SÃO MIGUEL
SEIA'25

Sensors and Electronic Instrumentation Advances:

Proceedings of the 11th International Conference on Sensors and Electronic Instrumentation Advances

24-26 September 2025

Ponta Delgada, São Miguel (Azores Islands), Portugal

Edited by Sergey Y. Yurish

IFSA Publishing

Sergey Y. Yurish, *Editor*
Sensors and Electronic Instrumentation Advances
SEIA' 2025 Conference Proceedings

E-mail (for orders and customer service enquires): ifsa.books@sensorsportal.com

Visit our Home Page on http://www.sensorsportal.com

ISBN: 978-84-09-76916-2
BN-20250917-XX
BIC: TJFC

Investigating the Ambiguity of SAM when Applied to Depth and RGB Images in Bin Picking Applications

M. Franaszek, P. Piliptchak, P. Rachakonda and K. S. Saidi
National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
E-mail: marek@nist.gov

Summary: The Segment Anything Model (SAM) can segment 2D images of novel objects not seen during training, making it particularly useful for robotic bin picking applications that involve diverse manufacturing parts. However, SAM's flexibility comes with inherent ambiguity: when prompted by a single point, it generates three binary masks with varying confidence levels. Our findings indicate that for complex parts that can be visually decomposed into simpler objects, the medium-confidence masks yield more accurate segmentations. In contrast, simpler, non-decomposable parts are more accurately segmented using the highest-confidence masks. This suggests that different mask lists can be sent to the robot's path planning module to accelerate the bin picking process.

Keywords: Foundation models, Segment anything model SAM, Robotic bin picking.

1. Introduction

To enable unsupervised, automated bin picking, a robot arm must be integrated with a vision system. The vision system must segment the acquired data to determine the accurate Six Degrees of Freedom (6DoF) pose of individual parts, allowing the robot's path planning module to select the best candidate for a successful pick [1].

Segmenting images acquired with RGB-D (red, green, blue, and depth) cameras in industrial settings is particularly challenging due to the presence of occluded and intermingled parts. The need to reprogram the vision system has been identified as a significant barrier to the widespread adoption of automated bin picking in small to medium-sized manufacturers (SMMs), which operate in low-volume, high-variability environments [2]. Early Convolutional Neural Network (CNN)-based segmentation techniques often required retraining the underlying models [3] and were primarily evaluated on the Common Objects in Context COCO dataset [4].

The development of foundation models in computer vision has been a significant breakthrough. The Segment Anything Model (SAM) excels at processing unknown objects without requiring retraining [5], enabling new approaches to instance segmentation of unseen objects [6, 7] and zero-shot 6DoF pose estimation of novel objects in cluttered scenes [8]. Subsequent research has focused on accelerating SAM's implementation [9, 10] and extending its capabilities to track moving objects in real-time [11]. Several public challenges have been launched to address perception in bin-picking applications [12, 13] and a novel approach to obtaining the 6DoF pose of industrial objects has been proposed [14].

However, SAM's zero-shot ability is tied to its inherent ambiguity, where a single point prompt yields three binary masks with different scores, as illustrated in Fig. 1.

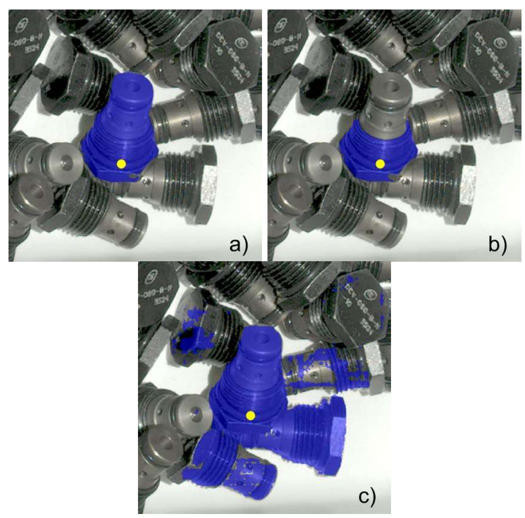

Fig. 1. Example of SAM ambiguity. All three masks (blue) initiated from the same seed (yellow dot), correspond to the scores a) 0.91; b) 0.86; and c) 0.58, respectively.

We investigated the performance of SAM applied to images of two different types of parts: type I shown in Fig. 2 and type II shown in Fig. 3. The first type contains parts with complex shapes that can be visually decomposed into simpler components. Type II groups parts with simpler shapes that cannot be naturally decomposed into sub-components. Our findings indicate that for bins containing type I parts, accurate masks are more frequently associated with the medium score. In contrast, simpler parts in the type II category are more often accurately segmented using the highest score. This insight suggests that the list of segmented masks and associated 6DoF poses sent to the robot's controller can be tailored to the specific parts in the bin, potentially improving the bin picking system's throughput.

2. Experiment

We employed a structured light camera to capture RGB images and corresponding organized 3D point clouds. The experiment involved filling a bin with multiple instances of a single part type, randomly arranged, and acquiring ten different random configurations. This process was repeated for four distinct parts, whose Computer Aided Design (CAD) models are shown in Fig. 2 and Fig. 3.

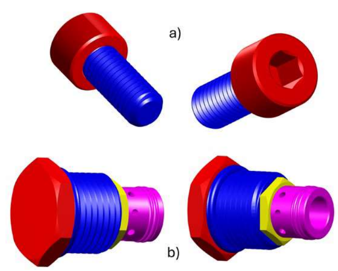

Fig. 2. Examples of two complex parts from the type I category: a) M8 hex screw; b) screw-in hydraulic valve. Different colors are used to highlight sub-components into which each part can be visually decomposed.

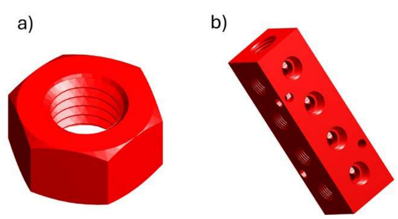

Fig. 3. Examples of two parts from the type II category: a) M12 nut; b) flow manifold.

3. Data Postprocessing

Each acquired depth or RGB image was segmented multiple times using SAM, initiated from numerous points (seeds) scattered across the image area. Initially, seeds were distributed on a regular grid, but some landed on pixels without depth data and were discarded. For each accepted seed, the two binary masks output by SAM, corresponding to the largest and medium scores, were stored. Masks associated with the third, smallest score were always grossly over-segmented and were discarded from further analysis.

For each seed, the resulting four binary masks output by SAM were visually evaluated and categorized into three classes C: 1) $C = 1$ when at least one mask (either in depth or RGB image) associated with the largest score was correct, and neither medium score mask was correct; 2) $C = 2$ when at least one mask associated with the medium score was correct, and neither largest score mask was correct; 3) $C = 3$ when at least one mask associated with the largest score and at least one mask associated with the medium score was correct, as illustrated in Fig. 4 – Fig. 7.

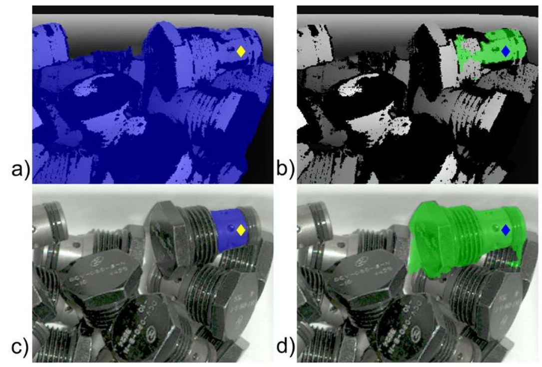

Fig. 4. Masks output by SAM applied to images of valves and classified in the class $\boldsymbol{C = 2}$, all masks prompted from the same seed (diamond mark), upper row: depth images, bottom row: RGB images, left column: largest score masks, right column: medium score masks.

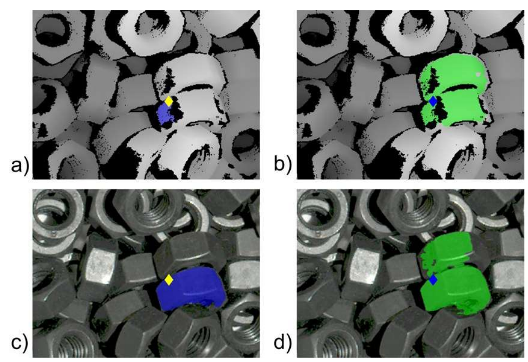

Fig. 5. Masks output by SAM applied to images of nuts and classified in class $\boldsymbol{C = 1}$, all masks prompted from the same seed (diamond mark), subplots a) – d) as in Fig. 4.

For each part used in the experiments, the fractions $f(C)$ were calculated as

$$f(C) \; = \; n(C)/\textstyle\sum_{C\,=\,1}^{3} n(C), \tag{1}$$

where $n(C)$ represents the total number of cases classified into the class C across all ten datasets, based on four SAM segmentation masks prompted from the same seed.

For each i-th seed and associated two SAM masks in the depth image, a difference Δ_i between the highest and medium scores was calculated as

$$\Delta_i = \; s_{Hi}(i) - \; s_{Mid}(i) \tag{2}$$

In addition, intersection over union fraction $\mathrm{IoU_i}$ was calculated as

$$\mathrm{IoU}_i = a\left(\boldsymbol{M}_{Hi,i} \cap \boldsymbol{M}_{Mid,i}\right)/a\left(\boldsymbol{M}_{Hi,i} \cup \boldsymbol{M}_{Mid,i}\right), \quad (3)$$

where $a(\dots)$ are the areas of intersection and union of two masks $\boldsymbol{M}_{Hi,i}$ and $\boldsymbol{M}_{Mid,i}$, prompted from the same i-th seed and associated with the highest and medium score. The same calculations were also repeated for the RGB images.

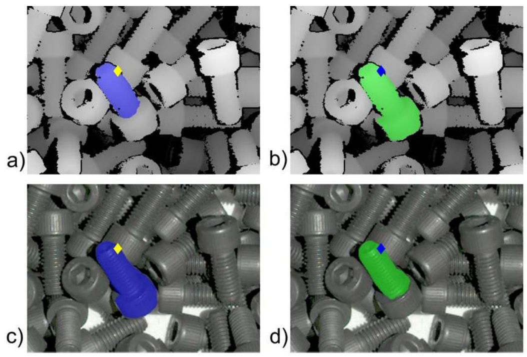

Fig. 6. Masks output by SAM applied to images of screws and classified in class $\boldsymbol{C = 3}$, all masks prompted from the same seed (diamond mark), subplots a) – d) as in Fig. 4.

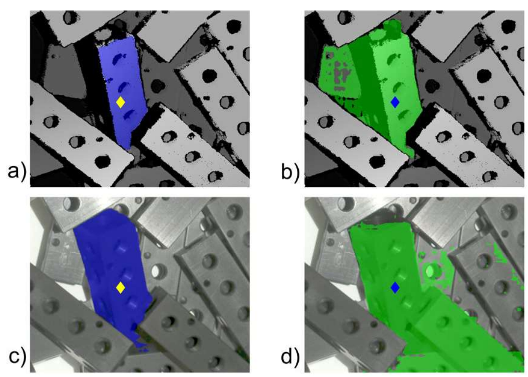

Fig. 7. Masks output by SAM applied to images of flow manifolds and classified in class $\boldsymbol{C = 1}$, all masks prompted from the same seed (diamond mark), subplots a) – d) as in Fig. 4.

4. Results

Fig. 8 illustrates the fractions $f(C)$ defined in (1), summarizing the SAM outcome classification for the four experimental parts.

Figs. 9 and 10 display plots of IoU versus Δ for depth and RGB images of screws, Fig. 11 and Fig. 12 show the corresponding results for nuts.

Fig. 13 presents the outputs of SAM applied to one depth image of screws and one RGB image of nuts. The top row displays complete masks associated with the highest score, but the corresponding masks generated with a medium score were incorrect for the same seed points. In contrast, the middle row shows complete masks associated with the medium score, while the strongest masks generated by the same seeds were incorrect. The bottom row illustrates cases where both the strongest and medium masks were accurate. Notably, multiple seeds often resulted in identical masks.

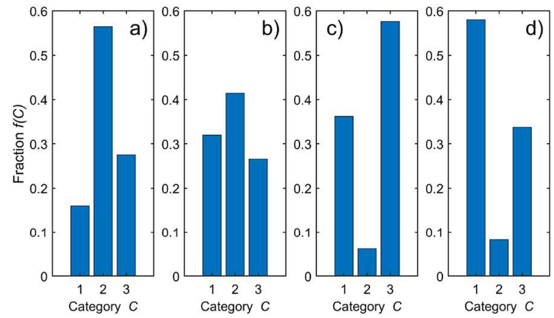

Fig. 8. The fraction $\boldsymbol{f(C)}$ which summarizes the classification of SAM results for: a) screws; b) screw-in valves; c) nuts; d) flow manifolds.

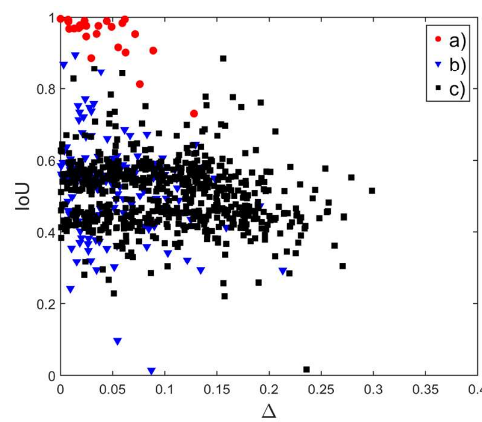

Fig. 9. IoU vs. Δ calculated for SAM outcomes obtained for depth images of screws: a) cases when both highest and medium score masks are correct; b) highest score mask correct and medium mask incorrect; c) highest score mask incorrect and medium mask correct.

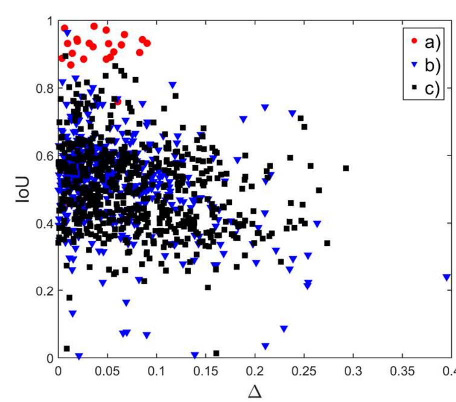

Fig. 10. IoU vs. Δ calculated for SAM outcomes obtained for RGB images of screws, labels a) – c) as in Fig. 9.

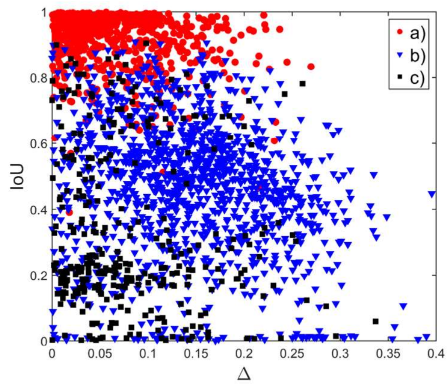

Fig. 11. IoU vs. **Δ** calculated for SAM outcomes obtained for depth images of nuts, labels a) – c) as in Fig. 9.

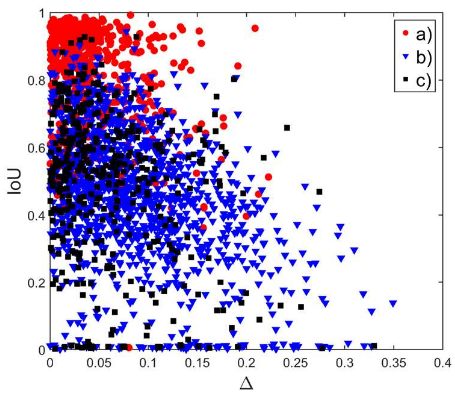

Fig. 12. IoU vs. Δ calculated for SAM outcomes obtained for RGB images of nuts, labels a) – c) as in Fig. 9.

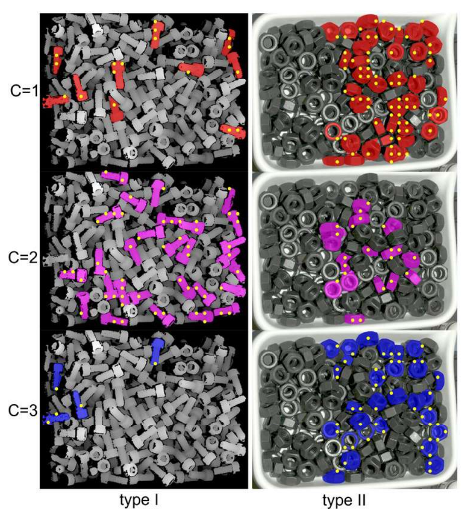

Fig. 13. Example of different SAM outputs for two types of parts and three categories *C*. Colored masks are prompted by seeds marked with dots.

Fig. 14 illustrates an example of SAM output applied to the same depth image as in Fig. 13, where the strongest score is associated with incorrect, under-segmented masks. Although some masks were generated by SAM in response to multiple seeds, only one seed per mask is displayed for clarity.

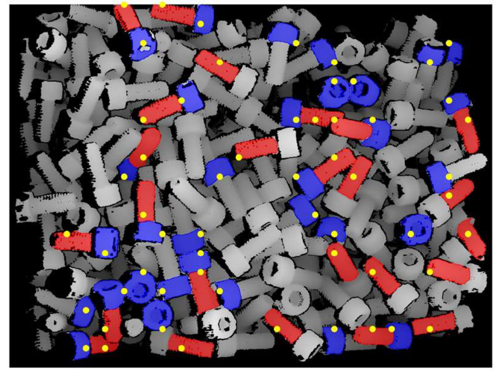

Fig. 14. Example output of SAM applied to a depth image. Incomplete masks associated with the strongest score are displayed. Two colors are used to mark two types of under-segmented masks; starting seeds are plotted with colored dots.

Fig. 15 presents examples of complete masks generated by SAM in response to different seeds, with nearly identical results (all pairwise Intersection over Union (IoU) values exceeding 0.99). Notably, the seeds were placed on distinct sub-components of the screw, such as the stem and head (as illustrated in Fig. 2a). Despite this, all seeds yielded complete masks, and no clear correlation was observed between the score associated with a particular mask and the seed's location.

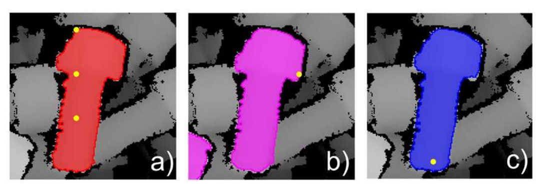

Fig. 15. Example of overlapping, nearly identical masks generated from SAM prompted by different seeds: a) three masks associated with the strongest scores; b) mask associated with the medium score; c) overlapping of the strongest and medium mask.

5. Discussion

The histograms in Fig. 8 clearly demonstrate differences in SAM performance when images of different parts are processed. For type I parts (screws and valves), more accurate masks were associated with medium scores than with the highest scores. In contrast, type II parts (nuts and flow manifolds) exhibited the opposite trend. This disparity is also apparent in Fig. 13, where the subplots within each row show notable differences. These findings suggest that a customized selection of masks from SAM-segmented images is necessary for further processing by the robot's controller. For type I parts, such as those in Fig. 2, choosing masks with medium scores appears to be more effective. Even when the prevalence of correctly segmented masks is moderate (as shown in Fig. 8b), this strategy remains beneficial, as the combined categories $C = 2$ and $C = 3$ will yield more accurate masks.

A closer examination of under-segmented masks associated with the highest score for type I parts reveals that they often coincide with simpler sub-components, as seen in Fig. 14 and Fig. 2a. This explains why more accurate masks for type I parts are associated with medium scores rather than the highest scores, as evident in Fig. 8 a, b and the left column of Fig. 13. The complexity of type I parts appears to exacerbate the ambiguity of SAM more than the simpler shapes of type II parts, as illustrated in Fig. 8 c, d and the right column of Fig. 13.

The Intersection over Union (IoU) metric provides a useful measure to assess the similarity between the strongest and medium masks. For screws, categorized as type I, the IoU values cluster into three distinct regions, as evident in Fig. 9 and Fig. 10. Cases labeled as a), where both masks are correct, exhibit high IoU values (IoU > 0.8). In contrast, the majority of cases labeled as c), with incorrect largest masks, are under-segmented and coincide with two simpler sub-components, displayed in Fig. 14. Consequently, the IoU values for these cases cluster around 0.45 and 0.58. In contrast, nuts (type II) exhibit IoU values scattered across the entire [0, 1] interval, with no discernible clustering, as Fig. 11 and Fig. 12 show. This is expected, as type II parts lack obvious sub-components, resulting in irregular variations in the area of incorrect, under-segmented masks. Notably, some cases labeled as b) or c), with incorrect strongest or medium masks, display IoU values close to zero, indicating minimal overlap between the two masks.

The difference Δ defined in (2) offers additional insight into the performance of SAM on type I and II parts. A small Δ indicates that SAM assigns comparable confidence to both output masks, whereas a larger Δ suggests a significant disparity between the confidence levels of the largest and medium masks. Interestingly, for nuts (type II), many cases with both masks correct and high IoU values exhibit relatively large Δ, as seen in the points labeled as a) in Fig. 11 and Fig. 12. In contrast, screws (type I) do not display a similar shift towards larger Δ for cases with correct masks, as evident in the points labeled as a) in Fig. 9 and Fig. 10.

The classification of correctly segmented masks into three C categories is influenced by the seed location, as demonstrated in Fig. 15. Despite all seeds being within the mask perimeter, different seed placements result in varying category assignments. It is unclear why seeds located on subcomponents of type I parts (like the stem or head of the screw) sometimes yield fully segmented correct masks, as shown in Fig. 15, or sometimes only subcomponents, as shown in Fig. 14. This unexplained variation requires further study and suggests that a more sophisticated seed selection strategy for SAM, rather than a regular grid of points, may be necessary for bin-picking applications.

In this study, the masks generated by SAM were evaluated visually, a subjective process susceptible to human error, similar to manual segmentation used to establish ground truth and calculate IoU. Moreover, the application of SAM to RGB or depth images is only the initial step towards the ultimate goal of accurately segmenting 3D points, fitting the corresponding CAD model, and obtaining a precise 6DoF pose of individual parts. However, even an accurate 2D mask can lead to inaccurately segmented 3D points, as illustrated in Fig. 16, potentially resulting in erroneous 6DoF pose estimation and increased risk of failed grasp attempts. Therefore, further refinement of SAM output is necessary to achieve reliable results.

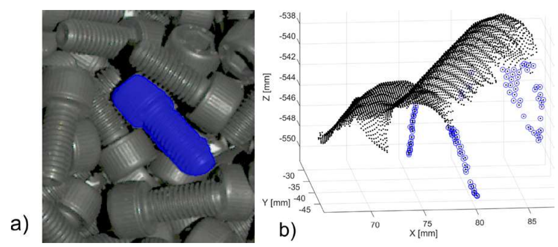

Fig. 16. Example of inconsistency between 2D and 3D segmentations: a) accurate 2D mask output by SAM applied to the RGB image; b) determined by 2D mask subset of the organized 3D point cloud. Colored circles in 3D graph mark outliers.

A fundamental question remains to be investigated: what characteristics of a 3D object determine whether it is classified as type I or type II? The parts shown in Fig. 2 and Fig. 3 were categorized based on a subjective criterion, namely, being visually decomposable into simpler sub-components, without a clear definition of simplicity or complexity. Establishing a precise definition of shape complexity would be beneficial, as it would enable the determination of part types in a bin and facilitate the implementation of an appropriate strategy for processing SAM outputs for a specific bin-picking task.

6. Conclusions

It is intuitive to assume that segmentation masks with the highest scores are generally more accurate than those with lower scores. However, due to SAM's inherent ambiguity, the highest-scoring masks may not always be the most accurate. Our results show that for complex parts like screws and valves, which can be visually decomposed into simpler subcomponents, the medium-scoring masks were often more accurate.

Inaccurate, under-segmented masks were usually associated with the largest scores, and such masks overlapped with subcomponents. In contrast, for simpler parts like nuts and manifolds, the highest-scoring masks were more frequently correct.

Therefore, to optimize bin picking applications, the choice of segmentation masks should be tailored to the specific parts being handled, using either the strongest or medium-scoring masks depending on the part's geometry.

References

[1]. H. Alzarok, S. Fletcher, A. P. Longstaff, Survey of the current practices and challenges for vision systems in industrial robotic grasping and assembly applications, *Advances in Industrial Engineering*, Vol. 9, Issue 1, 2020, pp. 19-30.

[2]. J. Horst, J. Marvel, E. Messina, Best practices for the integration of collaborative robots into workcells within small and medium-sized manufacturing operations, NIST Advanced Manufacturing Series 100-41, *National Institute of Standards and Technology*, 2021.

[3]. F. Sultana, A. Sufian, P. Dutta, Evolution of image segmentation using deep convolutional neural network: a survey, *Knowledge-Based Systems*, Vol. 201-202, 2020, 106062.

[4]. K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, Issue 2, 2020, pp. 386-397.

[5]. A. Kirillov, E. Mintun, N. Ravi, H. Mao, et al., Segment anything, *arXiv preprint*, 2023, arXiv:2304.02643.

[6]. R. Cao, C. Song, B. Yang, J. Wang, et al., Adapting segment anything model for unseen object instance segmentation, *arXiv preprint*, 2024, arXiv:2409.15481.

[7]. R. Schon, J. Lorenz, K. Ludwig, R. Lienhart, Adapting the segment anything model during usage in novel situations, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, 2024, pp. 21010-21019.

[8]. J. Lin, L. Liu, D. Lu, K. Jia, SAM-6D: segment anything model meets zero-shot 6D object pose estimation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'24)*, 2024, pp. 3379-3388.

[9]. X. Zhao, et al., Fast segment anything, *arXiv preprint*, 2023, arXiv:2306.12156.

[10]. C. Zhang, et al., Faster segment anything: towards lightweight SAM for mobile applications, *arXiv preprint*, 2023, arXiv:2306.14289.

[11]. L. Ke, et al., SAM 2: segment anything in images and videos, *arXiv preprint*, 2024, arXiv:2408.00714.

[12]. OpenCV, Perception Challenge for Bin-picking, https://bpc.opencv.org

[13]. V. N. Nguyen, et al., BOP challenge 2024 on model-based and model-free 6D object pose estimation, *arXiv preprint*, 2025, arXiv:2504.02812.

[14]. C. Zhuang, S. Liu, H. Ding, Instance segmentation based 6D pose estimation of industrial objects using point clouds for robotic bin-picking, *Robotics and Computer-Integrated Manufacturing*, Vol. 82, 2023, 102528.