

Anomaly Detection of Integrated Circuits Package Substrates Using the Large Vision Model SAIC: Dataset Construction, Methodology, and Application

Supplementary Material

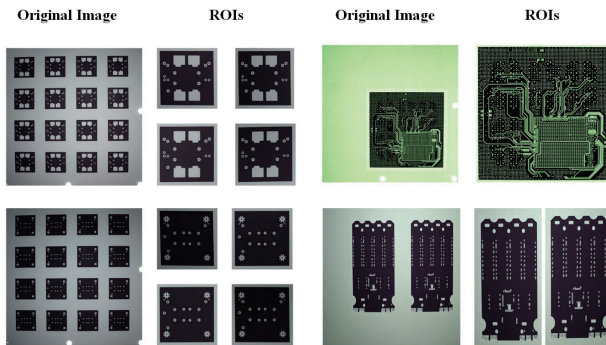


Figure 9. Examples of ROI extraction.

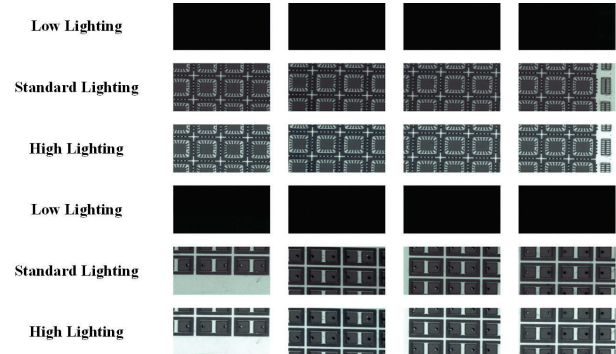


Figure 10. Examples of different lighting condition.

1. More Details of Data Collection

To conduct a high-quality annotation of each sample, we introduced the details of our data collection process in Sec. 3. Due to page limit, we did not introduce the details of ROI extraction, lighting adjustment, and annotation. In this section, we will further expand the details of the above three processes one by one.

1.1. ROI Extraction

As indicated in **Collection & Detection Process** of Sec. 3, in the samples we collected, the circuit shapes were evenly distributed among the four quadrants, but due to the different sizes of the circuit shapes, the number of shapes covered in each quadrant is also different. As shown in Fig. 9, there are three types of quantity distributions, namely 1, 2, and 16. In order to acquire accurate image of each circuit pattern, we introduce a ROI extraction method. Firstly, we apply adaptive thresholding to grayscale images. This method utilizes local neighborhood information to enable the generated binary image to adapt to uneven lighting conditions and improve the robustness of edge detection. Secondly, we perform morphological opening and closing operations on the binary graph to remove isolated small noise points. Thirdly, we use the Canny algorithm for edge detection, extracting salient edges from the image and obtaining accurate contour information. Fourthly, we use the findContours function in OpenCV to detect all contours in the binary image. By calculating the contour area, effective contours with an area greater than the set threshold are selected to reduce the processing of irrelevant information. Finally, we perform dilation on the filtered contours, merge similar regions, and extract the bounding rectangle of each contour to determine

if there is overlap. Through the above operation process, we can effectively extract each independent circuit pattern, thereby constructing more accurate annotation data. In addition, we can filter out defects outside of non-circuit pattern through ROI extraction, which will not affect the product quality.

1.2. Lighting Adjustment

As illustrated in **Collection & Detection Process** of Sec. 3, lighting is an important factor for anomaly detection. Due to the varying thickness of products from different batches, using a constant light source can result in uneven distribution of light during sample collection, which can have a significant impact on subsequent detection. In **AOI Equipment** of Sec. 3, we introduced that our collection system includes a scanning camera to recognize the QR code of each sample. Thus, we use the database of the upper computer to organize the thickness of different models of products. Then, the upper computer issues adjustment instructions to the PLC of the AOI equipment through the QR code information, thereby controlling the controller of the light source to change the lighting. As shown in Fig. 10, standard lighting can further improve the quality of collected samples.

1.3. Annotation

In **Annotation** of Sec. 3, we display the example of different defects. We use LabelMe to annotate each sample via specific defect ID from 1 to 6. We create polygon annotation (also can be called pixel-level annotation) and bounding box annotation, as shown in Fig. 11. The annotation process involved more than 20 personnel and took nearly a month to complete.

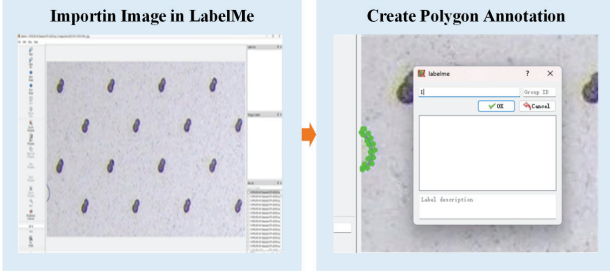


Figure 11. Examples of annotation accomplished by LabelMe.

2. More Details of Experiments

We introduce some experimental results in Sec. 4 and Sec. 5.3. Due to page limit, we did not introduce the details of data partitioning under different tasks and generalization of SAIC in other scenarios. In this section, we will further expand the details of the above two parts one by one.

2.1. Data Partitioning under Different Tasks

The CPS2D-AD dataset can be formulated as $\mathcal{D} = \{\mathcal{D}_{normal} + \mathcal{D}_{abnormal}\} = \{(x_i)_{i=1}^n, (x_j)_{j=1}^m\}$, where x_i is each normal sample, x_j is each abnormal sample. As indicated in Tab. 1, there are 11000 normal and 9781 abnormal samples in our dataset. The following is the data partitioning method for each task:

Unsupervised task aims to recognize abnormal samples only rely on normal samples, that means we do not require abnormal data to participate in the training process. In this task, we use 80% of normal data for unsupervised learning training, and the remaining 20% of normal data and abnormal data will be used for model evaluation, e.g. $\mathcal{D}_{unsup}^{train} = \{0.8 \times \mathcal{D}_{normal}\}$, $\mathcal{D}_{unsup}^{test} = \{0.2 \times \mathcal{D}_{normal} + \mathcal{D}_{abnormal}\}$.

Few-shot task uses a small amount of labeled data for model training. Specifically, we select 200 defect samples from each of the six types of defects to construct data suitable for few-shot task. We eliminated two of six types of defects sequentially to create three folds. By conducting numerous trials with various folds and calculating the mean, we effectively assessed the algorithm’s generalization capability in novel scenarios.

Semi-supervised task is a further extension of unsupervised task, which improves the performance of the algorithm by increasing the usage of annotated data. In this task, we use 50% of normal data and 50% of abnormal data for model training, and the remaining 50% of normal data and 50% of abnormal data will be used for model evaluation, e.g. $\mathcal{D}_{semi}^{train} = \{0.5 \times \mathcal{D}_{normal} + 0.5 \times \mathcal{D}_{abnormal}\}$, $\mathcal{D}_{semi}^{test} = \{0.5 \times \mathcal{D}_{normal} + 0.5 \times \mathcal{D}_{abnormal}\}$.

Fully-supervised task achieves much higher prediction

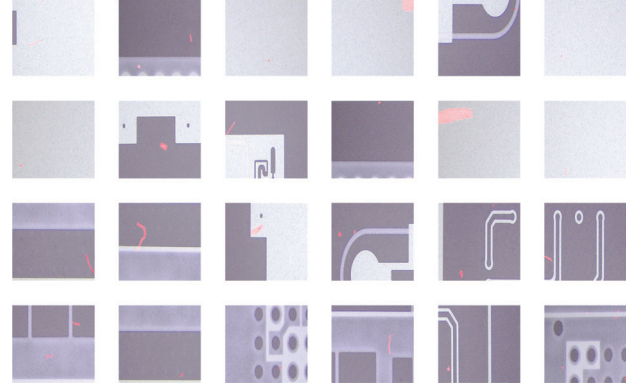


Figure 12. More visualization results.

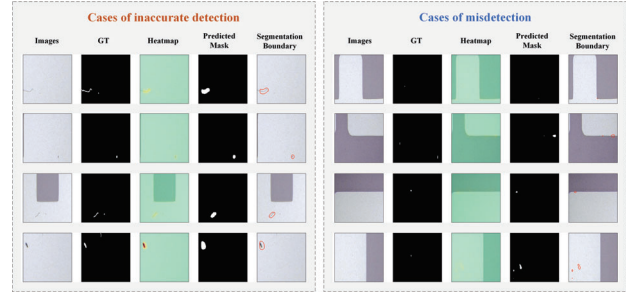


Figure 13. Failure cases of SAIC.

accuracy than the above tasks by using a large amount of annotated defect data. In this task, we do not apply normal data and only use 70% of abnormal data as training data, with the remaining 30% as testing data, e.g. $\mathcal{D}_{fully}^{train} = \{0.7 \times \mathcal{D}_{abnormal}\}$, $\mathcal{D}_{fully}^{test} = \{0.3 \times \mathcal{D}_{abnormal}\}$.

2.2. More Visualization Results

In order to provide a more intuitive display of the results, we have supplemented the detection results of SAIC, as shown in Fig. 12. It can be clearly seen that our algorithm has a good recognition effect on small surface anomalies.

2.3. Failure Cases

We have collected several failure cases of SAIC to further illustrate the challenges of our dataset, as shown in the Fig. 8. Although SAIC has achieved SOTA results, there are still some challenging images that may encounter problems such as missed detections and inaccurate detections. In addition, it can also be seen from the heat map that SAIC did not activate the defect area. This is because some defects in the defect samples are too small or have color and texture features that are close to the circuit.



Figure 14. More device details of AOI system.

3. Comparison with the Same Type Datasets

As indicated in 2D Industrial Anomaly Detection Datasets of Sec. 2, many industrial data have already been published in other publications. These datasets are all focused on expanding the boundaries of anomaly detection and providing additional support for algorithm evaluation. Another two famous anomaly detection datasets, MVTec AD and VisA, have better universality in covering sample types than ours, but the quantity of each sample is lower than ours. MVTec AD has 4096 normal samples and 1258 abnormal samples, covering five texture data and ten object data. VisA has 9621 normal samples and 1200 abnormal samples, including 12 subsets corresponding to 12 different objects. As a comparison, we divided the collected ceramic substrate objects into six categories based on the sources of different process flows and named them open circuit, mouse bite, overflow, foreign matter, poor pattern, and leakage, respectively. The types of defects in ceramic package substrates are more helpful for us in exploring anomaly detection at small sizes, which is also a more challenging and targeted task. Highly specialized research and datasets are crucial for driving the development of specific fields.

4. More Details of AOI System

The subject area of this paper we have selected is vision application and system. Therefore, in order to verify the authenticity of our AOI system, we have added more device details, as shown in Fig. 14. What’s more, we display the software interface as shown in Fig. 15. We also have uploaded a video of the device running in the attached file (02-Device Video.mp4).

5. Open-source License of CPS2D-AD

We construct a large-scale dataset about ceramic package substrates, which provides solid data support for researchers

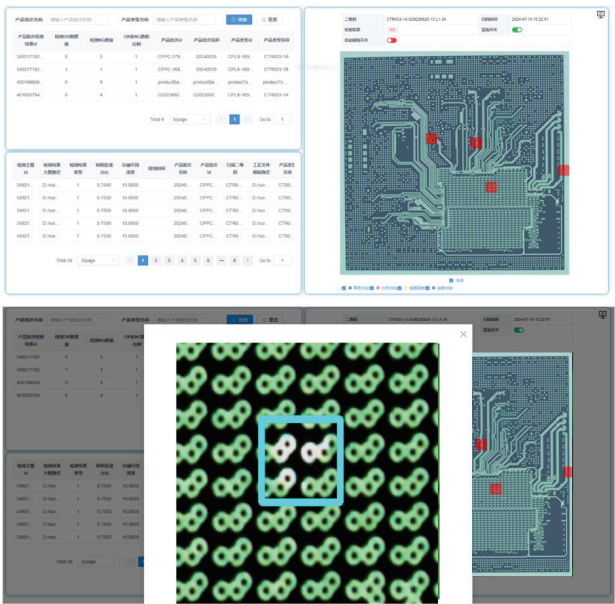


Figure 15. More device details of software interfacc.

associated with integrated circuits. Each sample of CPS2D-AD is obtained from actual industrial products, so no negative social impact will exist. Moreover, the data collection and open-source have been authorized by a resolution of the company’s board of directors. However, considering the possibility of data leakage during the review stage, our anonymous GitHub link does not fully cover all the data. If the paper is fortunate enough to be accepted, we will open source all the data samples and annotated files.