GRANADA SEIA'26

Sensors and Electronic Instrumentation Advances:

Proceedings of the XII International Conference on Sensors and Electronic Instrumentation Advances

30 September - 2 October 2026

Granada, Spain

Edited by Sergey Y. Yurish

IFSA Publishing

Sergey Y. Yurish, *Editor*
Sensors and Electronic Instrumentation Advances
SEIA' 2026 Conference Proceedings

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

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

ISBN: 978-84-09-90031-2
BN-20260922-XX
BIC: TJFC

(008)

Dependence of Segment Anything Model (SAM) Performance on Part Geometry in Bin Picking Applications

M. Franaszek, P. Piliptchak 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 novel, previously unseen objects from 2D images. This capability is especially important in automated bin picking applications, where low-volume and high-variability scenarios are common. When prompted by a point, SAM outputs three masks with varying confidence scores. Prior research indicates that SAM's performance depends on the type of parts in a bin: for parts that are visually decomposable into subcomponents, correctly segmented masks are more often associated with a medium confidence score. Conversely, for parts that are not readily visually decomposable into subcomponents, correct masks are typically associated with the highest confidence score. We investigated the transition from visually decomposable to non-decomposable part categories in simulation by using computer-aided design (CAD) models with gradually altered geometric features. Differences in the dimensions of these geometric features cause corresponding differences in grayscale images. We found that this transition occurs within a relatively narrow range of feature modifications and image intensities. This demonstrates the strong sensitivity of SAM to the specific types of parts within a bin. Consequently, prior knowledge of a part's complexity can be leveraged to better tune the processing of SAM's output.

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

1. Introduction

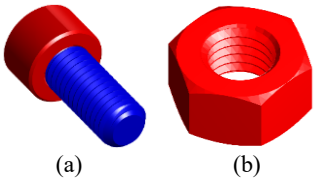
The successful deployment of bin picking systems to automate manufacturing tasks, such as assembly or kitting, depends strongly on vision system performance. The vision system must segment the acquired depth images to determine the six degree-of-freedom (6DoF) pose of individual parts, enabling the integrated robot to execute tasks successfully.

Accurate instance segmentation of unorganized piles of manufacturing parts remains especially challenging. While early segmentation algorithms required model retraining for different parts [1], the emergence of foundation models such as the Segment Anything Model (SAM) has made retraining for novel parts almost obsolete [2]. When prompted by a point in a query image or a bounding box, SAM outputs three binary segmentation masks with distinct confidence scores.

Continued research on prompts has led to the adoption of Large Language Models (LLMs), which allow the use of text-based prompts to obtain segmentation masks for specified objects [3-5]. These new Promptable Concept Segmentation (PCS) techniques have yielded promising results for semantic segmentation when tested on Common Objects in Context (COCO)-style databases [6]. However, extensive testing on images of industrial parts randomly filling a bin still needs to be conducted.

Another kind of prompt in the form of a computer-aided design (CAD) model was recently proposed [7]. It is especially useful for bin picking applications where CAD models of the parts filling the bins are usually available. A modified version of SAM (SAM3) [5] was trained on synthetic images created with the NVIDIA Isaac Sim simulator [8], which was used to build multi-object scenes utilizing CAD models from the ABC dataset of industrial parts [9]. It was tested on two standard industrial benchmark datasets, T-LESS and ITODD, which do not contain many heavy clutter scenes typical of bin picking applications [10, 11].

Currently, using point prompts in depth images remains the easiest way to obtain instance segmentation masks. To maximize the efficiency of robotic path planning, the vision system should provide as many correct masks as possible, prompted from points scattered across the entire image area. Previous work demonstrated that SAM's performance depends heavily on the type of parts in the bin [12]. For parts that are visually decomposable into subcomponents (as in Fig. 1(a)), accurate masks from SAM's multi-mask output were most often associated with the medium (second-highest) confidence score. For parts lacking obvious subcomponents (as in Fig. 1(b)), accurate masks typically correlated with the highest confidence score.



(a) (b)

Fig. 1. Examples of two types of parts: a) visually decomposable into subcomponents (drawn with different colors); b) not readily decomposable into subcomponents.

Previously, the distinction between the two types of parts was based on subjective, unquantifiable

evaluation. In this study, a more objective metric is formulated. We systematically investigate how SAM's performance shifts during the transition from one type of part to another by gradually altering the geometric features of the part within a simulated bin. Our results suggest that this transition occurs over a relatively narrow range of geometric changes. Those changes result in corresponding changes in the 8-bit grayscale intensity of depth images. We found that two compact, adjacent regions on a depth image whose intensities differ by less than 15 will most likely be segmented as one region that stretches over both adjacent areas. Thus, a part represented in a depth image as a patch where no compact subregions can be identified with intensity differences greater than $\Delta I = 15$, should be classified as non-decomposable part.

Patches consisting of subregions with intensity differences much larger ($\Delta I \gg 15$) correspond to decomposable parts for which binary masks associated with the medium score will be more accurate. Our results reveal a specific range of differences in geometric features, along with a corresponding range of intensities in a grayscale image, where no single strategy (strongest vs. medium score mask) is preferred. In such mixed regions, to maximize the number of correctly segmented parts in a bin, masks associated with both the strongest and medium confidence scores may need to be analyzed.

Thus, prior knowledge of a part's dimensions and shape can be utilized to optimize the processing of SAM's output, thereby improving overall bin picking system performance.

2. Simulation

We generated two types of grayscale images: one showing a bin filled with screws, and the other showing a mosaic of tiles with different intensities. The first type yielded a realistic-looking pile of screws, but to check SAM's performance, many images had to be visually evaluated. The second type shows artificial scenes composed of geometric primitives for which a metric gauging SAM's performance can be calculated by running autonomous image processing algorithms. An 8-bit grayscale was used for all images, and binary masks output by SAM associated with the largest confidence score were analyzed.

2.1. Generation of Screws Images

We use a physics-based simulator to simulate dropping parts into a bin and generate depth images derived from the known distances between the simulated parts and bin bottom. Using the CAD models of an M12 and M8 hex screw, we generated a series of $N = 20$ models for each with a gradually decreasing head diameter, $D_h(n)$, while keeping the stem diameter D_s, stem length L_s, and head length L_h fixed (see Fig. 2). The deformation rate, $\alpha(n)$, is defined as a ratio:

$$\alpha(n) = \frac{D_h(n) - D_s}{D_h - D_s}, \quad (1)$$

where $n = 0, \ldots, N$, and $D_h(0) = D_h$ corresponds to the head diameter in the original CAD model, while $D_h(N) = D_s$, as illustrated in Fig. 2, and $D_h(n)$ are equally distributed between D_s and D_h. In Table 1, the relevant dimensions of the unmodified CAD models for both screws are provided, where $L_{tot} = L_h + L_s$ is the total length of the screw.

Table 1. Screw dimensions in [mm].

	L_{tot}	L_s	L_h	D_s	D_h
M8	26	18	8	8	13
M12	32	20	12	12	18

For each n-th CAD model, we dropped a fixed number of parts into a bin and saved the resulting depth image in orthonormal projection. The initial pose of each part was randomized, and this step was repeated for $j = 1, \ldots, 10$ to obtain ten distinct instances of unorganized piles for each part.

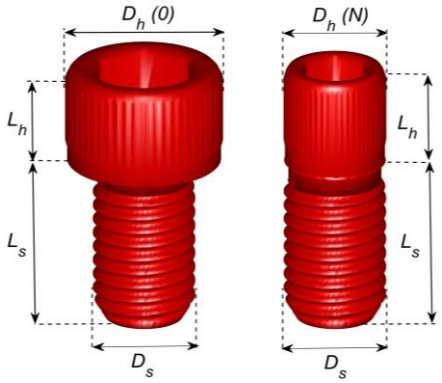


Fig. 2. Two limiting cases of the CAD models: the original with head diameter $\boldsymbol{D_h(0)}$ and the most reduced head diameter where $\boldsymbol{D_h(N) = D_s}$.

For each (n, j) saved depth image, we selected prompt points from a regular grid. We used the same bin for dropping M12 and M8 screws; therefore, the grid size and the number of dropped parts were set differently to ensure a similar complexity in the resulting unorganized pile. For M8 screws, the grid size was 12×14 and the number of dropped parts was 54; for M12 screws, the corresponding parameters were 11×12 and 48. Points falling on the background (the bin bottom) were ignored, as shown in Fig. 3. The accepted $M_j(n)$ points were used as prompts for SAM segmentation.

The resulting $M_j(n)$ binary masks were visually evaluated, and only those containing both the stem and the head of the screw were counted. This procedure enabled tracking the transition between the two categories of parts (visually decomposable into

subcomponents and with no obvious subcomponents) as the deformation rate $\alpha(n)$ was swept across its range. Accepted masks fell into two categories: the first, where the entire screw was unobstructed and the mask covered the full stem and head facing the camera; and the second, where only a portion of the screw was visible due to partial occlusion by neighboring parts. Example of masks from the first category are shown in Fig. 4 while masks belonging to the second category are shown in Fig. 5. While some masks in the second category may not always be good candidates for picking, they were included in the total count of correct masks since they represent a challenging test for segmentation. Not all masks containing both stem and head were accepted: in some cases, the segmentation successfully crossed the boundary between stem and head but ultimately failed to produce correct masks (e.g., due to under-segmentation or leaking). Examples of rejected cases are shown in Fig. 6. The masks in Fig. 6a) and d) are under-segmented while masks in Fig. 6b) and c) leaked.

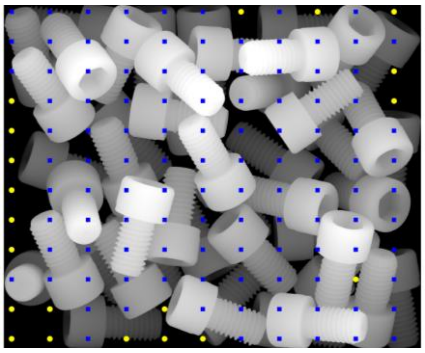


Fig. 3. Depth image displaying the grid of points: yellow circles denote ignored background points, while $\boldsymbol{M_j(n)}$ blue squares indicate active prompt points for SAM.

In this way, the number of accepted masks, $K_j(n)$, was determined. For each n-th deformed CAD model and each j-th instance of a pile, a fraction $\rho_j(n) = K_j(n)/M_j(n)$ was calculated from which the corresponding means $\bar{\rho}(n)$ and standard deviations $\delta(n)$ were calculated for M8 and M12 screws.

2.2. Generation of Images with a Mosaic of Tiles

We start with an image completely filled with the background (intensity 0) and all perimeter pixels set to maximum (intensity 255). Such a scene mimics the image of an empty bin. The entire background area is then divided into 11×9 squares, and the intensity in each square is set to a random number from $\langle 0{,}255 \rangle$, Then, each square tile is further divided into two rectangles of unequal areas, and the intensity inside the smaller rectangle is increased by $\Delta I(n) = \Delta I_{max}\alpha(n)$ where $0 \leq \alpha \leq 1$ plays the same role as deformation factor in (1), and ΔI_{max} corresponds to $D_h - D_s$ for screws. Finally, to simulate noise, a random integer from {-1,0,1} is added to the intensities of all pixels in all rectangles while keeping all final values clipped to the $\langle 0{,}255 \rangle$ interval Examples of such images are shown in Fig. 7.

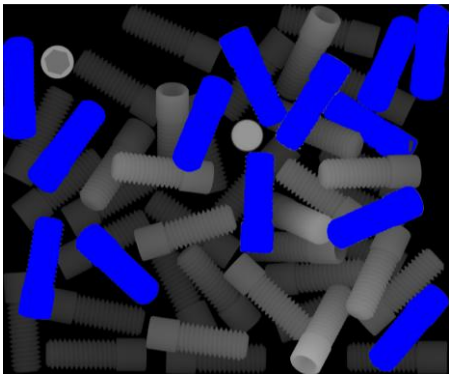


Fig. 4. Example of correctly segmented masks from the first category: unobstructed masks of M8 screws, deformation factor $\boldsymbol{\alpha = 0.098}$.

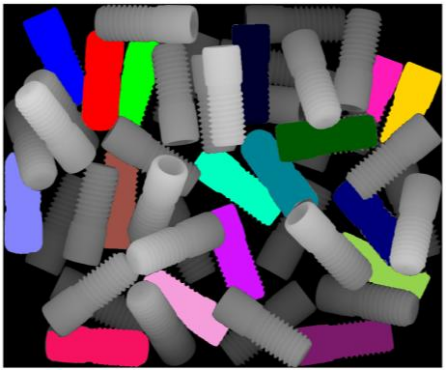


Fig. 5. Example of correctly segmented masks from the second category: obstructed masks of M12 screws, deformation factor $\boldsymbol{\alpha = 0.065}$.

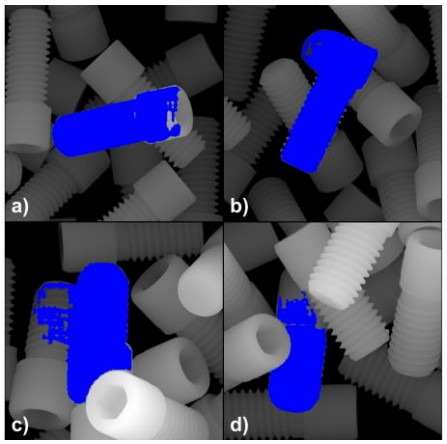


Fig. 6. Examples of discarded masks which contain both stem and head: a) under segmented mask from the first category; b) leaked mask from the first category; c) leaked mask from the second category; d) under segmented mask from the second category.

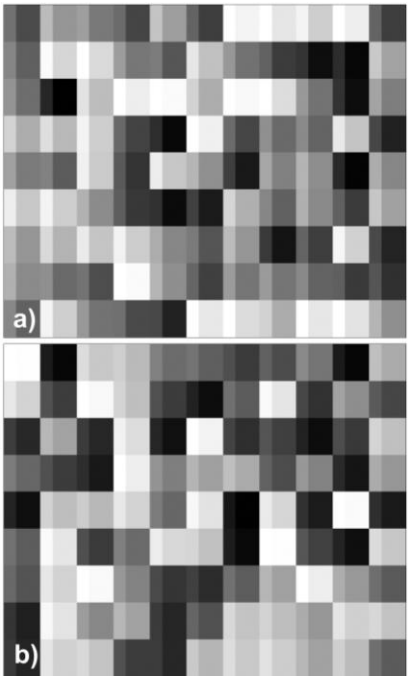


Fig. 7. Example of images with mosaic of tiles created for two different instances $\boldsymbol{j}$ and $\boldsymbol{\alpha = 1}$, $\boldsymbol{\Delta I_{max} = 38}$.

Each square tile may be viewed as a representation of a screw, with the smaller rectangle corresponding to the screw's head and the larger rectangle acting as its stem. At the center of each rectangle, we set a prompt point and perform segmentation using SAM. Since the boundaries of each square tile are known, the binary masks output by SAM can be directly compared with the corresponding tile. For each mask and square tile, the Intersection over Union (IoU) can be calculated. A mask output by SAM is accepted as correct if IoU > 0.95. For each given $\Delta I(n)$, we repeat the entire procedure for $j = 1, \ldots, 10$, assigning different randomly selected intensities to each square tile for each j, as shown in Fig. 7. The mean fraction of accepted masks $\bar{\rho}(n)$ and standard deviations $\delta(n)$ were calculated similarly to calculations for screws.

3. Results

Fig. 8 presents examples of two depth images of M12 screws corresponding to deformation factors $\alpha(0)$ and $\alpha(N)$, overlaid with their complete SAM masks. Examples of types of masks overlaid on the depth image of screws are shown in Fig. 9. Four types of masks are shown: masks labeled as a) are under-segmented and they overlap with only the heads of the screws; masks labeled as b) are also under-segmented and they overlap with the stems of the screws; the mask labeled as c) shows leaked segmentation; and the masks labeled as d) are correct and they overlap with the entire visible portions of the screws, including the heads and stems. Locations of prompt points resulting in the displayed masks are marked by dots. In Fig. 10, a similar plot is shown for the image of mosaic of tiles, labels a) – d) and color coding have the same meaning for square tiles and their subcomponents (rectangles) as for screws, heads and stems in Fig. 9.

Finally, the dependence of the mean ratio $\bar{\rho}$ of correctly segmented masks vs. deformation factor α for two sizes of screws is shown in Fig. 11. Similar plots for images of mosaics of tiles, generated for two different values of ΔI_{max}, are shown in Fig. 12. The mid-point α_{mid} for deformation factor is defined as such for which $\bar{\rho}(\alpha_{mid}) = (\bar{\rho}_{max} + \bar{\rho}_{min})/2$.

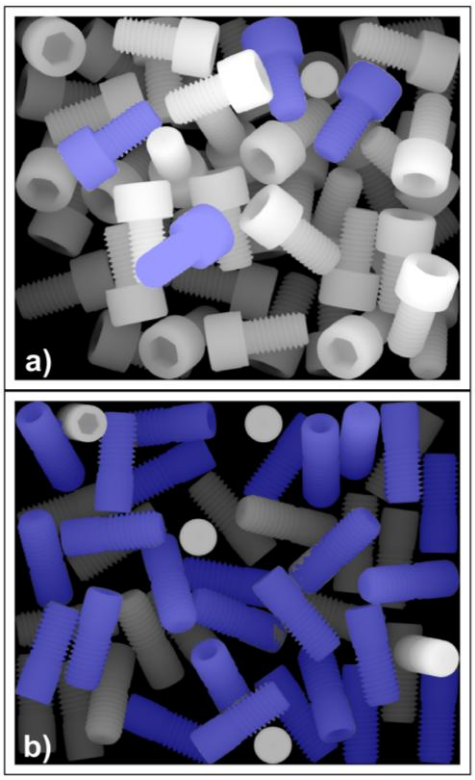


Fig. 8. Depth images of a pile of M12 screws for two limiting values of the deformation factor: a) $\boldsymbol{\alpha = 1}$; b) $\boldsymbol{\alpha \approx 0}$. Complete SAM output masks are displayed in color.

In Table 2, characteristics of screws and tiles at deformation factor equal to mid-point α_{mid} are displayed. $\Delta D_{max} = D_h - D_s$ is the largest difference between geometric features of the subcomponents of the decomposable part (diameters of head and stem for screws, see Fig. 2) and $\Delta D = D_h(\alpha) - D_s$ is calculated for a given deformation factor α. The corresponding difference in the image intensity (assuming eight-bit gray scale and the highest point in a scene determined by a bin height $h = 48$ mm) is in the rightmost column. The smaller ΔD_{max} is for M8 screws and the larger is for M12 screws. For mosaic of tiles, only the differences in image intensities ΔI are available since the intensities in the square tiles were already randomly selected from the $\langle 0{,}255 \rangle$ interval. Under these conditions, two ΔI_{max} values match closely the two ΔD_{max} values.

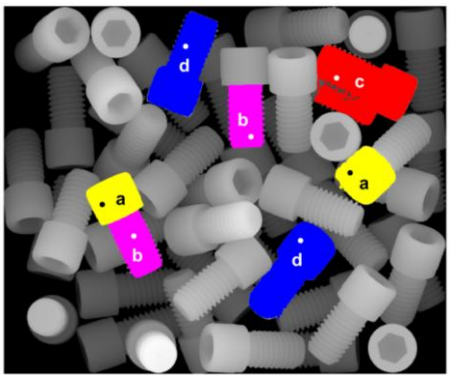


Fig. 9. Examples of four types of masks overlaid on depth image of screws: a) incomplete, only head segmented; b) incomplete, only stem segmented; c) leaked segmentation; d) complete, correct masks of entire screw. Dots represent locations of prompt points for SAM.

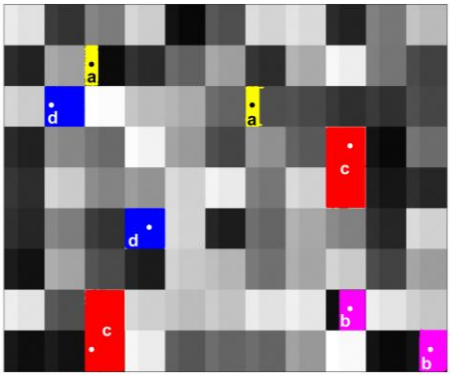


Fig. 10. Examples of four types of masks overlaid on image of mosaic with tiles, labels a) – d) and color coding have the same meaning as in Fig. 9. Dots represent locations of prompt points for SAM.

Table 2. Screws and tiles characteristics at mid-point.

	$\boldsymbol{\alpha_{mid}}$	$\boldsymbol{\Delta D(\alpha_{mid})}$	$\boldsymbol{\Delta I(\alpha_{mid})}$
ΔD_{max} = 5 mm	0.32	1.6 mm	10
ΔD_{max} = 6 mm	0.21	1.3 mm	8
ΔI_{max} = 32	0.42	--	13
ΔI_{max} = 38	0.39	--	15

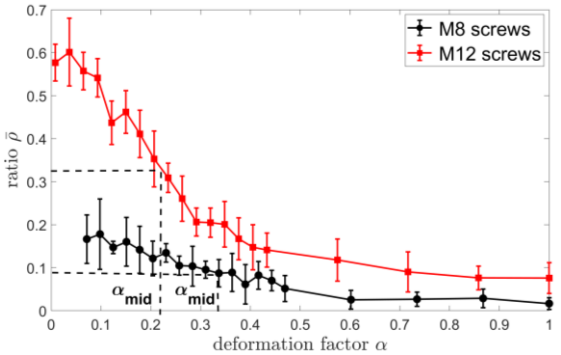


Fig. 11. Dependence of $\bar{\rho}$ on $\boldsymbol{\alpha}$ for images of screws, error bars given by $\boldsymbol{\delta}$. Dotted lines show location of mid-point $\boldsymbol{\alpha_{mid}}$.

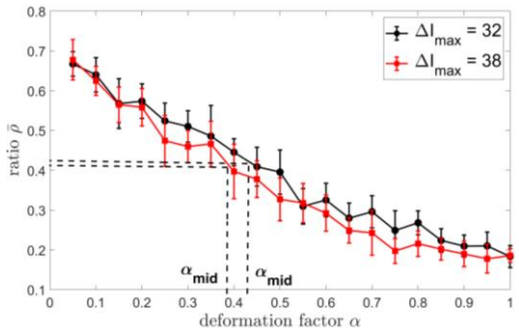


Fig. 12. Dependence of $\bar{\rho}$ on $\boldsymbol{\alpha}$ for images of mosaic of tiles, error bars given by $\boldsymbol{\delta}$. Dotted lines show location of mid-point $\boldsymbol{\alpha_{mid}}$.

4. Discussion

The distinction between the two types of parts – visually decomposable into subcomponents versus not readily decomposable, as shown in Fig. 1 – was originally based on a subjective, unquantifiable impression. By gradually altering the geometric features of the parts, we were able to quantitatively study the transition between both part types in a more systematic manner. The dependence of $\bar{\rho}$ on α plotted in Fig. 11 and Fig. 12 indicates that this transition is described by a smooth function, with no abrupt shift from maximum to minimum values of $\bar{\rho}$ at a single deformation factor α. Thus, within a certain range $[\alpha_{mid} - \beta, \alpha_{mid} + \beta]$, there is no single preferred selection strategy for the masks output by SAM: whether to select those associated with the highest confidence score (as shown in Fig. 8b) or those associated with the medium score when very few correct masks receive the highest score (as shown in Fig. 8a). To maximize the total number of correctly segmented masks and afford optimal robotic path planning options, processing masks associated with both the highest and medium confidence scores may be necessary, though this could lead to increased total cycle time in automated bin picking.

The width 2β of α interval inside which both strategies may be needed can be interpreted as a degree of separation between the two approaches: for $\alpha < \alpha_{mid} - \beta$, utilizing masks with the highest confidence is advantageous, whereas for $\alpha > \alpha_{mid} + \beta$, the opposite holds true, and a better strategy relies on masks with medium confidence scores. Larger values of β widen the α interval where masks from both confidence levels must be analyzed. Conversely, smaller β values expand the regions where a single strategy suffices, albeit with diminishing benefit because when α approaches α_{mid}, the contribution from the opposite strategy is increasing.

Dimensional differences among subcomponents of decomposable parts, ΔD (such as the difference between head and stem diameters in screws) induce corresponding intensity variations across subregions in depth images. Simplified grayscale tile-mosaic images, such as the one shown in Fig. 7, reproduced

SAM's performance on realistic bin-picking screw images remarkably well. The masks shown in Fig. 9 and Fig. 10 exhibit identical characteristics: for example, correctly segmented masks (type d) can be generated by a prompt point placed on either the head or the stem. The primary advantage of utilizing these synthetic images is that SAM's segmentation output can be evaluated both automatically and precisely.

The rightmost column in Table 2 lists ΔI values calculated at $\alpha = \alpha_{mid}$. All four values, derived from distinct grayscale images and scene configurations, fall within a relatively narrow interval of $\langle 8{,}15 \rangle$. The physical interpretation of ΔI is that when an image contains a compact region composed of two subregions with an intensity difference larger than ΔI, the mask with the highest confidence score will likely cover only one subregion rather than the composite entity. In such cases, a medium-confidence mask prompted from a single point is more likely to encompass the entire region. While both scenarios occur with high frequency, exceptions can occur, as is typical with point-prompted segmentation. In industrial bin picking, CAD models of the target parts are usually readily available, and recent developments – such as the CAD-prompted segmentation framework based on SAM3 – may resolve the ambiguity inherent to point prompts [7]. However, this emerging technique still warrants extensive evaluation on complex, heavily cluttered bin-picking scenes containing significant object occlusions.

5. Conclusions

Visual inspection of parts to be picked from a bin by a robotic arm integrated with a vision system may provide a first hint of whether they belong to the class of visually decomposable objects for which, paradoxically, not the largest score masks but the medium score masks provide more accurate segmentation. However, this subjective evaluation may lead to false conclusions. The knowledge of geometric dimensions of possible subcomponents, such as ΔD in Table 2, may not be sufficient to support the visual evaluation since what really matters is the resulting difference in the intensity ΔI in depth image. This difference depends not only on ΔD but also on the bit depth of the grayscale and the entire context of the scene, e.g., bin height. Other factors, like camera distance, lighting conditions, or part color and finish also affect formation of real depth image, but they were not accounted for in our simplified simulation. More testing is required, but the initial results obtained in this paper suggest that an intensity difference of $\Delta I \gtrsim 15$ may act as a barrier prohibiting the further expansion of a binary mask initiated from a point.

References

[1]. 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.

[2]. A. Kirillov, et al., Segment anything, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015-4026.

[3]. X. Lai, et al., LISA: Reasoning segmentation via large language model, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 9579-9589.

[4]. J. Li, et al., SAM3-I: Segment anything with instructions, in *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026, pp. 27232-27249.

[5]. N. Carion, et al., SAM 3: Segment anything with concepts, in *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

[6]. COCO dataset, https://cocodataset.org

[7]. Z. Tang, et al., CAD-prompted SAM3: Geometry-conditioned instance segmentation for industrial objects, *arXiv*, 2026, arXiv:2602.20551.

[8]. NVIDIA, Isaac Sim documentation, https://docs.isaacsim.omniverse.nvidia.com

[9]. S. Koch, et al., ABC: A big CAD model dataset for geometric deep learning, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9601-9611.

[10]. T. Hodan, et al., T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects, in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 880-888.

[11]. B. Drost, et al., Introducing MVTec ITODD – A dataset for 3D object recognition in industry, in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 2200-2208.

[12]. M. Franaszek, et al., Investigating the ambiguity of SAM when applied to depth and RGB images in bin picking applications, in *Proceedings of the 11th International Conference on Sensors and Electronic Instrumentation Advances (SEIA' 2025)*, ,24-26 September 2025, Ponta Delgada, São Miguel (Azores Islands), Portugal, 2025, pp. 64-69.