CircleSeg-XAI: Segmentation-Based Target Point Detection Method Using Circular Shape Label

Hiroyuki Tanabe, Mitsuji Ikeda, Yohei Minekawa, Yuichi Abe, Makoto Sato

Hitachi High-Tech Corporation

hiroyuki.tanabe.da@hitachi-hightech.com, mitsuji.ikeda.zg@hitachi-hightech.com, yohei.minekawa.dw@hitachi-hightech.com, yuichi.abe.hw@hitachi-hightech.com, makoto.sato.jz@hitachi-hightech.com

We propose a novel segmentation-based method for detecting target points in images. Utilizing fixed-shape circular labels, this method reduces annotation costs while simultaneously providing both the target point location and a confidence index within the image. When applied to identifying the position of a needle tip in focused ion beam scanning electron microscope (FIB-SEM) images, our method achieved a positional deviation of 2.16 pixels (σ=±1.24 pixels) compared with 3.12 pixels (σ=±1.49 pixels) for existing heatmap-based object detection methods, demonstrating superior positional accuracy. When comparing the effectiveness of confidence indexes for out-of-distribution (OOD) detection on heatmap based object detection methods, our method discriminated 89.5% of unlearned images, demonstrating the highest separation ratio between learned and unlearned images. Additionally, to evaluate data diversity, we assessed the effectiveness and limitations of the proposed method on scanning electron microscope (SEM) images related to semiconductor manufacturing.

Introduction

Deep convolutional neural networks (CNNs) have achieved excellent results in tasks like object detection (Ren et al. 2015) and segmentation (Long et al. 2014). However, applying these technologies to target point detection in images presents several challenges. In particular, alerting users is important when high positional accuracy cannot be guaranteed. This is because inaccurate positioning can significantly impact safety and work efficiency. Furthermore, in deep CNNs, confidence indexes are essential for inferring images containing out-of-distribution (OOD) features, as achieving high positional accuracy may not always be possible.

Previous studies have proposed methods using bounding boxes of objects with uncertainty or confidence indexes (Gasperini et al. 2021, Lee et al. 2022). However, common object detection models specialize in identifying the position and outline of objects using rectangular bounding boxes are not aimed at detecting the coordinates of a specific point with high accuracy.

Another approach uses segmentation, identifying object shapes at the pixel level for precise object coordinate detection through post-processing techniques like calculating the center of gravity. However, general segmentation has high annotation costs as pixel labeling of the target object is required.

In this study, we propose a novel segmentation-based position detection method using fixed shapes as target labels. Our method reduces annotation costs by using fixed circular shapes as labels at the center coordinate of the detection target. We confirm the effectiveness of our method by comparing its position detection performance and confidence score effectiveness against those of existing object detection models using images observed by a focused ion beam scanning electron microscope (FIB-SEM) system. Additionally, to evaluate data diversity, we assessed the effectiveness of the proposed method on scanning electron microscope (SEM) images related to semiconductor manufacturing.

Related works

Template Matching for Identifying Target Position

Template matching, widely used for identifying target positions on images, relies on finding the highest similarity between a template image and the input using such as normalized correlation (Rosenfeld, 1969). This rule-based method, classified as a white-box due to its transparency, does not require large amounts of data compared with deep learning models. However, while template matching based on the normalized correlation is robust to uniform lightning illumination variations, it struggles with partial shadows, local illumination variations, and diverse shapes of subjects. Approaches to improve robustness include using local textures for abstraction to compare features instead of pixels (Satoh et al. 2002) and matching based on the consistency of main gradient directions in region segmentation (Hinterstoisser et al. 2010). However, appropriate feature design and preprocessing rules are required depending on the object or scene

(measurement environment, lighting, temperature, etc.), and designing these features and preprocessing requires experience and knowledge. Since the mid 2010s, deep learning-based methods (Ren et al. 2015, Liu et al. 2016, Redmon et al. 2016, He et al. 2017) have surpassed rule-based methods in image recognition. Although deep learning requires data collection and training, it can eliminate the need for preprocessing and feature design that require certain knowledge and experience. In this paper, we focus on leveraging deep learning's benefits: robustness to diverse object shapes and the elimination of feature engineering. We aim to apply these strengths to achieve precise positioning tasks.

Object Detection for Identifying Target Position

Prominent object detection frameworks like Faster R-CNN (Ren et al. 2015), Cascade R-CNN (Cai and Vasconcelos 2017), and Focal Loss (Lin et al. 2018), identify objects in images using rectangular bounding boxes and calculate detection scores on the basis of classification precision. However, these scores do not reflect the uncertainty (confidence) of the object's location. Recent research endeavors like UAD (Lee et al. 2022) and CertainNet (Gasperini et al. 2021) aim to address this by incorporating measures of location uncertainty for object detection. UAD, based on FCOS (Tian et al. 2019), provides confidence ratings for bounding box edges in all four directions (top, bottom, left, right), proving beneficial for detecting occlusions and blurs. CertainNet, based on CenterNet (Zhou, Wang, and Krähenbühl 2019), assesses uncertainty related to objectness, location, and box dimensions. While these models contribute to detecting an object's point within an image, they primarily focus on ascertaining object positions and sizes with rectangular bounding boxes, rather than identifying the exact pixel-level location of an object.

Proposed method

This study proposes a novel segmentation-based position detection method using fixed shapes as target labels to reduce annotation costs. General segmentation models classify each pixel using information within its receptive field. From the results obtained through this segmentation, specific coordinates of objects can be detected through post-processing. For example, we can obtain the specific coordinate on the target object by calculating the centroid from the segmented target. However, general segmentation requires pixel labeling for creating ground truth (GT) images, resulting in high annotation costs. To address this issue, our proposed method uses fixed shapes as target labels, thereby reducing annotation costs.

Figure 1 shows an outline of the proposed method. This method uses U-Net (Ronneberger, Fischer, and Brox 2015) with a VGG16 base (Simonyan and Zisserman 2015) as the network structure for segmentation. The method is called CircleSeg-XAI as it provides position detection and a confidence index from segmentation results using circular shapes.

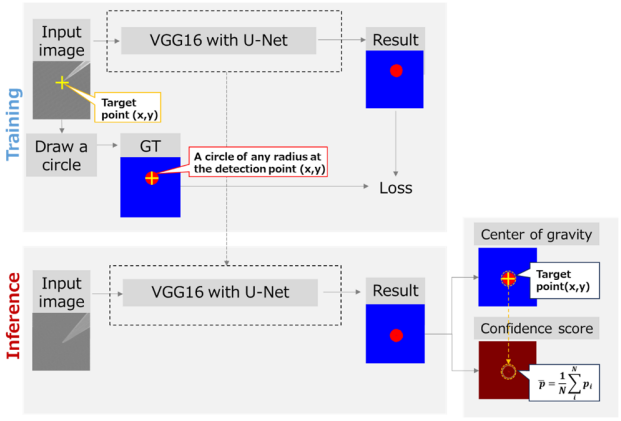

Fig. 1 Schematic diagram of the model training and inference.

During the **learning phase**, we use a ground truth (GT) image containing a circle annotated at the center of the target position that we aim to detect in advance. Unlike common segmentation methods, this method does not require pixel-level labeling. A specific circle radius is set in advance in accordance with the size of the target object, enabling GT images to be created with only the input of target coordinates within the object, reducing annotation costs. The model learns to generate circular outputs centered on the target coordinates for input images by comparing the output results with GT images.

During the **inference phase**, a circular shape centered on the target coordinates is expected to be output for the input image using the trained model. Each pixel determines whether it is within the range of the radius set by the GT for the features of the input image. By post-processing the shape of the obtained target label, the target coordinates are identified by calculating the center of gravity. In addition, this method uses the statistical values of the score values of each pixel obtained by segmentation as a confidence index.

Our **confidence index calculation** utilizes the probability data for each pixel in an image belonging to a specific class obtained by segmentation. The model evaluates the image's spatial information to classify each pixel. By aggregating these classification scores, we gain insights into the spatial attributes of the object, distinguishing between learned and unlearned data. We formalize this process with the following equation for the confidence index: $\bar{p}$

$$\bar{p} = \frac{1}{N}\sum_{i}^{N} p_i$$

Here, N is the number of pixels extracted as targets, and p_i is the score value of each pixel i. Averaging the scores from several pixels provides a more reliable confidence level. While the score attributed to a single pixel may be affected by local noise or misclassification, the process of averaging the scores across multiple pixels yields a more stable and reliable confidence index.

The confidence index is potentially useful for identifying OOD data. Low confidence values suggest that the features of the inference image are different from those of the training data, indicating potential OOD features. In addition, comparing high and low confidence regions can indicated how the model reacts to ID or OOD data and provide guidance for improving performance.

Experiment

In this paper, to evaluate the effectiveness of position detection and confidence indexes using segmentation with fixed-shape labels, we compared our proposed method with the CenterNet-based object detection method. While general object detection models identify object positions by analyzing their outlines, CenterNet utilizes a framework that directly calculates the center coordinates of objects using a heatmap (Zhou, Wang, and Krähenbühl 2019). Therefore, we selected heatmap-based methods such as CenterNet and CertainNet as our benchmark due to their excellence in identifying specific object positions within an image.

Dataset description

We used images taken by a FIB-SEM system as an example application for position detection as shown in Table 1. These grayscale images have a resolution of 1000×1000 pixels, and unlike natural images, they contain random noise throughout. These images are used to accurately extract and process micro-scale semiconductor devices by identifying the needle's tip position at the pixel level. The entire process of extraction and processing is automatically carried out inside the microscope, so incorrect position detection can lead to extraction/processing failure. Therefore, it is important to display the likelihood of failure to the user through the confidence index of the image, which includes features deviation from the training data. Furthermore, in this application, the same probe is used repeatedly, and the needle tip's shape changes significantly as the extraction/processing process progresses. Therefore, detecting OOD images using the confidence index is necessary to display potential failures for various shape deformations.

The images used in this experiment were divided into two groups as shown in Table 1. The first group comprises standard images, which show the needle tip in a sharp state, while the second group comprises deformed images, which feature needle tips with various forms, such as foreign objects attached, bent, or otherwise altered, as illustrated in the lower section of Table 1. We trained the model using only the standard images, thereby designating them as the ID training set, with the deformed images as the OOD images in this experiment. A total of 267 standard images and 256 deformed images were used.

Table 1 Example images with FIB-SEM

Standard images (In Distribution)	
Deformed images (Out of Distribution)	

Experimental condition

In this evaluation, we compared the proposed segmentation-based position detection method with two existing heatmap-based object detection methods in terms of position detection performance and the effectiveness of the confidence index for OOD detection. We used CenterNet (Zhou, Wang, and Krähenbühl 2019) and CertainNet (Gasperini et al. 2021) as the heatmap-based object detection methods. The CenterNet model was referenced from publically available code. CertainNet was constructed by modifying the CenterNet model to apply a radial basis function (RBF) to the heatmap generation part as described in the authors' paper (Gasperini et al. 2021). ResNet-DCN18 was used as the backbone for both models. For the confidence index, the peak values of the objectness and location uncertainty in the heatmap, which are intermediate characteristics of the model, were used. In the CertainNet paper, location uncertainty is defined as the uncertainty in the x direction u_x and the y direction u_y. In this experiment, we used the combined values of these uncertainties as the total location uncertainty, calculated as $u = \sqrt{{u_x}^2 + {u_y}^2}$. The hyperparameters of the experiment were as follows: batch size 16 (CenterNet) and 8 (CertainNet), learning rate 1.25e-4 (Adam), number of epochs 70 (CenterNet) and 80 (CertainNet), size regression loss weight: $\lambda_{size} = 0.01$ (CenterNet) and 0.1 (CertainNet), hyperspace regularization loss weight: $L_{reg} = 0.01$ (CertainNet). These parameters were set by parameter tuning in accordance with the references (Zhou, Wang, and

Krähenbühl 2019, Gasperini et al. 2021) and applying FIB-SEM images.

The proposed method uses U-Net (Ronneberger, Fischer, and Brox 2015) based on VGG16 (Simonyan and Zisserman 2015). The batch size was 4, the learning rate was 0.01 (SSD), the number of epochs was 600, and the circle radius size was 110 pix. Since the proposed method is segmentation-based, MC-Dropout (Kendall and Gal 2017) was used to improve segmentation accuracy, which can be easily integrated into the method. These parameters were set by parameter tuning in accordance with the FIB-SEM images. We trained and evaluated our models using TensorFlow on a single NVIDIA Quadro RTX 4000 8GB GPU. The inference time per image was 809 ms with MC-Dropout (N=50), which was not a significant issue for our application.

To evaluate the position detection performance, standard images were used as training images, and 30 images with sharper needle tips were extracted from the deformed images and evaluated as the test data of the position detection performance. The deviation: $d = \sqrt{(x - x')^2 + (y - y')^2}$ is measured as the Euclidean distance between the training coordinates (x, y) and the inferred coordinates (x', y') of the needle tip.

To evaluate the confidence index, that of the standard image group was set as a threshold, and the ratio of the deformed image group to the threshold was calculated. We compared the proposed method with the existing method from the viewpoint of how well the proposed method can discriminate unlearned deformed images (OOD) from standard images (in distribution, ID) using the confidence index.

To evaluate effectiveness in terms of data diversity, the proposed method was applied to five different images from an SEM system related to semiconductor manufacturing.

Experimental result

(1) Evaluation of position detection performance

Table 2 shows the statistical values of position deviation for the proposed method, CircleSeg-XAI, and the existing methods. In the standard images (ID), CircleSeg-XAI achieves an average positional deviation of 2.16 pixels, which is smaller than of CenterNet's 3.12 pixels and CertainNet's 3.85 pixels. In the deformed images (OOD), CircleSeg-XAI achieves an average positional deviation of 4.37 pixels, which is smaller than CenterNet's 4.97 pixels and CertainNet's 7.75 pixels.

Figure 2 shows the box plot of the positional deviation for the proposed method and the existing methods. Fliers in the box plot represent individual points that fall beyond 1.5 times the interquartile range above or below the first and third quartiles. The proposed method shows a smaller positional deviation for both ID and OOD images compared with the existing methods. Among the existing methods, CertainNet shows an increase in average positional deviation by 0.73 (=3.85–3.12) pixels (ID) and 2.78 (=7.75–4.97) pixels (OOD) compared with CenterNet. Although the backbone of CenterNet and CertainNet is common, the main difference is the reflection of uncertainty in the heatmap generation. This difference is considered to be the factor for the lower position detection performance of CertainNet.

Table 2 Positional deviation performance

	In Distribution (ID)		Out of Distribution (OOD)	
	Mean	Std	Mean	Std
CircleSeg-XAI (Ours)	2.16	1.24	4.37	2.28
CenterNet	3.12	1.49	4.97	2.47
CertainNet	3.85	1.99	7.75	3.86

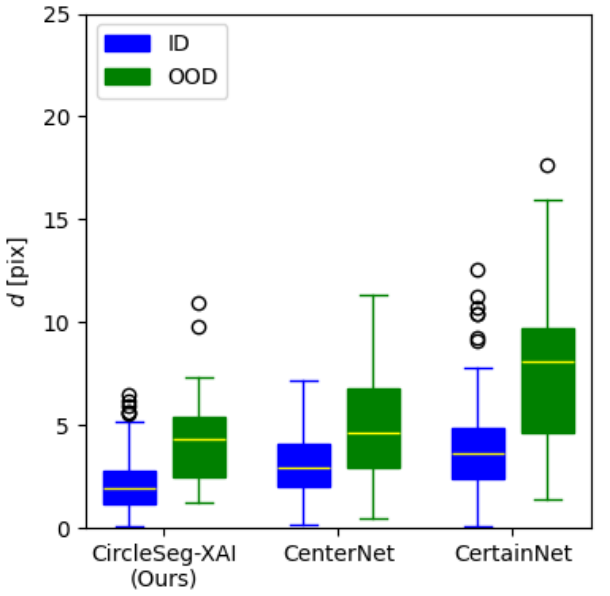

Fig. 2 Box plot of positional deviation

Figure 3 shows the top three images with the largest positional deviation for each method. While the deformed images group used in this evaluation extracted sharper probe tips compared with other images, they still showed deformation compared with the probe tip's sharpness in the standard images group. Therefore, both the proposed and existing methods showed increased positional deviation for the deformed images group compared with the standard images group. In the proposed method, the difference between the features of the ID and OOD images causes a change in the identification of the circular region of the target in each pixel, increasing the positional deviation. However, compared with the other methods, the proposed method showed a smaller positional deviation for both ID and OOD images, confirming the proposed segmentation-based method's superiority in position detection performance for this data.

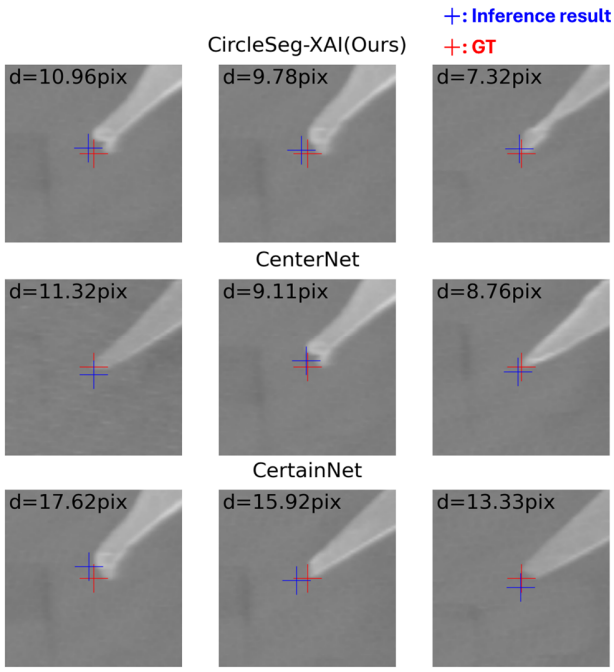

Fig. 3 Top3 images of position deviation for each method. Images cropped from the original images around the GT target positions.

(2) Validity evaluation of confidence index

Table 3 shows the results of the ratio of OOD images extracted for each model using the confidence index. The proposed method's confidence index, mean score $\bar{p}$, had the highest discrimination rate of 89.5%, followed by CenterNet's location uncertainty u at 57.4%.

Figure 4 shows the histograms of ID and OOD images for each model, extracting the highest discrimination rates among the confidence scores listed in Table 3. These results indicate that establishing a minimum confidence index value for the ID images as the threshold (as shown by the blue dotted line in Fig. 4) identifies the ratio of OOD images below this threshold, as shown as a numerical value on the histogram. In the proposed method, the numerical value of the ID image is locally distributed to 0.97~0.98, while that of the OOD image is widely distributed below 0.97. Conversely, the distributions of u in CenterNet/CertainNet tended to overlap for both ID and OOD images. These results show that the proposed method's confidence index is superior over existing methods in terms of discriminating OOD images.

Table 3 Discrimination rate of OOD images for each model and confidence score

Model	Confidence score	Extraction rate of OOD images[%]
CircleSeg-XAI (Ours)	Mean Score $\bar{p}$	**89.5**
	Circularity	53.5
CenterNet	Objectness	3.1
	Location uncertainty u	**34.0**
CertainNet	Objectness	5.1
	Location uncertainty u	**57.4**

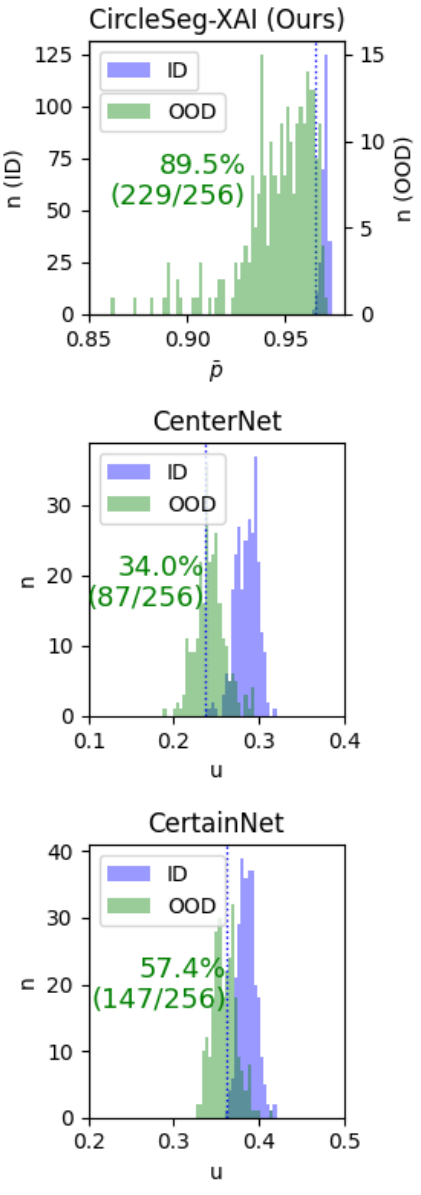

Fig. 4 Comparison of confidence scores for each model

Figure 5 shows the relationship between each image and the confidence index for each method. The proposed method demonstrates a strong correlation: the confidence index $\bar{p}$ tends to be higher when the needle tip closely resembles the trained needle (ID), and tends to be lower as the needle tip deforms from the trained image (ID). Conversely, the results of the confidence index u for CenterNet and CertainNet show a weak correlation between needle tip shape and u. The middle row values are smaller compared with those in the top and bottom rows and the u values are almost the same in the top and bottom rows. In the images, although the top row does not show deformation of the needle tip and the bottom row does show deformation, the u values are similar. These results show that the correlation between the shape and confidence index is higher in the proposed method, and the proposal is effective as an index showing the degree of deviation from the trained image (ID).

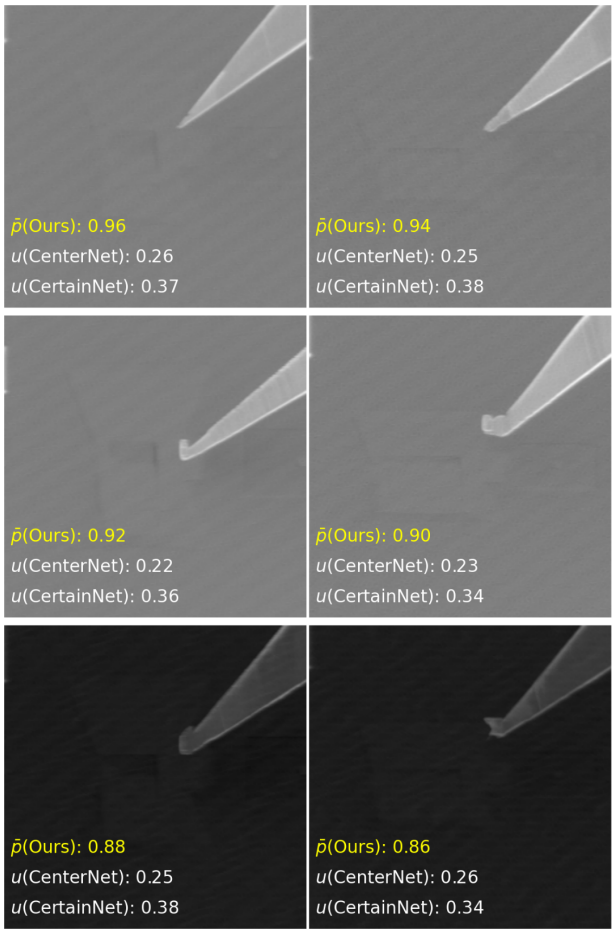

Fig. 5 Correspondence between each image and the confidence metrics for each method.

(3) Evaluation of data diversity for SEM images

Table 4 provides an overview condition of datasets observed using the FIB-SEM system for diverse datasets and use cases. The number of training data was set by randomly extracting from each dataset and determined it until there was enough data to detect the target. The original image sizes were resized to the model's input size (256 pixels) for training and inference. The model's output image is then resized back to match the original image size. GT radius is the circular radius of the target segmented image as ground truth. These radiuses were empirically determined based on the characteristic sizes around the target areas in the images.

Figure 6 shows representative images (left column) and inference results (middle and right columns) respectively. The proposed method demonstrates that the segmentation results are correctly performed at the same positions as specified by GT, including images of needle tips (A, B), corners (C, D), and holes (E). For example, even in images with shapes deviating from the training images, where the needle tips are more rounded (A2, B2), the targeted needle tip can be accurately detected. Additionally, in low contrast images (C2, D2), the target corner points are accurately detected. However, in D2, the target segmentation shows a distorted circle towards the adjacent corner. This suggests the model may al-so consider the adjacent corner as a potential target. This information indicates high uncertainty due to the deviation from the perfect circle specified by GT. Although the dataset E consists of images with different sizes, proposed method accurately detects holes in both E1 (150x150 pixels) and E2 (300x300 pixels). Therefore, the proposed method has been validated on five types of images measured using FIB-SEM, confirming its ability to accurately identify specific locations within the images.

Table 4 Overview conditions of the evaluated datasets observed using the FIB-SEM system.

Dataset	N-Total	N-Train	Image size [pix]	GT Radius[pix]
A	523	267	1000×1000	110
B	701	160	1000×1000	90
C	91	27	1000×1000	70
D	245	63	1000×1000	30
E	325	165	67×67 ∼ 300×300	18

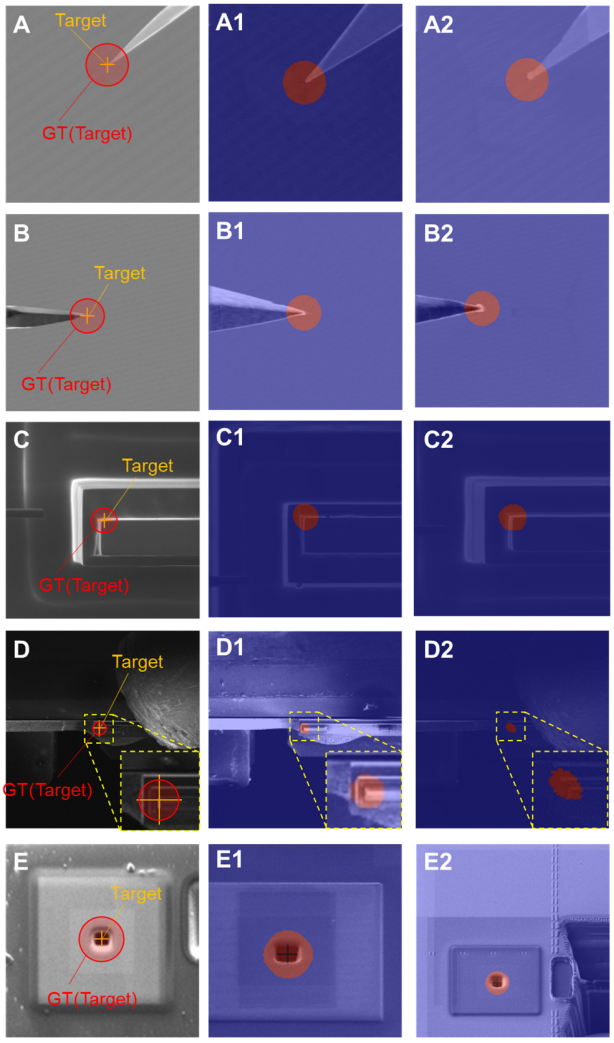

Fig. 6 Experimental results for datasets from five different scenes observed using the FIB-SEM system.

(4) Ablation study

As an ablation study of the proposed method, we examined the changes in performance of point detection with varying circle radius of the ground truth. We evaluated the position deviation d of the needle tip with changes in the circle radius for dataset A, keeping other hyperparameters fixed. The results, shown in Table 5, reveal a trend with a minimum value observed at 110 pixels. In this evaluation, we only confirmed on dataset A, the optimal circle radius may depend on each dataset. Optimizing the circle radius could be effective when applying the proposed method to applications requiring high-precision position identification.

Table 5 Results of position deviation d with varying radius.

GT radius[pix]	d[pix]
55	9.8
82.5	8.7
110	**7.4**
137.5	17.5
165	15.0

Conclusions

In this paper, we proposed a novel segmentation-based target point detection method. By using fixed-shape circular labels, our method reduces annotation costs while providing object position and a confidence index on the image. When applied to the task of identifying needle tip positions in FIB-SEM images, our method demonstrated superiority in terms of positional deviation, achieving 2.16 pixels (σ=±1.24 pixels) compared with the existing heatmap-based object detection method's 3.12 pixels (σ=±1.49 pixels). When comparing the effectiveness of the confidence index for OOD discrimination, our method had the highest ratio of separating trained and untrained images, confirming that our method can discriminate 89.5% of untrained images using our confidence index. Additionally, the evaluation of images from five different scenarios of semiconductor manufacturing measured using SEM, confirmed the effectiveness of our method. Future research will focus on a theoretical analysis of the proposed method to determine its effectiveness and limitations.

References

Cai, Z.; Vasconcelos, N. 2017. Cascade R-CNN: Delving into High Quality Object Detection. 2018. In Proceedings of Computer Vision and Pattern Recognition (CVPR), pp.6154-6162.

Gawlikowski, J.; Njieutcheu Tassi, C.R.; Ali, M.; Lee, J.; Humt, M. et al. 2021. A survey of uncertainty in deep neural networks. arXiv, 2107.03342.

Gasperini, S.; Haug, J.; Nikouei Mahani, M.A.; Marcos-Ramiro, A.; Navab, N.; Busam, B.; Tombari, F. 2021. CertainNet: Sampling-free Uncertainty Estimation for Object Detection. IEEE ROBOTICS AND AUTOMATION LETTERS. PREPRINT VERSION. ACCEPTED NOVEMBER, 2021.

He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. 2017. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961-2969.

Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Fua, P.; Navab, N. 2010. Dominant Orientation Templates for Real-Time Detection of Texture-Less Objects. Proceedings of Computer Vision and Pattern Recognition (CVPR), pp.2257-2264.

Kendall,A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. arXiv, 1703.04977.

Lee, Y.; Hwang, J.W.; Kim, H.I.; Yun, K.; Kwon, Y.; Bae, Y.; Hwang, S.J. 2022. Localization Uncertainty Estimation for Anchor-Free Object Detection. Computer Vision – ECCV 2022 Workshops: Proceedings Part VIIIOct 2022 Pages 27–42.

Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár .P. 2018. Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs.CV].

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. 2016. SSD: Single Shot MultiBox Detector. European Conference on Computer Vision, Springer, pp. 21–37.

Long, J.; Shelhamer, E.; Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In CVPR.

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. 2016. You Only Look Once: Unified, real-time object detection. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, pp. 779–788.

Ren, S.; He, K.; Girshick, R; Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 [cs.CV]

Ronneberger, O.; Fischer, P.; Brox, T. 2015. U-Net: Convolutional Networks for Bio-medical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 pp 234–241.

Satoh, Y.; Tanahashi, H.; Wang, C.; Kaneko, S.; Niwa, Y.; Yamamoto, K. 2002. Robust Event Detection by Radial Reach Filter (RRF) Proceedings of International Conference on Pattern Recognition (ICPR), Vol.2, pp.623-626.

Simonyan, K.; Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV].

Tian, Z.; Shen, C.; Chen, H.; He, T. 2019. FCOS: Fully Convolutional One-Stage Object Detection. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9627-9636.

Zhou, X.; Wang, D.; Krähenbühl, P. 2019. Objects as Points. arXiv:1904.07850.

Zeiler, M.D.; Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In ECCV 2014, Part I, LNCS 8689, pp. 818–833