Large-size Object Detection	9	8	1	Multiple detection	88.9%

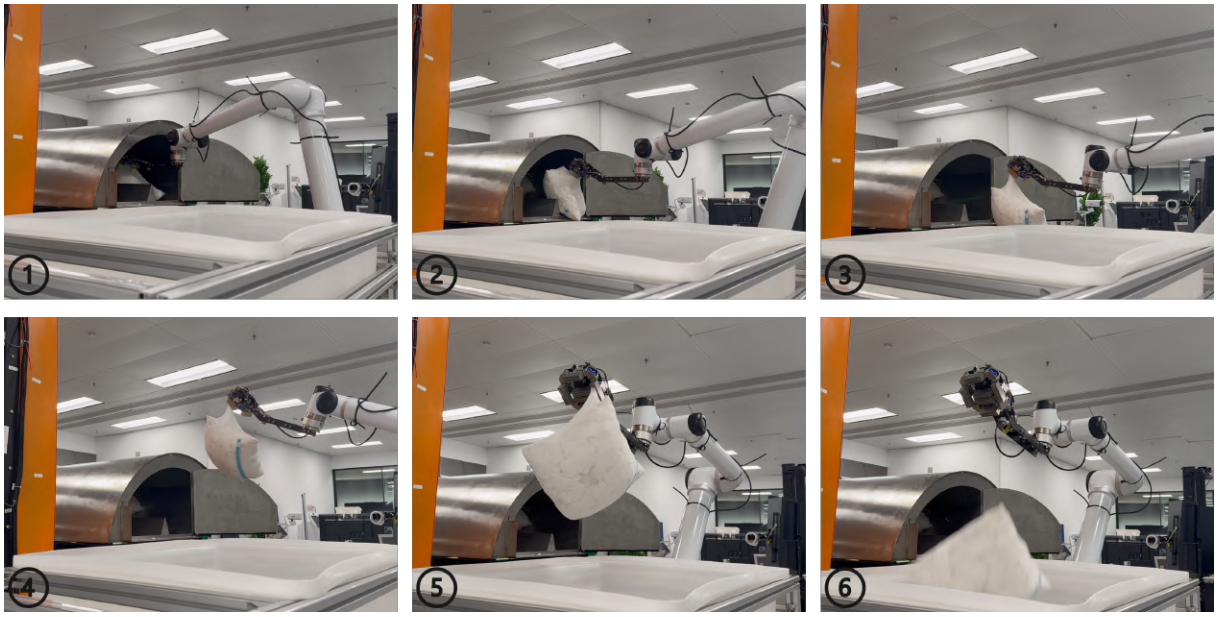

Fig. 8. Grasping of industrial laundry bag.

CONCLUSION

In this paper, a deep learning approach for large-size deformable objects grasping in clutter is proposed. It requires single RGB-D image as input and outputs the grasp pose for suction grasping with one-shot. We tackled the several limitations of large-size deformable objects grasping and conducted the experiments on the industrial washing machine mock-up. Experiment results demonstrate the effectiveness and accuracy of our method, which attains a detect success rate of 88.9% in real-world laundry grasping scenarios which performs better than the current state-of-art works. Considering of the spatial-constrained working environments, further research of us may be on the perception with partial observation and dynamic collision-awareness motion planning. Exploiting public available dataset holds the promise to refining the perception in those challenging cases [32].

REFERENCES

[1] S. Li, Z. Li, K. Han, X. Li, Y. Xiong, and Z. Xie, "An end-to-end spatial grasp prediction model for humanoid multi-fingered hand using deep network," in *2021 6th International Conference on Control, Robotics and Cybernetics (CRC)*, pp. 130–136, IEEE, 2021.

[2] Z. Li, S. Li, K. Han, X. Li, Y. Xiong, and Z. Xie, "Planning multi-fingered grasps with reachability awareness in unrestricted workspace," *Journal of Intelligent & Robotic Systems*, vol. 107, no. 3, p. 39, 2023.

[3] Z. Xie, S. Wang, W. Zhao, and Z. Guo, "A robust context attention network for human hand detection," *Expert Systems with Applications*, vol. 208, p. 118132, 2022.

[4] Z. Xie, S. Wang, W. Zhao, and Z. Guo, "Context attention module for human hand detection," in *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 555–560, IEEE, 2019.

[5] J. Liu, Y. Chen, Z. Dong, S. Wang, S. Calinon, M. Li, and F. Chen, "Robot cooking with stir-fry: Bimanual non-prehensile manipulation of semi-fluid objects," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5159–5166, 2022.

[6] J. Liu, Z. Li, W. Lin, S. Calinon, K. C. Tan, and F. Chen, "Softgpt: Learn goal-oriented soft object manipulation skills by generative pre-trained heterogeneous graph transformer," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4920–4925, IEEE, 2023.

[7] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.

[8] I. Huang, Y. Narang, C. Eppner, B. Sundaralingam, M. Macklin, R. Bajcsy, T. Hermans, and D. Fox, "Defgraspsim: Physics-based simulation of grasp outcomes for 3d deformable objects," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6274–6281, 2022.

[9] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao, "Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer," *arXiv preprint arXiv:2112.06374*, 2021.

[10] L. Yu and M. T. Orchard, "Single image interpolation exploiting semi-local similarity," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1722–1726, IEEE, 2019.

[11] Y. Gong, W. Tang, L. Zhou, L. Yu, and G. Qiu, "A discrete scheme for computing image's weighted gaussian curvature," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1919–1923, IEEE, 2021.

[12] L. Yu and M. T. Orchard, "Location-directed image modeling and its application to image interpolation," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 2192–2196, IEEE, 2018.

[13] L. Yu, D. Liu, H. Mansour, P. T. Boufounos, and Y. Ma, "Blind multi-spectral image pan-sharpening," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1429–1433, IEEE, 2020.

[14] L. Yu and M. T. Orchard, "Accurate edge location identification based on location-directed image modeling," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2971–2975, IEEE, 2019.

[15] L. Yu and M. T. Orchard, "When spatially-variant filtering meets low-rank regularization: Exploiting non-local similarity for single image interpolation," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 200–204, IEEE, 2019.

[16] L. Yu, D. Liu, H. Mansour, and P. T. Boufounos, "Fast and high-quality blind multi-spectral image pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[17] Z. Kou, G. Cui, S. Wang, W. Zhao, and C. Xu, "Improve cam with auto-adapted segmentation and co-supervised augmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3598–3606, 2021.

[18] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "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.

[19] 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*, pp. 11444–11453, 2020.

[20] J. Terven and D. Cordova-Esparza, "A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas," *arXiv preprint arXiv:2304.00501*, 2023.

[21] 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*, pp. 4015–4026, 2023.

[22] K. Wan, X. Liu, J. Yu, X. Zhang, X. Du, and N. Guizani, "Compiler-based efficient cnn model construction for 5g edge devices," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 8, pp. 5261–5274, 2021.

[23] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 922–928, IEEE, 2015.

[24] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1746–1754, 2017.

[25] K. Wan, S. Yang, B. Feng, Y. Ding, and L. Xie, "Reconciling feature-reuse and overfitting in densenet with specialized dropout," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 760–767, IEEE, 2019.

[26] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 598–605, IEEE, 2016.

[27] L. Xie, S. He, Z. Zhang, K. Lin, X. Bo, S. Yang, B. Feng, K. Wan, K. Yang, J. Yang, *et al.*, "Domain-adversarial multi-task framework for novel therapeutic property prediction of compounds," *Bioinformatics*, vol. 36, no. 9, pp. 2848–2855, 2020.

[28] X. Liu, K. Wan, Y. Ding, X. Zhang, and Q. Zhu, "Weighted-sampling audio adversarial example attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 4908–4915, 2020.

[29] W. Zhao, S. Wang, Z. Xie, J. Shi, and C. Xu, "Gan-em: Gan based em learning framework," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 4404–4411, International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[30] S. Wang, W. Zhao, Z. Kou, J. Shi, and C. Xu, "How to make a blt sandwich? learning vqa towards understanding web instructional videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1130–1139, 2021.

[31] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*, pp. 510–517, IEEE, 2015.

[32] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu, *et al.*, "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision," *arXiv preprint arXiv:2312.16256*, 2023.