FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 49 (2024) No. 3
ISSN 0867-6356
DOI: 10.2478/fcds-2024-0014 e-ISSN 2300-3405

CC-De-YOLO: A Multiscale Object Detection Method for Wafer Surface Defect

Jianhong Ma, Tao Zhang, Xiaoyan Ma, Hui Tian*

Abstract. Surface defect detection on wafers is crucial for quality control in semiconductor manufacturing. However, the complexity of defect spatial features, including mixed defect types, large scale differences, and overlapping, results in low detection accuracy. In this paper, we propose a CC-De-YOLO model, which is based on the YOLOv7 backbone network. Firstly, the coordinate attention is inserted into the main feature extraction network. Coordinate attention decomposes channel attention into two one-dimensional feature coding processes, which are aggregated along both horizontal and vertical spatial directions to enhance the network's sensitivity to orientation and position. Then, the nearest neighbor interpolation in the upsampling part is replaced by the CAR-EVC module, which predicts the upsampling kernel from the previous feature map and integrates semantic information into the feature map. Two residual structures are used to capture long-range semantic dependencies and improve feature representation capability. Finally, an efficient decoupled detection head is used to separate classification and regression tasks for better defect classification. To evaluate our model's performance, we established a wafer surface defect dataset containing six typical defect categories. The experimental results show that the CC-De-YOLO model achieves 91.0% mAP@0.5 and 46.2% mAP@0.5:0.95, with precision of 89.5% and recall of 83.2%. Compared with the original YOLOv7 model and other object detection models, CC-De-YOLO performs better. Therefore, our proposed method meets the accuracy requirements for wafer surface defect detection and has broad application prospects. The dataset containing surface defect data on wafers is currently publicly available on GitHub (https://github.com/ztao3243/Wafer-Datas.git).

Keywords: Surface defect detection on wafers, YOLOv7, coordinate attention, CAR-EVC, IDetect_Decoupled

* Zhengzhou University, Cyber Science and Engineering, ZhangZhou, China,
Email: majianhong@zzu.edu.cn, zhang_tao@gs.zzu.edu.cn, tianhui@zzu.edu.cn,
lyx_980715@163.com.

1. Introduction

With the continuous advancement of modern technology, micro-nano processing technology has been widely applied in various fields of life. The maturity and application of MEMS[7] technology provide important support for micro-nano processing technology in production and manufacturing, as well as open up broader application areas for MEMS technology development. Wafers are critical materials in micro-nano processing technology. In the manufacturing process, desired patterns are formed on the wafer through processes such as photolithography. However, the surface of wafers may exhibit certain defects due to factors such as room cleanliness, equipment performance, process parameters, and standardization issues with wafer handling. Theese defects have a significant impact on production yield, thereby seriously affecting wafer production efficiency and scrap rate. Additionally, with the increasing complexity of MEMS device design and wafer integration density, semiconductor wafer defects are becoming more complex, including mixed-type defects where multiple defect types overlap or intersect. Therefore, higher manufacturing accuracy and more reliable detection methods are required to ensure product quality and reduce economic losses. Once the correct defect type is identified, engineers can quickly locate abnormal faults during the manufacturing process, find the cause of the defect, adjust the manufacturing process, and improve production efficiency while reducing scrap rates. Therefore, accurate detection of surface defects on wafers is an indispensable step in the semiconductor chip manufacturing process. The aforementioned measures can assist engineers in promptly identifying defects and implementing corresponding solutions, thereby enhancing product quality and production efficiency, ultimately resulting in reduced cost expenditures.

In recent years, with the continuous development of convolutional neural networks, object detection methods based on this network have rapidly advanced and widely applied in the field of object detection due to their efficiency, high accuracy, and other characteristics. However, in early wafer surface defect detection, manual inspection was commonly used, which had issues such as low efficiency, poor accuracy, high cost, and strong subjectivity, which could not meet the requirements of modern industrial products. Therefore, image object detection systems based on deep learning have become a trend. Currently, deep learning-based object detection networks are mainly divided into one-stage networks represented by SSD[15] and YOLO series, and two-stage networks represented by Faster R-CNN[18] and Mask R-CNN[6]. The two-stage network generates several candidate regions that may contain objects through the region proposal network (RPN), converts these candidate regions into feature vectors, and then inputs them into classifiers and regressors for classification and positioning. The one-stage network is an end-to-end object detection algorithm that does not require the use of RPN, so it has faster detection speed and is easy to detect some clear edge contours and obvious defects. However, cross and overlapping defects in wafer surface defects present challenges for detection algorithms. To improve the detection performance of the above-mentioned wafer surface defects, this study proposes a method based on CC-De-YOLO, which improves the defect feature extraction ability and detection performance by introducing attention mechanisms, long dependency relations in upsampling, and decoupled heads. To verify the proposed method, we built our own wafer surface defect dataset, which includes six types of defects: open, edge bite, gray line, stains enbedded, scratch, and short circuit. Experimental results show that our proposed CC-De-

YOLO performs better than other object detection methods for detecting overlapping distributed defect types. Our main contributions are summarized as follows:

1) We created a wafer surface defect dataset containing six types of defects by using four typical micro-nano sensors as templates.
2) We proposed a CC-De-YOLO object detection model based on the YOLOv7 algorithm, which embeds defect locations into feature maps and introduces semantic upsampling and decoupled heads, significantly improving the detection ability of wafer surface defects.
3) We proposed a semantic upsampling module CAR-EVC that can capture long dependency relations.

This paper aims to explore a deep learning-based method for wafer defect detection. By establishing a suitable dataset and model, combined with traditional image processing and analysis methods, the automatic detection and classification of wafer surface defects can be achieved. This method will significantly improve the accuracy and efficiency of detecting surface defects on wafers, providing a broader range of application spaces for the production and manufacturing of micro and nanofabrication technologies.

2. Related Work

Defect detection is a hot topic in the industry, and image-based detection has shown great potential in surface defect detection, some of which have also been used in wafer defect detection. These studies can be divided into three main groups: image signal processing, machine learning, and deep learning methods.

In the early research of wafer surface defect detection algorithm, image signal processing methods were commonly adopted, mainly including wavelet transform methods and template matching methods. Y Kim et al. [11] proposed a surface detection method based on wavelet transform for detecting defects in solar wafers. Wavelet transform can decompose images into multi-resolution representations, presenting local sub-images of different spatial frequencies. It performs well in extracting texture features, but highly dependent on threshold selection, and with poor adaptability. Liu Xifeng [27] combined the SURF image registration algorithm with spatial position matching of test wafers and standard wafer patterns, and applied pattern recognition contour extraction technology for wafer defect detection. Through testing with a self-built dataset, accurate wafer defect matching was achieved.

Machine learning has achieved good results in wafer defect detection through supervised, unsupervised, and self-supervised algorithms. However, there are still some shortcomings. L. Xie et al. [26] proposed a wafer defect pattern detection scheme based on the support vector machine (SVM) algorithm. SVM is effective in classifying multivariate, multimodal, and inseparable data points, but it is not particularly friendly to multiple samples, and selecting the appropriate kernel function can be challenging. Through testing with simulated datasets and real wafer images, the accuracy achieved is over 90%. Huang [10] proposed an automatic wafer defect clustering algorithm that utilizes a self-supervised multilayer perceptron to detect defects and label all defective chips. Self-supervision can extract intrinsic feature information from large-scale unlabeled sample labels and generate sample labels. The algorithm was tested on the WM-811K dataset and significantly im-

proved the detection accuracy. However, a drawback is the requirement for a large number of data samples to enhance the accuracy of sample labeling.. Kong and Dong Ni [12] proposed a semi-supervised incremental modeling framework for wafer image analysis. First, a semi-supervised incremental model is used to classify wafer images, and then the model performance is improved through active learning and pseudo-labeling. The model performance is enhanced by active learning and pseudo-labeling, but it still depends on the amount of accurately labeled data for training the model.

With the development of deep learning technology, methods based on convolutional neural networks (CNN) have emerged, effectively solving the problems in traditional machine learning methods. Compared to traditional machine learning methods that require engineers to manually design features, CNN-based methods can automatically learn the optimal feature representation from raw data, with advantages such as adaptability and high generalization performance, bringing vitality to the field of wafer inspection. In traditional machine learning methods, different feature representations need to be constructed for different tasks, which requires professional domain knowledge and a large number of trial-and-error experiments. CNN-based methods, on the other hand, eliminate this limitation, requiring no manual intervention, reducing the impact of human factors, and improving the model's generalization performance. Wang J et al. [24] proposed a mixed DPR (MDPR) deformable convolution network (DC-Net) for wafer defect classification. The sampling region is focused on the defect feature area, which has good robustness for mixed defects. Simultaneously, the paper introduced a new dataset,"MixedWM38," containing 38 wafer defect patterns, with each pattern comprising 1000 wafer maps. The authors labeled the 38 wafer defect patterns as C1-C38 and categorized them into Single type, 2 mixed-type, 3 mixed-type, and 4 mixed-type. The experimental results indicate that the average accuracy can reach 93%. Ssu-Han Chen et al. [1] were the first to use a combination of generative adversarial networks and the YOLOv3 object detection algorithm for small-sample wafer defect detection. GAN was employed to enhance the diversity of defects and improve the generalization ability of YOLOv3. Testing on a proprietary dataset from a case company yielded an AP of 88.12%, but the accuracy still needs improvement.. P P Shinde et al. [20] proposed using the advanced YOLOv4 for wafer defect detection and localization. Compared to YOLOv3, the backbone extraction network was upgraded from Darknet-19 to Darknet-53, and the Mish activation function was adopted to enhance the network's robustness, significantly improving the detection capability and making it more effective in detecting and localizing complex wafer defect patterns. The model was tested on the WM-811k public wafer map dataset, achieving an accuracy of 92%. H Han et al. [5] designed a segmentation system based on an improved U-Net network. An RPN network was added to the original U-Net network to obtain defect region proposals, which were then input into the U-Net network for segmentation. Although two-stage networks have precise segmentation effects on wafer defects, the detection speed becomes slower.

Although inspection schemes for surface defects of wafers based on image processing, machine learning, and deep learning have made certain achievements, the detection accuracy of mixed defects on wafer surfaces is still not high. Therefore, there is still a vast research space in the field of wafer surface defect detection.

3. Materials and Methods

3.1. Data Acquisition and Processing

The focus of this study is on surface defects occurring during the manufacturing process of micro-nano sensors, which are influenced by environmental factors and production processes. The availability of representative datasets and accurate descriptions is crucial for advancing research in this field. To better simulate real-world defect generation in production, four typical micro-nano sensor components—Helical Excitation CoilCrossed Finger Electrode, Cantilever Beams, and PIN Photodiode Sensors—were used to create mask templates, as shown in Figure 1. These four micro-nano sensor components are widely used in the micro-nano manufacturing field and represent various manufacturing technologies and designs commonly found in the production process of micro-nano sensors. By choosing different designs, the research covers a broader range of manufacturing defect categories, and the observed defect types are typical and universal. In the lithography process, the mask and silicon wafer are first placed in specific positions on the lithography machine. Then, photoresist is coated onto the silicon wafer, followed by exposure processing using a lithography machine [17] to transfer the mask pattern onto the wafer. Subsequently, a development process is carried out to remove the exposed photoresist, leaving only the unexposed parts. The quality and clarity of the final chip pattern are also influenced by the thickness of the photoresist, exposure time and intensity settings, as well as the choice of developer and processing time. In this study, electron microscopy and an attached camera were used to collect samples. The collection process took place under yellow light in a cleanroom to ensure that the chip was not affected by external environmental interference. A total of 2278 photos were collected, with a sampling magnification of 1X and a sample resolution of 1280×720 pixels.

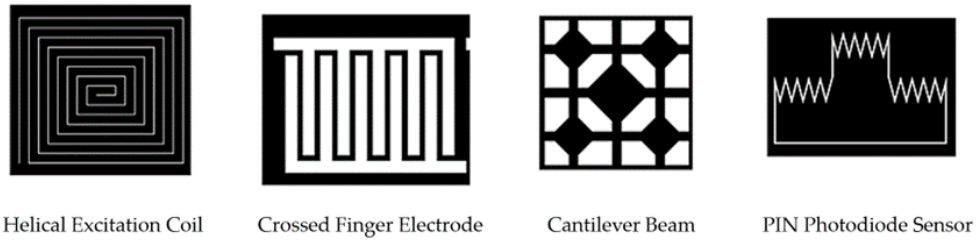


Figure 1. The structure of four Micro-nano Sensors are: Helical Excitation Coil, Crossed Finger Electrode, Cantilever Beam, and PIN Photodiode Sensor.

The collected defect samples underwent expert validation in the field of micro-nanotechnology to determine the types of defects. Based on variations in process errors during production, the defects were categorized into six different types. The first type is stains embedded defects, resulting from insufficient cleanliness in the production environment. The second to fourth types include open, edge bite, and gray line defects, which are caused by improper adjustments in exposure or development times. The fifth type is scratch defects arising from accidental contact between equipment and chips. The sixth type is short circuit defects, caused by the close proximity of lithographic lines or inaccuracies in the lithography machine. These six types of defects are illustrated in Figure 2. Defect samples were labeled by a specialized annotation team consisting of relevant graduate students

under the guidance of experts in the field of micro-nanotechnology. After annotation completion, the labels underwent further quality validation by experts to ensure professionalism and accuracy. The labeled files were saved in txt format, following the YOLO dataset format.

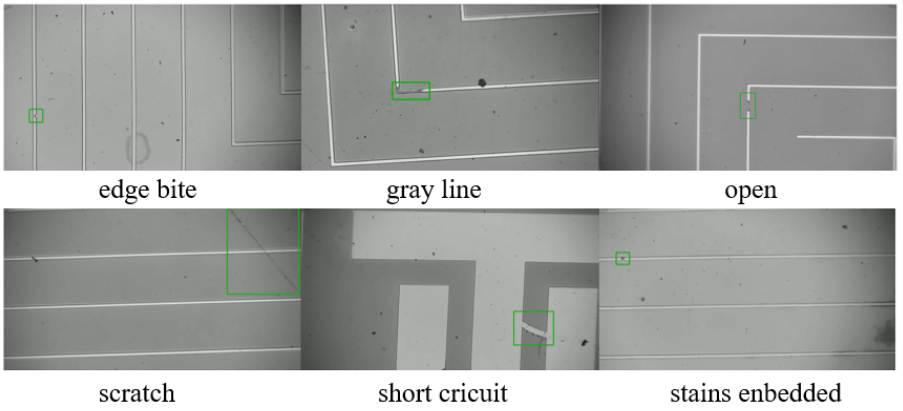


Figure 2. Six types of wafer surface defect images

Acquiring datasets for semiconductor chip defects is costly, and the limited number of samples poses a significant challenge to the semiconductor detection field. To overcome the issue of limited data samples, offline data augmentation techniques [21], along with the combination of MixUp and Mosaic, were employed to expand the dataset and improve the model's generalization ability. Offline data augmentation methods include rotation, random cropping, noise addition, and motion blur. Addressing the potential rotation of wafers during the actual inspection process, rotation data augmentation was applied to wafer samples. To better adapt the detection model to defects at different positions and orientations on wafers, random cropping was performed on wafer defect samples. Given that visual equipment during the actual detection process may experience external signal interference, generating noise in the samples, various types of noise were randomly added to the samples to help the model resist noise interference. To address potential vibration effects during the detection process, motion blur was added to sample images to help the model adapt to vibration interference. MixUp involves weighted blending of two randomly selected training images, mixing their labels in the same weighted proportion to generate new samples. The Mosaic technique combines four different images into a single large image, significantly expanding the diversity of target content exposed during model training. By using these techniques, a large number of new data samples could be generated from the original dataset without altering the sample distribution. This method effectively increases the diversity of data samples, reduces the sensitivity of the model to noise and complex situations, and improves the model's robustness and generalization ability. Additionally, data augmentation techniques can mitigate the risk of overfitting, preventing the model from overfitting to the training set and thereby enhancing model performance. Therefore, data augmentation is a critical technology in semiconductor defect detection. The dataset was ultimately expanded to 4,330 images, with 3,440 images used for training, 430 images for validation, and an additional 430 images for testing, in an 8:1:1 ratio. Each wafer sample has an average of 2.7 defects. The distribution of the six types of surface defects in the augmented dataset is shown in Figure 3.

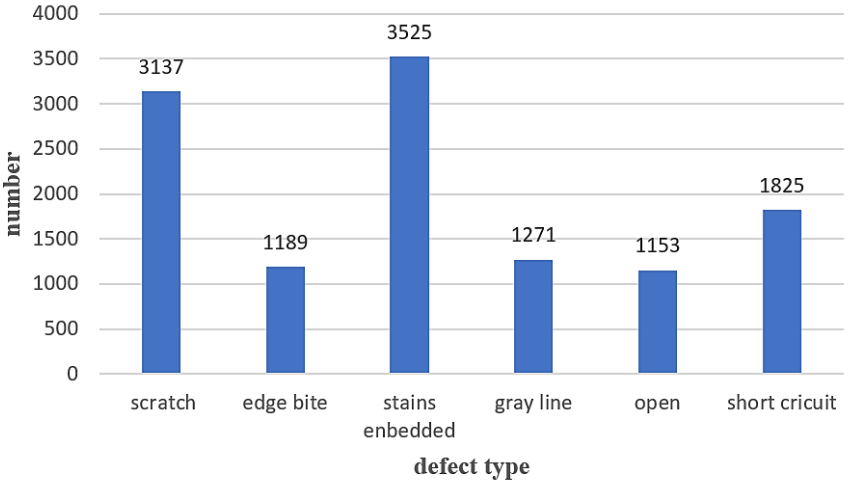


Figure 3. Distribution chart of the quantities of six types of wafer surface defects.

3.2. CC-De-YOLO

In order to improve the accuracy of surface defect detection on semiconductor wafers, this paper proposes a new detection network called CC-De-YOLO based on YOLOv7 [22]. As shown in Figure 4, the network is primarily composed of three parts: the Backbone (the feature extraction structure), the Neck (the feature enhancement structure), and the Detection Head. YOLOv7 as the baseline network, Coordinate Attention [8] which focuses on weighting different positions of input data to better capture key information in feature space and enhance feature extraction ability. The CAR-EVC module adopts a context-based sampling method to perform upsampling for strengthening semantic information and uses a feature extractor and a self-learning visual center mechanism to expand the receptive field, enhancing sensitivity to long-distance features and overcoming the problem of unclear feature information caused by nearest-neighbor upsampling. In order to effectively classify mixed-type wafer defects, an Efficient Decoupled Head with auxiliary head and implicit knowledge learning is introduced, which effectively increases the accuracy of localization and classification.

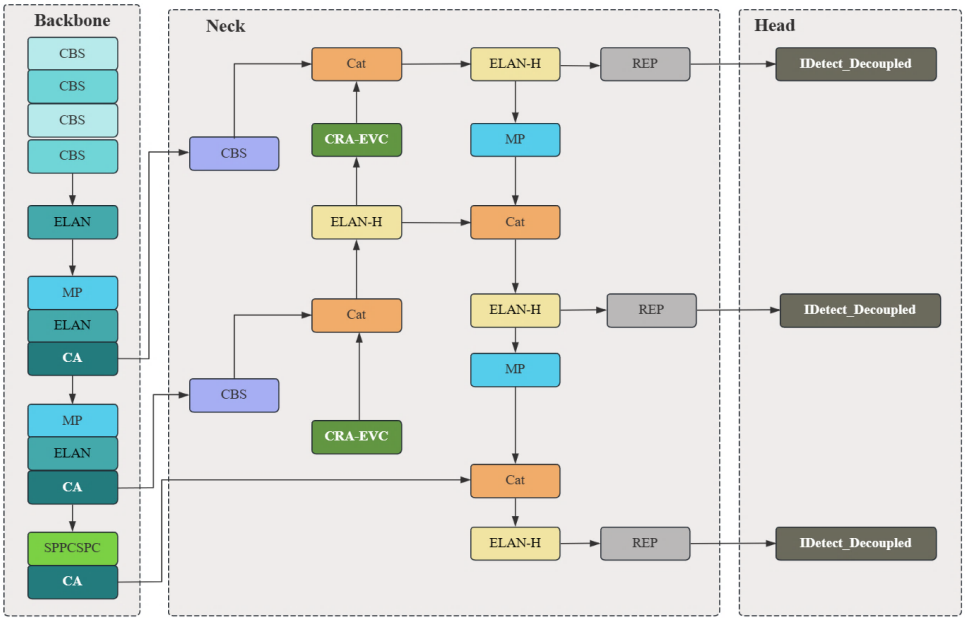


Figure 4. CC-De-YOLO network structure.

3.2.1. YOLOv7 Object Detection Network

YOLOv7 is the basic model of the YOLO series, with a detection approach similar to YOLOv5, but with improved optimization in the model architecture. As shown in Figure 5, this network mainly consists of three parts: Backbone, Neck, and Head. The input module scales the input image to a unified pixel size to meet the input size requirements of the backbone network. The backbone module consists of CBS convolutional layers, ELAN convolutional layers, and MP convolutional layers, where CBS is composed of convolutional layers, batch normalization (BN) layers, and SiLU activation functions [3], which are used to extract features from images of different scales. To better learn the features extracted by the backbone network, YOLOv7 uses a path aggregation feature pyramid network (PAFPN) [14] structure in the feature fusion region, which can separately learn and concentrate different granularity features, maximizing the learning effectiveness of image features. Additionally, YOLOv7 introduces REPVGG [2] reparameterization technology, which adjusts the channel numbers of different scale feature maps output by PAFPN, and outputs prediction confidence, category, and anchor frames through 1×1 convolution. Unlike previous versions, YOLOv7 incorporates an Auxhead, which uses intermediate layers in the network for loss calculation during training to improve performance. During inference, the auxiliary head can be removed to achieve faster inference speed. Because real-time and accurate detection of surface defects on wafers are required, we chose YOLOv7 as the baseline model to achieve a good balance between accuracy and speed.

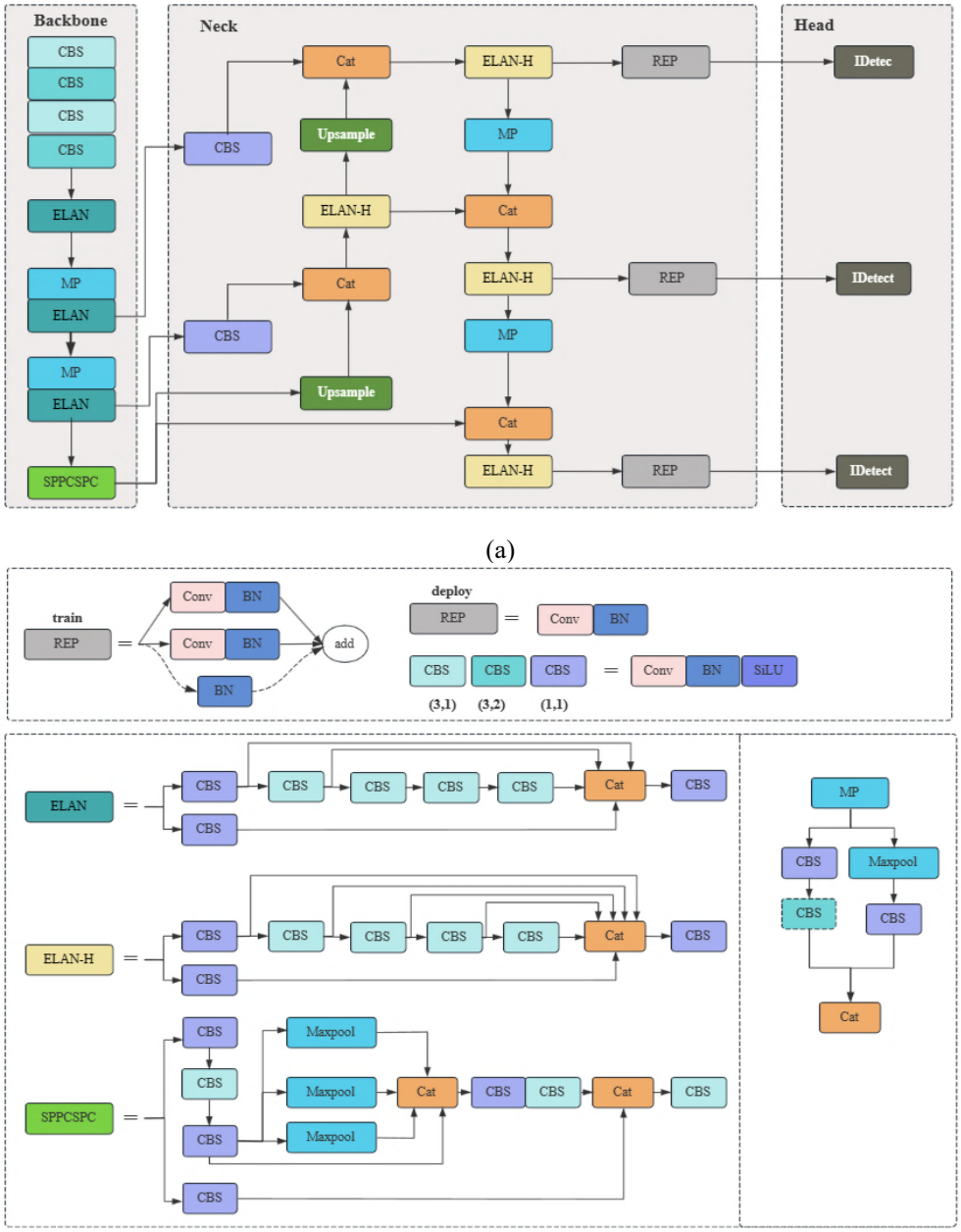


Figure 5. (a) The structure of YOLOv7 network, (b) YOLOv7 network structure components

3.2.2. Coordinate Attention

The attention mechanism was originally proposed in the context of human visual research, to simulate the phenomenon whereby humans selectively attend to certain visible information while ignoring others, in order to more efficiently utilize visual processing resources. In the field of deep learning, attention mechanisms have been introduced to address the problem of information redundancy, primarily by selecting a subset of input information or assigning different weights to different parts of the input information.

There are many excellent methods in computer vision that utilize attention mechanisms. For example, in 2017, Jie Hu et al. [9] proposed the Squeeze-and-Excitation（SE） channel attention method, which adaptively adjusts the weights of each channel to focus more on important features and improve model accuracy. However, this method cannot focus on spatial information. In 2018, Woo S et al. [25] proposed CBAM (Convolutional Block Attention Module), which combines spatial attention and channel attention to create a sequential attention structure from channel to space. This allows the network to focus more on pixel regions that are critical to the target object. However, due to the use of global pooling to map position relationships, this method is sensitive to local information and cannot capture large-scale dependency information. In 2021, Qibin Hou et al. proposed the Coordinate Attention (CA) method, which creatively incorporates positional coordinate information into channel attention. By embedding positional information into channel attention and aggregating features along two spatial directions, CA generates direction-sensitive and position-sensitive attention maps to enhance the representation of objects of interest. This method is simple and easy to use, and almost does not increase computational burden. It can be flexibly inserted into classical mobile networks.

In order to enhance the accuracy of object detection and localization, we applied the Coordinate Attention (CA) mechanism in the backbone network of YOLOv7. CA was separately applied to the three different sized feature output layers after the ELAN and SPPCSPC modules, and position information was incorporated. This enables the capture of remote dependencies between different channels in one feature space direction while preserving accurate position information in another feature space direction. Through this process, we addressed the insensitivity of the original network to long-range feature information, mitigating classification biases and effectively improving the network's feature extraction capabilities and localization precision.

The feature maps processed by the backbone network are fed into the CA module for further processing. The role of the CA module is to decompose the attention channels into one-dimensional features that encode channel relationships and long-range relationships. This process consists of two steps: envelope coordinate embedding and coordinate attention generation. The Coordinate Attention structure is shown in Figure 6. In the envelope coordinate embedding stage, given the input x, we use pooling kernels of size $(h, 1)$ and $(1, w)$ to perform global average pooling operations on the input feature map along the width and height directions, respectively. Specifically, we obtain the feature maps in the height and width directions by computing Equations (1) and (2).

$$Z_C^h(h) = \frac{1}{W} \sum_{0 \le i \le w} x_c(h, i) \tag{1}$$

$$Z_C^w(w) = \frac{1}{H} \sum_{0 \le i \le w} x_c(j, w) \tag{2}$$

During the coordinate generation phase, aggregation features are generated along two directions and concatenated using the transformation function F_1. The calculation formula is shown in equation (3), where $f \in c/r \times (h + w)$,c stands for dimension,δ denotes the non-linear activation function, and $[Z^h, Z^w]$ represents the concatenation operation along the spatial dimension.

$$f = \delta(F_1([Z^h, Z^w])) \tag{3}$$

Subsequently, a 1×1 convolution operation and Sigmoid function are applied to the features, followed by processing in the width and height directions. This process divides the feature tensor f into two independent tensors: $f^h \in c/r \times h$ and $f^w \in C/r \times w$. Next, g^h and g^w are expanded into attention weights, which are calculated using equations (4) and (5).

$$g^h = \sigma(F_h([f^h])) \tag{4}$$

$$g^w = \sigma(F_w([f^w])) \tag{5}$$

Finally, the input feature map $x_C(i, j)$ of the CA module is multiplied by the attention weights of the feature map in height and width to obtain the final feature map. The calculation formula is shown as equation (6).

$$y_C(i, j) = x_C(i, j) \times g^h(j) \times g^w(i) \tag{6}$$

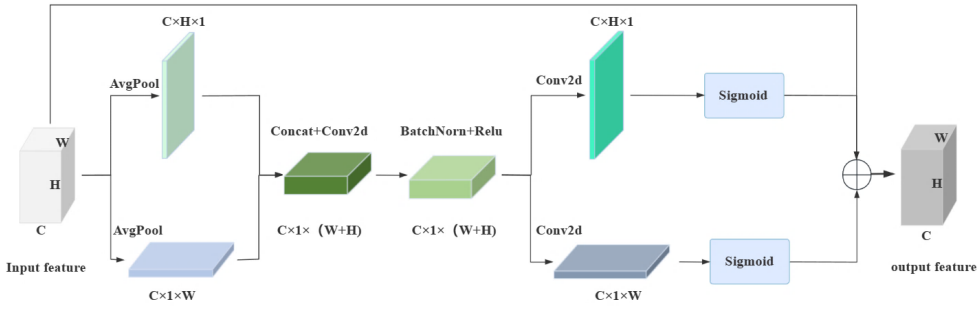


Figure 6. The structure of coordinate attention.

3.2.3. CAR-EVC

In the original YOLOv7, nearest-neighbor upsampling was utilized in the feature fusion network. This method only determines the upsampling core based on spatial position of pixel points, without utilizing semantic information within the feature map. Therefore, after resampling, the grayscale values are discontinuous and the perceptual field is small, resulting in unclear feature information for targets. To address this issue, this study introduces the CAR-EVC module to replace nearest-neighbor upsampling. As shown in Figure 7, the CAR-EVC structure consists of a lightweight upsampling operator CARAFE [23] and an EVCBlock [16] module. First, the input feature is associated with semantic information

using CARAFE to increase the receptive field. Then, the EVCBlock captures global long-term dependencies of the top-level features to obtain key feature information after upsampling. Finally, a learnable visual center mechanism is used to aggregate local key regions. The CAR-EVC module effectively addresses the distortion of targets caused by changes in feature size, making the detection model more refined and expanding the detection range.

The CARAFE model comprises two main modules: the upsample kernel prediction module and the feature recombination module. As shown in Figure 8, given an input feature map with shape $H \times W \times C$, the CARAFE model first a 1×1 convolution is applied to compress the channel number to $(H \times W \times C_m)$. Then, during the upsampling process,. During the upsampling process, in order to obtain a larger receptive field and more precise output feature map, different upsampling kernels are applied to each position. This requires predicting the shape of the upsampling kernel at each position. Specifically, a $k \times k$ convolutional kernel is applied to the compressed feature map, with the input channel number C_m and the output channel number $\sigma_2 \times k^2$. Following the expansion along the channel dimension, a upsampling kernel with dimensions of size $\sigma H \times \sigma W \times k^2$ is obtained. The upsampling kernel is then normalized using softmax. Finally, each position in the output feature map is mapped back to the corresponding position in the input feature map to complete the feature recombination process.

The EVC module mainly consists of two parallel-connected sub-modules: the Lightweight MLP and the Learnable Visual Center(LVC). As shown in Figure 6, each MLP is implemented by two residual structures composed of a deep convolution module and a channel MLP module, with channel scaling and DropPath operations applied. The calculation formulas for the output $\tilde{X}_i n$ of the two residual modules and $MLP(X_i n)$are shown in equations (7) and (8).

$$\tilde{X}_{in} = DConv(GN(X_{in})) + X_{in} \tag{7}$$

$$MLP(X_i n) = CMLP(GN(\tilde{X}_{in})) + \tilde{X}_{in} \tag{8}$$

In the equations, X_{in}represents the input, GN refers to Group Normalization, $DConv$ denotes the deep convolution module, and $CMLP$ represents the channel MLP module.

LVC is an encoder with an inherent dictionary. The output features X_{in} from the Stem module are first encoded through a combination of 1×1 convolution, 3×3 convolution, and 1×1 convolution. Then, they are processed by the CBR block, which consists of 3×3 convolution and ReLU activation. Furthermore, we employ a scaling factor $S = \{s_1, s_2, \cdots, s_k\}$to map the position information of $X_i$and the eigen-codebook $B = \{b_1, b_2, \cdots, b_k\}$. where the computation formula for the information of the whole image regarding the kth codeword is shown in 9.

$$e_k = \sum_{i=1}^{N} \frac{e^{-s_k \| \check{x}_i - b_k \|^2}}{\sum_{j=1}^{K} e^{-s_k \| \check{x}_i - b_k \|^2}} (\check{x}_i - b_k) \tag{9}$$

Then, we use φ (including BN, ReLU, and Average Layers) to fuse e_k, as shown in equation 10.

$$e = \sum_{k=1}^{K} \varphi(e_k) \tag{10}$$

After the fusion process, the output obtained through fully connected layers and 1×1 convolutions is channel-wise multiplied with the input feature X_{in} of the Stem block. The calculation formula is shown as 11. Finally, channel fusion is performed using the calculation formula shown as 12.

$$Z = X_{in} \oplus (\delta(Conv_{1\times1}(e))) \tag{11}$$

$$LVC(X_{in}) = X_{in} \oplus Z \tag{12}$$

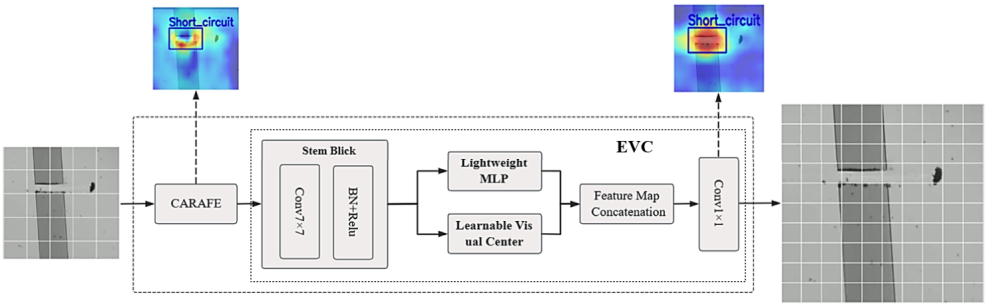


Figure 7. The structure of CAR-EVC

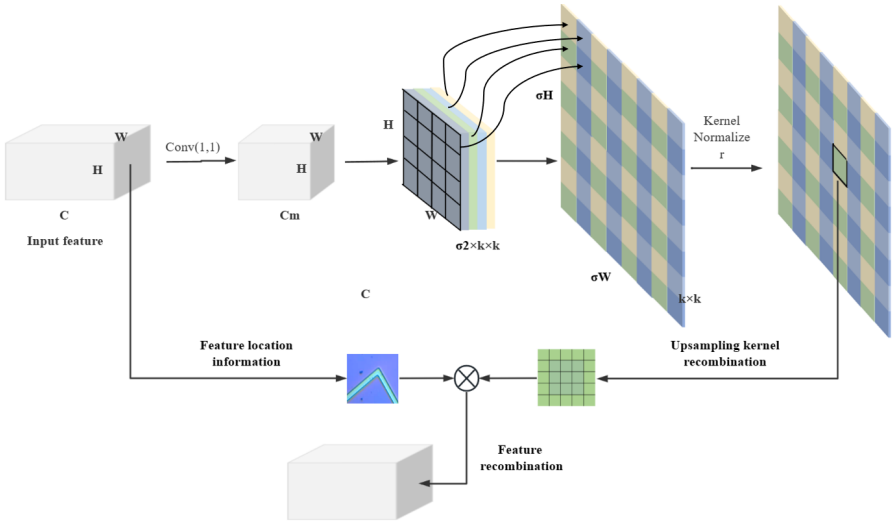
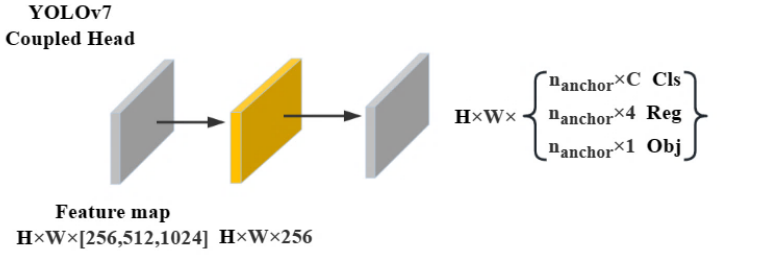


Figure 8. The structure of CARAFE.

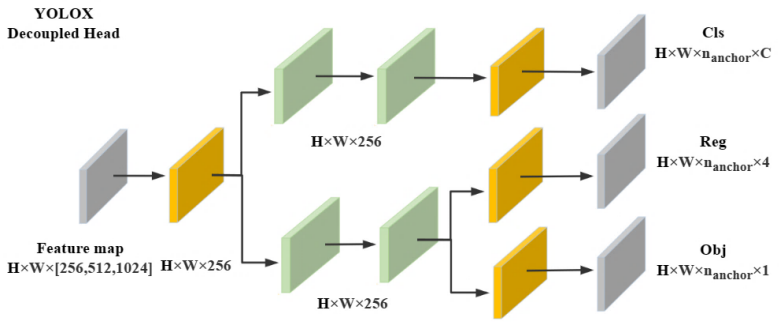
3.2.4. IDetect_Decoupled

Accurate classification and detection of mixed-type wafer defects is crucial for surface defect inspection, but the original YOLOv7 algorithm utilizes a unified detection head to

perform tasks such as classification, localization, and prediction of confidence, as shown in Figure 9(a). This coupled design may lead to interference and conflicts between classification and localization tasks. To address this issue, we adopt the Efficient Decoupled Head from YOLOv6[13] to separate classification and regression tasks, reducing interference between them. This approach helps the model better learn the features of the target and improves detection accuracy. The Efficient Decoupled Head uses implicit knowledge learning to improve model performance. Meanwhile, on the basis of balancing the performance of relevant operators and hardware computing costs, the Hybrid Channels strategy is used to redesign the decoupled head structure. Specifically, a higher efficiency is achieved by jointly scaling the width multiplier of the backbone and neck. The YOLOv6[4] decoupled head is shown in Figure 9(c), which first undergoes a 1×1 convolution, followed by two parallel 3×3 convolutional layers, with one used to predict the target category and the other used to predict the target position and confidence. The resulting tensor obtained through 1×1 convolution has a shape of $(H \times W \times n_{anchor} \times C)$, while the tensor obtained through 1×1 convolution for predicting the target position and confidence has a shape of $(H \times W \times n_{anchor} \times 4)$ and $(H \times W \times n_{anchor} \times 1)$, respectively, where n_{anchor} represents the three anchors for each target object in YOLOv7 and C represents six types of wafer surface defects. Finally, prediction boxes are filtered out that meet the confidence threshold and non-maximum suppression is performed to determine the final results. Compared to the YOLOX decoupled head[4], the YOLOv6 decoupled head reduces the number of 3×3 convolutions in the head to one, as shown in Figure 9(b). This improves the efficiency of the model while maintaining its accuracy. Additionally, this also alleviates the extra delay caused by the two 3×3 convolutions in the YOLOX decoupled head, further improving the classification speed of mixed-type wafer defects.



(a)



(b)

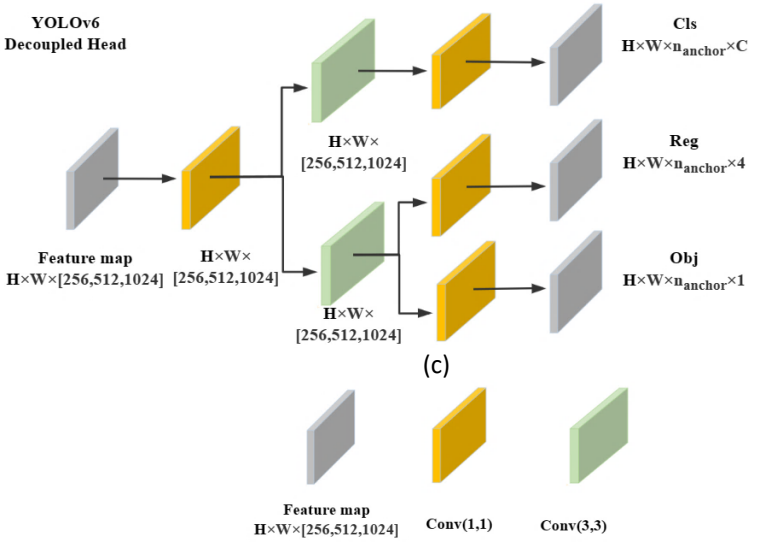


Figure 9. (a) The structure of YOLOv7 Coupled Head. (b) The structure of YOLOX Decoupled Head. (c) The structure of YOLOv6Decoupled Head

4. Results and Experiments

4.1. Analysis of Ablation Experiments

In this study, we adopted Coordinate Attention, CAR-EVC, and IDetect_Decoupled to enhance the detection performance of YOLOv7. We conducted ablation experiments to demonstrate the effectiveness of these methods. " √ " indicates the usage of specific modules, and the experimental results are shown in Table 1. The results indicate that each improvement module of CC-De-YOLO enhances the algorithm's performance. Compared with the original YOLOv7, adding Coordinate Attention in Solution 1 improved mAP@.5 by 1.2%, mAP@.5:.95 by 0.4%, and Precision by 4%. By introducing coordinate attention, the model can perceive the image more comprehensively, aiding in capturing objects of various scales and features. In Solution 2, replacing the upsampling operation with the CAR-EVC module, which incorporates semantic information into the feature extraction module for better understanding of long-range dependencies, improved mAP@.5 by 1.4%, mAP@.5:.95 by 0.6%, and Precision by 4.9%. In Solution 3, adding the IDetect_Decoupled decoupling head to the network, with its implicitly learned decoupling mechanism handling classification and localization separately, significantly increased the efficiency of object detection. This resulted in an increase in mAP@.5 by 1%, mAP@.5:.95 by 0.2%, and Precision by 8.6%. In Solutions 4 and 5, we added the CAR-EVC module and the IDetect_Decoupled decoupling head on top of the Coordinate Attention module, respectively, leading to various degrees of improvements in mAP, Precision, and Recall. Finally, in Solu-

tion 6, incorporating all three improvement modules resulted in an increase in mAP@.5 by 4.8%, mAP@.5:.95 by 2.9%, Precision by 7.7%, and Recall by 0.6%.

Table 1. Experimental results of CC-De-YOLO algorithm ablation

Item	Coordinate Attention	CAR-EVC	IDetect_ Decoupled	mAP@0.5	mAP@.5:.95	Precision	Recall
0				86.2%	43.3%	81.8%	82.6%
1	√			87.4%	43.7%	85.8%	82.3%
2		√		87.6%	43.9%	86.7%	82.1%
3			√	87.2%	43.5%	90.4%	80.4%
4	√	√		88.8%	44.8%	87.7%	82.7%
5	√		√	88.1%	44.4%	87.3%	82.9%
6	√	√	√	**91.0%**	**46.2%**	**89.5%**	**83.2%**

4.2. Algorithm Performance Analysis

According to the comparison results shown in Figure 10, the single-class average precision of the CC-De-YOLO algorithm is superior to that of the YOLOv7 algorithm, indicating that adopting a multi-scale method can improve the detection accuracy and reducing the rate of missed detections. This experiment further proves the effectiveness of the CC-De-YOLO algorithm in surface defect detection on wafers. Furthermore, as shown in Figures 11 and 12, we present the confusion matrices obtained by the YOLOv7 and CC-De-YOLO algorithms on the wafer surface defect dataset. Compared with the YOLOv7 algorithm, the CC-De-YOLO algorithm has smaller FP and FN values in the confusion matrix, which indicates that it has stronger classification and localization capabilities, and lower missed detection and false alarm rates. In addition, by comparing the PR curves of the CC-De-YOLO and YOLOv7 algorithms, as shown in Figure 13, we found that the CC-De-YOLO algorithm has higher precision and accuracy at the same recall rate, as evidenced by its PR curve clearly enclosing that of YOLOv7. In summary, the CC-De-YOLO algorithm outperforms in the detection of surface defects on wafers. Specifically, the added Coordinate Attention focuses on weighting different positions of input data to capture crucial information in the feature space and enhance feature extraction. The CAR-EVC module utilizes context-based sampling for upsampling, expanding the receptive field, improving sensitivity to distant features, and addressing the issue of unclear feature information caused by nearest-neighbor upsampling. The introduction of an Efficient Decoupled Head with implicit knowledge learning enhances localization and classification accuracy, further improving the recognition capability of mixed-type surface defects on wafers.

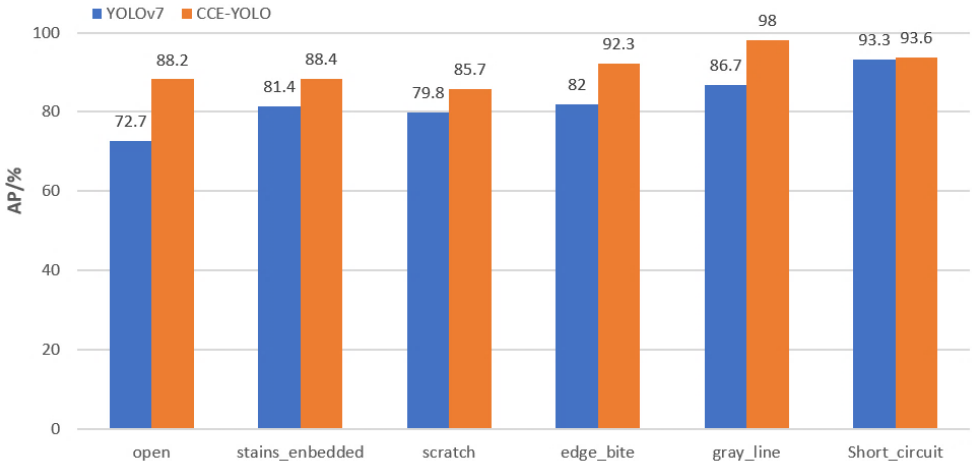


Figure 10. Single-class average precision comparison.

ID	Category
0	open
1	Stains embedded
2	scratch
3	edge bite
4	gray line
5	Short circuit

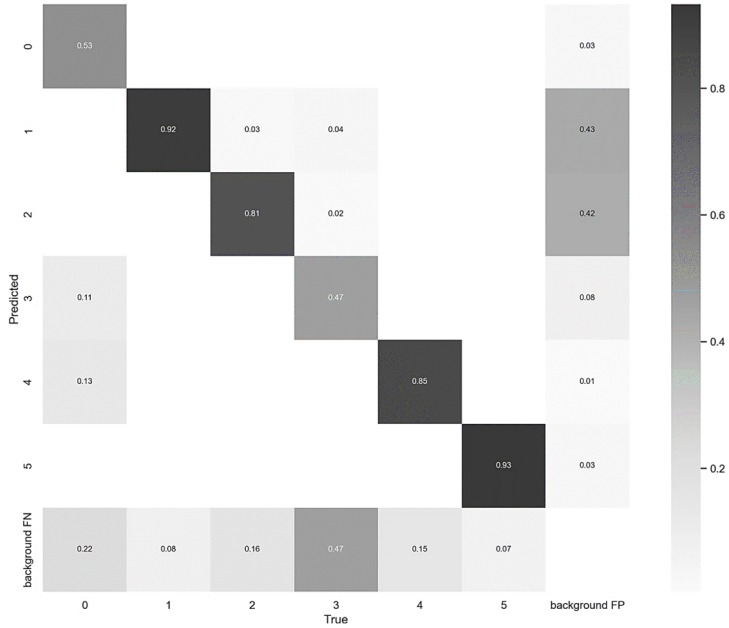


Figure 11. Confusion matrix for YOLOv7 network model.

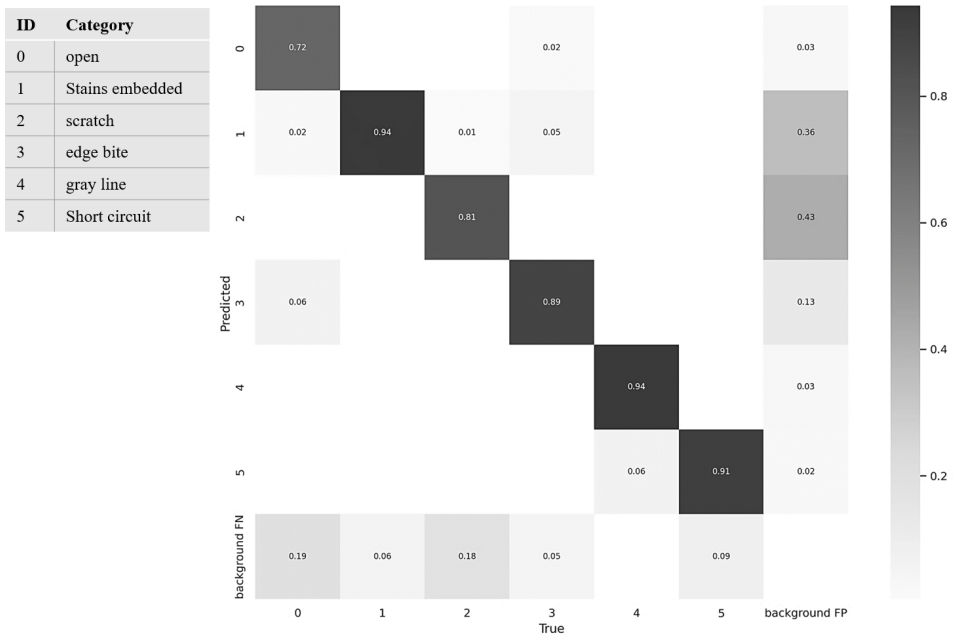


Figure 12. Confusion matrix for CC-De-YOLO network model.

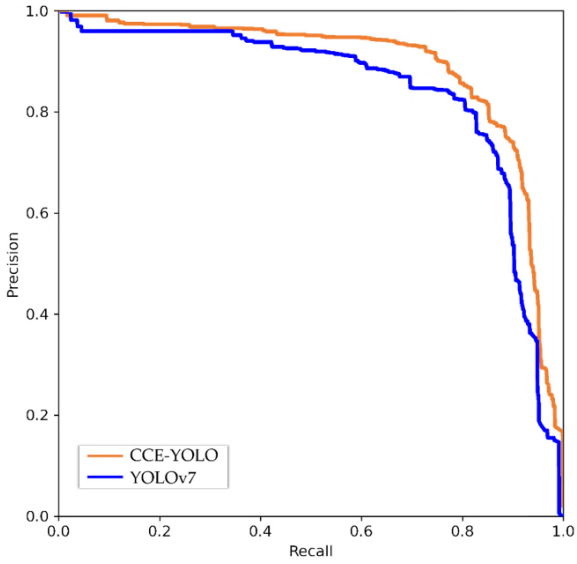


Figure 13. PR curve comparison.

To compare the focus of the CC-De-YOLO and YOLOv7 models on targets, we conducted GradCAM[19] heatmap visualization analysis. This analysis method calculates the gradient weights of each position on the feature map of each convolutional layer and multi-

plies these weights with the corresponding feature maps to obtain the GradCAM value at each position. By overlaying the GradCAM values onto the input image, we can obtain a color-coded heatmap, where darker colors indicate that the network is more focused on that region. In our study, we found that the CC-De-YOLO model performs better than the YOLOv7 model in terms of focusing on defect targets. In addition, as shown in Figure 14, the CC-De-YOLO model exhibits a more concentrated representation of defect regions, as compared to YOLOv7's relatively sparse representation. This is attributed to CC-De-YOLO's ability to extract low-level features such as texture and edges, thereby enhancing the network's sensitivity to defect contours and enabling it to focus more on defect regions. This also explains why CC-De-YOLO achieves higher prediction accuracy.

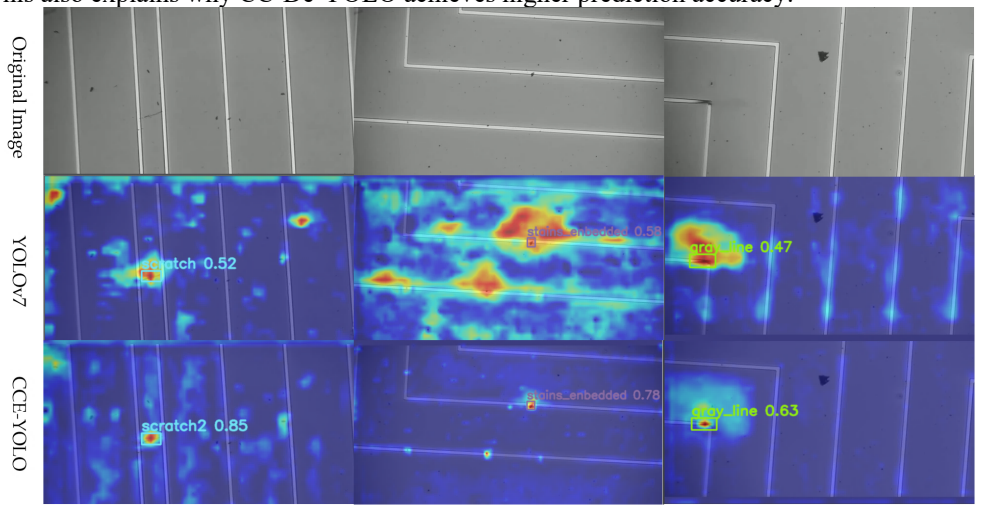


Figure 14. Model GradCAM heatmap visualization analysis

According to the detection results shown in Figure 15, the CC-De-YOLO model exhibits no missed detections and higher confidence levels. In particular, this model can detect overlapping and intersecting defects, which is mainly due to the powerful feature extraction capability introduced by the attention mechanism in the CC-De-YOLO algorithm, as well as the CAR-EVC module that enhances the network's focus on defect targets, thereby effectively improving the network's learning ability and robustness. Additionally, the IDetect_Decoupled decoupling head separately processes classification and localization tasks, avoiding mutual interference between the two tasks and improving the model's generalization ability and robustness. As each task has its own convolutional layer, the network can better learn the feature representations of each task, thereby improving the accuracy of prediction results. The IDetect_Decoupled decoupling head significantly improves the performance of object detection networks, especially in handling mixed-type defects, while also accurately detecting defects with defects with significant differences in size.

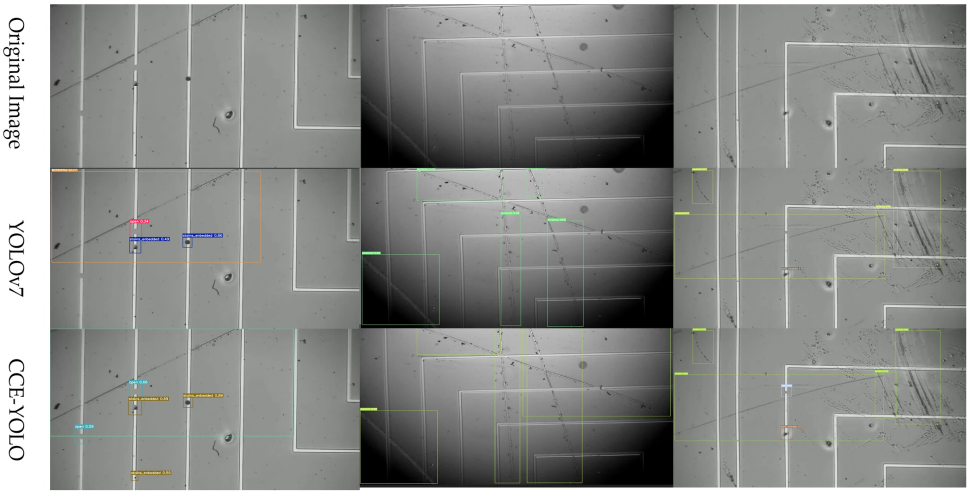


Figure 15. Comparison Experiment Results between CC-De-YOLO and YOLOv7

Wang J et al. proposed a wafer map detection model, DC-Net, which combines variability convolution and a multi-label output layer with one-hot encoding to enhance the recognition capability of mixed-type wafer map defects. The model achieved significant advantages in comparative experiments. In this study, the MixedWM38 wafer defect pattern dataset was utilized to train the model under the same experimental conditions. Performance evaluation of the CC-De-YOLO model and a comparison with the DC-Net model, reproduced from the original paper, were conducted using evaluation metrics Ap and mAp. Random sampling was performed in each training iteration, with 80% of the dataset used for training and 20% for validation.

As shown in Table 2, benefiting from CC-De-YOLO's powerful attention-based feature extraction module and a classification decoupling head with implicit knowledge learning, CC-De-YOLO achieved an mAp of 94.05%, significantly higher than DC-Net's 90.99%. As illustrated in Figure 16, for the Single type defect pattern (C1-C9), only two defect patterns had Ap lower than DC-Net. The mAp reached 97.28%, surpassing DC-Net by 1.24%. As illustrated in Figure 17, regarding the 2 mixed-type defect pattern (C10-C22), only one defect pattern had Ap lower than DC-Net, with an mAp of 96.94%, exceeding DC-Net by 1.18%. For the multi-defect patterns 3 mixed-type(C23-C34) and 4 mixed-type(C35-C38), CC-De-YOLO outperformed the DC-Net model, confirming the robust feature extraction capability from the attention mechanism and context-aware upsampling of CC-De-YOLO, combined with an efficient classification decoupling head, which is superior in handling mixed defect patterns compared to the variability convolution network DC-Net. As illustrated in Figure 18, the mAp for 3 mixed-type reached 92.04%, 3.2% higher than DC-Net. As illustrated in Figure 19, for 4 mixed-type, the mAp reached 89.95%, 3.06% higher than DC-Net. Each wafer defect pattern in these two categories had an Ap higher than DC-Net.

Table 2. Experimental results of CC-De-YOLO algorithm ablation

Defect Mode Classification	CC-De-YOLO	DC-Net
	mAp	
Single type	97.28%	96.04%
2 mixed-type	96.94%	95.13%
3 mixed-type	92.04%	88.84%
4 mixed-type	89.95%	83.95%
Average	94.05%	90.99%

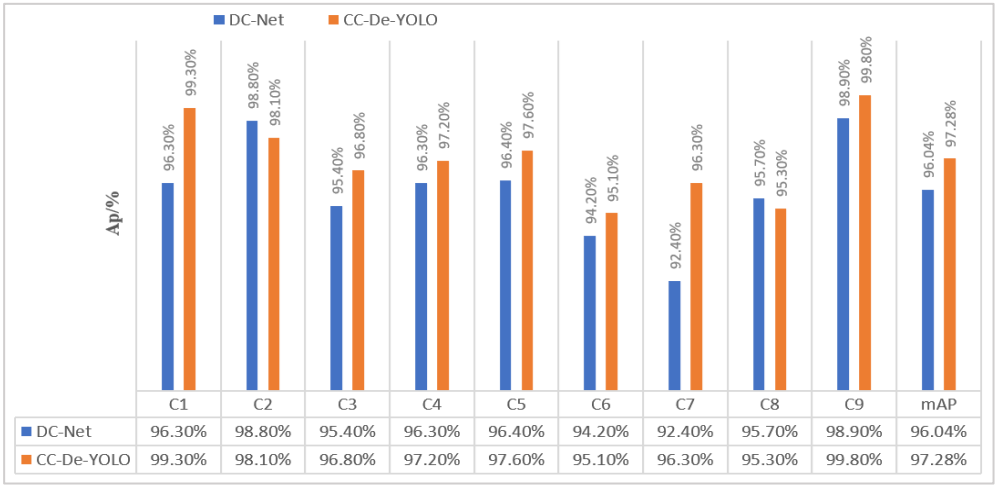


Figure 16. The average precision of Single type defects of the DC-Net, CC-De-YOLO

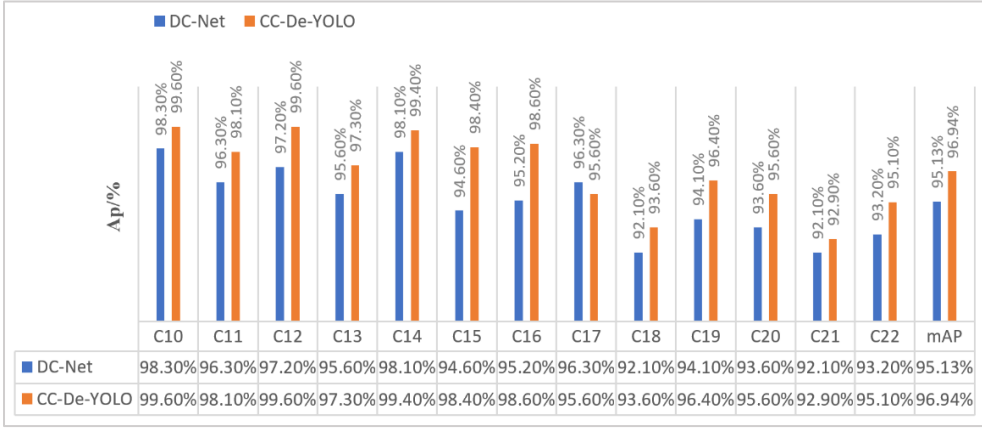


Figure 17. The average precision of 2 mixed-type defects of the DC-Net, CC-De-YOLO

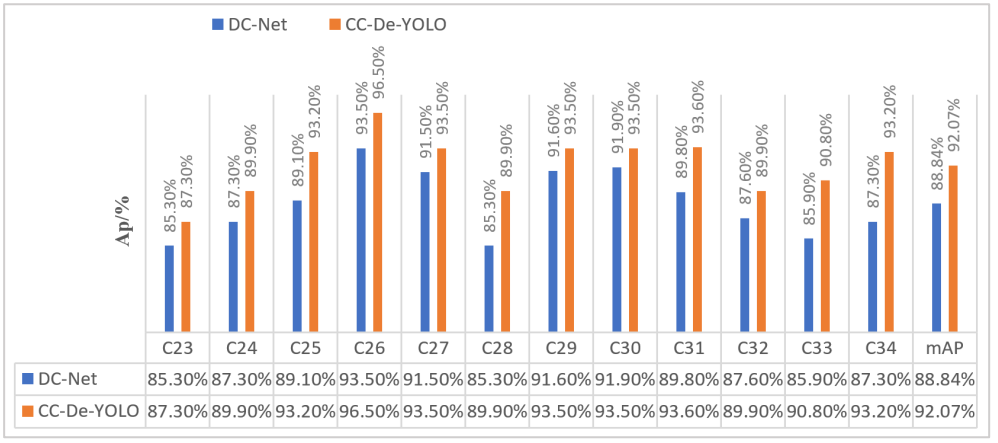


Figure 18. The average precision of 3 mixed-type defects of the DC-Net, CC-De-YOLO

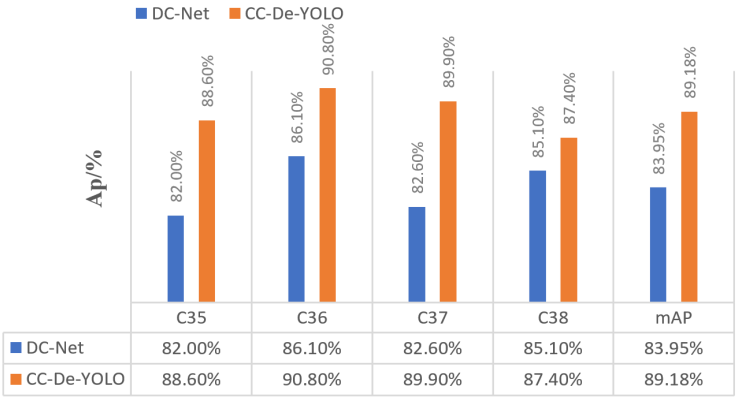


Figure 19. The average precision of 4 mixed-type defects of the DC-Net, CC-De-YOLO

4.3. Comparison of Different Object Detection Algorithms

To validate the effectiveness of our proposed method for detecting surface defects on wafers, we compared it with Faster R-CNN, YOLOv5m, YOLOX, DC-Net and YOLOv7 on the entire test dataset. To compare the results of each model fairly, we kept the testing environment and dataset consistent during the testing process, and set the input size of all images to 640*640. Table 2 shows the performance comparison results of different models in terms of precision, recall, mAP@0.5:95, and mAP@0.5. The experimental results demonstrate that the CC-De-YOLO model achieves the highest performance, with an mAP@0.5

of 91%, which is higher than that of Faster R-CNN (79.2%), YOLOv5m (83.4%), YOLOX (85.6%), DC-Net (85.9%) and YOLOv7 (86.2%). Additionally, CC-De-YOLO also achieves the highest mAP@0.5:95 of 46.2%, which is higher than that of Faster R-CNN (35.3%), YOLOv5m (39.7%), YOLOX (41.6%), DC-Net (40.2%)and YOLOv7 (43.3%). In terms of precision and recall, CC-De-YOLO outperforms all other models, achieving a precision of 89.5% and a recall of 83.2%. These results indicate that the Coordinate Attention and CAR-EVC modules added to CC-De-YOLO enhance the network's feature representation capability, while the improved IDetect_Decoupled module improves the target classification and localization capability. Overall, CC-De-YOLO demonstrates outstanding performance, achieving the highest mAP and precision while maintaining a high recall rate.

Table2. Experimental results of Comparison of different object detection algorithms

ALgorithms	mAP@0.5(%)	mAP@.5:.95(%)	Precision(%)	Recall(%)
FasterR-CNN	79.2%	35.3%	71.8%	72.6%
YOLOv5m	83.4%	39.7%	75.8%	74.3%
YOLOX	85.6%	41.6%	80.3%	78.1%
DC-Net	85.9%	40.2%	83.8%	80.3%
YOLOv7	86.2%	43.3%	81.8%	82.6%
CC-De-YOLO (our)	**91.0%**	**46.2%**	**89.5%**	**83.2%**

5. Conclusions

To address the low detection accuracy and missed detections caused by overlapping and intersecting mixed-type defects on wafer surfaces, we propose a multi-scale detection CC-De-YOLO model based on the YOLOv7 model. This model uses Coordinate Attention to enhance feature extraction and defect localization capabilities. The proposed CAR-EVC upsampling module captures long-range semantic dependencies, effectively enhancing the network's focus on defect targets. Finally, we introduce the Efficient decoupling head from YOLOv6 into YOLOv7, which improves classification and localization accuracy for stacked and intersecting defects. Additionally, we established a wafer surface defect dataset that includes six typical defect categories: edge_bite, gray_line, open, scratch, short_circuit, and stains_enbedded. Experimental results show that the CC-De-YOLO model achieves better performance than the original YOLOv7 model and other comparison models, with an mAP@0.5 of 91.0%, an mAP@0.5:.95 of 46.2%, a precision of 89.5%, and a recall of 83.2%. These results meet the accuracy requirements for wafer surface defect detection. In future work, we plan to further optimize the proposed method in terms of network structure and model parameters to meet the requirements of lightweight deployment.

References

[1] Chen S. H., Kang C. H., Perng D. B.: Detecting and measuring defects in wafer die using gan and yolov3. *Applied Sciences*, **10**, 23, 2020, 8725.

[2] Ding, X.H., Zhang, X.Y., Man, N.N., Han, J.G., Ding, G.G., Sun, J: RepVGG: Making VGG-style ConvNets great again. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville*, TN, USA, 20–25 June 2021;

[3] Elfwing S., Uchibe E., Doya K.: Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks,* **107**, 2018, 3-11.

[4] Ge Z., Liu S., Wang F., et al.: Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[5] Han H., Gao C., Zhao Y., et al.: Polycrystalline silicon wafer defect segmentation based on deep convolutional neural networks. *Pattern Recognition Letters*, 130, 2020, 234-241.

[6] He K., Gkioxari G., Dollár P., et al.: Mask r-cnn, *Proceedings of the IEEE international conference on computer vision*. 2017: 2961-2969.

[7] Ho C. M., Tai Y. C.: Micro-electro-mechanical-systems (MEMS) and fluid flows. *Annual review of fluid mechanics*, 30, 1, 1998, 579-612.

[8] Hou Q., Zhou D., Feng J.: Coordinate attention for efficient mobile network design, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2021, 13713-13722.

[9] Hu J., Shen L., Sun G.: Squeeze-and-excitation networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 7132-7141.

[10] Jin C. H., Na H. J., Piao M., et al.: A novel DBSCAN-based defect pattern detection and classification framework for wafer bin map. *IEEE Transactions on Semiconductor Manufacturing*, **32**, 3, 2019, 286-292.

[11] Kim Y., Cho D., Lee J. H.: Wafer map classifier using deep learning for detecting out-of-distribution failure patterns, *2020 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA)*. IEEE, 2020, 1-5.

[12] Kong Y., Ni D.: A semi-supervised and incremental modeling framework for wafer map classification. *IEEE Transactions on Semiconductor Manufacturing*, **33**, 2020, 62-71.

[13] Li C., Li L., Jiang H., et al.: YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976,* 2022.

[14] Liu S., Qi L., Qin H., et al.: Path aggregation network for instance segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018, 8759-8768.

[15] Liu W., Anguelov D., Erhan D., et al.: Ssd: Single shot multibox detector, *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands, October 11–14.

[16] Lou A., Loew M.: Cfpnet: channel-wise feature pyramid for real-time semantic segmentation, *2021 IEEE International Conference on Image Processing (ICIP)*. *IEEE,* 2021, 1894-1898.

[17] Miyajima H., Mehregany M.: High-aspect-ratio photolithography for MEMS applications. *Journal of microelectromechanical systems*, **4**, 4, 1995, 220-229.

[18] Ren S., He K., Girshick R., et al.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 2015, 28.

[19] Selvaraju R. R., Cogswell M., Das A., et al.: Grad-cam: Visual explanations from deep networks via gradient-based localiza tion, *Proceedings of the IEEE international conference on computer vision*. 2017, 618-626.

[20] Shinde P. P., Pai P. P,: Adiga S P. Wafer defect localization and classification using deep learning techniques. *IEEE Access*, 10, 2022, 39969-39974.

[21] Shorten C., Khoshgoftaar T. M.: A survey on image data augmentation for deep learning. *Journal of big data*, **6**, 1, 2019, 1-48.

[22] Wang C. Y., Bochkovskiy A., Liao H. Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023, 7464-7475.

[23] Wang J., Chen K., Xu R., et al.: Carafe: Content-aware reassembly of features, *Proceedings of the IEEE/CVF international conference on computer vision.* 2019, 3007-3016.

[24] Wang J., Xu C., Yang Z., et al.: Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition. *IEEE Transactions on Semiconductor Manufacturing*, **33**, 4, 2020, 587-596.

[25] Woo S., Park J., Lee J. Y., et al.: Cbam: Convolutional block attention module, *Proceedings of the European conference on computer vision (ECCV)*. 2018, 3-19.

[26] Xie L., Huang R., Gu N., Cao Z.: A novel defect detection and identification method in optical inspection, Neural *Computing and Applications* ,2013, 1–10.

[27] Xifeng L.: Image registration-based wafer surface defect detection (Chinese). *Instrumentation and analytical monitoring,2020*, 1-4

Received 17.09.2023, Accepted 19.05.2024