Hindawi
Journal of Sensors
Volume 2021, Article ID 2682286, 9 pages
https://doi.org/10.1155/2021/2682286

Research Article

A Fast Postprocessing Algorithm for the Overlapping Problem in Wafer Map Detection

Yang Li and Jianguo Wang

Novel Network and Detection Control Engineering Lab. of National and Local Joint, School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China

Correspondence should be addressed to Jianguo Wang; wangjianguo@xatu.edu.cn

Received 4 August 2021; Accepted 8 November 2021; Published 13 December 2021

Academic Editor: Haibin Lv

Mixed defects have become increasingly popular in defect detection and one of the hottest research areas in wafer maps. Postprocessing methods used to solve the overlapping problem in mass mixed defects have a poor detection speed, which is insufficient for rapid defect detection. In this paper, the fast-soft nonmaximum suppression (fs-NMS) method is proposed to solve this problem. The score of the detection box is updated by optimizing the penalty distribution function. Further, this paper analyzes the performance of the fs-NMS method in wafer defect detection. As a penalty, the logistic function is used, and experiments are conducted using single-stage and two-stage detectors. The final results show that, compared to the soft-NMS, the efficiency for the single-stage and two-stage detectors is increased on average by 9.63% and 21.72%, respectively.

1. Introduction

Defect detection is an important application of object detection that has received a lot of attention. For the semiconductor industry, wafer map defect detection has become a major defect detection problem. The semiconductor manufacturing process involves tens of complex steps, which can lead to defects due to numerous reasons [1, 2]. Visualizing and identifying defect patterns is essential for preventing defect generation. Defect pattern recognition (DPR) provides engineers with a reference for dealing with manufacturing-related problems by identifying wafer surface defects [3]. Currently, with a gradual reduction in wafer size and an increase in the complexity of production processes, the number of mixed complex defects (which combine multiple basic defects) has been increasing. When mixed defects are generated, defect detection becomes more complicated, especially when testing tens of millions of wafer maps in industrial production, which requires a high level of accuracy and speed, both online and offline.

Numerous approaches have been proposed in the literature to tackle the problem of hybrid wafer defect detection in recent years, ranging from manual feature recognition to deep learning networks for automatic feature recognition. Deep convolutional neural networks have demonstrated a very good performance in computer vision [4, 5]. When applied to the field of industrial image detection, deep learning-based object detection methods have proven to be very beneficial since engineers do not have to develop specific defect models and the data-driven approach does not require domain-specific prior knowledge.

Among deep learning-based detectors, two-stage detectors (e.g., R-CNN [6] series) improve accuracy while their efficiency is lacking, and single-stage detectors (e.g., YOLO [7], SSD [8], and RetinaNet [9]) redesign the overall network structure but partly lose precision. In the stage of regression of the candidate box, these approaches are subject to post-processing. The main purpose of postprocessing is to remove redundant candidate boxes. The extracted candidate boxes will produce cluttered detections during the refinement of localization, such as multiple extracted candidate boxes will be regressed to the same region of interest (RoI) in the postprocessing stage. The detector uses a greedy non-maximum suppression (greedy NMS) algorithm to reduce

the number of false-positive boxes. Greedy NMS was presented by Dalal and Triggs [10], and a bounding box with maximum detection fraction is selected and suppresses its neighboring boxes using a predefined intersection over union (IoU) threshold. In detecting complex mixed wafer defects based on deep convolutional neural networks, greedy NMS drastically reduces the screening of false-positive boxes in the postprocessing stage, but mixed defects are still difficult to detect. This causes the detector lose mass positive boxes at a certain threshold, while causes a decrease in the average precision. During mixed defect detection, soft-NMS [11] can eliminate more false-positive boxes, increasing accuracy enormously. In industrial detection, however, detecting thousands of wafer defects is very inefficient and often insufficient.

To solve this problem, this paper proposes an improved fast-soft nonmaximum suppression (fs-NMS) postprocessing algorithm to improve detection efficiency by optimizing the distribution of penalty terms in soft-NMS [11], so as to better apply to large quantities of industrial production. Experiments are performed on some baseline detectors. The results show the effectiveness of object detection in wafer map detection and the efficiency and precision of the postprocessing stage after replacing fs-NMS. It is concluded that our approach is effective for both single-stage and two-stage detectors.

2. Related Work

This section mainly introduces the shortcomings of wafer map detection based on deep learning and general object detection algorithm (Section 2.1) and then expounds the problems existing in the traditional NMS and its improved algorithm (Section 2.2).

2.1. Wafer Map Detection and General Object Detection. *Wafer map detection*. Recently, many studies have attempted to classify wafer maps based on convolutional neural networks (CNNs). Nakazawa and Kulkarni [12] proposed a CNN method for wafer map pattern classification and image retrieval and studied the classification of 22 types of mixed defects. Mixed defects have a large degree of mutual occlusion, and the average recognition accuracy only reached 91%. However, the accuracy used in defect detection is far from adequate. Kyeong and Kim [1] applied CNNs to classify mixed defect patterns of wafer maps and established a separate model for each single defect pattern (whether there is a corresponding model when multiple defect patterns are mixed on a wafer), which contains 16 defect types. On the test set of mixed defects, the detection efficiency of each wafer map is 0.13 s and the accuracy is 98%. However, for the defect detection of a large number of wafer maps, Kyeong's method improves accuracy and reduces detection efficiency.

General object detection. In recent years, object detection is popularized by both two-stage and single-stage detectors. Two-stage detectors divide a detection task into two phases, namely, the extraction RoI phase and the classification and regression phase for RoIs. R-CNN [8] used a selective search method [13] to locate RoIs in the input image and then a classifier to classify them. SPP Net [14], Fast R-CNN [15], and Faster R-CNN [16] are gradually developed. With the emergence of the region proposal network (RPN) [16], the efficiency of the detector has been greatly improved, and the detector can be trained end-to-end. The anchor-based approach is widely used in object detection, and the proposed R-CNN is a milestone. Since then, FPN [17] combined ResNet [18] and ResNeXt [19], which is essential for small object detection, and the performance of small object recognition has been greatly improved, the detection efficiency can reach 5 fps under a single GPU. R-FCN [20] replaces the full-connection layer with a position-sensitive fraction graph, doubling the detection efficiency compared with [16]. Cascade R-CNN [21] explored the cascade architecture of R-CNN and extended it to multistage detectors, which train a series of detectors with increasing IoU thresholds to tackle the problem of overfitting in training and quality mismatch in inference. However, such cascade detectors generate more parameters, resulting in a decrease in detection efficiency. Mask R-CNN [22] added the mask branch based on [16], refined the detection results using multitask learning, and predicted its mask while detecting the bounding box, so that its detection efficiency can still reach 5 fps with a single GPU.

On the other hand, single-stage detectors (such as YOLO [7, 23] and SSD [8]) reduce the stage of RoI extraction and directly predict the bounding box and classification probability with the deep convolutional neural network, which is simpler and faster than the two-stage detector. After the introduction of focal loss [10], its precision is improved. At the same time, it is aimed at solving the problem of a serious imbalance between positive and negative samples, but the overall network detection efficiency of RetinaNet is far inferior to that of the YOLO series and SSD.

2.2. Nonmaximum Suppression. NMS is widely used in computer vision postprocessing algorithms. In the general object detection methods (Section 2.1), manual processing and greedy NMS are still used as postprocessing methods. Recently, soft-NMS [11] proposed an improved NMS, which reduces the score of the adjacent candidate box by adding a penalty rather than discarding the candidate box whose score is lower than the threshold. The algorithm is satisfactory in improving AP, but there are still candidates with high overlap false positives, and the algorithm efficiency is insufficient. Learning NMS [24] designed a complex deep neural network, which requires only box and score as input to implement NMS. Fitness NMS [25] proposed the regression loss of the object box matching IoU maximization, which is combined with [11] to improve precision, and the loss converges well. Adaptive NMS [26] considered the relationship between sparse and dense objects in crowd detection. Increasing the NMS threshold to retain neighboring detection boxes with high overlap based on [11] is an effective solution for crowded scenes, and a module for density prediction is designed for learning density scores. KL loss [27] presented a bounding box regression loss for learning the difference between transformation and location of bounding boxes, estimated the confidence of localization as well as the

location on the baseline, and predicted its complex probability distribution to guide the NMS to retain more accurate localized bounding boxes.

The above postprocessing methods are effective means in general object detection, but as the complexity of the parameters or network structure increases, the inference efficiency will instead reduce. At present, most networks still use greedy NMS as the postprocessing method, which requires a fast postprocessing algorithm to solve the efficiency problem and ensure that accuracy is not lost.

3. Proposed Fast-Soft Nonmaximum Suppression Algorithm

In this section, the proposed wafer map postprocessing algorithm is presented in detail. For the problem faced, the problems of soft-NMS are first analyzed (Section 3.1), then improvement ideas and methods are elaborated (Section 3.2), and finally, the training and inference processes are introduced (Section 3.3).

3.1. Problems with Soft-NMS. In the wafer map detection task, the postprocessing stage is essential, and the detection effect is unsatisfactory because the greedy NMS pruning branch is very strict. As shown in Figure 1, when the object overlaps, the score of Scratch will be insufficient. Although some of the extracted detection boxes cover the parts that are not covered by the highest-scoring box, they can still extract the object and the scores of the extracted detection boxes are very low. Then, some positive samples of Scratch will be filtered by the greedy NMS threshold.

To solve this problem, the soft-NMS [11] algorithm presents a rescoring formula as shown below [26].

$$s_i = \begin{cases} s_i, & \text{iou}(M, b_i) < N_t, \\ s_i f(\text{iou}(M, b_i)), & \text{iou}(M, b_i) \geq N_t, \end{cases} \tag{1}$$

$$f(\text{iou}(M, b_i)) = 1 - \text{iou}(M, b_i), \tag{2}$$

$$f(\text{iou}(M, b_i)) = e^{-\text{iou}(M,b_i)^2/\sigma}. \tag{3}$$

The penalty is added to the score in greedy NMS when the IoU is greater than the threshold N_t. The score of other detection boxes b_i with high overlap with the highest scoring box M needs to be reduced, which is a promising way to improve greedy NMS, and the scores of detection boxes with higher overlap with M should be decayed more because they have higher false alarm rates. In soft-NMS, a linear attenuation term (2) and a Gaussian attenuation term (3) are designed. Since the IoU is not a continuous value, the linear attenuation term generates an abrupt penalty, while the Gaussian attenuation term adds redundant parameters, and the algorithm time complexity reaches $O(n^2)$. Wafer map detection that produces dense stacking will be very unfriendly to the detector, especially that it will cause the false detection rate to be very high and the detection rate of positive samples does not meet certain requirements, which will have a greater impact on industrial production. Soft-NMS effectively solves the problem of dense overlap in wafer maps. However, when it is applied to industrial production, the efficiency of soft-NMS in the test phase may be far more enough. The computational complexity increases with the increase of parameters, resulting in the inefficiency of processing a large number of samples. For this problem, the following requirements are imposed on the postprocessing process:

(1) The number of wafer map defects in industrial inspection is too large, and the detection speed is improved under the premise of ensuring the detection precision

(2) The sample elimination process cannot affect the distribution of positive and negative samples, where a noncontinuous penalty will lead to sudden changes in the ranking queue, to ensure that the penalty imposed on the score needs to be a continuous penalty value

(3) For the imposed penalty, when the overlap between the highest score extraction box within a range and other boxes is high, the penalty needs to be increased and vice versa

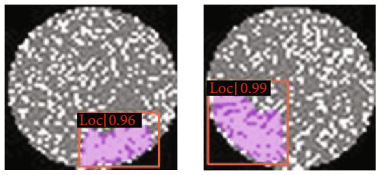
3.2. Improved fs-NMS. According to the above analysis, the logistic function is used to solve this problem, as in Equation (4).

$$s_i = \begin{cases} s_i, & \text{iou}(M, b_i) < N_t, \\ s_i g(\text{iou}(M, b_i)), & \text{iou}(M, b_i) \geq N_t, \end{cases} \tag{4}$$

where in Equation (1), $f(\text{iou}(M, b_i)) = 1/1 + e^{-\text{iou}(M,b_i)}$ is set here, $g(\text{iou}(M, b_i)) = 1 - f(\text{iou}(M, b_i)) = 1 - 1/1 + e^{-\text{iou}(M,b_i)} = 1/1 + e^{\text{iou}(M,b_i)}$, from this transformation.

$$s_i = \begin{cases} s_i, & \text{iou}(M, b_i) < N_t, \\ s_i \dfrac{1}{1 + e^{\text{iou}(M,b_i)}}, & \text{iou}(M, b_i) \geq N_t, \end{cases} \tag{5}$$

When the IoU score of the candidate box falls into a certain threshold range, this distribution belongs to an exponential distribution, which is a generalized linear model, such as the Bernoulli distribution and the Poisson distribution. The logistic function based on the generalized model is used as a penalty term. The logistic function is a continuous function with a range of 0 to 1. This property of the logistic function ensures that the probability estimated by the logistic model will never be greater than 1 or less than 0, which can be used as a penalty function. It is worth noting that, as stated in Section 3.1, IoU is not a continuous function, but rather a nonlinear function composed of explanatory variables $\text{iou}(M, b_i)$. If it is nonlinear, then imposing a sudden penalty can lead to a change in the ranking list, which can be transformed into a linear function. It would be transformed into a linear function, the resultant variable and independent variable would be transformed into a linear relationship, and the penalty value would become a



(a) (b)

Figure 1: Hybrid wafer map defects detected by Mask R-CNN with the traditional greedy NMS as the postprocessing method. Each map includes two objects, one for "Loc" and one for "Scratch" (these two categories belong to six basic wafer maps, as defined in [28]). When the image passes through the detector network, a category score is obtained, and postprocessing is performed based on the score. (a) "Loc" score 0.96 and (b) "Loc" score 0.99.

continuous value, which would not affect the rankings. If the highest extraction box M in each candidate range has a high degree of overlap with b_i in this range, the penalty needs to increase gradually, and the penalty term is close to 0.5, $\lim_{N_t \le \text{iou}(M,b_i) \longrightarrow 1} g(\text{iou}(M, b_i)) = 0.5$. The highest score box M in the candidate range has a low rate of overlapping with b_i in this range. On the contrary, the penalty should be reduced and the penalty tends to be 0. As a result, the positive samples with low scores and difficulty to detect cannot be easily removed from the ranking queue.

In addition, when the variance reaches a certain degree, the results of the standard normal distribution become similar to the logistic function in [29]. When compared with Gaussian penalty, the logistic function has fewer parameters. At the same time, the dimension of the distribution function is reduced, as is the number of calculations. Especially in inference for object detection, it will have good results as well as an improved speed.

Algorithm 1 shows the fs-NMS algorithm. D represents the final detection set, which means that the detection boxes screened by the algorithm will be sorted in D, and the final output target box and score S will be obtained. Starting from a set of extracted boxes B with corresponding scores, the top score M is chosen first and moved from set B to set D. Then, we calculate the overlap between b_i (in set B) and M, and compare it with N_t. A penalty is set to the score s_i of extraction box with degrees of overlap greater than N_t. Other scores remain unchanged, and all extraction boxes can be sorted in D. Among them, $g(\text{iou}(M, b_i))$ of fs-NMS optimizes the penalty distribution of soft-NMS according to (5), which causes it to be more efficient. The overall algorithm flowchart is shown in Figure 2.

3.3. Implementation Details. This subsection elaborates on datasets, evaluation metrics, and experiment-specific parameters and describes in detail the training and inference processes of the network. A model pretrained on ImageNet [30] is used in the experiments to initialize the detection network. The complete training and testing code was built on Pytorch [31] and mmdetection [32]. The settings of mmdetection are followed if some hyperparameters are not mentioned in this experiment.

Input: $B = \{b_1 \cdots, b_N\}, S = \{s_1 \cdots, s_N\}, N_t$
B is the list of initial detection boxes
S contains corresponding detection score
N_t is the NMS threshold
begin
  $D \longleftarrow \{\}$.
  while $B \neq$ empty do
    $m \longleftarrow$ argmax S
    $M \longleftarrow b_m$
    $D \longleftarrow D \cup M$; $B \longleftarrow B - M$
    for b_i in B do
      if $\text{iou}(M, b_i) \geq N_t$ then
        $s_i \longleftarrow s_i g(\text{iou}(M, b_i))$
      end
    end
  end
  return D, S
end

Algorithm 1: fs-NMS algorithm.

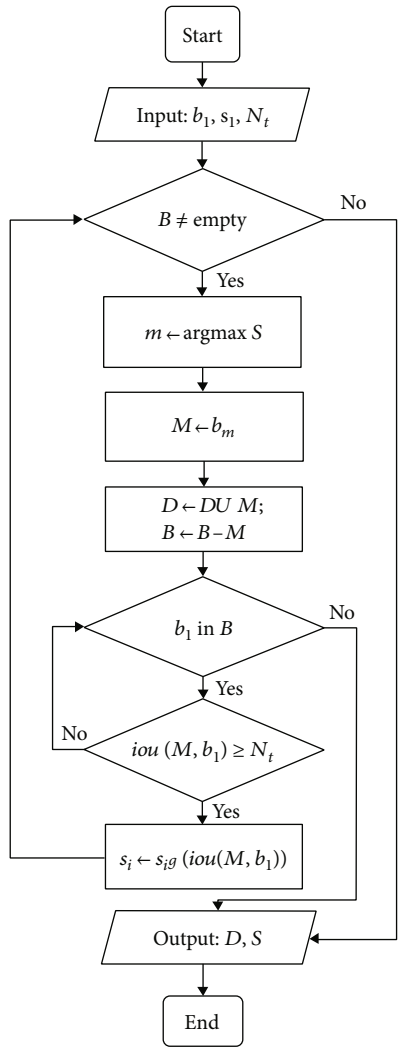


Figure 2: Flowchart of the proposed fs-NMS algorithm.

Datasets. Experiments on the 6 category wafer map datasets [28] used 2 k training images and 382 validation sets and 429 test images. There are 3.1 k wafer bin maps (WBMs), including 6.2 k objects. In wafer testing, the final test results on the wafer are stored in the WBMs, which consist of binary values, and the WBM has six classical defect patterns (a–f) in our experiment. According to [33], these generally divided into these categories, i.e., "Center," "EdgeLoc," "Loc," "Donut," "Scratch," and "EdgeRing." The size of each wafer map is 52×52, and some specific image information is shown in Figure 3.

Evaluation metrics. Two evaluation metrics are used in our experiments. The AP and time are used to evaluate the applicability of the network and apply a statistical parameter of precision (P) to our experiment. Next, if the defect detection effect is evaluated and the detection positioning performance is evaluated, the detection index and four parameters are defined as follows:

True positive (TP): predicting positive, the actual is positive.

False positive (FP): predicting positive, the actual is negative.

False negative (FN): predicting negative, the actual is negative.

True negative (TN): predicting positive, the actual is negative.

(i) AP. P is the ratio of the number of correctly predicted WBMs to the number of WBMs tested, and R is the ratio of the number of correctly predicted WBMs to all ground truths of WBMs. After each object is classified, a confidence level is an output, and a confidence level threshold is set to obtain a pair of P-R. Taking different confidence level thresholds, more pairs of P-R can be obtained, and the maximum value of P corresponding to all the recall R greater than the specified recall r is used as the maximum P under the currently specified recall r.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{6}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{7}$$

Next, the WBM classification task is a multiclassification task, and mAP is the average precision for all categories. Thus, the mAP is used to evaluate the overall effect as follows:

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i. \tag{8}$$

With AP being the AP_i in the i-th class and N is the total number of classes of WBMs being evaluated. The average precision (AP) metric averages the AP across IoU thresholds from 0.5 to 0.95 with an interval of 0.05. For box AP, AP_{50}, AP_{75} (AP at different IoU thresholds), and AP_S, AP_M (AP at different scales, where small (S) size is 32×32 and medium (M) size is 52×52) are also reported.

(ii) *Time*. Time to detect all wafer map test sets. The runtime is measured on a single NVIDIA Tesla P100 GPU.

Experimental setting. Experiments use a single GPU to train the detector for 24 epochs, and the other model parameters are listed in Table 1. For the two-stage detectors, the baselines in the experiments are to use heuristic methods for model initialization and optimization. To avoid the model oscillation caused by a large learning rate (lr), a warm-up strategy is used in the initial 500 iterations, which caused the model to stabilize slowly. After the model is relatively stable, a preset lr of 0.02 is used for training. Among them, the initial lr for Cascade Mask R-CNN [21] and Grid R-CNN [34] is set to 0.002 because of the divergence of the loss function caused by the gradient explosion in the training process. After the 16th and 22nd epochs, the lr is reduced by 0.1, respectively. For each type of detector, the image size is adjusted to 52×52, and the aspect ratio uses the same design parameters. The detectors use a stochastic gradient descent (SGD) optimizer with a weight decay of 0.0001 and momentum 0.9. The batch size of the dataset is set to 8. For single-stage detectors, the specific parameters of the baseline are listed in Table 1. Due to the nature of the feature extraction part of the network, the image size and aspect ratio follow the setting of mmdetection.

Training. Experiments are trained on some baseline detectors, as shown in Figure 4. For the backbone part of the object detection network in two-stage (RPN in the RoI extraction stage and the classification localization stage) using a model pretrained on ImageNet [30] by ResNeXt-101 [19], the experiments are compared to the performance of ResNet [18] and ResNeXt [19] network, and the ResNeXt network is higher than ResNet in terms of accuracy. For the extraction of the RoI phase, the experiment explores the feature pyramid network (FPN) [17]. The R-CNN network with FPN backbone can extract RoI features from different levels of the feature pyramid. The backbone of ResNeXt based on ResNet-FPN is used for feature extraction, which has better improvement in precision and speed [22]. In the training stage, greedy NMS is used for postprocessing after RPN extraction, and the threshold value is 0.7 [32] in the RPN extraction stage. All detectors are trained for 24 epochs, and the average accuracy of baseline detector classification reached 98.89% on the validation sets of the wafer map. For the backbone part of the single-stage detector, Darknet [23], VGG [8], and ResNet [9] are used for feature extraction.

Inference. In inference, the greedy NMS is replaced with soft-NMS and fs-NMS for postprocessing, respectively. Compared with the two-stage network structure, it is worth noting that in the two-stage shown in Figure 4, the NMS with a threshold of 0.7 is used in the RoI extraction stage (RPN) and the fs-NMS with a threshold of 0.5 in the classification and localization stage (after fully connectional layer). On the same amount of wafer map test sets, the

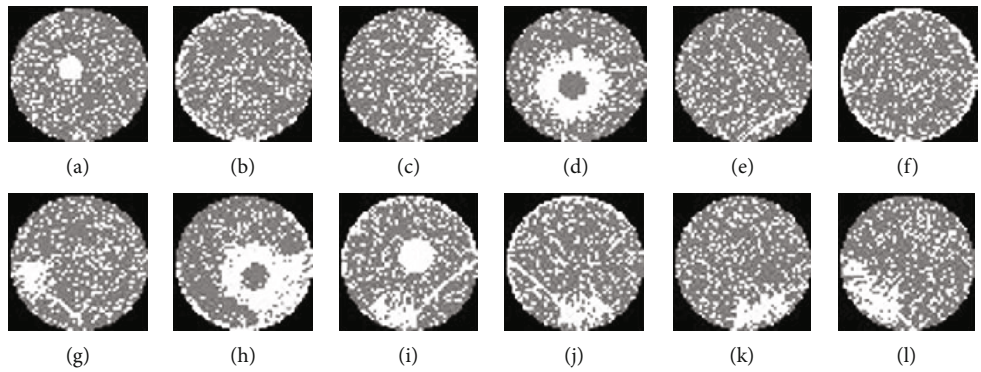


Figure 3: Basic types of wafer maps. (a–f) and mixed defect types (g–l). (a) "Center"; (b) "EdgeLoc"; (c) "Loc"; (d) "Donut"; (e) "Scratch"; (f) "EdgeRing"; (g) "Loc"+"Scratch"; (h) "Donut"+"Loc"+"EdgeLoc"; (i) "Center"+"EdgeLoc"+"Scratch"+"Loc"; (j) "Loc"+"EdgeLoc"+"Scratch"; (k) "Loc"+"Scratch"; (l) "Loc"+"Scratch".

Table 1: The basic experimental parameter settings of the baseline.

Baseline	Learning rate	Weight decay
[22, 35, 36]	0.02	0.0001
Cascade Mask R-CNN [21]	0.002	0.0001
Grid R-CNN [34]	0.002	0.0001
SSD 300 [9]	0.002	0.0005
YOLO v3 [23]	0.001	0.0005
RetinaNet [10]	0.01	0.0001

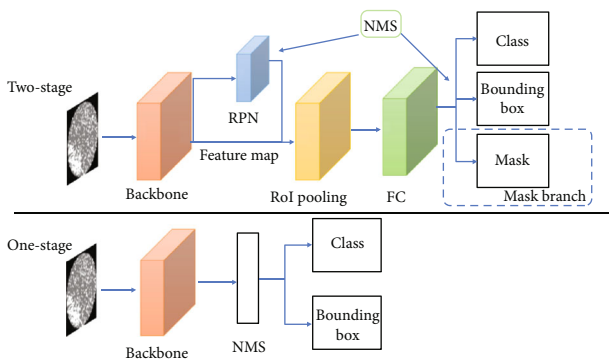


Figure 4: Network structures of the single-stage and two-stage detectors. These include the overall network structure of the general single-stage (e.g., YOLO [23], SSD [8], and RetinaNet [9]) and two-stage (e.g., Faster R-CNN [16] and Mask R-CNN [22]) detectors.

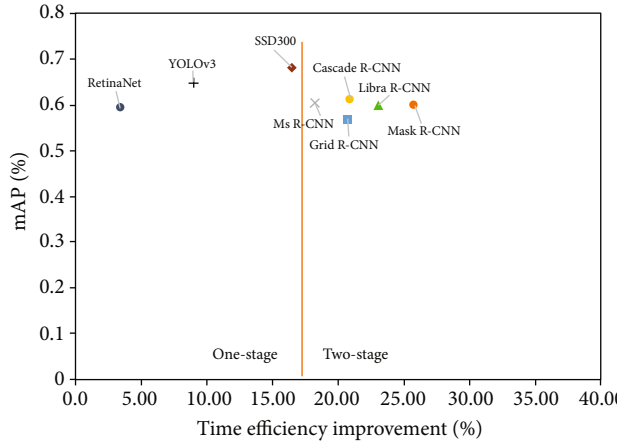
detection speed will be improved compared to soft-NMS [11], and the precision will not be affected. For single-stage detectors, the threshold of 0.45 for NMS in inference is better. Since the standard Gaussian penalty considered in Section 3.2 is similar to the logistic function to some extent, the imposed penalty will have the same effect, and the AP is unaffected. Secondly, compared with the linear function, the logistic function has the same efficiency as the general linear function, and the time complexity of our algorithm reaches $O(n)$, which is the same as that of the linear penalty in (3).

4. Experimental Results

This section mainly describes the results of the comparative experiments, including the comparison of detection efficiency in inference of soft-NMS and fs-NMS (Section 4.1) and the comparison of detection precision between greedy NMS and fs-NMS (Section 4.2). The experiments are carried out on some baseline detectors.

4.1. Efficiency Comparison Test. Our algorithm is first tested against some baseline detectors to better understand how it affects efficiency. In inference, since the time efficiency of soft-NMS is far more enough, the processed method improves the detection efficiency by optimizing the distribution model of penalty. As shown in Figure 5, for the two-stage detector, comparison experiments are performed on Mask R-CNN [22], Mask scoring R-CNN [36], Cascade Mask R-CNN [21], Grid R-CNN [34], and Libra R-CNN [35]. The performance of the proposed method is validated by replacing the postprocessing component. In inference, the greedy NMS is used in the RoI extraction stage, while the fs-NMS is used in the classification and localization stages. Regardless of the relative position or angle of the object or different image features, the detector can find accurate objects in the key area. The detection precision would not be affected; this method improved the detection speed. As shown in Figure 5, the efficiency of the two-stage detector based on the extraction box increases by 21.72% on average, especially in the Mask R-CNN [22] detector by 25.8%. For the single-stage detector, the experiment was carried out on YOLO v3 [23], SSD 300 [8], and RetinaNet [9]. The filtering thresholds were set to 0.45, 0.45, and 0.5, respectively. The detection efficiency of single-stage detectors based on extraction box improved by 9.63% on average.

The overall time of the single-stage detectors using the improved algorithm is increased compared with that of the two-stage detector. Due to a large number of extraction boxes of the single-stage detector, the postprocessing algorithm of rescore will cause the detector efficiency not as good as that of the greedy NMS of the strict pruning branch, and the efficiency of detection (Figure 5) and the detection effect (more clutter extraction boxes will be



	mAP	Time (%)
Mask R-CNN	0.598	25.8%
Ms R-CNN	0.605	15.2%
Cascade R-CNN	0.612	20.9%
Grid R-CNN	0.566	20.7%
Libra R-CNN	0.600	23.0%
YOLOv3	0.645	9.0%
SSD300	0.681	16.5%
RetinaNet	0.594	3.4%

Figure 5: Efficiency improvement (compared to soft-NMS) versus the average precision (mAP) on the baseline and wafer map test sets.

generated) is not as good as the two-stage detector using the improved algorithm.

4.2. AP Comparison Test. In terms of the overall detector precision, the experiment replaced only the postprocessing of inference, comparing the detector precision using traditional greedy NMS and fs-NMS, where threshold settings remain the same. The ablation study is shown in Figure 6, where the postprocessing algorithm is charged to perform the comparison experiments. The abscissa of Figure 6 represents the time of fs-NMS, and the ordinate represents the growth of mAP. From the perspective of inference time, the time of the two-stage detector is almost within 60 ms. The overall efficiency is greater than one-stage detector. On the test sets, according to Section 3.3, AP evaluates the correct positioning of the object. The regression of AP_{75} relative to AP_{50} is more accurate, and AP_s and AP_m find different details of the object for the boxes of different scales. In postprocessing, the precision of the baseline detector is improved, as well as the location of the object at different scales. According to (8), the fs-NMS algorithm improves the average precision of the detector for each category of object location, and the two-stage detector on average mAP is increased by 1.76%.

Some recent single-stage detectors all have better detection precision and efficiency compared with the two-stage detectors, showing a 2.7% improvement in mAP. RetinaNet extracts a large amount of anchor in the extraction stage, reaching 100 k [9]. For such two-stage detectors [16], the total number of extracted boxes is only 20 k, and only a small

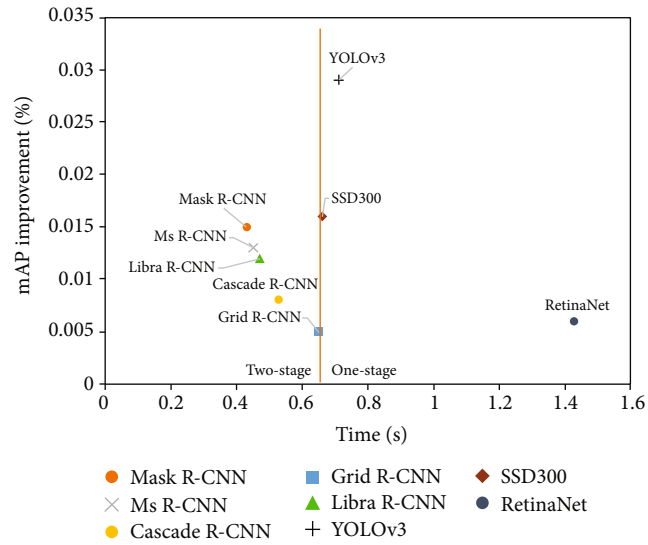


Figure 6: Time (fs-NMS) versus average precision (compared to greedy NMS) on the baseline and wafer map test sets.

fraction of them eventually coincide with the ground truth box. Therefore, anchor-based single-stage detectors are dependent on the postprocessing algorithm, and the improvement in precision will be higher than the two-stage detectors. Overall, the fs-NMS algorithm is effective in both the single-stage and two-stage detectors and achieves the same effect as soft-NMS in improving precision.

Table 2: Greedy NMS vs. fs-NMS in mAP using Mask R-CNN [22] detector.

Baseline	Greedy NMS	fs-NMS	mAP	AP_{50}	AP_{75}	AP_s	AP_m
Mask R-CNN [22]	√		0.584	0.830	0.648	0.472	0.760
		√	0.598	0.831	0.676	0.485	0.777

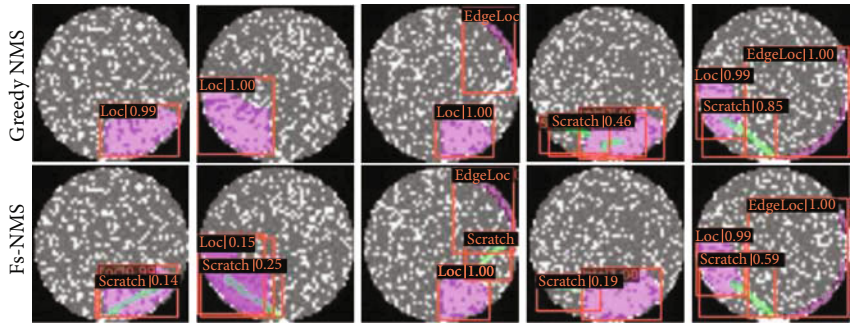


Figure 7: (a) Greedy NMS vs. (b) fs-NMS using Mask R-CNN [22] detector.

According to Table 2, Mask R-CNN uses two types of postprocessing methods (i.e., greedy NMS and fs-NMS). And fs-NMS obtains an improvement of 0.014 on mAP compared to greedy NMS. Secondly, the experimental results show that some hard-to-detect (e.g., too much overlap between ranges) objects appear with lower scores, as shown in Figure 7. At the same time, it can be clearly seen that the algorithm rescores the results after imposing penalties on scores.

5. Conclusion

This paper proposed a novel fs-NMS algorithm for the postprocessing stage of wafer map detection. Firstly, we discussed several key issues relating to the inefficiency of the traditional NMS algorithm and proposed an improved fs-NMS algorithm to solve this problem. The algorithm rescores the score of the detection box by optimizing the penalty distribution model of soft-NMS, with the objective of improving the detection efficiency of inference and ensuring the stability of the precision. Meanwhile, the object detection method was explored and applied to the defect detection of the wafer map to improve the efficiency of industrial detection.

The experiments used base on wafer map datasets. The results show that in inference, the fs-NMS algorithm outperforms traditional NMS in anchor-based detection precision. From the test results, highly overlapped defect objects will produce many false-positive boxes (Figure 7) that cannot be completely eliminated. This provides a direction for our future research.

Data Availability

The original data are https://github.com/Junliangwangdhu/WaferMap.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. Kyeong and H. Kim, "Classification of mixed-type defect patterns in wafer bin maps using convolutional neural networks," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 3, pp. 395–402, 2018.

[2] K. B. Lee, S. Cheon, and C. O. Kim, "A convolutional neural network for fault classification and diagnosis in semiconductor manufacturing processes," *IEEE Transactions on Semiconductor Manufacturing*, vol. 30, no. 2, pp. 135–142, 2017.

[3] F. Adly, O. Alhussein, P. D. Yoo et al., "Simplified subspaced regression network for identification of defect patterns in semiconductor wafer maps," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 6, pp. 1267–1276, 2015.

[4] Y. Xia, H. Yu, and F. Y. Wang, "Accurate and robust eye center localization via fully convolutional networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 5, pp. 1127–1138, 2019.

[5] P. M. Kebria, A. Khosravi, S. M. Salaken, and S. Nahavandi, "Deep imitation learning for autonomous vehicles based on convolutional neural networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 7, no. 1, pp. 82–95, 2020.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, Columbus, USA, 2014.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, pp. 779–788, Las Vegas, USA, December 2016.

[8] W. Liu, D. Anguelov, D. Erhan et al., "SSD: single shot multibox detector," in *Proceedings European conference on computer vision (ECCV). LNCS*, pp. 21–37, Amsterdam, The Netherlands, October 2016.

[9] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, Venice, Italy, October 2017.

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886–893, San Diego, USA, June 2005.

[11] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS-improving object detection with one line of code," in *Proceedings of the IEEE international conference on computer vision*, pp. 5562–5570, Venice, Italy, October 2017.

[12] T. Nakazawa and D. V. Kulkarni, "Wafer map defect pattern classification and image retrieval using convolutional neural network," *IEEE Transactions on Semiconductor Manufacturing*, vol. 31, no. 2, pp. 309–314, 2018.

[13] J. R. R. Uijlings, K. E. A. Van De Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision (ECCV)*, pp. 346–361, Zurich, Switzerland, 2014.

[15] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, pp. 91–99, Montreal, Canada, 2015.

[17] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Honolulu, USA, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, USA, 2016.

[19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 5987–5995, Honolulu, USA, 2017.

[20] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Advances in Neural Information Processing Systems 29 (NIPS 2016).*,, pp. 379–387, Barcelona, Spain, December 2016.

[21] Z. Cai and N. Vasconcelos, "Cascade R-CNN: delving into high quality object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6154–6162, Salt Lake City, USA, 2018.

[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[23] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," pp. 1–6, 2018, http://arxiv.org/abs/1804.02767.

[24] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6469–6477, Honolulu, USA, 2017.

[25] L. Tychsen-Smith and L. Petersson, "Improving object localization with fitness NMS and bounded IoU loss," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6877–6885, Salt Lake City, USA, 2018.

[26] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: refining pedestrian detection in a crowd," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, pp. 6452–6461, Long Beach, USA, 2019.

[27] Y. He, C. Zhu, J. Wang, M. Savvides, and X. Zhang, "Bounding box regression with uncertainty for accurate object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2883–2892, Long Beach, USA, 2019.

[28] J. Wang, C. Xu, Z. Yang, J. Zhang, and X. Li, "Deformable convolutional networks for efficient mixed-type wafer defect pattern recognition," *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 4, pp. 587–596, 2020.

[29] F. E. Harrell, *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer, Heidelberg, 2nd edition, 2016.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 248–255, Miami, USA, June 2009.

[31] A. Paszke, S. Gross, S. Chintala et al., "Automatic differentiation in PyTorch," in *Proceedings of 31st Conference on Neural Information Processing Systems*, pp. 1–4, Long Beach, USA, 2017.

[32] K. Chen, J. Wang, J. Pang et al., "MMDetection: open MMLab detection toolbox and benchmark," 2019, http://arxiv.org/abs/1906.07155.

[33] M. J. Wu, J. S. R. Jang, and J. L. Chen, "Wafer map failure pattern recognition and similarity ranking for large-scale data sets," *IEEE Transactions on Semiconductor Manufacturing*, vol. 28, no. 1, pp. 1–12, 2015.

[34] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid R-CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7355–7364, Long Beach, USA, 2019.

[35] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: towards balanced learning for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 821–830, Long Beach, USA, 2019.

[36] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*, pp. 6402–6411, Long Beach, USA, June 2019.